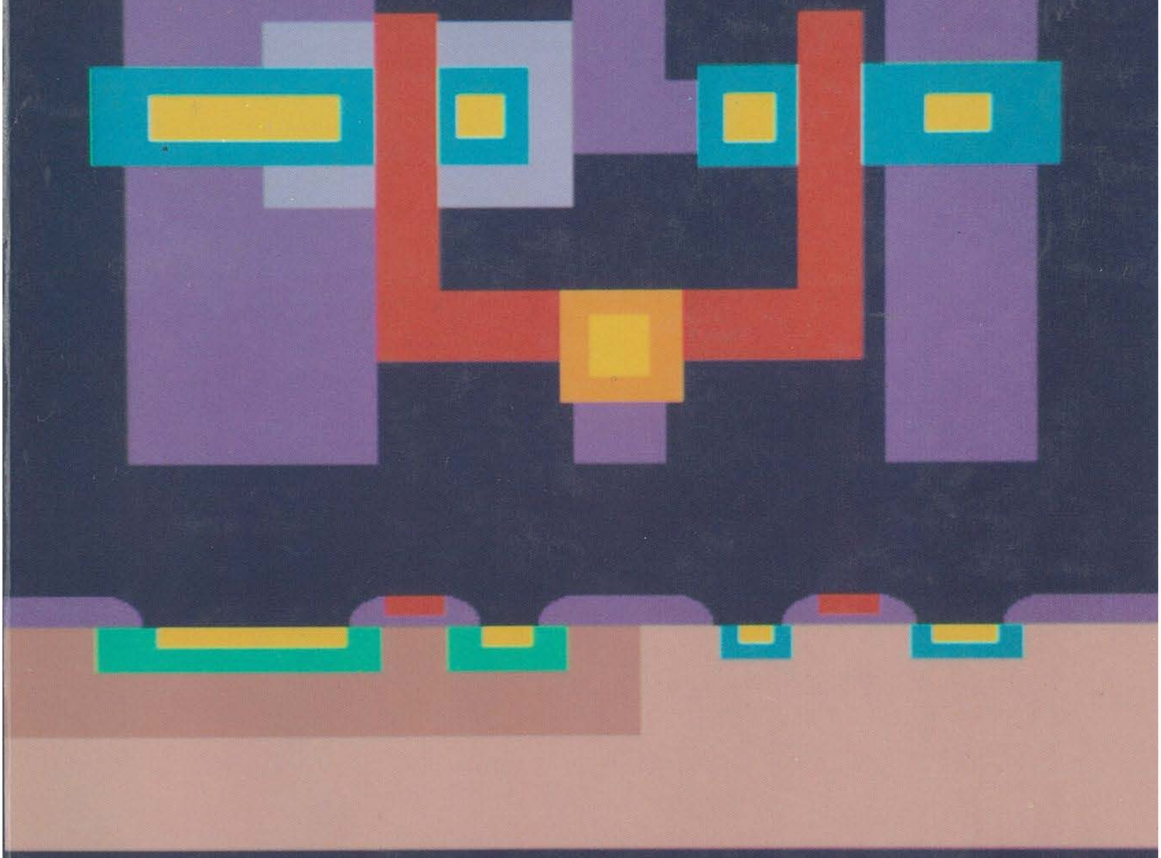
Silicon Processing
for the VLSI Era
Volume 2 - Process Integration



S. Wolf

Silicon Processing
for the VLSI Era

Volume 2
Process
Integration

S. Wolf

THE COMPANION VOLUME TO THIS BOOK

Silicon Processing

for the VLSI Era
Volume 1 – Process Technology

By STANLEY WOLF and RICHARD N. TAUBER

"Arguably the best available text on the subject"
IEEE Transactions on Education

"The authors of this volume, have succeeded admirably. The book is a uniquely valuable reference for the professional microelectronic engineer.... Volume 2 of this text will be eagerly awaited."
Semiconductor Science and Technology

"Well organized, and presented in a logical and clear fashion. Lucid writing style, with abundant and well-placed illustrations... Overall, an excellent book on silicon processing... I highly recommend it!
IEEE Circuits and Devices Magazine

TABLE of CONTENTS - Ch. 1 Silicon: Single-Crystal Growth and Wafering; **Ch. 2** Crystalline Defects, Thermal Processing; and Gettering; **Ch. 3** Vacuum Technology for VLSI Applications; **Ch. 4** Basics of Thin Films; **Ch. 5** Silicon Epitaxial Film Growth; **Ch. 6** Chemical Vapor Deposition of Amorphous and Polycrystalline Films; **Ch. 7** Thermal Oxidation of Single-Crystal Silicon; **Ch. 8** Diffusion in Silicon; **Ch. 9** Ion Implantation for VLSI; **Ch. 10** Aluminum Thin Films and Physical Vapor Deposition in VLSI; **Ch. 11** Refractory Metals and Their Silicides in VLSI; **Ch.12** Lithography I: Optical Photoresists-Material Properties and Process Technology; **Ch. 13** Lithography II: Optical Aligners and Photomasks; **Ch. 14** Advanced Lithography: E-Beam and X-Ray; **Ch. 15** Wet Processing: Cleaning, Etching and Lift-Off; **Ch. 16** Dry Etching for VLSI; **Ch. 17** Material Characterization Techniques for VLSI Fabrication; **Ch. 18** Structured Approach to Design of Experiments for Process Optimization.

1986 LATTICE PRESS 660 pp. ISBN 0-9616721-3-7

Order using form on last page of this book

CONTENTS

LIST OF TECHNICAL REVIEWERS

Each of the chapters was reviewed for technical correctness. The following persons graciously undertook the review task for the chapters indicated:

Chapter 2	Dr. Joseph R. Monkowski Lam Research Corp. - CVD Division Fremont, CA	Dr. Haiping Dun Intel Corp. Santa Clara, CA
Chapter 3	Dr. Robert S. Blewer Sandia National Laboratories Albuquerque, NM	Dr. Stan Swirhun Honeywell SSPL Bloomington, MN
Chapter 4	Dr. Farhad K. Moghadam Intel Corp. Santa Clara, CA	Dr. Terry Herndon MIT - Lincoln Laboratory Lexington, MA
Chapter 5	Mr. Andrew R. Coulson TRW Electronic Systems Group Redondo Beach, CA	
Chapter 6	Dr. John Y. Chen Boeing Electronics Seattle, WA	Dr. Samuel T. Wang International CMOS Technology, Inc. San Jose, CA
Chapter 8	Professor Al F. Tasch, Jr. University of Texas Austin, TX	
Chapter 9	Dr. Michael Kump Technology Modeling Associates, Inc. Palo Alto, CA	

PREFACE

SILICON PROCESSING FOR THE VLSI ERA is a text designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The text will consist of three volumes, of which this book is the second, subtitled, *Process Integration*. Volume 1, subtitled *Process Technology*, was published in 1986. Volume 3, to be subtitled *Assembly, Packaging and Manufacturing Technology* is scheduled for publication in 1993. In Volume 1, the individual processes utilized in the fabrication of silicon VLSI circuits are covered in depth (e.g., epitaxial growth, chemical vapor and physical vapor deposition of amorphous and polycrystalline films, thermal oxidation of silicon, diffusion, ion implantation, microlithography, and etching processes).

In this volume, we undertake to explain how the individual processes described in Volume 1 are combined in various ways to produce silicon integrated circuits. This task is referred to as *process integration*. The first part of the book deals with *sub-process integration;* that is, the effort involved in forming circuit structures that can be implemented into a variety of circuit types. These structures include *isolation structures* (Chap. 2), *metal-silicon contacts* (Chap. 3), and *device-interconnect structures* (Chap. 4).

The second part of the book covers the process integration tasks of full-device-type technologies, including *NMOS* (Chap. 5), *CMOS* (Chap. 6), *bipolar* and *BiCMOS* (Chap. 7), and *semiconductor memories* (Chap. 8). Chapter 9 describes the process simulation tools that are available for aiding in the process integration and development efforts.

Volume 3 will cover process control, VLSI manufacturing issues and facilities, contamination and yield, automation, assembly, packaging, and parametric testing.

The purpose of writing this text was to provide professionals involved in the microelectronics industry with a single source that offers a complete overview of the technology associated with the manufacture of silicon integrated circuits. Other texts on the subject are available only in the form of specialized books (i.e., that treat just a small subset of all of the processes), or in the form of edited volumes (i.e., books in which a group of authors each contributes a small portion of the contents).

Such edited volumes typically suffer from a lack of unity in the presented material from chapter-to-chapter, as well as an unevenness in writing style and level of presentation. In addition, in multi-disciplinary fields, such as microelectronic fabrication, it is difficult for most readers to follow technical arguments in such books, especially if the information is presented without defining each technical "buzzword" as it is first introduced.

In our books such drawbacks are avoided by treating the subject of VLSI fabrication from a unified and more pedagogical viewpoint, and by carefully defining technical terms when they are first introduced. The result is intended to be a user friendly book for workers who have come to the semiconductor industry after having been trained in but one of the many traditional technical disciplines.

An important technical breakthrough has occurred in publishing that the author also felt could be exploited in creating a unique book on silicon processing. That is, revolutionary electronic publishing techniques have recently become available, which can cut the time required to produce a published book from a finished manuscript. This task traditionally took 15-18 months, but can be now reduced to less than 3 months. If traditional techniques are used to produce books in such fast-breaking fields as VLSI fabrication, these books automatically possess a built-in obsolescence, even upon being first published.

We have taken advantage of these rapid production techniques, and have been able to successfully meet the reduced production-time schedule. As a result, information contained in technical journals and conferences which was available within three months of the book's publication date has been included.

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

The book is an outgrowth of several intensive seminars conducted by the author through the Engineering Extension of the University of California, Irvine. Over one thousand engineers and managers from more than 75 companies and government agencies have enrolled in these short courses since they were first presented in 1984.

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

Each of the chapters was reviewed after the writing was completed. The engineers and scientists who participated in this review were numerous. The main reviewers are

listed on the page before the preface, and we would like to thank them once again at this point for their contributions. That is, the following professionals (each one an expert in the topic covered by the chapter they reviewed), read an entire chapter for technical correctness, and provided appropriate comments and corrections: Robert S. Blewer, John Y. Chen, Andrew R. Coulson, Haiping Dun, Terry Herndon, Michael Kump, Farhad K. Moghadam, Joseph R. Monkowski, Stan Swirhun, Al F. Tasch, Jr., and Samuel T. Wang. In addition, J. B. Price, of Spectrum CVD, Chris A. Mack, of the National Security Agency, and Sidney Marshall, Editor of Solid State Technology, provided other valuable technical and editorial input. Robert Shier, of VTC Corp., and Jim Cable of TRW also kindly provided the author with timely and valuable technical literature.

The copy editing of the book was undertaken by Mary Nadler, and the clarity and grammatical correctness of the prose owes a great deal to her professional efforts. The aesthetically pleasing graphics of the cover were designed by Roy Montibon of Visionary Art Resources, Inc., Santa Ana, CA.

Stanley Wolf

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

Lattice Press
P.O. Box 340-W
Sunset Beach, CA, 90742

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

CHAPTER 1

PROCESS INTEGRATION FOR VLSI and ULSI

Since the creation of the first integrated circuit in 1960, the density of devices that can be fabricated on semiconductor substrates has steadily increased. In the late 1970s the number of devices manufactured on a chip exceeded the generally accepted definition of *very large scale integration*, or *VLSI* (i.e., more than 100,000 devices per chip) (Fig. 1-1a). By 1990 this number had grown to more than 32 million devices per chip (16-Mbit DRAMs), and it is generally acknowledged that the era of *ultra-large-scale integration (ULSI)* has begun. The increasing device count has been accompanied by a shrinking minimum feature size (Fig. 1-1b), which as of 1990 had decreased to ~0.5 μm in the most advanced commercially available chips.

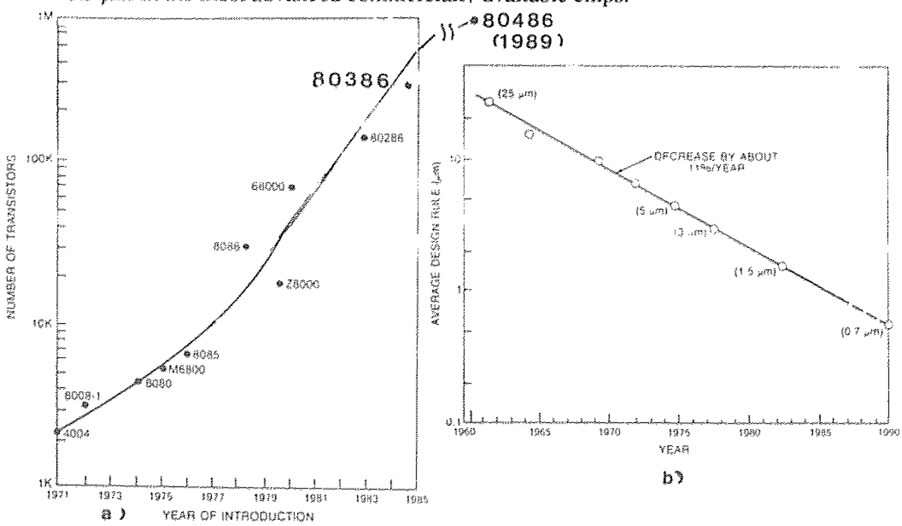


Fig. 1-1 (a) Increase in the number of transistors per microprocessor chip versus year of introduction, for a variety of 8-bit and 16-bit microprocessors, and (b) The decrease in minimum device feature size versus time on integrated circuits.

Progress seems likely to continue at a rapid pace, with even further reductions in the unit cost per function and in the power-delay product of integrated circuits projected. Silicon processing has been the dominant technology of IC fabrication and is likely to retain this position for the foreseeable future. The entire adventure of silicon device manufacturing represents a remarkable application of scientific knowledge to the requirements of technology. Our books* are intended to serve as a comprehensive and cohesive report on the state of the art of this technology, as practiced at the time of publication.

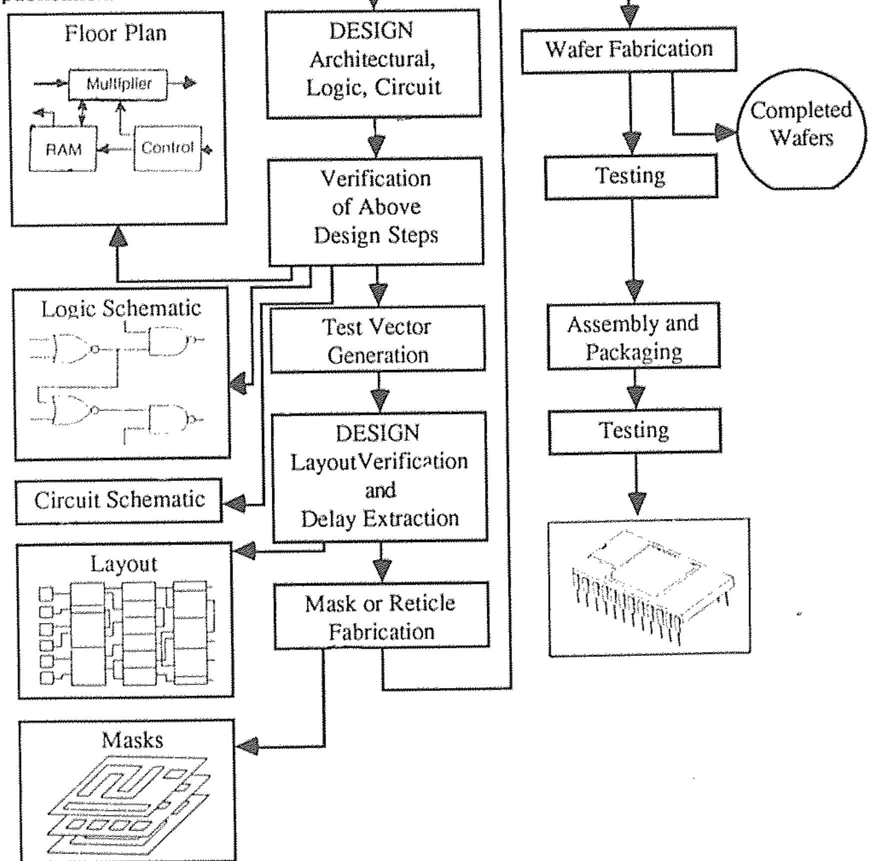


Fig. 1-2 Steps required for the manufacture of very large scale integrated circuits (VLSI).

* SILICON PROCESSING FOR THE VLSI ERA - Volume 1: Process Technology; Volume 2 - Process Integration; and Volume 3 - Assembly, Packaging, and Manufacturing Technology (the latter is scheduled to to be published *approximately* in 1992)

Figure 1-2 illustrates the sequence of tasks followed in the manufacture of an integrated circuit. These tasks can be grouped into three phases: *design*, *fabrication*, and *testing*. Our books are concerned primarily with the fabrication phase. We describe the IC manufacturing steps that occur from the point at which the circuit design has been completed and the necessary design information has been rendered into the form of a circuit layout. In this form, the layout information is ready to be used to generate a set of masks (or reticles) that will serve as the tools for specifying the circuit patterns on silicon wafers. For VLSI and ULSI circuits, the layout information is stored in a computer.

The details concerning the individual fabrication processes (e.g., those associated with creating patterns, introducing dopants, and depositing films on silicon substrates to form integrated circuits) are the subjects of Volume 1. In this second volume, we

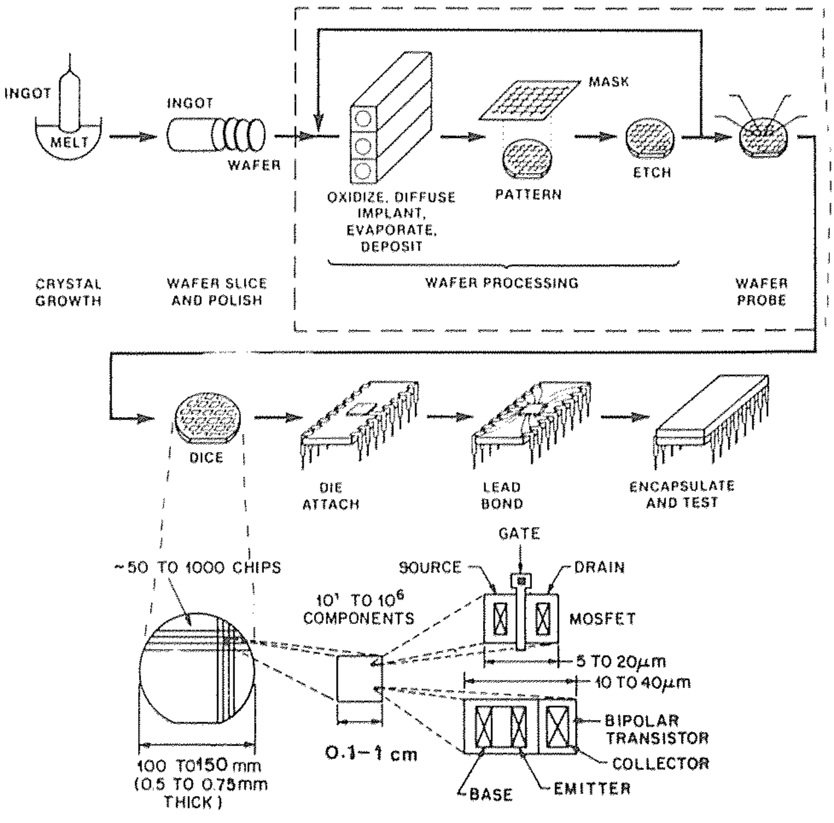


Fig. 1-3 The fabrication process sequence of integrated circuits.

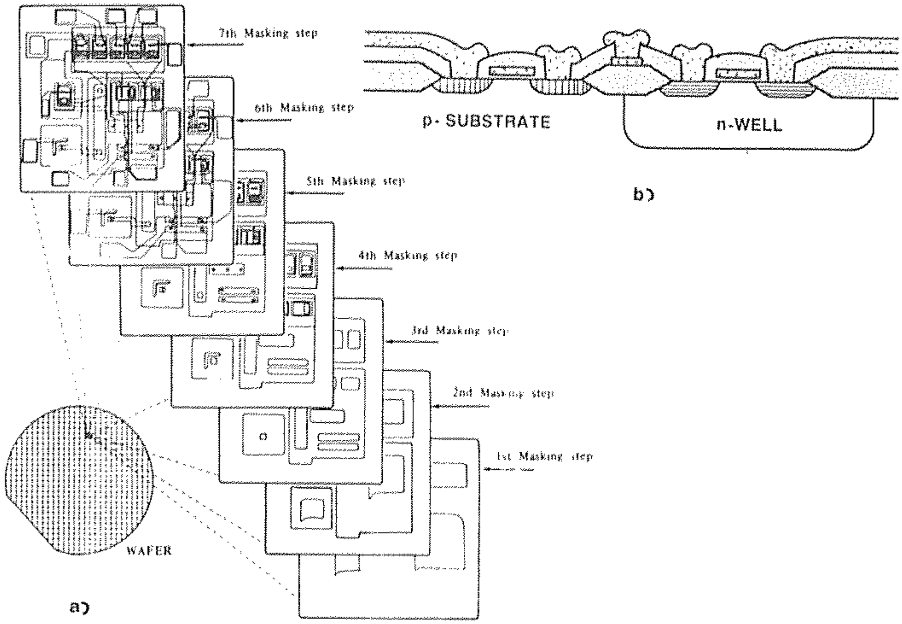


Fig. 1-4 (a) Example of the patterns transferred to a wafer during a seven-mask process sequence, and (b) Cross section of completed devices in a basic CMOS process.

undertake a discussion of the *sequences* of the steps performed to produce the device structures of the circuits. The text describes how the various individual processes are integrated together so that the end result is a completely realized integrated circuit.

One perspective of such a *process-integration* effort is illustrated in Fig. 1-4. It can be seen that a series of masking steps must be sequentially performed for the desired patterns to be created on the wafer surface. The other processing procedures performed between the masking steps serve to create the desired device structures. An example of the end result of a complete process sequence is shown in Fig. 1-4b as an IC cross-section.

This book is roughly divided into two parts. The first part deals with what we refer to as *subprocess integration* (Chapters 2, 3, and 4), while the second covers *complete process integration* (Chapters 5, 6, 7, and 8). Chapter 9 covers the subject of *process simulation*, an approach applicable to both the subprocess and complete process-integration development efforts.

The function of subprocess integration is to produce device structures that can be used in a variety of integrated-circuit technologies (i.e., isolation, contact, and interconnect structures). The purpose of the process-integration task is to design the process sequences used to manufacture complete IC technologies (i.e., NMOS, CMOS, bipolar, BICMOS, and IC memory devices).

CHAPTER 5

MOS DEVICES AND NMOS PROCESS INTEGRATION

This chapter deals with NMOS process integration. To provide background information for this topic, a review of the physics of long-channel MOS devices (i.e., whose channel lengths are greater than 2 μm) and the basics of MOS circuit design is presented. Also included is a discussion of the relationship between desired device performance and process technology. (More rigorous treatments of MOS device physics are found in references 1 - 5.)

The history of NMOS processing is also presented, emphasizing how the obstacles to fabricating reliable NMOS ICs were overcome. A detailed example of the fabrication sequence of a typical NMOS inverter circuit culminates this discussion. The chapter concludes with a description of short-channel and hot-carrier effects in MOS devices, together with processing techniques developed to combat the problems they cause.

5.1 MOS DEVICE PHYSICS

5.1.1 The Structure and Device Fundamentals of MOS Transistors

A perspective view of an *n-channel MOSFET* is shown in Fig. 5-1a, while additional details of its structure and its circuit symbol are given in Figs. 5-1b and 5-1c, respectively. The device has a *gate terminal* (to which the input signal is normally applied), as well as *source* and *drain* terminals (across which the output voltage is developed, and through which the output current flows, i.e., the drain-source current, I_D). The gate terminal is connected to the gate electrode (a conductor), while the remaining terminals are connected to the heavily doped source and drain regions in the semiconductor substrate.

A *channel* region in the semiconductor under the gate electrode separates the source and drain. The channel (of length L and width W) is lightly doped with a dopant type opposite to that of the source and drain. The semiconductor is also physically separated

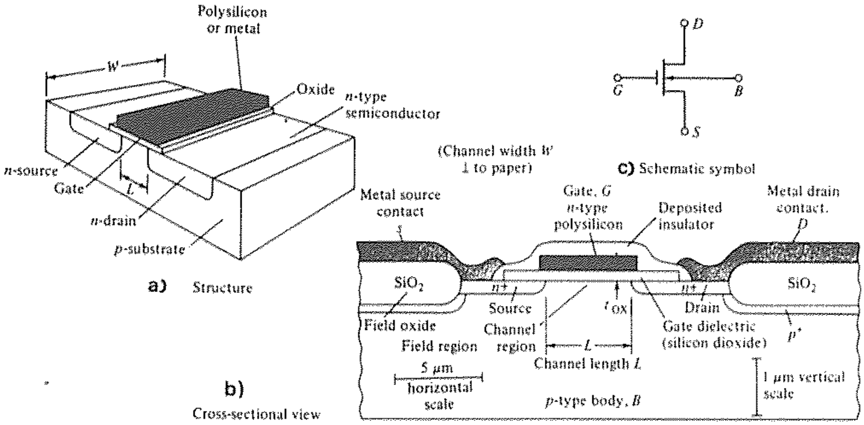


Fig. 5-1 (a) Structure of an MOS device. (b) Cross sectional view. (c) Schematic symbol.[7] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

from the gate electrode by an insulating layer (typically, SiO_2), so that no current flows between the gate electrode and the semiconductor.

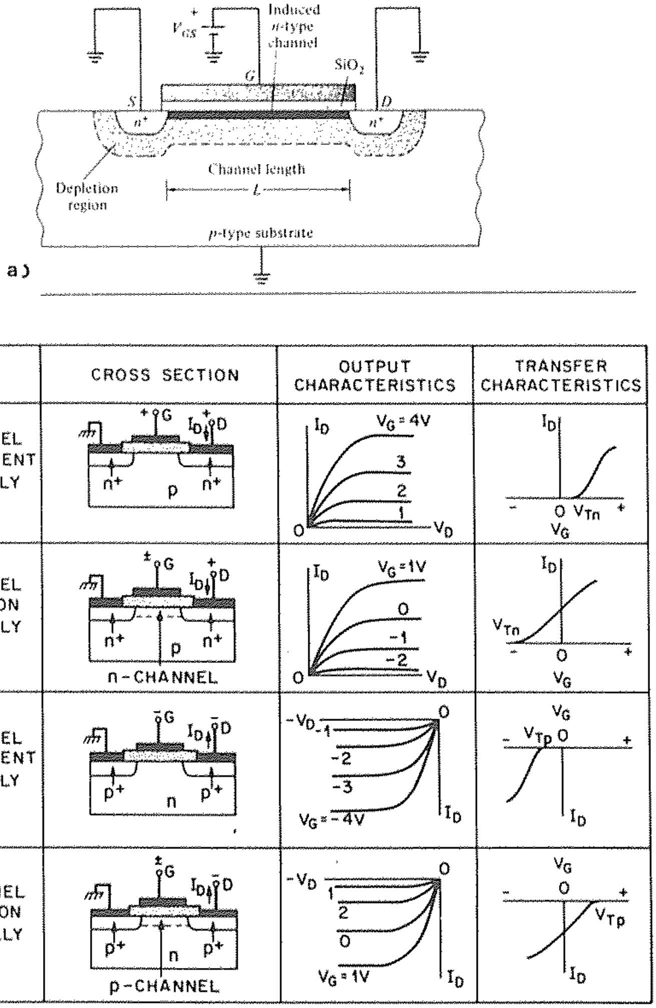
As shown in Fig. 5-1b, MOS transistors are symmetrical devices, which means that the source and drain are interchangable. In NMOS circuits, however, *the more positive of these two electrodes is normally defined to be the drain*, and thus the input signal is defined as the voltage between the gate and source terminals ($V_{in} = V_G$).

In simplest terms, the operation of an MOS transistor involves the application of an input voltage to the gate electrode, which sets up a transverse electric field in the channel region of the device (5-2a). By varying this transverse electric field, it is possible to modulate the longitudinal conductance of the channel region. Since an electric field controls current flow, such devices are termed *field-effect transistors (FETs).* They are further described as *metal-oxide-semiconductor (MOS) FETs* because of the thin SiO_2 layer that separates the gate and substrate.

The substrate (or body) of the MOS transistor is a silicon wafer; this wafer also provides mechanical support for the finished circuit. In addition, an external electrical connection (or *terminal)* can be made to the body, making the MOS transistor a four-terminal device. (In later sections we will see how the transistor behavior is impacted if a bias is applied between the source and body terminals of the device.)

The top surface of the body consists of *active* or *transistor* regions as well as *passive* or *(field)* regions. The active regions are those in which transistor action occurs; i.e., the channel and the heavily doped source and drain regions. Conduction between separate active regions must be prevented (see chap. 2). A thick oxide layer (0.5-1.0 μm) is normally grown over the field regions as one of the measures to achieve this goal.

If no gate bias is applied, the electrical path between source and drain consists of two back-to-back two *pn* junctions in series. If a drain bias is applied such that the source-



b)

Fig. 5-2 (a) Idealized NMOS cross section with positive V_{GS} applied showing depletion regions and the induced channel.[7] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission. (b) Cross sections and output and transfer characteristics of four types of MOSFETs. [7] From S. M. Sze, *Semiconductor Devices - Physics and Technology*, Copyright, 1985, John Wiley & Sons. Reprinted with permission.

body and drain-body junctions remain reverse-biased, I_D will consist of only the reverse-bias diode leakage current and hence will be considered negligibly small.

When positive bias is applied to an NMOS transistor gate, electrons will be attracted to the channel region and holes will be repelled. (Holes are the majority carriers in the channel of the *p*-type body when no gate bias exists.) Once enough electrons have been drawn into the channel by the positive gate voltage to exceed the hole concentration, the region behaves like a *n*-type semiconductor. Under these circumstances, an *n*-type channel connects the source and drain regions (Fig. 5-2a). Current will flow if a voltage, V_{DS}, is applied between the source and the drain terminals. The voltage-induced *n*-type channel does not form unless the voltage applied to the gate exceeds the *threshold voltage,* V_T.

MOS devices such as those just described, in which no conducting channel exists when $V_G = 0$, are referred to as *enhancement-mode* (or *normally OFF*) transistors (see Fig. 5-2b). With NMOS enhancement-mode transistors, a positive gate voltage, V_G, greater than V_T must be applied to create the channel (or to turn them *ON*), while to turn *ON* PMOS enhancement-mode devices, a negative gate voltage (whose magnitude is $>V_T$) must be applied. Note that in NMOS transistors a positive voltage must also be applied to the drain to keep the drain-substrate reverse-biased, while in PMOS devices this voltage must be negative.

On the other hand, it is also possible to build MOS devices in which a conducting channel region exists when $V_G = 0$ V (see Fig. 5-2b), and such MOS devices are described as being *normally ON*. Since a bias voltage to the gate electrode is needed to deplete the channel region of majority carriers (that is, the channel is eradicated as long as the bias is applied), such devices are also commonly called *depletion-mode* devices. NMOS depletion-mode devices require a negative gate voltage to be turned *OFF*, while corresponding PMOS devices require positive gate voltages.

5.1.2 The Threshold Voltage of the MOS Transistor

If the source and body of an MOS transistor are both tied to ground ($V_{SB} = 0$), the threshold voltage, V_T, of the transistor can be found from the following equation (note that $V_T = V_{T0}$ when $V_{SB} = 0$):

$$V_{T0} = \varphi_{ms} - 2\varphi_f - Q_{tot}/C_{ox} - Q_{BO}/C_{ox} \qquad (5-1)$$

where φ_{ms} is the workfunction difference (in V) between the gate material and the bulk silicon in the channel, φ_f is the equilibrium electrostatic potential in a semiconductor (in V), Q_{BO} is the charge stored per unit area (C/cm^2) in the depletion region (when the voltage between source and body is zero), C_{ox} is the gate-oxide capacitance per unit area (F/cm^2), and Q_{tot} is the total positive oxide charge per unit area present at the interface between the oxide and the bulk silicon (see section 5.3.3).

Expressions have been established for these various quantities in terms of doping concentrations in the material, physical constants, device structure dimensions, and temperature. They are:

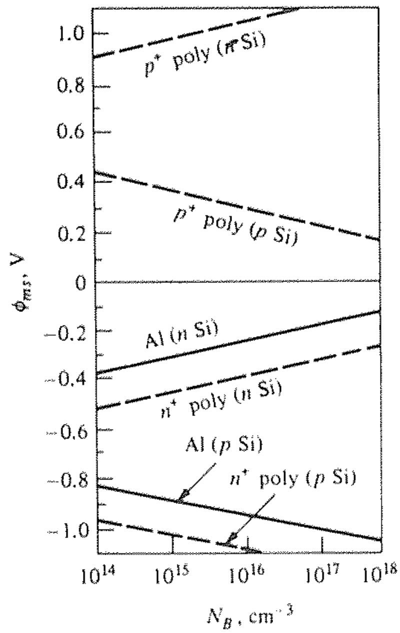


Fig. 5-3 Work-function difference ϕ_{ms} versus doping for degenerate polysilicon and Al electrodes.[80] Reprinted with permission of Solid-State Electronics.

$$\varphi_f = kT/q \ \ln (n_i / N_A) \quad (p\text{-type semiconductor}) \qquad (5 - 2a)$$

$$\varphi_f = kT/q \ \ln (N_D / n_i) \quad (n\text{-type semiconductor}) \qquad (5 - 2b)$$

where N_A is the acceptor concentration in a p-type semiconductor (cm^{-3}), N_D is the donor concentration in an n-type semiconductor, and n_i is the intrinsic carrier concentration in the semiconductor. Note that n_i is a strong function of temperature. For silicon at 300 K, however, $n_i = 1.45 \times 10^{10}$ cm^{-3}. Therefore, at 300 K:

$$\varphi_{ms} \text{ (metal gate, with Al as gate electrode)} = \varphi_{f(sub)} - \varphi_{f(Al)} = \varphi_{f(sub)} - (+0.6 \text{ V}) \qquad (5 - 3a)$$

$$\varphi_{ms} \text{ (Si gate)} = \varphi_{f(sub)} - \varphi_{f(gate)} \qquad (5 - 3b)$$

Figure 5-3 shows the value of φ_{ms} for various substrate-doping values if an aluminum gate electrode is used. Note also that in silicon-gate NMOS, the doping of the polysilicon gate material is usually n-type, with $N_D \cong 10^{20}/cm^3$, so that $\varphi_{f(gate)}$ in this case is +0.59 V. Figure 5-4 shows the value of the threshold voltage of n^+ silicon gate MOS transistors (both PMOS and NMOS) versus substrate doping concentration, assuming either a 65-nm, a 25-nm, or a 15-nm thick gate oxide, and $Q_{tot} = 0$.

Q_{BO} is found from:

$$Q_{BO} = -\sqrt{2\, q\, N_A\, \varepsilon_{si}\, |-2\varphi_f|} \qquad \text{NMOS} \qquad (5-4a)$$

$$Q_{BO} = +\sqrt{2\, q\, N_D\, \varepsilon_{si}\, |-2\varphi_f|} \qquad \text{PMOS} \qquad (5-4b)$$

and C_{ox} is calculated from:

$$C_{ox} = \varepsilon_{ox} / t_{ox} = 3.9\, \varepsilon_o / t_{ox} \qquad (5-5)$$

where $\varepsilon_{si} = 11.8\, \varepsilon_o$, ε_o is the permittivity of vacuum, and ε_{ox} and t_{ox} are the permittivity and thickness of the gate oxide.

It is easy to become confused about the signs of the various terms in the threshold voltage equation. Equation 5-1 gives correct results for NMOS and PMOS if the signs shown in Table 5.1 are used for each of the parameters. When calculating the various terms, the table can be used to insure that the value of each parameter is entered into the equation with the correct sign (e.g., φ_{ms} will have a negative value for n^+ Si gate NMOS devices).

Table 5.1 Signs in the Threshold-Voltage Equation[7]

Parameter	NMOS	PMOS
Substrate	*p*-type	*n*-type
φ_{ms}		
Metal gate	-	-
n^+ Si gate	-	-
p^+ Si gate	+	+
φ_f	-	+
Q_{BO}	-	-
Q_{tot}	+	+
γ	+	+
C_{ox}	+	+
Source-body voltage, V_{SB}	+	+

EXAMPLE 5-1: Find the threshold voltage of an NMOS silicon-gate transistor that has substrate doping $N_A = 10^{15}/\text{cm}^2$, gate doping $N_D = 10^{20}/\text{cm}^3$, gate-oxide thickness $t_{ox} = 100$ nm, and $Q_{tot} = q\,(1 \times 10^{11}/\text{cm}^2)$. Note that the values for t_{ox} and Q_{tot} are typical of the values that existed in early NMOS transistors.

SOLUTION:

$$\varphi_{f(sub)} = kT/q \ln [n_i / N_A] = -0.026 \ln [10^{15} / 1.4 \times 10^{10}] = -0.29 \text{ V}$$

$$\varphi_{ms} = \varphi_{f(sub)} - \varphi_{f(gate)} = -0.29 - kT/q \ln [10^{20} / 1.4 \times 10^{10}] = -0.88 \text{ V}$$

$$\varepsilon_{ox} = 3.9\varepsilon_o = 3.5 \times 10^{-13} \text{ F/cm}; \quad C_{ox} = \varepsilon_{ox} / 1 \times 10^{-5} = 35 \times 10^{-9} \text{ F/cm}^2$$

$$Q_{BO} = -[2 \times 1.6 \times 10^{-19} \times 10^{15} \times 1.04 \times 10^{-12} \times |-0.58|]^{1/2} = -1.4 \times 10^{-8} \text{ C/cm}^2$$

$$Q_{BO} / C_{ox} = -1.4 \times 10^{-8} / 35 \times 10^{-9} = -0.4 \text{ V}$$

$$Q_{tot}/C_{ox} = 1.6 \times 10^{-19} \times 1 \times 10^{11} / 35 \times 10^{-9} = 0.46 \text{ V}$$

$$V_{T0} = -0.88 - (-0.58) - (-0.4) - (0.46) = -0.36 \text{ V}$$

Note that a negative value of V_{T0} is yielded, implying that this transistor would be *ON* at $V_G = 0$ V (would therefore behave as a *depletion-mode* device). The parameters used in this calculation of V_T are typical of the NMOS device parameters encountered in the early days of MOS ICs. Hence, at that time it was not possible to easily or reliably fabricate enhancement-mode NMOS transistors. We shall later show how this problem was overcome so that enhancement-mode NMOS devices could be manufactured.

5.1.3 Impact of Source-Body Bias on V_T (Body Effect)

As mentioned earlier, the MOS transistor is a four-terminal device, insofar as a contact can also be made to the body region. A bias, V_{SB}, can also be applied between the source and body (e.g., with the source being tied to ground, as shown in Fig. 5-5) and such a bias will have an impact on V_T. If $V_{SB} = 0$, inversion will, of course, occur when the voltage drop across the semiconductor equals $2\varphi_f$. If $V_{BS} < 0$ V, the semiconductor still attempts to invert when the voltage drop across it reaches $2\varphi_f$. However, any inversion-layer carriers that do appear at the semiconductor surface migrate laterally into the source because this region is at a lower potential. Thus, the surface potential must be lowered to $2\varphi_f - V_{SB}$ in order for inversion to occur, implying that the threshold voltage required to achieve inversion must be increased.

Hence, for either enhancement-mode or depletion-mode MOSFETs, V_T becomes *more positive* for *n*-channel transistors and *more negative* for *p*-channel transistors as V_{SB} is increased. A simple, quantitative way to view this effect is to assume that biasing changes the inversion point in the semiconductor from $2\varphi_f$ to $2\varphi_f - V_{SB}$. The threshold-voltage equation (Eq. 5-1), which was applicable when $V_{SB} = 0$, is in turn modified to

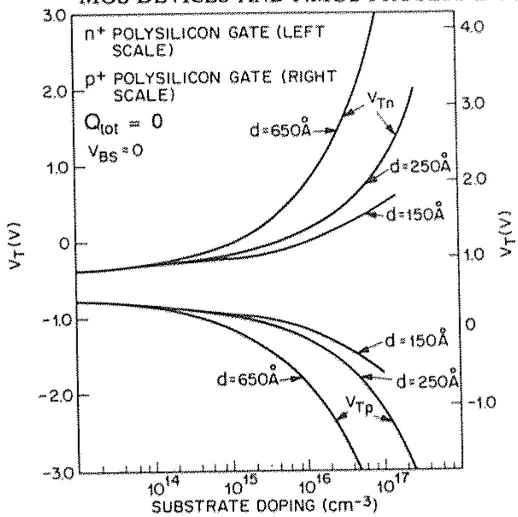


Fig. 5-4 Calculated threshold voltages of n-channel (V_{Tn}) and p-channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon (right scale). Curves for gate oxide thicknesses d of 150 Å, 250 Å, and 650 Å are shown. From S. M. Sze, Ed., *VLSI Technology*, 2nd. Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Laboratories. Reprinted with permission of McGraw-Hill.

$$V_T = V_{T0} + \gamma\left(\sqrt{|-2\varphi_f + V_{SB}|} - \sqrt{2|\varphi_f|}\right) \qquad (5-6)$$

where the parameter γ (referred to as the *body-effect coefficient* or *body factor*, with units of $V^{1/2}$) is given by

$$\gamma = \frac{\sqrt{2 q \varepsilon_{si} N_A}}{C_{ox}} \qquad \text{NMOS} \qquad (5-7a)$$

and

$$\gamma = \frac{\sqrt{2 q \varepsilon_{si} N_D}}{C_{ox}} \qquad \text{PMOS} \qquad (5-7b)$$

5.1.4 Current-Voltage Characteristics of MOS Transistors

We will now present the equations that describe the large-signal current-voltage (I-V) characteristics of *long-channel* MOS transistors, assuming an NMOS device with its source grounded and with bias voltages V_{GS}, V_{DS}, and V_{BS} applied as shown in Fig.

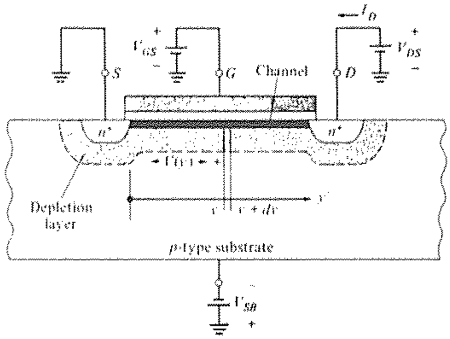


Fig. 5-5 NMOS device with bias voltages V_{GS}, V_{DS}, and V_{BS} applied.[7] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

5-5. (Any modifications that occur in the I-V characteristics due to short-channel effects will be described section 5.5.)

In the simplest MOS model, if V_G is smaller than V_T, no channel exists, and no current is assumed to flow between the source and drain. If V_G is greater than V_T, a conducting channel is present, and V_{DS} causes a drift current (I_D) to flow from drain to source. For small values of V_{DS}, the drain current I_D is linearly related to V_{DS}. In this so-called *linear region* of operation, the equation that describes I_D is

$$I_D = k/2\,[2(V_G - V_T)\,V_{DS} - V_{DS}^2] \qquad V_G \geq V_T,\ V_{DS} \leq (V_G - V_T) \qquad (5-8)$$

where k (the so-called *device transconductance parameter*) is defined as

$$k = \mu_n\, C_{ox}\,(W/L) \qquad (5-9)$$

and μ_n is the surface mobility of electrons in the channel.

As the value of V_{DS} is increased, the induced conducting-channel charge decreases near the drain. When V_{DS} equals or exceeds $V_G - V_T$, the channel is said to be pinched off. Increases above this critical voltage produce little change in I_D, and Eq. 5-8 no longer applies. The value of I_D in this region is then given by the following

$$I_D = k/2\,(V_G - V_T)^2 \qquad V_G \geq V_T;\quad V_{DS} > (V_G - V_T) \qquad (5-10)$$

This is the so-called *saturation region* of operation.

A plot of I_D versus V_{DS} (with V_G as a parameter) for a long-channel NMOS transistor as described by Eqs. 5-8 and 5-10, is shown in Figure 5-6. If the value of V_G is smaller than V_T, the device is said to be in *cutoff*. In the model given here, I_D is assumed to be zero in cutoff (i.e., no *subthreshold current*, I_{Dst}, flows). We shall see in a later section that this assumption is not completely correct (even in long-channel devices), and that in short-channel devices such currents can cause severe problems.

5.1.5 The Capacitances of MOS Transistors

It can be shown that the switching speed of MOS digital circuits is not limited by the channel transit time (i.e., the time required for a charge to be transported charge across the channel), but by the time required to charge and discharge the capacitances that exist between device electrodes and between the interconnecting lines and the substrate. Figure 5-7 shows the significant capacitances between the electrodes of an MOS transistor. The capacitance from gate to other electrodes [C_G] is, to a first approximation, the sum of C_{GB}, C_{GS}, and G_{GD}. Furthermore, since C_{GS} and C_{GD} are small in silicon-gate technology (as the gate and source/drain regions are self-aligned), we can treat C_G as a constant, essentially determined by

$$C_G = W\,L\,C_{ox} = W\,L\,\varepsilon_{ox} / t_{ox} \qquad (5-11)$$

The capacitance per unit area of the source/body and drain/body junctions (C_{SB} and C_{DB}, respectively) are calculated using the parallel-plate capacitance formula with a plate spacing of W.[1] The larger the doping of the substrate, the larger the value of these capacitances. As an example, in the case of zero bias and a doping concentration of $10^{16}/cm^3$, the junction capacitance is approximately 10 nF/cm^2.

5.2 MAXIMIZING DEVICE PERFORMANCE THROUGH DEVICE DESIGN AND PROCESSING TECHNOLOGY

Having identified the parameters that determine V_T and the I-V characteristics of MOSFETs, we next discuss how to link device design and fabrication procedures in

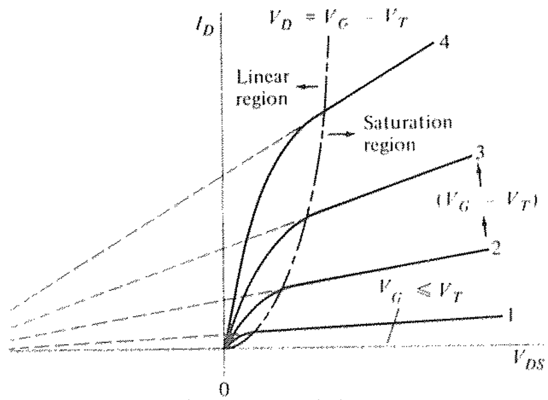


Fig. 5-6 NMOS device with I_D - V_{DS} characteristics.

order to achieve optimum device and circuit performance.

The desired device properties of MOSFETs include the following: high-output current drive (large value of I_D), high and stable transconductance (g_m), predictable and stable threshold voltage (V_T), fast switching speed, very small subthreshold current (I_{Dst}), high gate-oxide breakdown voltage, high drain/body breakdown voltage, low source/drain-to-body capacitances, high field-region threshold voltage and punchthrough voltage values, and high reliability of device operation. Note that the models used to select the process conditions that provide such optimum device behaviors are Eqs. 5-1, 5-8, and 5-10. In practice, these equations accurately describe the dc circuit behavior of *long-channel MOS devices.* (In section 5.5 we'll see how this behavior is modified in short-channel devices, as well as what device design and processing constraints result.)

5.2.1 Output Current (I_D) and Transconductance (g_m)

Equations 5-8 and 5-10 predict how I_D in MOSFETs can be impacted by various device parameters. Since increasing the *gate width* linearly increases I_D, one option when large drive-currents are needed, is to increase the dimension of the gate width. However, when minimum-sized devices must be used (e.g., for maximum packing density), this option cannot be implemented, and the other parameters that can influence I_D in Eqs. 5-8 and 5-10 must be considered.

From the dependence of I_D on $k' = \mu\, C_{ox}$, both the mobility of the carriers in the channel and the gate oxide capacitance should be as high as possible. Since electron mobility is greater than hole mobility, circuits using NMOS devices will exhibit higher performance than those built with PMOS devices. In fact, *NMOS transistors of the same width as PMOS transistors will indeed provide roughly 2.5x the current drive.* In addition, the mobility of carriers decreases as the doping concentration of the channel increases. Hence, lightly doped channel regions are also favored.

Because the value of C_{ox} is inversely proportional to t_{ox}, as thin a gate oxide as possible (commensurate with oxide breakdown and reliability considerations) is normally used. I_D is also inversely proportional to channel length, and minimum channel lengths are therefore desirable. On the other hand, if the channel lengths become too short, they adversely impact device operation in other ways, as described in section 5.5.

Although the equations indicate that (V_G - V_T) should be as large as possible, V_G is usually fixed by system specifications and material limitations, and cannot be changed by the device or circuit designer. Similarly, V_T is selected primarily by other circuit considerations (such as adequate noise margin in digital circuits).

The transconductance in saturation, $g_{msat} = dI_D/dV_G$, can be expressed simply by way of the following equation

$$g_{msat} = 2\,\mu_n\, C_{ox}\, (W/L)\, [V_G - V_T] \qquad (5\text{-}12)$$

Hence, we can maximize g_{msat} by varying the process and device parameters in the same manner as discussed for I_D.

5.2.2 Controlling the Threshold Voltage through Process and Circuit Design Techniques

In many MOS IC applications, it is critical to be able to establish and maintain a uniform and stable value of V_T (i.e., the value of V_T should not vary with time or with device-operating conditions). An example of the importance of being able to control this parameter involves semiconductor memory devices. In these circuits, charge flows from the memory cells to the sense amplifier. The sense amplifier is a delicately balanced flip-flop whose voltage-sensing capability is directly related to the threshold-voltage variation between the transistors (see chap. 8). Hence, such circuits would not function reliably without the presence of a highly uniform and stable threshold voltage in the circuit devices.

The factors that impact threshold voltage (when $V_{SB} = 0$) are given in Eq. 5-1. Examining each term of this equation will enable us to determine which device parameters can be adjusted to provide practical control of V_T.

The φ_{ms} term depends on the work-function difference between the gate, $\varphi_{f(gate)}$, and the semiconductor, $\varphi_{f(sub)}$. For metal and heavily doped silicon gates, $\varphi_{f(gate)}$ is constant while the parameter $\varphi_{f(sub)}$ depends on the substrate doping, but only in a logarithmic manner. Hence, each factor-of-10 increase in substrate doping will change the φ_{ms} term by only 2.3 kT, or ~0.06 V (kT/q = 0.026 V at 300 K). Thus, changes in the substrate-doping concentration produce changes in V_T through the φ_{ms} term which are too small to provide the required degree of threshold-voltage control. The next term ($2\varphi_f$) also changes only slightly as a result of changes in the substrate doping concentration (for the same reason given for the φ_{ms} term); thus, the $2\varphi_f$ term is also not of much use in controlling V_T.

Since every attempt is made to keep Q_{tot} as low as possible through various processing procedures, the Q_{tot}/C_{ox} term is very small in modern MOS devices. Hence, it must also be ruled out as a candidate for controlling V_T.

While is true that C_{ox} can be varied (primarily by changing the gate oxide thickness), this parameter is not a practical vehicle for controlling V_T in active devices, since the gate oxide is normally made as thin as possible to achieve maximum I_D. In the field regions, however, large V_T values are needed to prevent inversion under the field oxide. A thick field oxide makes C_{ox} small, allowing V_T to be increased Thus, the C_{ox} _is_ one of the parameters normally used to control V_T in the field regions of the circuits.

This leaves the Q_{BO} term as the remaining candidate for controlling V_T in active devices. Equation 5-4 indicates that the doping concentration can indeed provide a large change in V_T. The signs of Eq. 5-1 indicate that increasing the substrate doping (i.e., N_A in *n*-channel devices, and N_D in *p*-channel devices) will increase V_T. Thus, to increase V_T in Example 5-1 in this manner, it would be necessary to increase the substrate doping concentration above $N_A = 10^{15}$ cm^{-3}. Figure 5-7 shows how V_T can be controlled by changing the substrate doping concentration for various gate-oxide thicknesses (assuming that Q_{tot} is kept small enough that it is not a significant contribution to V_T).

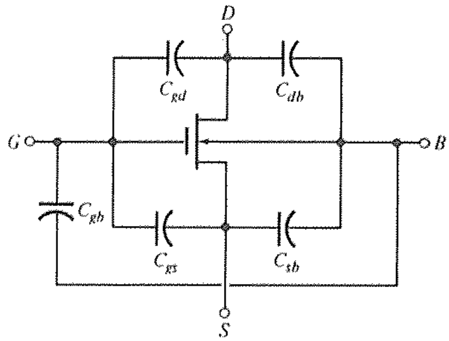


Fig. 5-7 MOS transistor capacitances. $C_G = C_{GS} + C_{GB} + C_{GD}$.[7] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

Significant increases in substrate-doping concentration give rise to lower junction-breakdown voltages, larger junction capacitances, and lower carrier mobilities, making such substrate doping concentration increases undesirable. Yet, prior to the development of ion implantation in the early 1970s, adjustment of substrate doping was the only practical *processing approach* for significantly controlling V_T in active devices.

A *circuit approach* known based on applying a bias between the source and body (and hence known as *body biasing)* can also control V_T. When body biasing is used V_T is given by Eq. 5-6. Figure 5-8 gives an example of how the application of V_{SB} can change the V_{TO} value in an NMOS device.[14] The intercept of the channel conductance and the V_G axis when $V_{SB} = 0$ corresponds to V_{TO}, which in this example is +0.99 V.

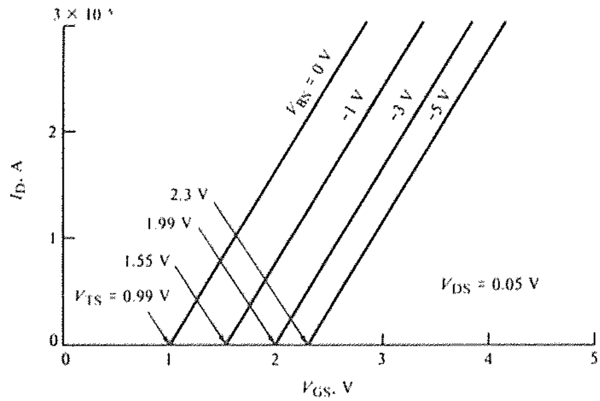


Fig. 5-8 Threshold voltage adjustment using substrate bias.

As V_{SB} is increased from 0 V to -5 V, V_T increases up to +2.3 V. However, body biasing is an added complication, and it is now avoided as a technique for adjusting V_T whenever possible in favor of the newer *ion implantation V_T-adjust method.*

We nevertheless include a discussion of body biasing because it pre-dated ion implantation as a method of controlling V_T. C. T. Sah points out that the availability of the technique allowed IBM to implement MOS memory devices instead of core memory for the first time in the IBM-370/158 mainframe computer in 1973.[60] The access time of NMOS RAMs at that time was competitive with magnetic core memory (~1 μs), whereas PMOS RAMs were slower. Body biasing allowed the higher-performance NMOS devices to function as enhancement-mode devices, even though their threshold voltages without body biasing would have caused them to behave like depletion-mode devices.

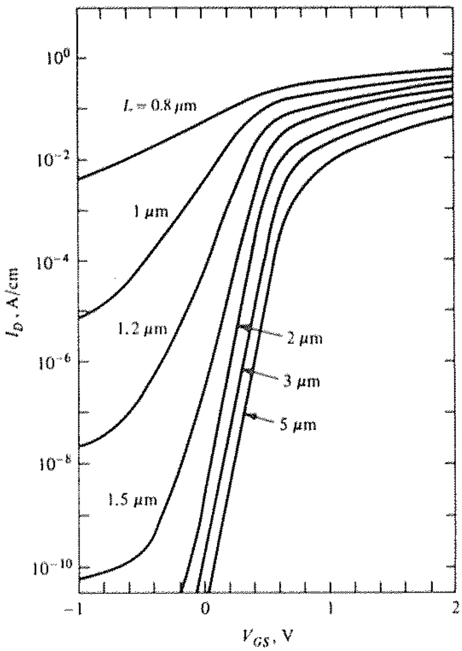
Another important reason for describing body biasing is that the body effect impacts devices in MOS IC logic circuits even when no deliberate attempt is made to apply an external bias. For example, when a logic gate containing a transistor as the load device has a logic 1 output (see for example Fig. 5-13), the voltage at the source of this device is different from that of the body. Hence, the device is subject to a non-zero V_{SB}, and V_T no longer equals V_{TO} value. The degree to which the change occurs depends on the *body factor*, γ. Since the smallest change in V_T is generally desired, small values of γ are preferred. Equation 5-7 indicates that γ depends on substrate doping, and that lower doping concentrations provide smaller values of γ, providing further impetus for lightly doped substrates in MOS ICs.

The value of γ also decreases with channel length below 1 μm. For example, γ in NMOS devices decreases by about 50% as L_{eff} is decreased from 1 μm to 0.6 μm, while for PMOS the decrease is about 30%.[8]

5.2.3 Subthreshold Currents (I_{Dst} when $V_G < |V_T|$)

The basic MOS device model neglects all free charges in the channel until the magnitude of the gate voltage exceeds V_T. This is a valid approximation for most subthreshold bias conditions because the free-charge densities in the channel change exponentially with the channel voltage. When this approximation is used, it implies that *no current flows between the drain and source* if $V_G < |V_T|$. As V_G approaches the value of V_T, (corresponding to a condition of *weak inversion)*, however, the magnitude of I_D is not well defined by this simple approach. In fact, it is observed that $I_D \neq 0$ if V_G is close to (but still less than) the value of V_T. The current which is observed in such cases is therefore referred to as *subthreshold current*, I_{Dst}.

The values of I_{Dst} can be well predicted in long-channel MOSFETs by modifying the basic MOS model to take into account the fact that the minority carrier concentration at the Si surface is greater than the value at equilibrium, but still less than the bulk doping concentration.[19] The results of this modified model indicate that the magnitude of I_{Dst} is essentially independent of V_{DS} but is exponentially proportional to V_G.



Log current vs. gate bias characteristics for various channel lengths, L (from full two-dimensional computer calculation), $N_a = 10^{15}\ cm^{-3}$, $x_o = 0.5\ \mu m$, $r_j = 0.33\ \mu m$, $V_{DS} = 2.0$ V, $V_{BS} = 0$ V.

Fig. 5-9 Example of subthreshold MOSFET I_D - V_{GS} curves for various channel lengths.[19] (Reprinted with permission of Solid State Electronics).

Since I_{Dst} can be accurately modeled in long-channel MOSFETs (Fig. 5-9), circuit designers can readily calculate the gate bias required to ensure a given allowable subthreshold leakage current. *Typically, to ensure that I_{Dst} will be negligibly small, the bias applied to the gate should be 0.5 V below V_T.*

Since I_{Dst} is exponentially proportional to V_G, if log I_{Dst} is plotted versus V_G the result on a semilogarithmic plot will be a straight line for values of V_G below V_T. The slope of the I_{Dst} versus V_G line, when plotted in this manner, is characterized by the *subthreshold swing*, S.S., where

$$\text{S.S.} = \Delta V_G / \Delta \log I_{Dst} \qquad (5-13)$$

Hence, S.S. is the change in V_G that produces a decade increase in I_{Dst}. A small value of S.S. is desirable, since it indicates maximum control of the gate over the channel current. Typical values of subthreshold swing in long-channel MOSFETs are around 90 mV/decade. Such values can be achieved by building devices that have a low value of substrate doping concentration (which gives a larger depletion width) and a thin gate oxide. (A more detailed treatment of the derivation of the I_{Dst} dependence on V_G can be found in references 1 and 3.)

Note that when the channel length gets small, the values of I_{Dst} are larger than than those predicted by the modified long-channel MOSFET model. This is due to so-called *short-channel effects* (discussed in section 5.5). However, measurements of the subthreshold swing can be used to detect the onset of these short-channel effects (*punchthrough* and *drain-induced barrier lowering*). Since the measurable I_{Dst} is the sum of both the normal subthreshold and the short-channel subthreshold current components, an increase in the value of S.S. will signal the onset of these effects.

5.2.4 Switching Speed

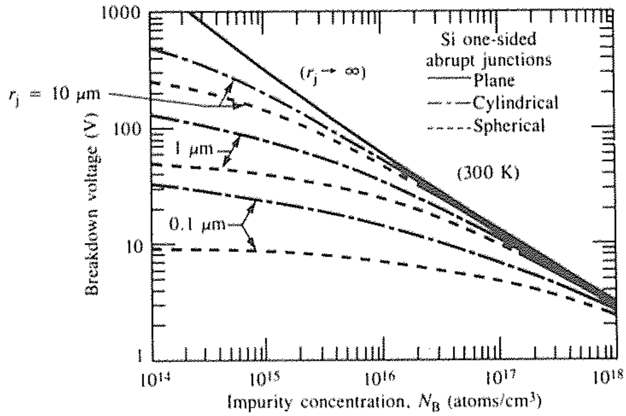
The switching speed of the logic gates in an MOS IC is limited only by the time required to charge and discharge the capacitances between device electrodes and between the interconnecting lines and ground (or other lines). At the circuit level the propagation delay is frequently limited by the interconnection-line capacitances. At the device level, however, the gate delay is determined primarily by the channel transconductance, the MOS gate capacitance (C_G) and the other two MOS parasitic capacitances, C_{DB} and C_{SB} (as defined in the previous section). If these capacitance values can be reduced, the device switching speed will be increased.

The gate capacitance is decreased by decreasing the gate area (although decreasing the gate oxide thickness increases its value). The dominant parasitic capacitance on the device level, however, is that due to C_{SB} and C_{DB} (i.e., junction capacitances). An analytical study has shown that these junction capacitances account for up to 50% of the total capacitance in logic gates.[9] Therefore, reductions in these capacitances should produce corresponding decreases in the gate delay. In general, *the lower the doping concentration in the body, the lower the junction capacitances.*

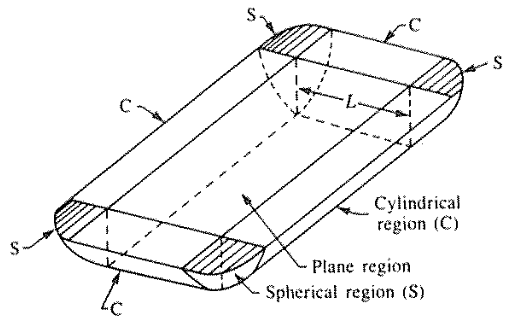
5.2.5 Junction Breakdown Voltage (Drain-to-Substrate)

The source and drain regions are very heavily doped to minimize their resistivities. Thus, the breakdown voltage of the drain-to-substrate junction will be determined by the lighter doping concentration of the body. As seen in Fig. 5-10a (which shows breakdown voltage of a one-sided *pn* junction as a function of the lighter doping concentration), the breakdown voltage decreases as the doping increases. Thus, *lightly doped substrates also yield high junction breakdown voltages.*

Junction curvature enhances the electric field in the curved part of the depletion region (Fig. 5-10b), and this effect reduces the breakdown voltage below that predicted by one-dimensional junction theory. A rectangular source or drain region (formed either by diffusion or ion implantation) has regions with both cylindrical and spherical curvature.[10,11] Figure 5-10a also shows the effect of junction curvature on the breakdown voltage of a one-sided step junction in silicon. It can be seen that as the junction depths get shallower (making the radius of curvature, r_j, of the spherical and cylindrical structures smaller), the breakdown voltage is significantly reduced, especially for low substrate impurity concentrations.



(a)



(b)

Fig. 5-10 (a) Abrupt *pn* junction breakdown voltage versus impurity concentration on the lightly doped side of the junction for both cylindrical and spherical structures. r_j is the radius of curvature.[10] (© 1957 IEEE). (b) Formation of cylindrical and spherical regions by diffusion through a rectangular window. From S. M. Sze, *Semiconductor Devices - Physics and Technology*, Copyright, 1985, John Wiley & Sons. Reprinted with permission.

5.2.6 Gate-Oxide Breakdown Voltage

High-quality SiO_2 films will typically break down at electric fields of 5-10 MV/cm (the exact value is a function of oxidation and anneal conditions, oxide charges, surface crystallographic orientation, surface preparation and a number of other factors). This corresponds to 50-100 V across a 100-nm-thick oxide. Present day 5 V processes use gate-oxide thicknesses of 15-100 nm. Below around 5 nm (at less than 3 V), there is a finite probability that electrons will pass through the gate by means of a quantum-mechanical tunneling effect. For proper device operation, the tunneling current must be small. This effect therefore sets a fundamental lower limit of about 5 nm for the thickness of the gate oxide. A search for alternative gate dielectric materials to mitigate this limitation is being conducted; this issue will be discussed in the CMOS chapter.

Oxide breakdown may also occur at electric-field values smaller than those given above, as a result of process-induced flaws in the gate oxide. Such defects include: metal precipitates on the silicon surface prior to oxide growth (see Vol. 1, chap. 2); high defect density in the silicon lattice at the substrate surface (e.g., stacking faults and dislocations, see Vol. 1, chap. 2); pinholes and weak spots created in the gate oxide by particulates; thinning of the oxide during growth caused by the Kooi effect (see chap. 2); and oxide wearout due to failure mechanisms related to hot-electron injection (this topic will be covered in Vol. 3).

5.2.7 High Field-Region Threshold-Voltage Value

A high value of threshold voltage in the field region is needed to keep the parasitic field channels between adjacent active devices from being turned on. This topic is described in great detail in chapter 2.

5.3 THE EVOLUTION OF MOS TECHNOLOGY (PMOS AND NMOS)

Following the introduction of MOS integrated circuits in the 1960s, MOS technology evolved through several stages. At first, PMOS was the dominant technology, but it was supplanted by NMOS in the early 1970s. NMOS in turn was largely replaced by CMOS in the mid-1980s. (The evolution of CMOS will be described in chap. 6.) In addition, the drive to shrink feature dimensions led to smaller device sizes in all IC technologies. In MOS technology, the main dimension that summarized the shrinkage in each new generation was the *minimum gate length* of the transistors. NMOS overtook PMOS roughly when gate lengths reached ~6 μm, while CMOS became the dominant technology at gate lengths of ~1.5-2 μm.

Although the *physical gate length*, L, of the MOS device is commonly used to identify each generation of technology, this practice can be misleading in that it may convey the impression that this dimension is also the minimum dimension of all other device features fabricated in the technology. In fact, the physical gate length does not necessarily reflect the dimensions of the other design rules (many, if not most, of which are larger than the physical gate length dimension).

In addition, the physical gate length does not accurately represent the electrical or *effective* channel length (L_{eff}), which in fact is given by

$$L_{eff} = L - 2x_{jl} \qquad (5-14)$$

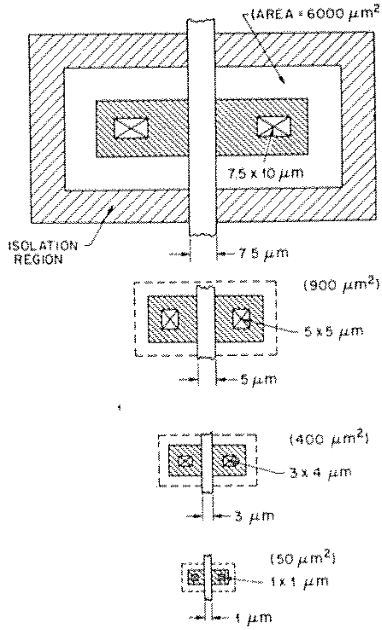


Fig. 5-11 Reduction in the area of the MOSFET as the gate length (minimum feature length) is reduced.[12] Reprinted with permission of Solid State Technology.

where x_{jl} is the lateral distance that the source or drain junction extends under the gate. For example, in a 1.2 μm MOS technology, the source and drain junctions both extend ~0.125 μm under the gate region, so L_{eff} in this technology is actually 0.95 μm. Furthermore, in CMOS technologies, the minimum physical length of PMOS devices is generally longer than that of NMOS devices, for reasons that will be discussed later. Figure 5-11 shows the reduction in MOS area as the gate length is reduced.[12] A longer and more detailed history of the evolution of the MOS transistor has been published by C. T. Sah, who describes both the evolution of the field-effect devices themselves and the technology that has been developed to fabricate these devices.[60]

5.3.1 Aluminum-Gate PMOS

The first MOS ICs built in the mid-1960s were implemented with *p*-channel enhancement-mode devices whose threshold voltages were approximately -4 V. (An example of an early PMOS IC was the Intel 256-bit SRAM, introduced in the late-60s with devices having gate lengths of 12 μm.) Such early MOS integrated circuits were built on silicon wafers of <111> orientation and used Al as the gate electrode material. These choices grew out of the experience gained in manufacturing bipolar ICs prior to the development of MOS ICs. When MOS IC technology was first being implemented, semiconductor manufacturers transferred their bipolar fabrication know-how to the newer technology.

Since precise control of dopant diffusion in <111> Si was a mature bipolar process, <111> wafers were also the logical choice for building MOS ICs. Similarly, since aluminum metallization was already being implemented, it was natural to adapt Al as the MOS gate electrode material. Unknowingly, these choices worsened the problems which prevented NMOS devices from being used to produce early MOS ICs, and instead, forced the use of PMOS devices in these circuits.

The earliest MOS circuits exhibited two serious limitations, partly as a result of being implemented with *p*-channel devices. First, the PMOS devices had a V_T of -4 V, which required a power-supply voltage of -12 V for the drain supply (a value that was incompatible with the +5 V power-supply voltage used in bipolar digital [TTL] ICs). Second, the circuits were very slow (e.g., a PMOS flip-flop could operate at 500 kHz to 1 MHz, while a bipolar flip-flop could operate at 5-10 MHz). It was known that the latter problem was due to the low surface mobility of holes in the channel and that electron mobility in silicon is nearly three times as large. Therefore, NMOS circuits would have been able to provide significantly improved performance. Nevertheless, the decision to manufacture ICs with PMOS devices was dictated primarily by the existence of large (and quite variable) oxide-charge densities in the early MOS technologies.

These oxide charges were generally positive, and positive voltages on the gate tend to accumulate *n*-type surfaces (but will deplete or invert *p*-type surfaces). The oxide charges present in early MOS devices were often large enough to cause inversion in NMOS devices fabricated with reasonably thin gate oxides, even when $V_G = 0$. A thicker gate oxide could have been used to increase V_T, but this would also have degraded g_m to such a degree that circuits built with such devices would have been even slower than those built with PMOS devices. Thus, only *depletion*-mode high performance NMOS devices could be reliably manufactured. Since *enhancement*-mode devices are needed for most applications, this presented a major difficulty.

In addition, because the oxide charges were also capable of inverting *p*-doped substrate regions under field oxides, it was difficult to reliably isolate *n*-channel devices. This problem was made worse by the depletion of boron during thermal oxidation, because it reduced the boron concentration at the *p*-type surface. In summary, at the outset of the MOS era, it was not possible to reliably manufacture integrated circuits with *n*-channel enhancement-mode MOSFETs.

On the other hand, since positive oxide charges tend to accumulate an *n*-type surface, they merely increase the negative voltage that is required to turn on a *p*-channel device. Hence, the manufacture of enhancement-mode PMOS devices was possible despite the presence of the oxide charges. Nevertheless, it was clear that processing innovations were needed if the potential benefits of MOS (i.e., increased packing density, lower power consumption, TTL power-supply voltage compatibility, and process simplicity) were to be fully realized.

5.3.2 Silicon-Gate MOS Technology

One of the key process innovations for MOS ICs was the use of heavily doped polysilicon as the gate electrode in place of Al (see Vol. 1, chap. 6 for more information on polysilicon thin films). The development of this *silicon gate technology* improved the fabrication of MOS ICs in the following ways:

- Since aluminum must be deposited following completion of all high-temperature process steps (including drive-in of the source and drain regions), the gate electrode must be separately aligned to the source and drain. This alignment procedure adversely affects both packing density and parasitic overlap capacitances between the gate and source/drain regions. Since polysilicon has the same high melting point as the silicon substrate, it can be deposited prior to source and drain formation. Furthermore, the gate itself can serve as a mask during formation of the source and drain regions (by either diffusion or ion implantation). The gate thereby becomes nearly perfectly aligned over the channel, with the only overlap of the source and drain being that due to lateral diffusion of the dopant atoms. This *self-alignment* feature simplifies the fabrication sequence, increases packing density, and reduces the gate-source and gate-drain parasitic overlap capacitances.

- The threshold voltage of PMOS devices is reduced by the use of a polysilicon gate, since the φ_{ms} is less negative (see Eqs. 5-2). For PMOS devices on <111>- Si, the threshold voltage is reduced from roughly -4 V to -2 V. This smaller threshold voltage value enabled PMOS ICs to become compatible with TTL (bipolar) ICs, allowing MOS to be designed into many digital systems that operated at TTL-defined power supply voltage levels (i.e., 0 V to 5 V).

- The ability of polysilicon to withstand high temperatures also permits it to be completely encapsulated by an SiO_2 layer. This allows the polysilicon film to function as an interconnect path, in addition to serving as the gate electrode. By taking advantage of this new interconnection structure (without having to use a second layer of metal, as was necessary with bipolar ICs), it was possible to give MOS ICs an additional level of interconnection that could be crossed by the usual metal layer, or even by another polysilicon layer. This eased the problem of routing the electrical paths among the devices of an IC, thereby facilitating the layout of compact digital integrated circuits. (Techniques for establishing contact between the polysilicon layer and substrate are described in chap. 3, section 3.11.1). The ability of polysilicon to withstand high temperatures was also exploited to allow the dielectric (e.g. phosphorus-doped SiO_2) that covers it to be flowed, thereby making a significantly smoother surface topography for metallization layers.

The greatest disadvantage of polysilicon as a gate material compared to Al is its significantly higher resistivity. Even when doped at the highest practical concentrations, a 0.5-μm-thick polysilicon film has a sheet resistance of about 20 Ω/sq (compared to ~0.05 Ω/sq for a 0.5-μm-thick Al film). The resulting high values of interconnect line resistance can lead to relatively long RC time constants (i.e., long propagation delays) and severe dc voltage variations within a VLSI circuit. Consequently, the formation of refractory metal silicide layers on top of polysilicon layers (which results in so-called *polycide* films) was developed to reduce the severity of this drawback. Such polycide films can provide sheet resistances of 1 Ω/sq, at the expense of more complex processing (see Vol. 1, chap. 11 for more information on polycides). Despite of the above limitation, the development of silicon-gate technology proved to be the most important contribution to MOS technology during the reign of PMOS.

5.3.3 Reduction of Oxide-Charge Densities

Another set of important advances allowed the magnitude of the positive oxide-charge densities to be reduced. These oxide charge reduction techniques can be summarized as those involving *cleanliness, gettering, annealing,* and *replacing <111> wafers with <100> wafers.*

Cleanliness and gettering techniques reduced the densities of *mobile ionic charge,* which are due to the incorporation of ionized alkali metal atoms (Na^+, K^+) in the gate oxide. Na contamination can be controlled through clean gate oxidation processing. Gettering is used to prevent any Na^+ that enters the gate oxide from significantly degrading the V_T (see Vol. 1, chap. 7). Although tests for Na contamination of MOS devices must be routinely performed to ensure that accidental contamination does not occur, it is possible to establish fabrication procedures in which the instability of V_T due to mobile ionic charge is less than 0.1 V.

Another source of positive oxide charge is the *interface trap charge.* Again, as described in Volume 1, chapter 7, two annealing techniques were discovered that reduced this charge to acceptably low levels (i.e., to the low 10^{10} cm^{-2} eV^{-1} range). At these levels the contribution to the threshold voltage of the device is acceptably small for modern MOS devices.

The final source of positive oxide charge is the so-called *fixed oxide charge* which is located in the transition region between the silicon and the SiO_2 (see Vol. 1, chap. 7). It was learned that this charge can be reduced by the use of proper annealing techniques and that the lowest fixed oxide charge densities (~10^{10} cm^{-2}) are obtained on <100> wafers. As a result, the production of MOS ICs was shifted from <111> to <100> starting material.

In summary, the threshold voltage is negatively shifted by an amount proportional to the sum of these three oxide charge densities, Q_{tot}. When Q_{tot} was quite high this had an extremely important influence on MOS production. Through an enormous effort by the semiconductor community (industry and universities), each of the positive oxide charge density values was reduced. As a result, Q_{tot} can now be kept to less than 5 x 10^{10} charges /cm^2, and in currently used MOS devices, the oxide charge contribution to threshold voltage is minimal. As an example, the change in threshold voltage (ΔV_T) due to a Q_{tot} of q ($5x10^{10}$ cm^{-2}) in an MOS device with a 20-nm-thick oxide is:

$$\Delta V_T = Q_{tot}/C_{ox} = Q_{tot}\, x_o / 3.9\, \varepsilon_o$$

$$= (1.6x10^{-19}\ C)\ (5x10^{10}\ cm^{-2})\ (2x10^{-6}\ cm)\ /(3.5x10^{-13}\ F/cm) \approx 0.05\ V.$$

The primary reason that thermal SiO_2 is used as the gate insulator in almost all MOSFETs is that it exhibits the best interface with silicon (where "best" means that the interface has a very low concentration of interface fixed charges and traps). In fact, if a phenomenon such as the tying up of the dangling silicon bonds at the silicon surface by SiO_2 did not exist, and if an annealing process was not discovered for reducing the remaining bonds and traps to an acceptable level, MOS devices would have remained merely a laboratory curiosity.[61] Experimentation, however, is still being conducted to determine the suitability of other materials as gate insulators; this is discussed in chapter 6 which deals with CMOS.

EXAMPLE 5-2: Recalculate the threshold voltage of the NMOS transistor considered in Example 5-1 of section 5.1.2 when the oxide thickness, t_{ox}, is reduced to 15 nm, and the total oxide-charge density is reduced to $5x10^{10}$ cm^{-2}.

SOLUTION: $\varphi_{f(sub)} = -0.29$ V; $\varphi_{ms} = -0.88$ V;

$$\varepsilon_{ox} = 3.9\varepsilon_o = 3.5x10^{-13}\ F/cm; \quad C_{ox} = \varepsilon_{ox}/15x10^{-6} = 2.3x10^{-7}\ F/cm^2$$

$$Q_{BO} = -\ [2\ x\ 1.6x10^{-19}\ x\ 10^{15}\ x\ 1.04\ x10^{-12}\ x\ |-0.58\ |\ ]^{1/2} = -1.4x10^{-8}\ C/cm^2$$

$$Q_{BO}/C_{ox} = -\ 1.4x10^{-8}/2.3x10^{-7} = -0.06\ V$$

$$Q_{tot}/C_{ox} = 1.6x10^{-19}\ x\ 5x10^{10}/2.3x10^{-7} = 0.46\ V$$

$$V_{TO} = -0.88 - (-0.58) - (-0.06) - (0.03) = -0.27\ V$$

Note that the threshold voltage is still negative, and that this device would still be *ON* if the applied gate bias was $V_G = 0$ V.

EXAMPLE 5-3: Repeat the threshold-voltage calculation in Example 5-2 for an NMOS transistor whose oxide thickness (t_{ox}) is increased to 500 nm. This oxide thickness is typical of the field-oxide thickness between MOS devices on an IC. Hence, it allows us to calculate V_T for the parasitic NMOS field-region device.

SOLUTION: $\varphi_{f(sub)} = -0.29$ V; $\varphi_{ms} = -0.88$ V;

$$\varepsilon_{ox} = 3.9\varepsilon_o = 3.5x10^{-13}\ F/cm; \quad C_{ox} = \varepsilon_{ox}/5x10^{-5} = 7x10^{-9}\ F/cm^2$$

$$Q_{BO} = -[2 \times 1.6\times10^{-19} \times 10^{15} \times 1.04\times10^{-12} \times |-0.58\ |]^{1/2} = -1.4\times10^{-8}\ C/cm^2$$

$$Q_{BO}/C_{ox} = -1.4\times10^{-8}/7\times10^{-9} = -2\ V$$

$$Q_{tot}/C_{ox} = 1.6\times10^{-19} \times 5\times10^{10}/7\times10^{-9} = 1.14\ V$$

$$V_{TO} = -0.88 - (-0.58) - (-2) - (1.14) = 0.66\ V$$

This shows that an increase in the oxide thickness in the field will increase the V_T of the NMOS device in Example 5-2 so that it is now an enhancement-mode device. Unfortunately, if $V_G = 5$ V, this device would still turn on, and thus the parasitic field device would conduct.

5.3.4 Ion Implantation for Adjusting Threshold Voltage

The development of ion implantation for V_T adjustment removed the last obstacle to reliable production of *n*-channel devices for MOS ICs, because this procedure made it possible to select the substrate-doping value without having to consider its impact on V_T. Substrate doping could now be selected strictly on the basis of optimum device performance, since V_T became separately adjustable by means of ion implantation. In addition, since dopants could be selectively implanted into the field regions, high-performance NMOS circuits could also be reliably fabricated on lightly doped substrates (i.e., without the possibility of inadvertent inversion of the surrounding field regions).

This technique of adjusting V_T involves implantation of boron, phosphorus, or arsenic ions into the regions under the oxide of a MOSFET. The implantation of boron causes a positive shift in the threshold voltage, while phosphorus or arsenic implantation causes a negative shift. For shallow implants, the procedure has essentially the same effect as placing an additional "fixed" charge at the oxide-semiconductor interface. To first order, the threshold-voltage change (ΔV_T) is thereby estimated from[1]

$$\Delta V_T = q\, N_I / C_{ox} \qquad (5-15)$$

where N_I is the dose of the implanted ions (atoms/cm^2) introduced into the silicon near its surface. For example, Eq. 5-15 predicts that when $N_I = 5\times10^{11}$ ions/cm^2 and t_{ox} = 25 nm, a shift in V_T of 0.58 V will be produced. Exact modeling can be performed to calculate the actual threshold voltage shift more accurately. Figure 5-12 graphically shows the results of such modeling calculations.[16,75]

The V_T-adjust implant is usually done through the gate oxide layer. When the correct implant energy for the gate-oxide thickness being used is selected, the peak of the implant will occur at the oxide-silicon interface. After the implant-activating anneal, the implanted distribution is broader than the as-implanted profile. Calculating the effect of the implant on V_T is greatly simplified by approximating the actual distribu-

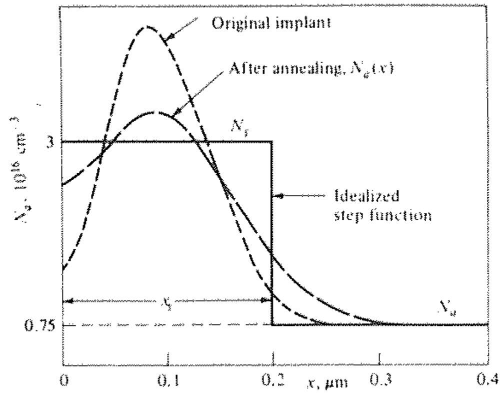


Fig. 5-12 Doping profile of implanted region beneath the gate oxide. The original implant is broadened by thermal annealing. The step doping is used to estimate the threshold voltage shift achieved using ion implantation.[75] Reprinted with permission of IBM J. of Research and Development.

tion via a "box" distribution (in which the implanted dopant is assumed to have a constant density from the surface to a depth of x_i.) Figure 5-12 shows that V_T does not change greatly as x_i is varied. Thus, the first-order approximation of the threshold-voltage change, ΔV_T (which ignored the depth of the implant as a parameter), will therefore give reasonably good estimates of ΔV_T. This calculation is often accepted in practice.

Ion implantation can also be used to fabricate depletion-mode MOSFETs. Depletion-mode NMOS devices (*i.e.* in which $V_T < 0$ V) are commonly used in NMOS logic circuits (see the following section). In order for the required negative threshold voltage for a depletion-mode NMOS device to be produced, *n*-type impurities are implanted to form a built-in channel between the source and drain. The dose required to shift the threshold voltage by the desired value may also be estimated using Eqs. 5-1 and 5-15.

EXAMPLE 5-4: If the NMOS transistor of Example 5-2 is to be used in an application that requires a $V_T = 1.0$ V, calculate the boron dose needed to adjust V_T to this value.

SOLUTION: The threshold voltage of the device in Example 5-2 is -0.27 V. We wish to have a V_T of 1.0 V. Thus V_T must be shifted (ΔV_T) by 1.27 V. Using Eq. 5-15, we see that the boron dose needed to cause this ΔV_T is

$$N_I = \Delta V_T \; C_{ox}/q \; = \; [\, 1.27 \text{ V x } 2.3\text{x}10^{-7} \text{ F/cm}^2 \,]/1.6\text{x}10^{-19} \text{ C}$$

$$= 1.8\text{x}10^{12} \text{ boron atoms /cm}^2.$$

5.3.5 Isolation Technology for MOS

Although MOS transistors are inherently self-isolating devices it is still necessary to prevent the formation of spurious channels between MOS devices (see chap. 2). This can be accomplished with the combination of a thick field oxide between the devices and a high surface concentration under the field oxide.

Prior to about 1970 the process for obtaining thick oxide regions in the field involved growing an oxide to the desired thickness on the wafer surface and then etching windows into it. This approach caused severe steps in the wafer topography, which were difficult to cover with subsequent metal layers. The introduction of LOCOS isolation in 1970 substantially overcame this problem (see chap. 2). The smoothly tapered step from the edge of the active region to the top of the field oxide in LOCOS permits overlying conductors to be easily deposited on such steps without the occurrence of significant thinning. In addition, with the development of the threshold-voltage adjustment process via ion implantation, high surface concentrations of boron could also be selectively placed under the field oxide regions. These two advances made it feasible to reliably isolate devices in NMOS circuits.

EXAMPLE 5-5: If the parasitic field NMOS transistor of Example 5-3 is implanted with the boron implant dose calculated in Example 5-4 before the field oxide is grown, determine V_T of this device (assume that no boron is lost because of segregation effects during the oxide growth).

SOLUTION: The threshold voltage of the device in Example 5-3 is 0.66V. Using Eq. 5-15, we see that a boron dose of 1.8 x 10^{12} atoms /cm^2 would cause a ΔV_T of

$$\Delta V_T = q\, N_I/C_{ox} = 1.6\text{x}10^{-19}\ \text{C x}\ \ 1.8\text{x}10^{12}\ \text{cm}^{-3}/7\text{x}10^{-9}\ \ \text{F/cm}^2$$

$$= 41\ \text{V}$$

Thus, the V_T of the parasitic field device would be (0.66 V + 41 V), or almost 42 V. This example shows that a combination of a thicker field oxide and a channel-stop implant dose can increase the threshold voltage of the parasitic field device to sufficiently large values.

5.3.6 Short-Channel Devices

As MOS channel lengths got smaller than about 3 μm, so-called short-channel effects began to become increasingly significant. As a result, device design and, consequently, process technology had to be modified to take these effects into account so that optimum device performance could continue to be obtained. Short-channel effects and their impact on processing in will be discussed in section 5.5.

5.4 PROCESS SEQUENCE FOR FABRICATING NMOS INVERTERS WITH NMOS DEPLETION-MODE LOADS

This section will describe the process sequence used to fabricate silicon-gate NMOS digital integrated circuits. A simple inverter circuit with enhancement-mode pull-down and depletion-mode pull-up NMOS transistors is used in the example. Many other logic circuits can also be implemented with this *enhancement-depletion* (E-D) NMOS process sequence.

A brief outline of how the inverter functions from a circuit point of view is given at the outset to help define some of the characteristics of MOS IC circuits. The process flow described in this section represents a relatively simple NMOS technology. In fact, many techniques have been developed to improve the performance and packing density of MOS circuits beyond those that can be produced by this process sequence. These include alternatives to LOCOS isolation structures; thin gate oxides (and alternative gate oxide materials); shallow source/drain junctions; spacers (for forming lightly-doped drain structures and salicides); punchthrough prevention implants; double polysilicon; 2-level (or more) metallization; and self-aligned contact structures. While most of these techniques are discussed in this chapter; others are discussed elsewhere in the book (e.g., isolation technology in chap. 2, salicides and self-aligned contact structures in chap. 3, and multilevel metallization technology in chap. 4).

5.4.1 Operation of an NMOS Inverter with a Depletion-Mode Load

The most common applications of MOS transistors are in integrated circuit digital logic gates and memory arrays. Several types of circuits have been developed to implement logic gates in MOS ICs, with each circuit type characterized by the type of *load device* it utilizes. The class of logic-gate circuits that has become standard in most NMOS digital ICs is based on enhancement-depletion (E-D) NMOS technology, and such E-D logic gates are the basis for most NMOS microprocessors, microprocessor peripheral devices, and static NMOS memories.

The inverter is the fundamental logic gate. E-D NMOS inverters are composed of two transistors; an enhancement mode MOSFET called the *driver*, which is switched *ON* and *OFF* by the input signal; and a depletion-mode MOSFET, called the *load*. The circuit diagram of this inverter is shown in Fig. 5-13a, and an example of a layout (as it would appear on the completed wafer) is shown in Fig. 5-13b. The cross-sectional view of this structure is shown in Fig. 5-13c.

The load connects the power supply voltage V_{DD} and the output of the inverter. The gate of the load transistor is electrically connected to its source region, so that V_G in the load transistor always equals zero. Since depletion-mode devices are always *ON* when $V_G = 0$, tying together the gate and drain ensures that the device is always *ON*. The driver transistor has its source connected to ground, while its drain region is electrically connected to the source region of the load transistor and the output. The threshold voltage of the driver transistor, V_{TE}, is selected to be between 0 and V_{DD},

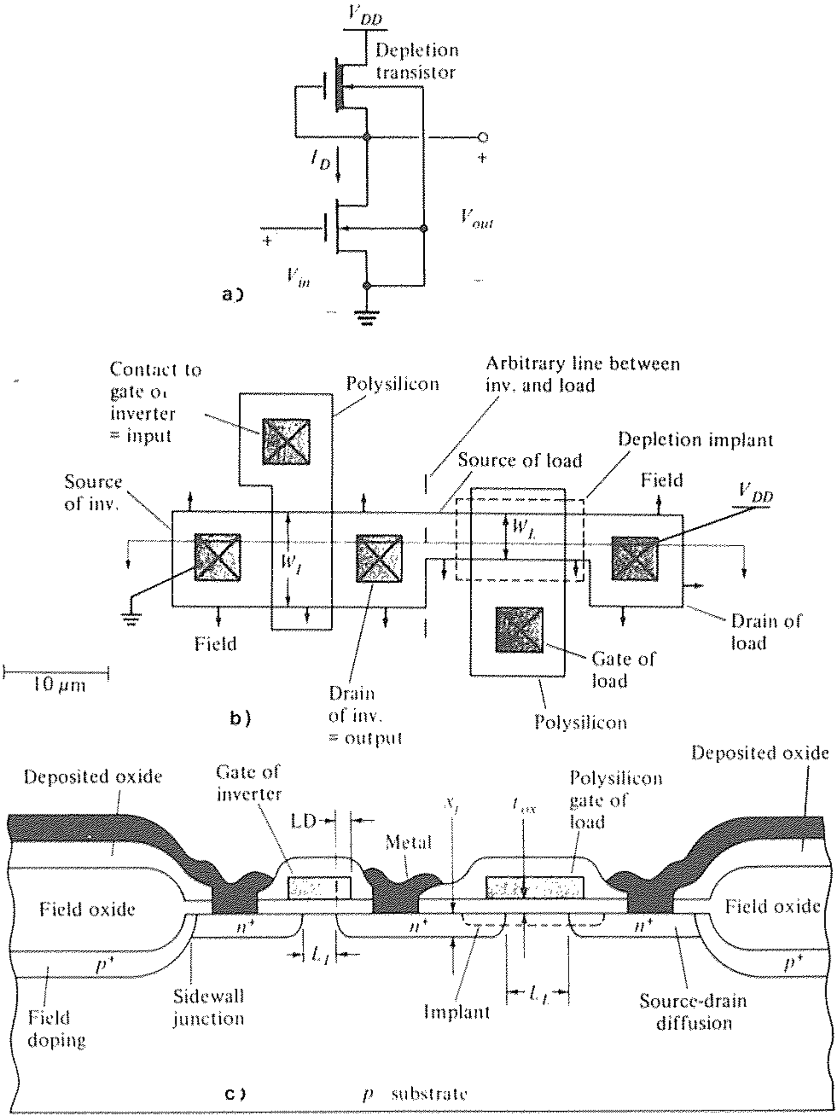


Fig. 5-13 NMOS inverter, depletion MOSFET load. (a) Schematic representation. (b) Layout. (c) Cross sectional view.[7] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

while the threshold voltage of the load transistor (since this is a depletion-mode device), V_{TD}, is selected to have a negative voltage value.

The input signal to the inverter (a voltage, V_{in}) is fed to the gate of the driver transistor, and the output signal is the voltage level at the output node, V_o. When there is a logic 0 input signal (low voltage at V_{in}), V_G at the driver transistor is $<V_{TE}$. In this case, there is no conducting channel between source and drain, so the impedance between the output and ground is very high. Since the depletion transistor is *ON*, however, the output is electrically connected to V_{DD}, and V_o rises (or is *pulled up*) to logic 1 (close to V_{DD}).

When the input voltage to the inverter is logic 1 (close to V_{DD}), the gate voltage applied to the driver is greater than V_{TE}, thus turning the driver transistor *ON*. A low-impedance path then exists between the output node and ground. Hence, the driver transistor can conduct a large current with a small voltage drop across it, allowing the output to go to logic 0. This demonstrates how the logic level at the input is inverted at the output of the circuit.

The desired characteristics of digital logic gates as IC elements are the following: fast switching speed (i.e., small propagation-delay time); low power dissipation; small size (i.e., minimum silicon area); and high noise margin. We'll now see how the NMOS inverter just described is designed to meet these requirements. (Note: the circuit design considerations used to arrive at the configuration presented here can be found in texts dealing with IC design.[7,17,18] Since our interest here is in how these design choices impact the processing of the device, readers interested in their justification are advised to consult the references mentioned.)

In order to obtain maximum packing density (i.e., most logic gates per area of substrate) with minimum power consumption* an E-D NMOS inverter would be designed in the following manner:[17] The driver transistor would have a gate area whose dimensions would be the minimum that could be fabricated in that generation of technology. The depletion load transistor would have an effective *channel length* (L_{effD}) that would be four times the effective channel length of the driver device (L_{effE}) and a minimum gate width.

As a simple example, let us assume an inverter is to be fabricated with E-D NMOS technology, and that the minimum manufacturable feature dimension is 5 μm. Let us also assume that a 1-μm lateral diffusion occurs under each side of the gate. In this case, L_{effE} is 3 μm and thus L_{effD} should be 12 μm. This means that the drawn gate lengths of the driver and load devices would be 5 μm and 14 μm respectively. The drawn gate widths of both devices would be 5 μm (since these are the dimensions between the walls of the field oxide).

If a 5 V power supply is assumed, the threshold voltages of the driver and depletion load transistor (V_{TE}, and V_{TD}) would be selected to be approximately 1.0 V and -3.0 V, respectively. (These two V_T values are chosen in order to give a high noise margin without severely impacting switching speed.) To conserve space, *buried contacts* between the polysilicon gate and the silicon substrate would be employed. (The procedure used for fabricating such contacts is presented in chap. 3.)

* But as a result, with less than maximum switching speed,

While the E-D NMOS technology has some important advantages over other IC technologies (particularly high packing density), it does have some drawbacks that become extremely serious as the number of devices on the chip gets very large. The most important of these is the high total power that is consumed. The origin of this power consumption arises from the operation of the NMOS inverter. When the inverter (and similarly other logic gates) has a low output state, both driver and load are *ON*, allowing current to flow from V_{DD} to ground. The power consumed by each inverter in a low-output state is the product of this current and V_{DD}. Thus, if a series of inverters are connected together, 50% of them will be drawing power at all times. When the devices get small enough, the power density on the chip becomes so large that it becomes necessary to replace NMOS with CMOS, since this technology consumes much less power per logic gate.

5.4.2 Process Sequence of a Basic E-D NMOS IC Technology

Figure 5-14 is a flow-chart representation of the sequence of steps that were used to fabricate typical E-D NMOS digital integrated circuits for gate lengths down to about 3 μm.[76] Figures 5-15a through 5-15j show what occurs on the wafer as this sequence of process steps is followed. The process being illustrated is a seven mask process (including the passivation pad mask, even though this final pad mask is not shown). The E-D NMOS inverter described in the previous section is used here as a vehicle for showing how device features on the wafer surface are created during the course of the process flow.

5.4.1.1 Starting Material. The starting material is a lightly doped (~5×10^{14}-10^{15} atoms/cm^2) *p*-type <100> silicon wafer (substrate). As described earlier, the lightly doped substrate is chosen to provide low source/drain-to-substrate capacitance, high source/drain-to-substrate breakdown voltage, high carrier mobility, and low sensitivity to source-substrate bias effects. A backside gettering process, such as implanting with Ar, to create crystalline-damaged regions that will trap mobile impurities during subsequent heat steps during the process may be used prior to the next step (see Vol. 1, chap 2).

5.4.1.2 Active Region and Field Region Definitions. The first task in this processing sequence is to define the active device and field regions on the wafer surface. This is done by selectively oxidizing the field regions so that they are covered with a thick field oxide, using the LOCOS process. The steps involved in this task are those of boxes 3-9 in Fig. 5-14, and are illustrated in Figs. 5-15a and 5-15b.

A thin pad oxide (20-60 nm thick) is first thermally grown or CVD-deposited on the wafer surface as a stress-relief layer. This is followed by the deposition of a CVD nitride layer (100-200 nm thick). *Mask #1* is then used to expose a resist film that was spun on after the nitride deposition (Fig. 5-15a2). After exposure and development, the resist layer remains behind only in the regions that will be the active device regions

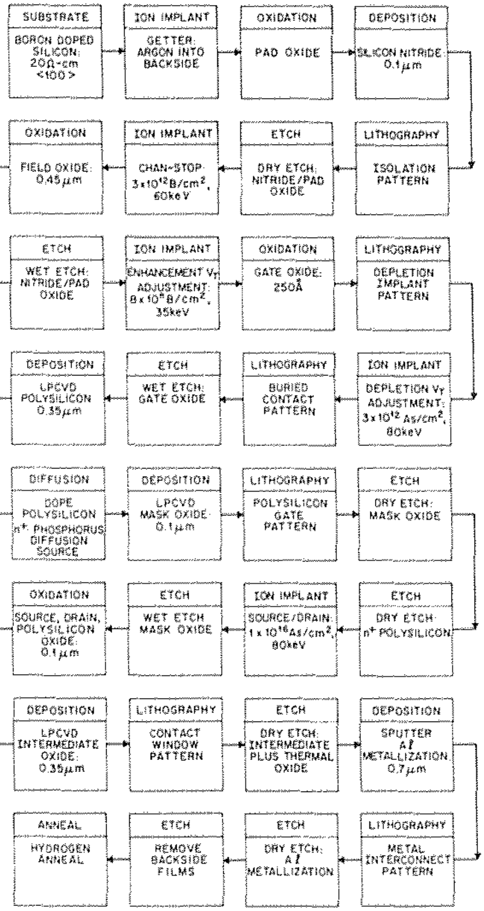


Fig. 5-14 Main steps in an *n*-channel, polysilicon-gate MOS IC process flow.[76] (© 1980 IEEE).

(Fig. 5-15b1). Next the nitride and pad oxide are anisotropically dry-etched away in the regions not covered by the resist (field regions). Thus, after the removal of the resist, the active areas are covered with the nitride/pad-oxide layer (Fig. 5-15b3).

In the next step, a boron implant (10^{12}-10^{13} atoms/cm^2, 40-80 keV) is performed to create *channel stops* in the field regions. The nitride/pad oxide layer now acts as a mask (Fig. 5-15c1) to prevent the boron from penetrating the silicon in the active areas. (Note that in some processes the resist is not removed until after the channel-stop

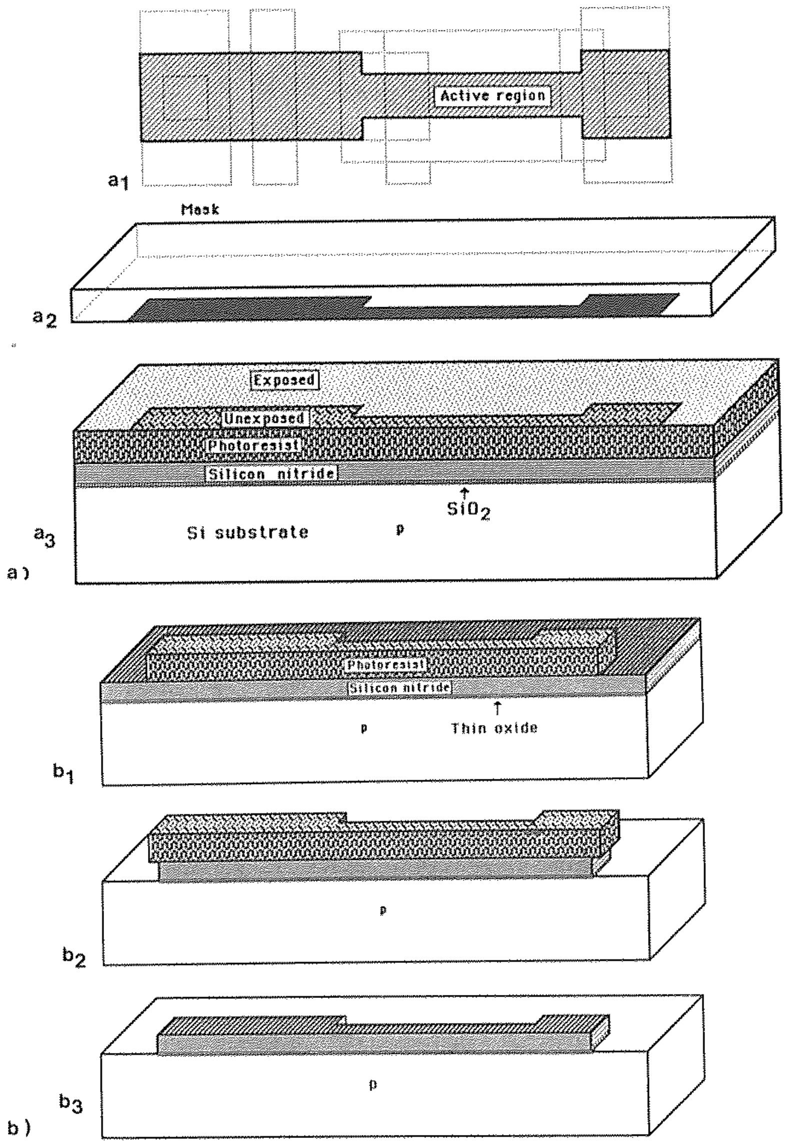


Fig. 5-15 NMOS E-D inverter fabrication sequence. (a) Patterning of the active region. (b) Patterning the silicon nitride-pad oxide layers. From W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission.

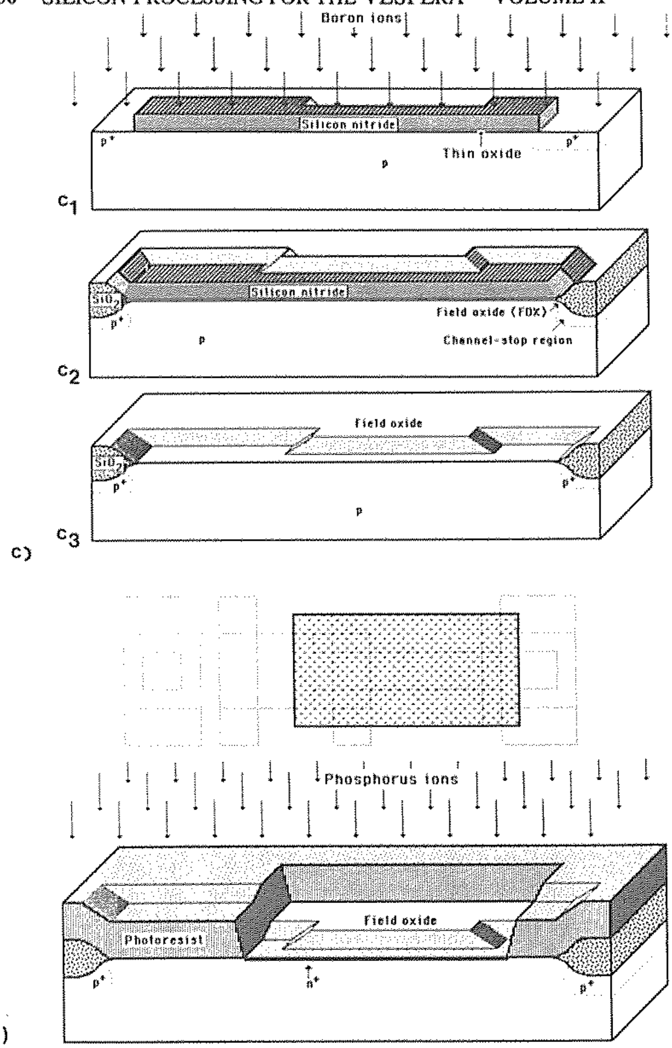


Fig. 5-15 (c) Active region formation and channel-stop implant. (d) Implantation of the channel of the depletion-mode transistor. From Maly, *Atlas of IC Technologies.*

implant, as the nitride/pad oxide layer may be too thin to act as an effective implant mask. In fact, with the patterned resist still in place, the channel stop may be implanted through the nitride/pad oxide layer, which in this case would not be etched until the implant is performed.)

A thermal-oxidation step is then performed to grow a thick (0.5-1.0 μm) field oxide over the regions where the nitride has been etched away. In this process, the field oxide is self-aligned to the channel stops. During the oxide growth, some lateral oxidation also occurs under the nitride edges, forming the *bird's beak* structure (see chap. 2 for more details concerning this effect, as well as other problems that arise in connection with the field-oxide growth step). After the field oxide has been grown, the remaining nitride and pad oxide are stripped, leaving the active areas with exposed silicon surfaces for further processing (Fig. 5-15c3).

5.4.1.3 Gate-Oxide Growth and Threshold-Voltage Adjust Implant In the next major step the gate oxide is grown, and the threshold voltages of the enhancement-mode and depletion-mode transistors of the inverter are adjusted through ion implantation. The growth of the gate oxide is a critical step, as a defect-free, very thin (15-100-nm), high-quality oxide without contamination is essential for proper device operation. The gate oxide is grown only in the exposed active region (the field-oxide thickness is actually increased slightly as a result of this oxide-growth step). As noted earlier, the drain current in an MOS transistor is inversely proportional to the gate-oxide thickness (for a given set of terminal voltages). As a result, the gate oxide is normally made as thin as possible, commensurate with oxide breakdown and reliability considerations.

In order for a high-quality gate oxide to be obtained, the surface of the active area is wet-etched to remove any residual oxide. A sacrificial oxide is often deliberately grown on the exposed active areas after field oxidation to remove any dry-etch induced damage or unwanted nitride (due to the Kooi effect, see chap. 2).[65] After such oxides have been stripped, the gate oxide is grown slowly and carefully, usually through dry oxidation in a chlorine ambient (see Vol. 1, chap. 7).

The threshold-voltage adjust implant of the enhancement-mode devices is performed next. In this step, boron is implanted though the gate oxide (10^{12}-10^{13} atoms/cm^2, 50-100 keV), but the ions are not given enough energy to penetrate the field oxide. No mask is used in this step. (Note that in many processes, another pre-gate oxide oxide is grown, through which this implant is performed. It is again stripped off following the implant, and the gate oxide is then grown.)

Next, the depletion-mode devices of the circuit are given their threshold-voltage implant dose (Fig. 5-15d). The areas of the depletion-mode transistor channels are implanted with phosphorus or arsenic ions (~10^{12} atoms/cm^2, 100 keV) to give a threshold voltage of about -3.0 V. The implant dose is adjusted so that it over-compensates for the previous boron threshold-voltage-adjust implant, thus making the surfaces *n*-type. A negative threshold voltage is thus yielded, as required to establish a depletion-mode device. Photoresist (patterned by the use of *Mask #2*) is used to selectively allow the depletion-mode transistor channels to be implanted. The ions cannot penetrate the resist to reach active areas below. Likewise, the ions cannot penetrate any field oxide that is exposed by the resist opening. Hence, the location of the depletion transistor channel is defined by the intersection of the *Mask #2* window and the active region.

Buried contacts are then opened in the gate oxide using *Mask #3* (Fig. 5-15e). This opening in the gate oxide must be provided wherever it is desired to have polysilicon electrically contact the active silicon area (details of buried-contact formation are described in chap. 3). Since the polysilicon is deposited on the gate oxide, it will remain isolated from the substrate below unless a special opening is cut in the gate oxide. With *Mask #3*, resist covers the entire wafer except in those areas where the buried contact is desired. The gate oxide can then be etched from these regions, uncovering the silicon below.

5.4.1.4 Polysilicon Deposition and Patterning. A layer of polysilicon (typically 0.4-0.5 μm thick) is next deposited by CVD over the whole wafer (see Vol. 1, chap. 6 for more information on the properties of polysilicon and its deposition process). Either ion implantation or diffusion with phosphorus is then used

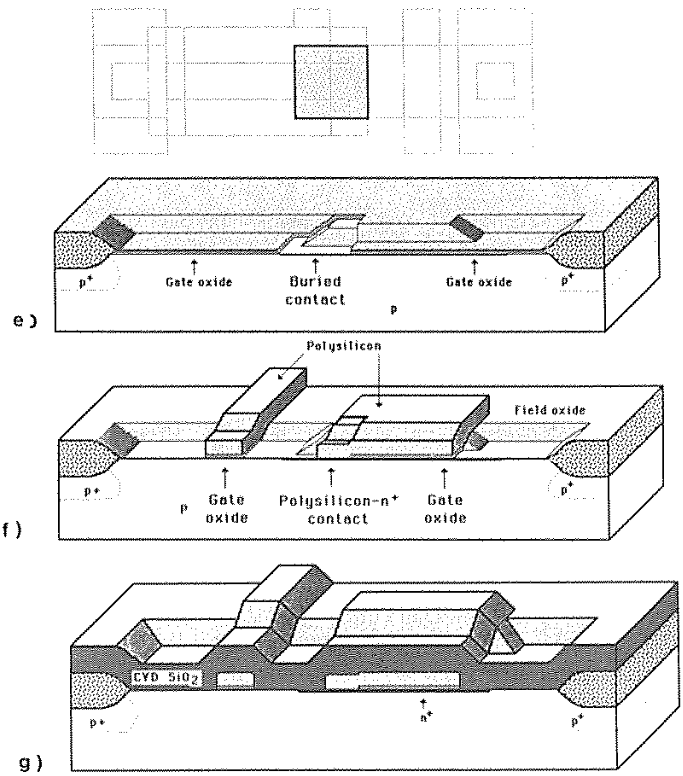


Fig. 5-15 (e) Buried contact etching. (f) Patterning of the polysilicon layer followed by gate oxide etching. (g) Deposition of the CVD SiO_2 layer followed by the diffusion of the drain and source regions. From Maly, *Atlas of IC Technologies*.

to dope the polysilicon to a sheet resistance of 20-30 Ω/sq. This resistance is adequate for MOS circuits with gate lengths ≥3 μm. For smaller devices, polycide layers (i.e., composite layers of refractory metal silicides and polysilicon) can be used to reduce the sheet resistance to ~1 Ω/sq (see Vol. 1, chap. 11). Using a polycide gives us the benefits of both silicon-gate and metal-gate technologies.

The gate structure and polysilicon interconnect structures are then patterned using *Mask #4* (Fig. 5-15f). Following exposure and development of the resist, the polysilicon film is etched (in current technology this is done by means of a dry-etch process). This is a critical etch step for several reasons. First, the channel length of the device depends on the gate length, because of the self-aligned nature of the silicon gate technology. Hence, the gate-length dimension must be precisely maintained across the entire wafer, and from wafer to wafer. Second, the profile of the etched poly gate structure should be vertical; this will prevent variation of channel lengths by the penetration of the ions of the thinner regions of the gate sidewalls during formation of the source/drain regions by ion implantation. Third, to achieve the above goals, an anisotropic polysilicon etch process must be employed. This type of process, however, requires overetching to remove the locally thicker regions of polysilicon that exist wherever it crosses steps on the wafer surface. During the overetch time, areas of the thin gate oxide are exposed to the etchants. Thus, it is necessary to use a polysilicon etch process that is highly selective with respect to SiO_2.

5.4.1.5 Formation of the Source and Drain Regions. Once the gate has been fabricated, the source and drain regions can be formed. This is normally done by ion implantation without the use of a lithography step (Fig. 5-15g). The gate and the field oxide act as masks to prevent the ion implantation from penetrating to the silicon substrate below. Therefore, only the active regions covered by the gate oxide (and no gate polysilicon), are implanted. An n^+ implant is used, with an energy that is insufficient to penetrate the gate-poly or field-oxide layers (arsenic is typically used, with a dose of ~10^{16} atoms/cm^2 and an energy of 30-50 keV). As noted earlier, the source and drain are thereby *self-aligned* to the gate, and the dimension of the polysilicon gate thus plays a major role in the defining of the MOS gate length.

Following the source/drain implant, an anneal (or drive-in) step is performed to activate the implanted atoms and to position the source/drain junctions as desired. During this step, some of the phosphorus doping of the polysilicon outdiffuses into the silicon substrate wherever a buried contact opening the gate oxide has been cut. This diffusion (which occurs both vertically and laterally into the silicon below) forms a heavily doped n^+ region under the polysilicon in the buried-contact exposed region. The lateral diffusion of the implanted source/drain dopant thereby becomes electrically connected to the n^+ region under the polysilicon buried-layer region. In this manner, an electrical connection between the polysilicon and the silicon is established at the buried contact locations. In some processes, the junction formed by the buried-contact dopant outdiffusion from the polysilicon is deeper than the source/drain junctions, while in others it is not as deep.

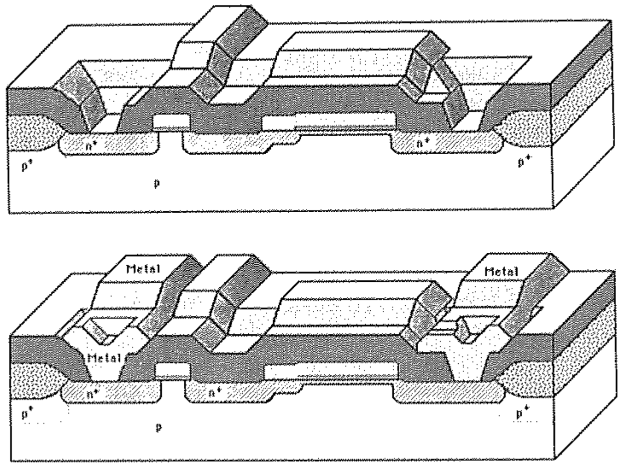


Fig. 5-15 (h) Contact cuts of the E-D inverter. Metallization of the E-D inverter. (i) Completed NMOS E-D inverter structure. Figures 15a through 15i from W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission.

The source/drain drive-in step also plays a part in determining the effective channel length (L_{eff}). That is, if the lateral junction depth is x_{jl} (which is primarily determined by the the lateral diffusion during the drive-in step, because the lateral straggle of arsenic at 30 keV is only ~5 nm, see Vol. 1, chap. 9), L_{eff} will be decreased by $2x_{jl}$ from the gate length at the mask level. Note that the channel *width* is also reduced by the bird's beak encroachment into the active area (see chap. 2). Thus, the actual width, W_w, of an MOS device is $W_w = W - \Delta W$, where W is the width at the mask level and ΔW is the channel-width shrinkage during processing.

The depth of the source and drain is thus a critical dimension, but the doping concentration is not as important. (A discussion of shallow source/drain junction formation techniques is presented in chap. 3, section 3.10.) To a first approximation, the device characteristics will not depend on the doping concentration value, provided it is sufficiently heavy.

A diffusion step may be used to dope the source/drain regions. In some of these cases the dopant source of the diffusion is the CVD oxide layer that is deposited after the gate has been defined (see next section).

5.4.1.6 Contact Formation. After the source/drain regions have been formed, a CVD process is used to deposit a layer of doped SiO_2 (glass), about 1 µm thick, onto the wafers (see Vol. 1, chap. 6). The dopant in the SiO_2 is either phosphorus (in which case the material is referred to as *phosphosilicate glass),* or both phosphorus and boron (making it a layer of *borophosphosilicate glass*). In some processes a thin thermal oxide is grown on the polysilicon prior to deposition of the glass layer. Nevertheless, a thick layer of SiO_2 cannot be thermally grown because of the excessive redistribution of the impurities that would take place during such growth. Hence, a lower-temperature CVD process must be used to get a sufficiently thick oxide.

The doped CVD glass layer plays several roles in the fabrication and operating aspects of the circuit. First, it acts as an insulating layer between the polysilicon and the metal to be deposited. Second, it reduces the parasitic capacitance of the interconnect metallization layer. Third, the addition of the phosphorus to the glass makes the layer an excellent getter of Na ions (recall that contamination by Na can cause instabilities in the V_T of the MOS devices). The phosphorus-doped glass immobilizes the otherwise mobile Na atoms within the CVD layer, preventing them from reaching the gate oxide and altering the threshold voltage. Finally, the dopants in the glass make it viscous at elevated temperatures (1000-1100°C for PSG, and 800-950°C for BPSG, see Vol. 1, chap. 6), allowing the layer of doped glass to be flowed after it is deposited. Through this procedure, a rounding of the contours of the glass and a smoothing out of any sharp steps is achieved. This produces better step coverage of the metal (which is deposited next) over the otherwise severe wafer topography. The high-temperature glass-flowing step also serves to activate the source/drain implanted junctions and drive them to their desired positions.

Contact openings are next created by a lithography-and-etch step (Fig. 5-15h). *Mask #5* is used to define contact opening patterns in a photoresist film, and a dry-etch process is then used to open the contact windows through the CVD SiO_2 to the underlying polysilicon and the n^+ regions in the silicon.

The contact-opening step can be critical, as the contact size and alignment limit the minimum size of the device. The source and drain regions must be large enough for the contact to fit, with allowance for alignment tolerance. If the contact opening exposes a part of the substrate, the drain or source will be shorted to the substrate. Likewise, any overlap of the source/drain contact opening and the gate will cause the gate to be shorted to the source or drain.

To keep the transistor as small as possible, the contact window is usually made at the minimum size achievable with the given process. In some processes, the exposed silicon in the contact is redoped to prevent shorting between the source/drain areas and the substrate (see chap. 2). Also note that the gate contact (i.e., Al to polysilicon) is often made outside the active device area to avoid possible damage to the thin gate oxide.

After the contact etch is completed and the resist is stripped, the doped CVD glass is again subjected to another flowing step. This procedure rounds the corners of the top of the oxide windows so that metal step coverage into the contact windows is improved. The process is called *reflow* of the contact windows (see chap. 3 and Vol. 1, chap. 6).

5.4.1.7 Metallization Deposition and Patterning. After the contacts have been opened, the metallization layer is deposited (~1 μm thick). Because the metal layer is highly conductive, it is used whenever possible to interconnect circuit elements and to carry large amounts of supply current. The metal interconnect lines that are fabricated must have sufficient thickness, width, and step coverage to keep the current density in each line below the value that could produce electromigration failure (see chap. 4). In addition, the spacing between adjacent metal lines must be kept large enough that the lines will never touch, even under worst-case process variations.

Although evaporation was the method employed to deposit Al in the early days of MOS, it has generally been replaced by sputtering. To a great extent, the change was made because Al alloys with tightly controlled compositions became the materials of choice for the metal layer. Sputtering allows alloys to be deposited with much better compositional fidelity (see Vol. 1, chap. 10).

The metal alloy that was eventually chosen for NMOS is Al:1wt% Si. The silicon is added to the aluminum film to prevent spiking of the contacts during subsequent annealing steps (see chap. 3). In CMOS, such Al:Si alloys are being phased out as the metallization material for reasons that are discussed in the CMOS chapter and in chapters 3 and 4. Either wet-chemical etching or dry etching is used to pattern the Al film, using *Mask #6* (Fig. 5-15i and 5-15j).

Following the patterning of the metal, the Al-silicon contacts are alloyed. This step brings the Al and the n^+ silicon into intimate contact, since it allows the thin native SiO_2 layer that is likely to exist at the Al-Si interface to be reduced by the Al (see chap. 3). Such intimate contact between Al and n^+Si establishes a low-resistance ohmic contact. The anneal process exposes the wafer to a 375-500°C temperature in an H_2 or

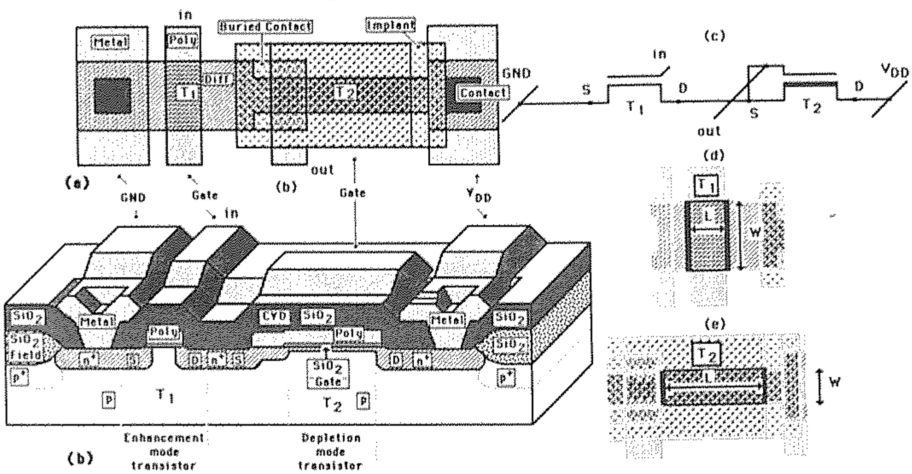


Fig. 5-15 (j) E-D inverter. 1) Composite drawing of the layout. 2) Cross section of complete structure. 3) Electrical diagram. 4) The enhancement transistor. 5) The depletion transistor. From Maly, *Atlas of IC Technologies*.

$N_2 + H_2$ (5%) ambient for about 30 minutes. As a result, this step may also be used as the annealing process for reducing the interface trap density in the gate oxide that was introduced by earlier processing steps (see Vol. 1, chap. 7).

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

Openings are etched into this layer so that a set of special metallization patterns under the passivation layer is exposed. These metal patterns are normally located in the periphery of the circuit and are called *bonding pads* (Fig. 5-16). Bonding pads are typically about 100 x 100 μm in size and are separated by a space of 50 to 100 μm. Wires are connected (bonded) to the metal of the bonding pads and are then bonded to the chip package. In this way connections are established from the chip to the package leads.

The bonding-pad openings are created by patterning the passivation layer with *Mask #7*. If a PSG layer is used, the phosphorus (2-6 wt%) in the glass not only causes the PSG to act as a getter for Na but also prevents the glass film from cracking. Care must be taken to ensure that not more than 6% phosphorus is incorporated into the PSG, as this can cause corrosion of the underlying metal if moisture enters the circuit package (see Vol. 1, chap. 10). When silicon nitride is used, care must be taken to ensure that the deposited nitride film exhibits low stress (either tensile or compressive), so that it will not crack, since cracking would compromise the sealing capability of the film.

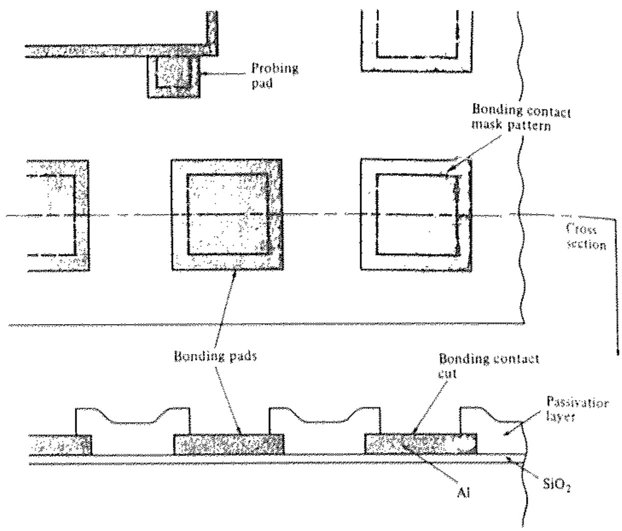


Fig. 5-16 Passivation layer and bonding pad openings. (Note, cross-section not to scale.)

5.5 SHORT-CHANNEL EFFECTS AND HOW THEY IMPACT MOS PROCESSING

The device characteristics of MOSFETs (such as threshold voltage, subthreshold currents, and I-V characteristics beyond threshold) are well predicted by Eqs. 5-1 through 5-10, if the channel lengths of the transistors are "long" (i.e., if they exceed 2 μm in length). A guide to the design of such long-channel MOSFETs is given in reference 70.

For shorter channel devices, however, a series of effects arise that result in significant deviations from the values predicted by the long-channel models. Such *short-channel effects* become a dominant part of MOS device behavior when channel lengths decrease below 2 μm. The effects are briefly described here to provide readers with a basis for understanding the processing steps that have been developed to mitigate their adverse impact on device performance. So-called *hot-carrier effects* will be described separately in the following section, even though these have only been observed in short-channel devices. More details on short-channel effects in *p*-channel CMOS devices are also given in chapter 6. A guide to the design of submicron channel MOSFETs (those with channel lengths ≤ 1 μm), is presented in reference 62.

Short-channel effects can be divided into the following categories: (a) those that impact V_T; (b) those that impact subthreshold currents; and (c) those that impact I-V behavior beyond threshold.

5.5.1 Effect of Gate Dimensions on Threshold Voltage

5.5.1.1 Short Channel Threshold Voltage Effect. V_T becomes less well predicted by Eq. 5-6 as the dimensions of the gate are reduced, and the error becomes significant when the dimensions are reduced to less than 2 μm. To get good agreement with measured data, a term $|\Delta V_T|$ must be subtracted from the V_T value obtained from Eq. 5-6. Thus, as the device length is reduced, the measured value of V_T of *n*-channel enhancement-mode devices becomes *less positive* than that given by Eq. 5-6, while for *n*-channel depletion-mode devices, V_T becomes *more negative*. For *p*-channel (enhancement-mode) devices, V_T becomes *less negative*.

The discrepancy arises because the equations for V_T given earlier in the chapter are based on one-dimensional theory.[20] It is assumed that the space charge *under* the gate is a function of only the vertical electric field, E_x (and is thereby influenced only by the charge *on the gate)*. If the channel length is long, this is a reasonable assumption, as the influence of the drain and source junctions on the quantity of charge in the channel can be neglected. However, as the channel length approaches the dimensions of the widths of the depletion regions of the source and drain junctions, these depletion regions become a greater part of the *channel-depletion region* (Fig. 5-17a). Thus, some of the channel-depletion region charge is actually linked to the charge in the depletion region within the source and drain structures, rather than being linked to the gate charge. Hence, some of the channel region is partially depleted without any influence of the gate voltage. (In the extreme, if the two built-in depletion regions spanned the entire channel length when no voltage existed between source and drain, they could deplete all

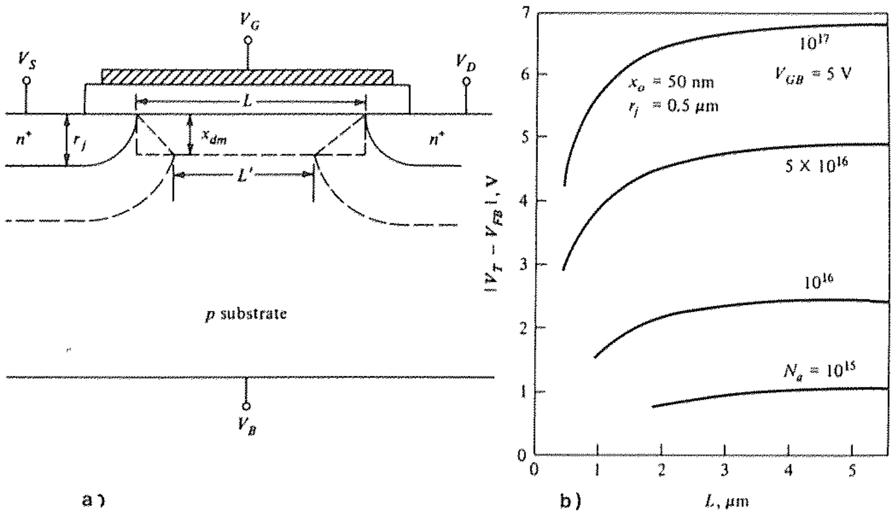


Fig. 5-17 (a) Yau's model of charge sharing.[73] (b) Theoretical threshold voltage as a function of channel length for various substrate doping concentrations.[73] Reprinted with permission of Solid-State Electronics.

of the channel region at the Si-SiO_2 interface.) Since some of the channel is depleted without the need to apply a gate bias, *less gate charge is required to invert the channel in short-channel devices than in a long-channel device with comparable substrate doping.*

A relatively simple equation that predicts the threshold lowering (ΔV_T) in terms of the device parameters is

$$\Delta V_T = \frac{q\, N_A\, W_{max}\, x_j \left(\sqrt{\left\{ 1 + \frac{2\, W_{max}}{x_j} \right\}} - 1 \right)}{C_{ox}\, L} \qquad (5-16)$$

where W_{max} is the maximum depth of the depletion region in the channel, and x_j is the junction depth of the source drain regions.[74]

This effect is important because in order to be able to establish slightly positive V_T values (e.g., < +1 V) in long-channel NMOS transistors with lightly doped channels, it is necessary to increase the doping concentration at the surface of the channel. Consequently, in order to allow short-channel enhancement-mode NMOS devices to be fabricated with the same V_T value, the substrate doping concentration must be further increased. Since the magnitude of this effect increases as the device length is reduced (as is predicted by Eq. 5-16 and illustrated in Fig. 5-17b), it will be necessary to progressively increase the substrate doping concentrations as devices are made smaller, in order to maintain suitably positive values of V_T.

Application of a drain voltage causes the drain depletion region to extend into the channel region, where it acts as an additional substrate bias, and reduces V_T.[21]

5.5.1.2 Narrow Gate-Width Effect on Threshold Voltage. In contrast to the short-channel-length effect, devices with narrow channel-widths require that such a positive-value correction term be *added* to Eq. 5-6 to give good agreement with the calculated values.[22] This effect is primarily due to the encroachment of the channel-stop dopants under the edges of the sides of the gate (see chap. 2). This has the effect of doping the channel at these edges more heavily than at the center (Fig. 5-18a). Thus, it requires more charge on the gate to invert the channel than if such encroachment did not occur, and this causes a shift in V_T from its predicted value (Fig. 5-18b).[23] On the other hand, if the voltage on the gate is held constant, the edges of the channel will have a higher V_T than the center of the channel. Because (V_G - V_T) is smaller, narrow-channel devices will thus conduct less drain current. Several schemes are also described in chapter 2 for reducing the channel-stop encroachment and thus reducing the narrow-width effect.

Narrow-width effects are still observed even if there are no channel-stop implants. These arise because the relatively thinner depletion region under the field-oxide device distorts the depletion region under the gate oxide and prevents the formation of an

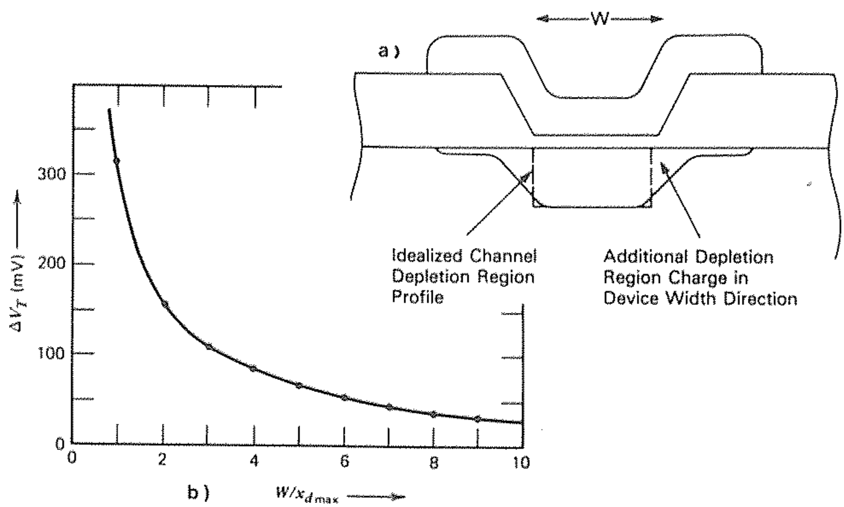


Fig. 5-18 (a) Schematic depiction of the narrow-channel effect. (b) Threshold voltage shift ΔV_T caused by the narrow-channel effect for a MOSFET with $N_A = 10^{15}$ cm^{-3}, and t_{ox} = 50 nm.[22]

inversion layer at the two edges. Although this leads to a slightly higher V_T, the effect is much less severe than that observed in devices with heavy channel-stop implants.

5.5.2 Short-Channel Effects on Subthreshold Currents (Punchthrough and Drain-Induced Barrier Lowering)

In section 5.2.3 we described the nature of subthreshold current flow (I_{Dst}) in MOSFETs, noting that a specific value of the *subthreshold-swing* parameter (S.S.) can be attributed to such "normal" I_{Dst} currents in long-channel devices. In short-channel MOSFETs, however, larger I_{Dst} values are observed at lower voltages than predicted by long-channel device models: one manifestation is an increase in the value of S.S. Note that even relatively small values of I_{Dst} can limit the transistor's ability to isolate nodes in a dynamic circuit or can allow excess current in static inverters. Hence, care must be taken to minimize I_{Dst}. Two of the primary causes of increased I_{Dst} are *punchthrough* and *drain-induced barrier lowering* (DIBL).

Punchthrough is normally observed when the gate voltage is well below V_T. It occurs as a result of the widening of the drain depletion region when the reverse-bias voltage on the drain is increased. The electric field of the drain may eventually penetrate into the source region and thereby reduce the potential energy barrier of the source-to-body junction (Fig. 5-19).[23] When this occurs, more majority carriers in the source region have enough energy to overcome the barrier, and an increased current then flows from source to body. Some of this current is collected by the drain, thereby increasing I_{Dst}. In general, punchthrough current begins to dominate I_{Dst} when the drain and

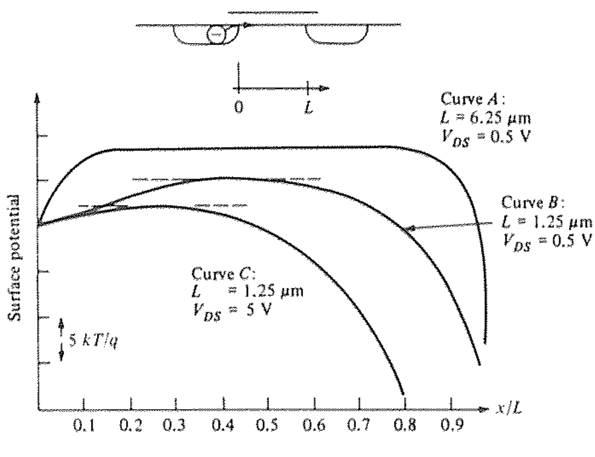


Fig. 5-19 Surface potential in the channel for devices with different channel lengths.[23] (© 1979 IEEE).

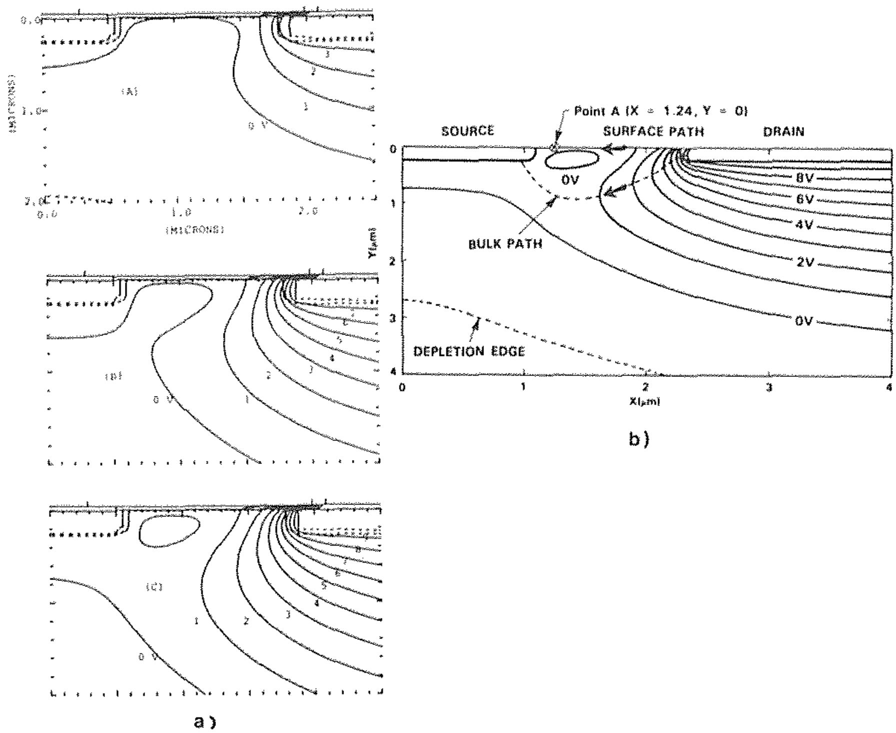


Fig. 5-20 (a) 2-D potential profile of an *n*-channel MOSFET with a drain bias of: 1) 3 V; 2) 7 V; 3) 9 V. Channel length = 1 μm. (b) Simulation of the potential profile of an *n*-channel MOSFET with a gate and drain bias of 0 and 9 V respectively. The surface DIBL and bulk punchthrough paths are indicated. From K. M. Cham, *et al., Computer Aided Design and VLSI Device Development*. Copyright 1986 Kluwer Academic Publishers. Reprinted with permission.

source depletion regions meet, and it can be suppressed by keeping the total width of the two depletion regions smaller than the channel length.[24]

Calculations of the potential in the bulk channel region in devices that use ion implantation to adjust V_T indicate that the barrier is lowest away from the Si-SiO_2 interface (usually at almost the same depth as the source/drain junction depths). That is, the V_T-adjust implant increases the doping concentration near the surface of the channel, causing the drain depletion region to be wider in the bulk than it is near the Si-SiO_2 interface. As a result, punchthrough current flows *below* the surface (Fig. 5-20). Consequently, the gate voltage has less control over the subthreshold current (i.e., even with sufficient gate voltage to turn off the channel, I_{Dst} can still flow in such devices).

An enhancement-mode device which is not turned off when $V_G = 0$ loses its ability to function as a switch.

Similarly, the application of a drain voltage in short-channel devices can also cause drain-induced barrier lowering (DIBL). That is, the drain voltage can cause the *surface potential* to be lowered (Fig. 5-21).[23, 71] As a result, the potential energy barrier at the *surface* will be lowered, and the subthreshold current in the channel region at the Si-SiO_2 interface can be increased (5-21b). This implies that I_{Dst} at the surface due to DIBL is expected to become larger as the gate voltage approaches V_T.

These two effects illustrate the complexity involved in modeling the overall subthreshold I-V behavior of short-channel MOSFETs. That is, both punchthrough current (in the bulk), as well as DIBL-induced current (at the surface), may simultaneously contribute to I_{Dst}.

To prevent punchthrough current in short-channel devices, the substrate doping can be increased to decrease the depletion-layer widths. These widths can be estimated using the formula for the width of a one-sided step junction:

$$W = \sqrt{\frac{2 K_s \varepsilon_o (|V_A| + V_{bi}}{q N_B}} \qquad (5 - 17)$$

where the built-in voltage, V_{bi}, given by

$$V_{bi} = 0.56 + (kT/q) \ln (N_B/n_i) \qquad (5 - 18)$$

and where V_A is the total applied voltage and N_B is the doping concentration of the body. Figure 5-22 gives the depletion-layer width of *pn* junctions as a function of doping and applied voltage.

However, increasing the substrate doping also increases the source-to-body and drain-

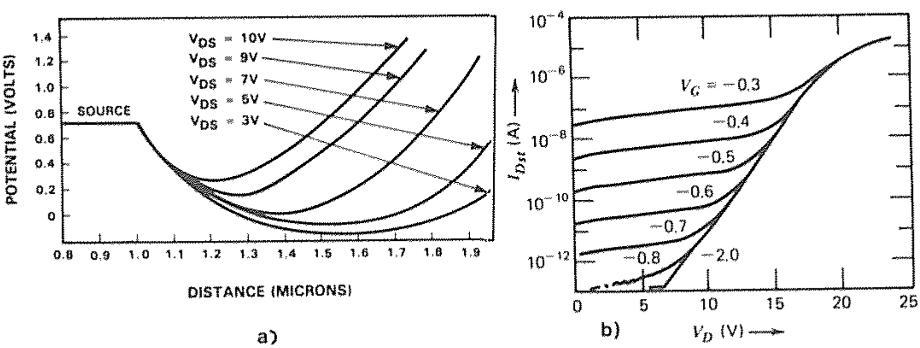


Fig. 5-21 (a) DIBL versus drain bias for short-channel MOSFET. (b) Experimental low-current characteristics for a MOSFET with L = 2.1 μm, $V_{SB} = 0$.[81] (© 1974 IEEE).

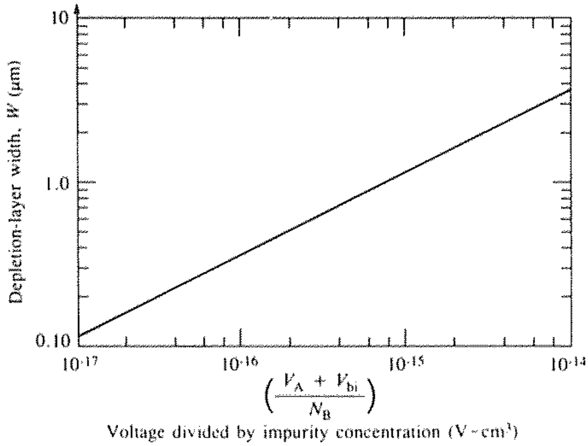


Fig. 5-22 Depletion-layer width of a one-sided step function as a function of doping and applied voltage calculated from Eqs. 5-17 and 5-18.

to-body junction capacitances, as well as the body factor. In addition, it reduces the breakdown voltages of the source/drain junctions. To avoid these drawbacks, an additional boron implant (whose peak concentration is located at a depth near the bottom of the source-drain regions) can be performed. This additional doping reduces the lateral widening of the drain-depletion region below the surface without increasing the doping under the junction regions. With such implants, the component of the punchthrough current can be suppressed to well below the normal I_{Dst} current of the device.

For example, in Fig. 5-23, a 1.2-μm device with a body doping of 1.9×10^{15} cm^{-3} without such a punchthrough-stopping implant shows a large value of I_{Dst} even when $V_G = 0$ V (curve A). This indicates that the device is already exhibiting punchthrough.[26] Implants of boron with a dose of 8×10^{11} atoms/cm^2 and different energies are then performed in an attempt to reduce I_{Dst} to the values exhibited by a long-channel device (curve B). If the implant is too shallow, the extra implant has the effect of shifting the V_T of the device to well beyond the desired value. When the energy is increased so that the implant is sufficiently deep, the value of I_{Dst} drops to that exhibited by the long-channel device. At the same time, the surface concentration remains essentially unchanged, so that V_T is not appreciably shifted.

In another example, the S.S. of a device without a punchthrough-prevention implant is measured as its length is varied (Fig. 5-24a). At an L_{eff} of ~0.85 μm the S.S. starts to increase, indicating that punchthrough current begins to dominate I_{Dst}. By adding an implant step that places boron atoms in the dashed subsurface region shown in Fig. 5-24b, the punchthrough component of I_{Dst} is suppressed so that it is not observed until L_{eff} becomes nearly as small as 0.5 μm.

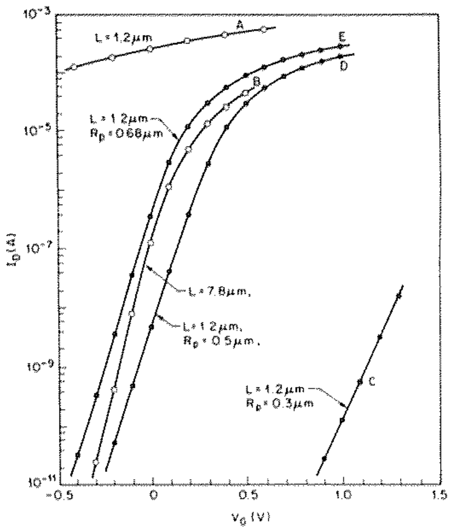


Fig. 5-23 Drain current versus gate voltage for *n*-channel devices with a substrate doping of 1.9×10^{15} atoms/cm^3, source/drain junctions 0.47-μm deep, 575 Å gate oxide, drain voltage of 5 V, and V_{BS} of 0 V. Devices A and B have no channel implant, and devices C and E have a boron channel implant of 8×10^{11} atoms/cm^2 at various energies.[26] (© 1978 IEEE).

5.5.3 Short-Channel Effects on I-V Characteristics

The I-V characteristics of short-channel devices are significantly altered in three ways. First, the combined effects of reduced gate length and gate width produce a change in V_T. Second, the channel length is modulated by the drain voltage when the device is in saturation (i.e., $V_{DS} > [V_G - V_T]$), causing an increase in device gain over that predicted in an ideal long-channel device (*channel-modulation effect*). Third, the mobility of the carriers in the channel is reduced by two effects, which also reduces I_D. (The two effects are the *mobility-degradation factor*, due to the gate field, and the *velocity-saturation factor*).

Figure 5-25 shows the I-V characteristics of an MOS device.[28] The curves in Fig. 5-25a are those of an ideal long-channel MOS device, while those in Fig. 5-25b show the effect of adding the channel modulation factor. Figure 5-25c shows the combined effect of adding the mobility degradation factor to those of Fig. 5-25b. The velocity saturation factor also has the effect of making the both I_D and g_m independent of channel length in silicon MOS transistors for $L_{eff} \leq 1.25$ μm. More details on these effects are provided in suitable device physics texts.[1,3,28]

A general guide to the design of short-channel MOSFETs is given in references 62 and 69. In addition, a simple *engineering model* for short-channel devices has also been developed.[68] Its purpose is to provide a simple picture of the essential electrical behaviors of the short-channel MOSFETs from the perspective of a circuit designer. That is, this engineering model relates the terminal voltages to the drain current, much as Eqs. 5-8 and 5-10 yield the I-V characteristics for long-channel MOS devices. Consequently, device designers who need to relate device and process parameters to circuit parameters should also find this model useful.

5.5.4 Summary of Short-Channel Effects on the Fabrication of MOS ICs

In the first section of this chapter, we showed that the use of lightly doped substrates generally produced optimum device behavior in long-channel MOS transistors. In this

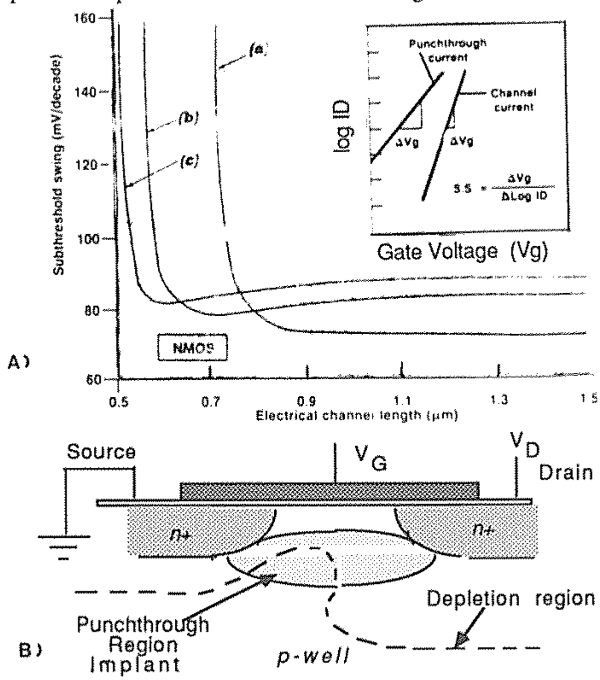


Fig. 5-24 (A) Subthreshold slope versus electrical channel length for NMOS devices (V_{Tn} = = 0.7 V), having a common threshold adjustment implant and punchthrough implant doses of: of: (a) zero; (b) 2×10^{11} cm^{-2}; and (c) 3×10^{11} cm^{-2}.[27] Reprinted with permission of Semiconductor International. (B) NMOS cross-section with implant placed into punchthrough region.

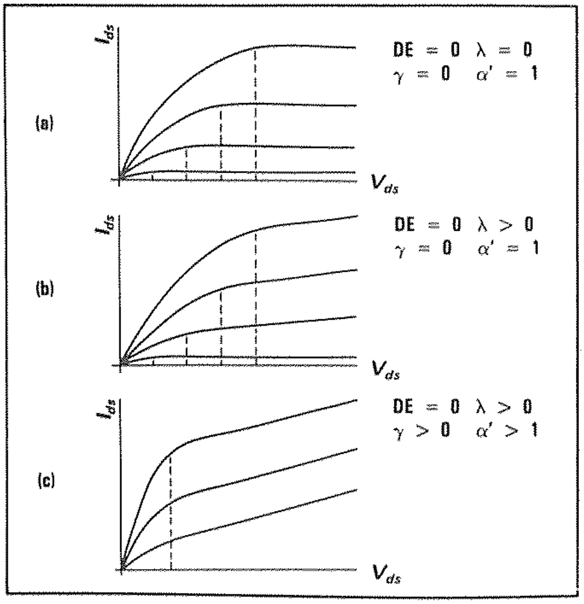


Fig. 5-25 The I-V curves of an MOS device showing the the effects of progressively increasing short-channel behavior. (a) Long-channel behavior. (b) With channel-length modulation. (c) Addition of velocity saturation.[28] (© 1986 IEEE).

section we noted that higher substrate doping is needed to overcome some of the detrimental impacts of short-channel effects. Thus, trade offs need to be made in selecting the proper substrate doping-concentration values to achieve optimum short-channel MOS device performance. Some of these trade offs are discussed by Kakumu,[29] who points out that a higher substrate doping concentration produces decreased ring-oscillator gate delay in submicron CMOS because of increased junction capacitances and decreased carrier mobility (due to increased impurity scattering).

5.6 HOT-CARRIER EFFECTS IN MOSFETs

If device dimensions are reduced and the supply voltage remains constant (e.g., 5 V), the lateral electric field generated in MOS devices increases. If the electric field becomes strong enough, it can give rise to so-called *hot-carrier* effects in MOS devices. This has indeed become a significant problem in NMOS devices with channel lengths smaller than 1.5 μm (and in PMOS devices with submicron channel lengths).[30] Hot-electron effects are more severe than hot-hole effects because of the higher electron mobility. Therefore, we begin our discussion by considering hot-electron effects in *n*-channel devices. At the end of this section we will also discuss the impact of hot-carrier effects on *p*-channel devices.

The maximum electric field, E_M, in a MOSFET occurs near the drain during saturated operation. A rigorous calculation of the field near the drain is a complex procedure, requiring a computer-aided solution of the two-dimensional Poisson equation, with the results of one such analysis being shown in Fig. 5-26. Nevertheless, the value of E_M can be estimated from[32]

$$E_M = (V_{DS} - V_{Dsat}) / m \tag{5-19}$$

where

$$m = 0.22\, t_{ox}^{1/3} / x_j'^{1/2} \tag{5-19a}$$

and t_{ox} is the gate oxide thickness and x_j' approximately corresponds to the source/drain junction depth. Although V_{Dsat} depends on L_{eff}, the dependence is weaker than a linear

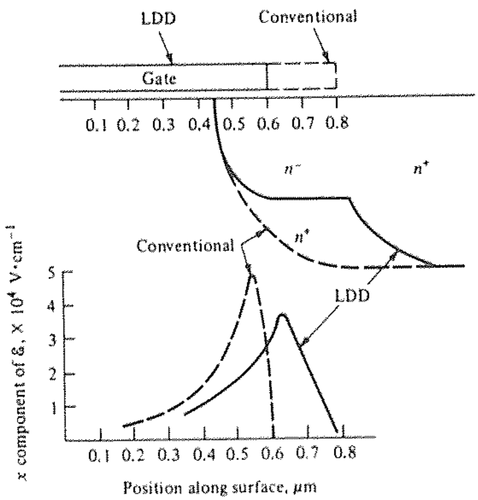


Fig. 5-26 Magnitude of the electric field at the Si-SiO$_2$ interface as a function of distance: L = 1.2 μm, V_{DS} = 8.5 V, $V_{GS} = V_T$.[38] (© 1980 IEEE).

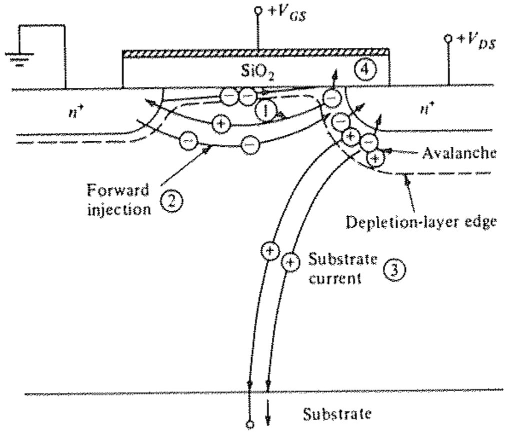


Fig. 5-27 Hot-carrier generation and current components. (1)- Holes reaching the source. (2)- Electron injection from the source. (3) - Substrate hole current. (4) - Electron injection into the oxide.

relationship (especially if $L_{eff} > 1$ μm). Thus, we can see that E_M increases as device dimensions shrink, but this is due to thinner gate oxides and shallower junctions, as well as to the reduction in L_{eff}.

Regardless of the factors that increase their magnitude, such high electric fields cause the electrons in the channel to gain kinetic energy and become "hot" (i.e., their energy distribution is shifted to a much higher value than that of electrons which are in thermal equilibrium with the lattice). Such hot electrons (which become hot near the drain edge of the channel, since that is where E_M exists) can cause several effects in the device. First, those electrons that acquire ≥1.5 eV of energy can lose it via *impact ionization*, which generates electron-hole pairs. The total number of electron-hole pairs generated by impact ionization is *exponentially dependent on the reciprocal of the electric field*, -$1/E_M$. In the extreme, this electron-hole pair generation can lead to a form of *avalanche breakdown* (*Process 1*, shown in Fig. 5-27). Second, the hot holes and electrons can overcome the potential energy barrier between the silicon and the SiO_2 (~3.1 eV), thereby causing *hot carriers to become injected into the gate oxide*. Each of these events brings about its own set of repercussions.

5.6.1 Substrate Currents Due to Hot Carriers

When electron-hole pairs are created by impact ionization, the electrons are normally attracted to the drain, and they add to the drain current. The *holes*, on the other hand, enter the substrate and constitute a part of the parasitic substrate current (*Process 3* in Fig. 5-27). This substrate current, I_{sub}, can itself produce several problems:

- If a substrate bias-generator circuit is included on-chip, the output of the bias generator will be less negative with increasing substrate current.

- If some of these holes are collected by the source (instead of by the body contact), *and* this collected hole current causes a voltage drop in the substrate material on the order of 0.6 V, the substrate-source *pn* junction will conduct significantly. Electrons will then be injected from the source to the substrate, just like electrons injected from emitter to base of an *npn* transistor (the *forward injection* shown in Fig. 5-27). These electrons can, in turn, gain sufficient energy as they travel toward the drain to cause additional impact ionization and create new electron-hole pairs. A positive-feedback mechanism thus exists, one that can sustain itself if the drain voltage exceeds a certain value. This is observed externally as a form of breakdown, referred to as a *snapback breakdown*. A particularly clear explanation of this effect, including the reason for the negative-resistance, or "snapback," portion of the curves, is given in reference 3.

- As some of the holes are accelerated through the drain-substrate depletion region, they may acquire enough energy to cause secondary impact ionization. This will create electrons far from the drain region. Some these electrons, instead of being collected by the drain, may escape the drain field and instead travel (sometimes hundreds of microns) to other nodes on the chip. This may lead to a reduction of the storage time of dynamic circuit nodes in DRAMs (i.e., manifested as a degradation of the refresh time).[33] This excess electron current is reported to be around 10^{-4} times smaller than the substrate ionization impact current itself.

- I_{sub} may induce latchup in CMOS circuits.

The magnitude of the substrate current depends exponentially on the value of E_M. An example of how I_{sub} depends on decreasing channel length is seen in Fig. 5-28.[34] This shows the maximum I_{sub} generated at a voltage of 5 V versus the effective channel length for MOS transistors processed with the same technology. The magnitude of I_{sub} would increase even more rapidly with shrinking L_{eff} if the oxide thickness and junction depth were scaled together with the channel length.

5.6.2 Hot-Carrier Injection into the Gate Oxide

The hot holes and electrons that are injected into the gate oxide cause another set of problems. First, some of these carriers pass to the gate electrode (mostly hot holes) and thus constitute a gate current, I_G, typically in the fA (10^{-15} A) range. For higher, but still nondestructive, biases, the gate current can grow rapidly to become several pA (10^{-12} A).

However, some fraction of the hot carriers injected into the gate oxide do not reach the gate electrode. This is because the gate oxide contains empty electron states (also known as *traps*), which can be filled by the injected hot carriers. Such occupied traps generally become electron traps, even if they are initially filled with holes. As a result, there is a negative charge density caused by the trapping of the hot carriers in the oxide. This *trapped charge* acts like a contribution to the fixed oxide charge term in the expression for the device threshold voltage. Furthermore, the trapped charge accumulates with time. Due to the polarity of the trapped charge, the resulting shift in the *n*-channel device threshold voltage is *positive* (i.e., V_T increases in NMOS devices). If I_G becomes of the order of pA, the trapping of a fraction of the electrons injected into the oxide can become a significant effect (the fraction that gets trapped is ~1 in 10^6 injected electrons). As a result, this increase in *negative* stored charge can lead to a permanent change in V_T in the MOSFET. (The means by which charge in the oxide impacts V_T was described earlier.)

EXAMPLE 5-6: V_T Shift due to Hot-Electron Effects. Given a short-channel NMOS device with t_{ox} = 20 nm, a gate *width* of 5 μm, and a gate current of 1 nA that is momentarily caused by hot carrier injection. The current flows through a region of the gate oxide near the drain end which is 5 μm wide, and 0.2 μm long. Assume that 1 in 10^6 electrons of the gate current becomes trapped at an average distance of 0.1 t_{ox} from the Si-SiO_2 interface. How long does the gate current have to flow to change V_T by 0.1 V in the region where the injection is occurring?

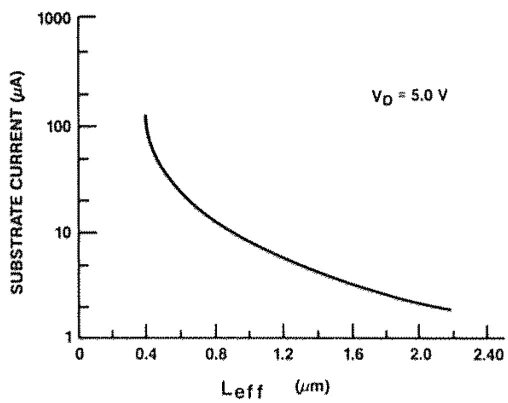


Fig. 5-28 The maximum substrate current due to impact ionization produced at a drain voltage of 5 V vs effective channel length for 250Å gate oxide transistors.[34] (© 1986 IEEE).

SOLUTION: The gate current of 1 nA produces a current density across the injection area of

$$10^{-9} \text{ nA} / 1\times10^{-8} \text{ cm}^{-2} = 0.1 \text{ A cm}^{-2}$$

Since 1 in 10^6 electrons become trapped, the rate of charge trapping, J_{ot}, in the gate oxide is $0.1 \times 10^{-6} = 1 \times 10^{-7}$ C sec^{-1} cm^{-2}. The shift in V_T due to the trapped charge in the oxide can be found from

$$\Delta V_T = (1/C_{ox})\ (0.9\ t_{ox}/t_{ox})\ \Delta Q_{tot}$$

or, solving for ΔQ_{tot},

$$\Delta Q_{tot} = C_{OX}\ \Delta V_T / 0.9 = [1.7\times10^{-7} \times 0.1]/0.9$$

$$= 1.89 \times 10^{-8} \text{ C/cm}^2$$

The time, t, necessary to trap this quantity of charge is

$$t = \Delta Q_{tot}/J_{ot} = 1.89\times10^{-8}/1\times10^{-7} = 0.19 \text{ sec}$$

5.6.3 Device-Performance Degradation Due to Hot-Carrier Effects

The increase that occurs in V_T leads to other changes in the MOS characteristics. First, the saturation current decreases because (V_G - V_T) becomes smaller (Fig. 5-29a).[77] Second, as the substrate current increases, the transconductance decreases. Finally, since the trapped charge accumulates with time, the device performance will become unacceptable for a given application after a certain time of device operation.

A device lifetime, τ, can therefore be defined by selecting the maximum percentage of allowed degradation in the critical device parameter. It has been found, however, that this lifetime (regardless of whether V_T or g_m is monitored) can be related to I_{sub} by a power-law relationship (i.e., τ is observed to be inversely proportional to I_{sub} when plotted on a log-log plot). Figures 5-29b and c show τ as a function of I_{sub} for: (a) τ is defined as a 10-mV shift in V_T; and (b) τ is defined as a 10% degradation in g_m.

This has led to accelerated testing techniques that stress the devices with higher drain voltages than would be used in normal operating conditions.[35] A larger I_{sub} is thereby produced, leading to shorter times to device failure (i.e., possibly defined by a 10-mV shift in V_T or a 10% degradation in g_m). The time to failure under normal operating drain voltages (τ) is then extrapolated from these data using curves as shown in Figs. 5-29b and 5-29c.[36,37] Device design can then be modified to yield a desired lifetime (typically, 10 years) under normal operating conditions.

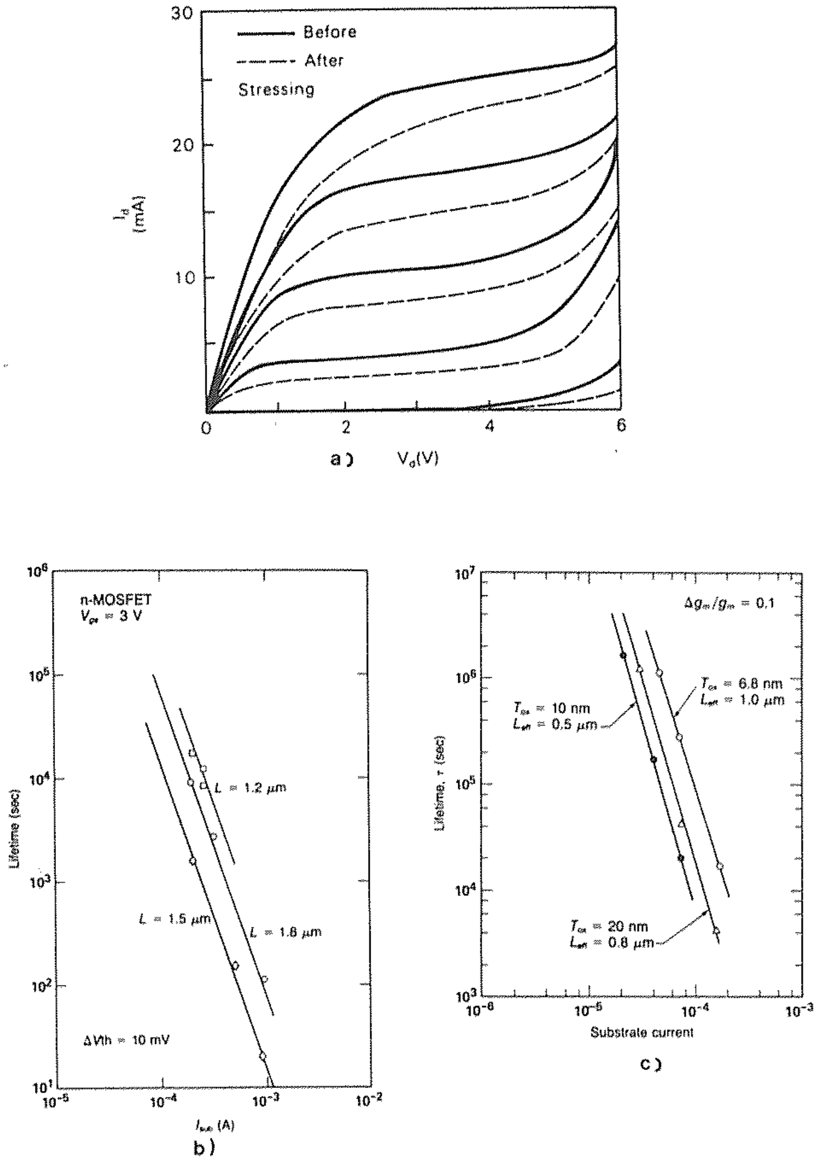


Fig. 5-29 (a) Degradation of I_{DSsat} after stressing.[77] (© 1984 IEEE). Degradation of other device-performance parameters as a function of substrate current (I_{sub}): (b) Lifetime defined as a 10-mV shift in V_T; (c) Lifetime defined as a 10-percent degradation in g_m.

5.6.4 Techniques for Reducing Hot-Carrier Degradation

Device performance degradation from hot electron effects can be reduced by a number of techniques. These include the following:

- The voltages applied to the device can be decreased (e.g., by lowering the power-supply voltage from 5 V to 3.3 V). The decision to implement this reduction, however, is not in the hands of the device designer or fabricator. The issue of reduced supply voltages for submicron MOS technologies is covered in chapter 6, section 6.7.2.
- The time the device is under the voltage stress can be shortened (e.g., by using a lower duty cycle and clocked logic).
- Appropriate drain engineering design techniques, which results in special drain structures that reduce hot electron effects in MOS devices (i.e., double-diffused drains and lightly-doped drains), can be implemented.
- The density of trapping sites in the gate oxide can be reduced through the use of special processing techniques.
- Thin, lightly doped epitaxial-layers on low-resistance substrates can be employed to shunt away substrate current and help eliminate the problem of impact-ionization-induced latchup.

We will next discuss the details of the latter three approaches since their implementation is the task of device and process engineers.

5.6.5 Lightly Doped Drains (LDD)

It has been determined that hot-carrier effects will cause unacceptable performance degradation in NMOS devices built with conventional drain structures if their channel lengths are less than 2 μm. To overcome this problem, such alternative drain structures as *double-diffused drains* and *lightly doped drains* (LDDs) must be used. The purpose of both types of structures is the same – namely, to absorb some of the potential into the drain and thus reduce E_M. Double-diffused drains, however, are less effective for short-channel devices (i.e., $\leq$1.25 μm) because they cause deeper junctions and more overlap capacitance. We will thus restrict our discussion to LDDs. Table 5-2 shows the evolution in AT&T's Twin-Tub CMOS technology, with respect to such drain structures.[30]

In the LDD structure, the drain is formed by two implants (Fig. 5-30). One of these is self-aligned to the gate electrode, and the other is self-aligned to the gate electrode on which two oxide *sidewall spacers* have been formed. (An extensive report on the details of sidewall spacer technology for both MOS and bipolar devices can be found in ref. 66.) The purpose of the lighter first dose is to form a lightly doped section of the

Table 5.2 Evolution of Device Structures in AT&T's Twin-Tub CMOS Technology Development[30] (Twin-Tub VI announced in 1989)[83]

	Twin-Tub I	Twin-Tub II	Twin-Tub III	Twin-Tub IV	Twin-Tub V	Twin-Tub VI
Design Rule	3.5 μm	2.5 μm	1.75 μm	1.25 μm	0.9 μm	0.6 μm
L_{eff}	2 μm	1.5 μm	1.3 μm	1.0 μm	0.75 μm	0.4 μm
t_{ox}	600 Å	350 Å	250 Å	200 Å	150 Å	125Å
Device Structure	Conventional	Conventional	DDD	LDD	N&P LDD	N&P LDD

drain at the edge near the channel. In NMOS devices, this dose is normally $1\text{-}2\times10^{13}$ atoms/cm^2 of phosphorus.

The value of E_M is reduced by this structure because the voltage drop is shared by the drain *and* the channel (in contrast to a conventional drain structure, in which almost

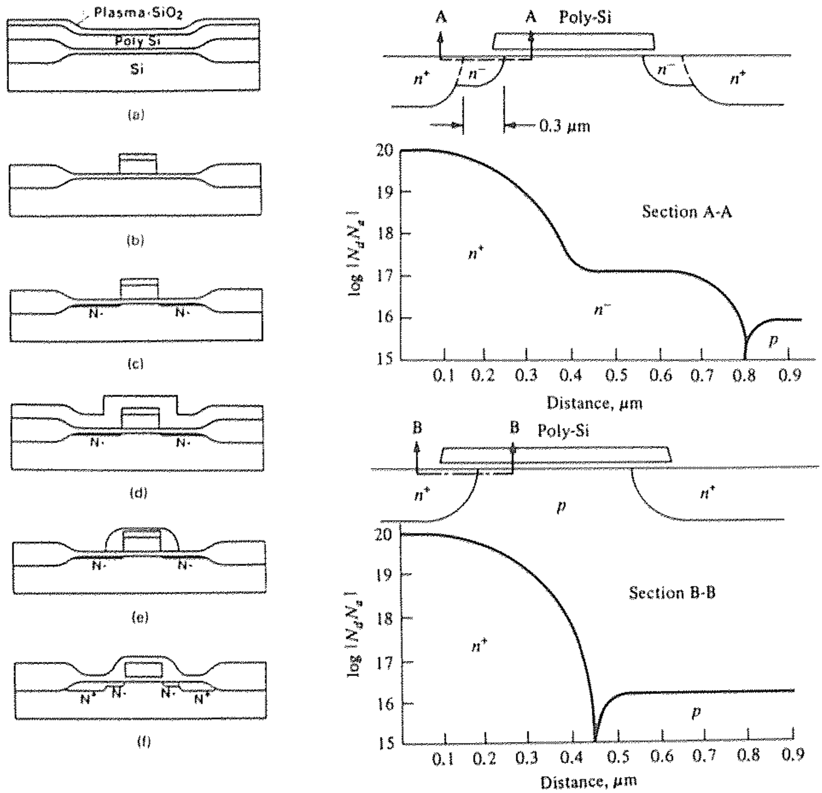


Fig. 5-30 (a) Process sequence used to form lightly doped drain (LDD) structures. (b) Doping profile in an LDD structure taken through section A – A. (c) Doping profile in a conventional drain structure taken through section B – B.

the entire voltage drop occurs across the lightly doped *channel* region). Figure 5-25 shows the electric-field profile at the drain of a MOSFET, both with and without an LDD structure.[38] We see that the electric field can be reduced by about 30-40%. Since the hot-electron-induced gate currents are exponentially dependent on E_M, this is sufficient to reduce currents by many orders of magnitude. As a result, the stability of the device is greatly increased. To ensure a high quality interface under the sidewall-spacer oxide, it is important that the thermally grown gate oxide remain in place after the polysilicon gate etch. This requires a polysilicon etch process with high selectivity to oxide, since the gate oxide is typically less than 20 nm.

The heavier, second dose forms a low resistivity region of the drain region, which is also merged with the lightly doped region. (In NMOS devices this implant is typically arsenic at a dose of about $1x10^{15}$ atoms/cm^2.) Since it is further away from the channel than would be the case in a conventional drain structure, the depth of the heavily-doped region of the drain can be made somewhat greater without adversely impacting the device operation (e.g., 0.3 μm deep versus 0.18 μm deep). The increased junction depth lowers both the sheet resistance and the contact resistance of the drain (see chap. 3). Deeper junctions also provide better protection against junction spiking.

The disadvantages of LDD structures are their increased fabrication complexity compared to conventional drain structures and the increased parasitic resistance of the source and drain regions caused by the lightly doped regions of the drain. This increase in parasitic channel resistance results in devices that dissipate more power for a constant applied voltage.

The effect on the I_D-V_{DS} curves of the MOSFET due to the additional voltage drop across the two lightly doped regions is that I_D does not saturate until a higher value of V_{DS} is applied. Due to the extra series resistance, the total channel conductance is appreciably lower in the linear region, but only slightly lower in the saturation region. This is because the channel region in saturation already has a high channel resistance, while in the linear region the resistance is much lower. The additional series resistance of the LDD therefore does not significantly increase the total resistance of the MOSFET in saturation, and I_D remains less affected. Consequently, the drain current is reduced more significantly in the linear region than in the saturation region.)

The I-V curves for LDD devices are consequently recognizable by their characteristically round shape, as shown in Fig. 5-31, curve 6. These curves also illustrate progressively the other short-channel effects described in section 5.5, beginning with the I-V curve of an ideal long-channel MOS device (Curve 1). Thus, curve 6 shows the I-V curve of a short-channel MOSFET with an LDD.

5.6.5.1 Drain Engineering for Optimum LDD Structures. A proper LDD design should provide adequate hot-carrier protection while not introducing excessive source/drain resistance, which would degrade the device performance. The design and fabrication effort undertaken to achieve such an optimum LDD structure is referred to as *drain engineering*; this as we shall see, is a relatively complex task. A thorough approach to designing adequate hot-electron protection would entail addressing all of the following objectives:

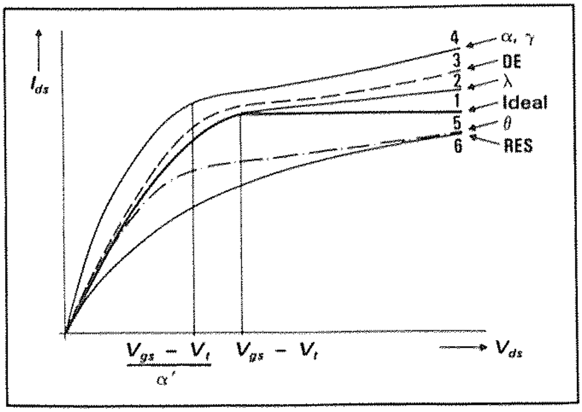


Fig. 5-31 Progressive influence of short-channel effects and LDD on the characteristics of an ideal, long channel MOSFET (Curve 1). Curve 6 shows the I-V curve of a short channel MOSFET that is also fabricated with an LDD, while Curve 5 shows the characteristic of the same device without an LDD.[28] (© 1986 IEEE).

1. Reducing the maximum electric field in the silicon as much as possible.
2. Ensuring that the injection position (i.e., the E_M point) is located under the gate edge.
3. Ensuring that the impact ionization region is pushed far below silicon surface to reduce the possibility of hot carriers reaching the $Si\text{-}SiO_2$ interface.
4. Separating the point where the electric field in the silicon is a maximum from the point of maximum current flow in the channel.
5. Minimizing the increase in the parasitic resistance due to the LDD structure.

The degree by which E_M is reduced (*Objective #1*) depends primarily on the doping value of the lightly doped extension of the drain, as well as on its length. If the doping level is too high, the value of E_M is not sufficiently reduced, and hot-electron protection is not provided. If the level is too low, the drive-current capability of the device is severely reduced, and the surface will easily be depleted by the hot carriers that do get trapped in the gate oxide above it (i.e., the device will again be vulnerable to hot-electron degradation). A model which calculates the electric field in LDD structures has been published.[72] It concludes that the lightly doped region should be long enough to attenuate the electric field to a value that is below the critical ionization field, but should still be short enough to keep the series resistance from becoming excessive. In early LDD structures in NMOS this meant that the primary parameters selected were the n^- length and its doping concentration.[39]

A process for forming LDD structures in *either* or *both* PMOS and NMOS structures in CMOS with only two photomasks was reported by Parrillo, et al.[40] An extension of this process also uses removable TiN gate sidewall spacers and incorporates self-alignment of the lightly doped regions to their respective gates without overcompensation (Fig. 5-32). Another example LDD structure in 0.8-μm CMOS is given in reference 41.

To keep the location of the E_M point under the gate (*Objective #2)* and yet maintain high drivability requires the proper combination of gate-drain (g/d) overlap length and spacer length. The g/d overlap length should be long enough to let the E_M point lie under the gate, and the spacer length should be short enough to let the n^+ region reach under the gate. In general, for process controllability, if a 0.3-μm-thick polysilicon film is used for the gate, a 0.2-μm-wide spacer is selected.

To attain *Objective #3*, a metal-coated LDD structure has been developed.[42] It uses a deeper n^- phosphorus profile than the n^+ source-drain (arsenic), in order to steer the maximum current path away from E_M (Fig. 5-33b).

A buried LDD structure has also been proposed to reach *Objectives #3 and #4*. This structure uses a retrograde LDD profile, with the peak concentration below the Si surface (Fig. 5-33c). Besides a reduction in substrate current similar to the metal-coated LDD, this profile also suppresses hot-carrier injection by driving the current away from the surface and shifting the avalanche region further into the bulk silicon.

The buried LDD approach together with an additional shallow arsenic implant (Fig. 5-34d), may allow NMOS transistors of 0.6 μm to be built with adequate hot-electron protection and with minimal the increase in parasitic resistance due to the LDD structure (*Objective #5*).[30,43,44] The idea is to form an abrupt profile at the silicon surface underneath the spacer to reduce the series resistance. The buried LDD profile is

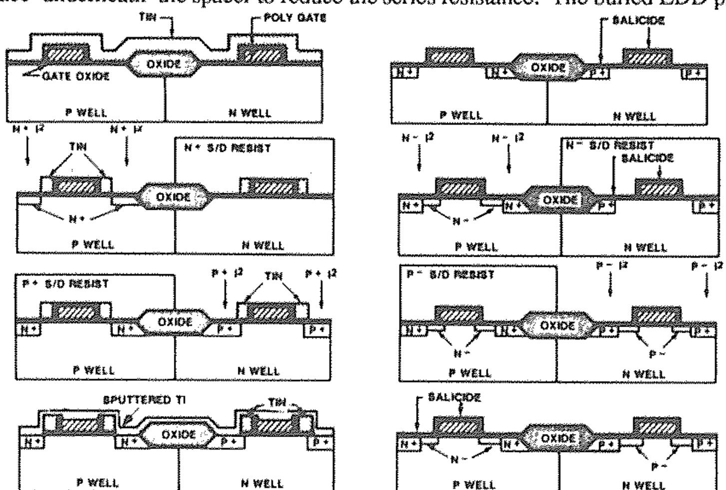


Fig. 5-32 Process sequence showing the TiN removable spacer process.[82] (© 1989 IEEE).

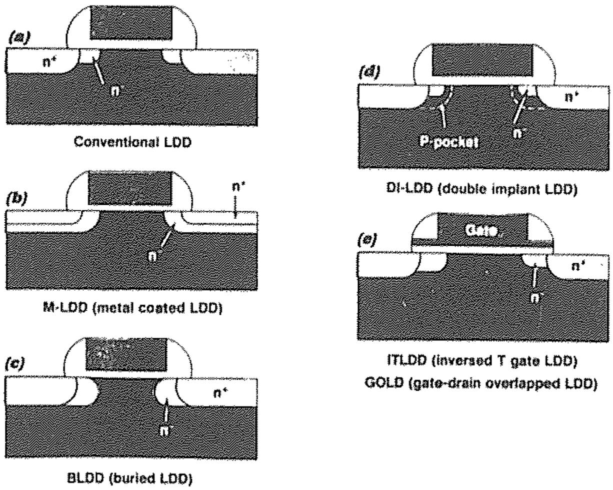


Fig. 5-33 Conventional LDD drain engineering and some of its variations – all designed to reduce hot-carrier effects – are shown.[30] (© 1986 IEEE).

used to force the main current path and the impact-ionization region deep into the silicon. There are two peaks in the lateral electric-field distribution, and the maximum current path is through the saddle point of these two high field regions. Thus, this device structure improves both the current drive and the hot-carrier resistance. A further modification of the buried LDD is the *sloped-junction LDD (SJLDD)*, shown in Fig. 5-33e.[45] In this structure a 165-keV phosphorus implant is used to form the lightly doped drain region; this provides improved device lifetime under high field stress. The main reason for the improvement is that hot-carrier generation is driven further away from the Si-SiO_2 interface by the junction of the SJLDD.

5.6.5.2 Asymmetrical Characteristics of LDD MOSFETs. Formation of the lightly doped regions of LDDs is accomplished by means of ion implantation, with the polysilicon gate used as the implant mask. This can cause asymmetric doping of the source and drain regions,[46,64] with such asymmetry occurring as a result of implanting off-axis (typically, at an angle of 7°) to avoid channeling (see Vol. 1, chap. 9). This produces lateral shadowing (S in Fig. 5-34) of the substrate on one side of the poly gate, and penetration of dopants through the leading corner of the poly on the other side. The problem is compounded by etch processes that produce a re-entrant angle in the poly gate sidewall (Fig. 5-34 - poly etch process A.) The impact of this effect on the device characteristics can be examined by considering the implants to the lightly doped and the heavily doped regions separately.

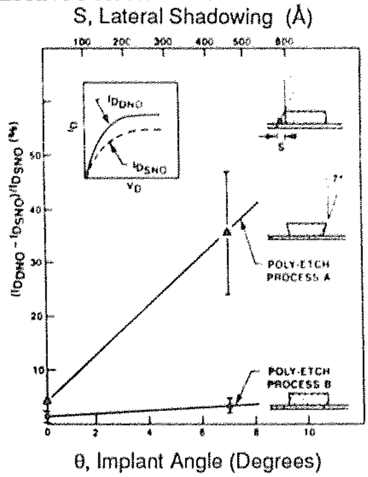


Fig. 5-34 Percent asymmetry in saturated drain currents ($V_{DS} = V_{GS} = 5$ V) versus source/drain implant angle for LDD NMOS devices having reentrant- and vertical-wall polysilicon gates. Insert schematically shows source-non-overlap (SNO) and drain-overlap (DNO) I-V behavior.[40] (© 1986 IEEE).

If the implant that forms the lightly doped region of the drain (usually phosphorus in NMOS devices) is done off-axis, the as-implanted region in the shadow of the gate will have its edge displaced from the gate edge. As a result, there will be *no overlap* of the implanted region and the gate edge. Even after a drive-in step, the overlap of this edge with the gate will be less than the overlap of the nonshadowed implant region edge and the gate. If the less-overlapped side is used as the source, the extra resistance introduced will reduce the drain current (I_{Dsno} in Fig. 5-34), thus degrading circuit speed. However, if the drain side has less overlap, the extra resistance will not affect I_D (curve I_{Ddno} in Fig. 5-34). The asymmetry in drive current when the same device is connected in these two ways can be as much as 40%; this can be catastrophic in circuits that require closely matched device characteristics.

The problems arising from such n^- implant-shadowing can be alleviated by one of the following measures:

- Insuring that the poly-etch process produces vertical sidewalls in the poly gates.

- Avoiding excessive reoxidation of the polysilicon after its definition (note that this also prevents a gap from being established between the drain region edge and the edge of the gap). A dry-oxidation step is often employed to reduce the accelerated oxidation rate on n^+ polysilicon.

- Implanting with a vertical implant through a screen oxide[27] (which reduces the channeling that would otherwise occur when a vertical implant is performed).

- Using a poly-etch process that produces a slight positive bevel to the poly sidewall profile, and combining this etch with a smaller angle (3° vs 7°) off-axis implant.[46]

Let us next examine how the implantation of the heavily doped region impacts the structure of the LDD. If an off-axis implant is again used with the heavily doped implant, asymmetric implant effects can again alter the LDD structure. That is, on the side of the gate that is *not* shadowed, the arsenic atoms penetrate the leading corner of the spacer and thereby enter the substrate under the spacer to some extent. If the spacer is too narrow, this can wipe out the lightly doped drain-extension region. When this happens, the device can lose the hot-carrier protection that was to be provided by the LDD structure, and it will become vulnerable to degradation by hot-carrier effects. To avoid this, the spacer must be sufficiently wide. Some guidelines for choosing the proper spacer width to deal with this problem are given in reference 47.

The general problem involving the gate-to-drain overlap with respect to its impact on the MOSFET characteristics was reviewed by Ko, et al.[48] It was observed that the critical dimensions in drain structures having weak overlap (WO) to the gate were only tens of nm in devices in which $L_{eff} = 1\ \mu m$. In general, the drain current decreases as the overlap is weakened, a double hump appears in the substrate current of asymmetrical WO devices, and the reliability of the WO devices can be lower than devices having adequate overlap. To avoid random degradation of device performance by WO effects that arise as a result of process variations, stringent process-control measures must be exercised when submicron devices are fabricated.

5.6.6 The Impact of IC Processing on Hot-Carrier Device Degradation

The lifetime of a device in which the hot-electron degradation impacts device performance can be increased by keeping the number of trapping centers in the gate oxide to a minimum. In essence, this reduces the density of states that the hot electrons injected into the oxide can occupy. Maintaining a minimum number of trapping centers can be achieved in several ways during the device-fabrication process sequence. First, the gate oxide should be grown by a process that produces low interface-trap densities (such processes are described in Vol. 1, chap. 7). It has also recently been reported that the incorporation of optimized amounts of fluorine during the gate-oxide growth process suppresses interface-state generation during injection.[49] Second, the degradation of the oxide caused by damage during process steps carried out in a plasma environment should be minimized (damaging processes include plasma-enhanced CVD of oxides and nitrides, RIE, and sputter deposition of metal). Finally, the amount of hydrogen that is incorporated at the $Si\text{-}SiO_2$ interface should be reduced. The latter two topics will be considered in more detail here.

In typical CMOS processing, most of the damage created by RIE or ion implantation is annealed out at a high temperature (above 800°C) before metallization. Once the first layer of aluminum is deposited, however, the maximum annealing temperature is limited to ~450°C. It has therefore been reported that the RIE of the second layer of metal in multilevel metal processes deteriorates the device-aging characteristics. It has been found that while a subsequent anneal performed at 375°C is ineffective in annealing out this damage,[50] an anneal at 450°C improves device-aging characteristics.

Excess hydrogen at the Si-SiO_2 interface is also reported to be a culprit in increasing the density of the interface states. In most cases, hydrogen is used to fill the dangling bond, forming Si-H at the Si-SiO_2 interface. However, the Si-H bond can be easily broken by injected hot electrons. Furthermore, excess hydrogen introduced during processing can diffuse to the interface and lead to enhanced bond-breaking behavior. Such excess hydrogen can arrive during the final sinter step before passivation (i.e., in H_2 or N_2 + 5% H_2), or during the deposition of a conventional silicon nitride passivation layer. It has therefore been reported that sintering in pure N_2 produces devices with a lower device degradation rate.[51] Similarly, it has been reported that when a fluorinated nitride (F-SiN) is used as a passivation layer,[50,52] devices exhibited slower degradation rates than those seen in devices with conventional SiN passivation layers. F-SiN films can be produced by incorporating NF_3 in the deposition process to form an Si-F bonding structure instead of Si-H in this film.

The presence of fluorine in the gate oxide (possibly inadvertently originating from a BF_2^+ source/drain implant implant) has recently been reported to increase the hot electron resistance of devices fabricated with such oxides.[74] Other reports confirm that a deliberate incorporation of fluorine into the gate oxide (from implanting fluorine into a polysilicon film, and then diffusing it into the gate oxide) produces a more hot-electron resistant interface.[78,79]

5.6.7 Hot-Carrier Effects in PMOS Transistors

Hot-carrier effects are not significant in PMOS transistors at channel lengths greater than 1 μm because the impact-ionization rate of holes is 3 to 4 orders of magnitude lower than that of electrons at a given electric field.[53] At submicron channel lengths, however, hot-electron effects in PMOS start to become important. It has been reported that two hot-carrier effects predominate in such PMOS devices. Both are caused by hot electrons that are generated by impact ionization and are then injected into the oxide, becoming trapped. Hot holes do not appear to cause significant effects unless a device is stressed at large magnitudes of V_G. In PMOS devices, the gate-bias polarity favors electron injection (which is opposite to that in NMOS) but retards the injection of holes.

In the first degradation effect, these electrons are trapped near the drain and shorten the effective channel length. This reduction in L_{eff} is even more severe in the buried-channel PMOS devices (used in CMOS technologies that have an n^+ polysilicon gate), because their buried-channel nature makes them more vulnerable to source-drain punchthrough,[54] and consequently increases subthreshold leakage. This effect is called *hot-electron-induced punchthrough* (HEIP).

The second effect was observed in PMOS devices with p^+, as well as n^+, polysilicon gates. In these devices, the injected and trapped hot electrons also produce V_T shifts (which reduce the magnitude of V_T) and increases in the transconductance, g_m. Nevertheless, contrary to what occurs in NMOS, in PMOS these effects tend to saturate. This is explained by the fact that the electrons trapped near the drain reduce the electric field present there in PMOS devices.[55] As a result, this effect does not appear to limit PMOS devices fabricated with p^+ polysilicon gates as long as L_{eff} is ≥0.6 μm.[56] Still, LDD structures appear to be needed for *p*-channel devices fabricated with *n^+ polysilicon gates* when L_{eff} gets to be 0.8 μm or less.[54,57]

The LDD structure also helps to reduce subthreshold leakage in PMOS devices, where excessive junction depths caused by lateral diffusion have been a problem. The use of LDD basically increases L_{eff}, hence compensating for the L_{eff} reduction caused by hot-electron injection. A new LDD structure for PMOS devices, called a *halo LDD*, is described in reference 63. In this structure, a deeper phosphorus implant is placed below the lightly doped drain-extension *p*-type implant. The punchthrough resistance of the PMOS device is reported to be significantly improved by this LDD structure.

5.6.8 Gate-Induced Drain-Leakage Current

Another type of leakage current between drain and substrate in thin gate-oxide (12-28 nm) MOS devices has been observed at drain voltages much lower than the breakdown voltage.[58,59] This is not really a hot-electron-induced effect, but since it uses LDD structures to reduce its magnitude, we describe it in this section. The basis of this current is band-to-band tunneling that occurs through the gate oxide into the deep depletion layer in the gate-to-drain overlap region. It has been reported that in order for this leakage current to limited to less than 0.1 pA/μm^2, the oxide field in the gate-to-drain overlap region must be limited to 1.9 MV/cm. LDD structures are effective in reducing this leakage current, but only if the doping in the n^- extensions is less than $10^{18}/cm^3$. Buried LDD structures with the peak of the *n* concentration at several hundred angstroms beneath the Si-SiO_2 interface appear to be even more effective than conventional LDD structures in reducing the effect. This effect is predicted to become more of a problem as devices continue to be scaled - that is, if a 10-nm-thick gate oxide is used, the E-field will be less than 1.9 MV/cm in the oxide only if a 3.1 V power supply is used.

REFERENCES

1. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed., New York, John Wiley & Sons, 1986.
2. R. F. Pierret, *Field-Effect Devices*, Vol. IV in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1983.
3. D. K. Schroder, *Advanced MOS Devices*, Vol. VII in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1987.
4. Y. P. Tsividis, *The Operation and Modeling of the MOS Transistor*, New York, McGraw-Hill Book Co., 1987.

5. E. H. Nicollian and J.R. Brews, *MOS Physics and Technology,* Wiley-Interscience, New York, 1982.
6. S. M. Sze, *Physics of Semiconductor Devices,* 2nd Ed. Wiley, New York, 1981.
7. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits,* McGraw-Hill Book Co., New York, 1983.
8. R. A. Chapman et al., *Tech. Dig. IEDM* , 1987, p. 362.
9. R. J. Bayruns et al., *IEEE J. Solid-State Circuits,* **SC-19**, 1984, p. 755.
10. H. L. Armstrong, "A Theory of Voltage Breakdown of Cylindrical P-N Junctions," *IRE Trans. Electron Dev.*, **ED-4**, 15 (1957).
11. A. S. Grove, *Physics and Technology of Semiconductor Devices,* Wiley, New York, 1967, Chap. 6
12. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Tech.*, **24**, May 1981, p. 65.
13. C. M. Osborn and E. Bassous, *J. Electrochem. Soc.*, **122**, 89, (1975).
14. S. M. Sze, *Semiconductor Devices: Physics & Tech.,* Wiley, New York, 1985.
15. N. Kotani and S. Kawazu, *Solid-State Electronics,* **22**, p. 63, (Jan. 1979).
16. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, **ED-31**, 1984, p. 205.
17. C. Mead and L. Conway (Eds.), *Introduction to VLSI Systems,* Reading, MA, Addison-Wesley, 1980.
18. A. Mukherjee, *NMOS and CMOS VLSI Systems Design,* Prentice-Hall, Englewood Cliffs, N .J., 1986.
19. N. Kotani and S. Kawazu, *Solid-State Electronics,* **22**, 63 (January 1979).
20. L. D. Yau, *Solid-State Electronics,* **17**, 1059, (1974).
21. L. M. Dang, *IEEE J. Solid-State Circuits,* **SC-14**, 358 (1975).
22. G. Merkel, *Process and Device Modelling for IC Design* (F. Van de Wiele, W. L.Engle, and P.G. Jespers, Eds.), Noordhoff, Leyden (1977), p. 705.
23. R. R. Troutman, "VLSI limitations from Drain-Induced Barrier Lowering," *IEEE Trans. Electron Dev.*, **ED-26**, p.461, 1979.
24. J. Zhu et al., "Punchthrough Current for Submicron CMOS VLSI," *IEEE Trans. Electron Dev.,* Feb. 1988, p. 145.
25. K. M. Cham, *et al., Computer Aided Design and VLSI Device Development,* Kluwer Academic Publishers, Boston, 1986, p. 194.
26. H. Nihara, et al., "Anomalous Drain Current in NMOS and its Suppression by Deep Ion Implantation," *IEDM Tech. Dig.*, p. 487 (1978).
27. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconduct International,* April 1988, p. 64.
28. C. Duvvury, "A Guide to Short-Channel Effects in MOSFETs," *IEEE Circuits and Systems Magazine,* Nov. 1986, p. 6.
29. M. Kakamu et al., "Power Supply Voltage for Future CMOS VLSI in Half- and Sub Micrometer," *Tech. Dig. IEDM,* 1986, p. 399.
30. M. L. Chen, "CMOS Hot-Carrier Protection with LDD," *Semiconductor Internat.,* April 1988, p. 78
31. P. K. Ko, "Hot Electron Effects in MOSFETs," Doctoral Thesis, Dept. of EECS, University of California, Berkeley, June 1972.

32. T. Y. Chan, P.K. Ho, and C. Hu, *IEEE Electron Dev. Letts.*, **EDL-6**, p. 551, 1985.
33. P. K. Chatterjee, *Tech. Dig. IEDM*, 1979, p. 14.
34. M. L. Woods,"MOS VLSI Reliability and Yield Trends," *Proceedings IEEE*, Dec. 1986, p. 1715.
35. P. Yang and S. Aur, "Modelling of Device Lifetime due to Hot Carrier Effects," *Proc. of 1985 Internat. Symp. on VLSI Tech.*, p. 227.
36. C. Hu et al., *IEEE J. Solid-State Circuits*, **SC-20**, 1985, p. 295.
37. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation due to Hot Carrier Injection," *IEEE Trans. Electron Dev. Letts.*, **EDL-4**, p. 111, 1983.
38. S. Ogura et al., *IEEE Trans. Electron Dev.*, **ED-27**, 1980, p. 1359.
39. H. Mikoshiba, "Comparison of Drain Structures in n-Channel MOSFETs," *IEEE Trans. Electron Dev.*, Jan. 1986, p. 140.
40. L. C. Parrillo et al., *IEDM Tech. Dig.*, 1986, p. 244.
41. R.A. Chapman et al., *IEDM Tech. Digest*, 1987, p. 362.
42. Y. Tsunashima et al., *Proc. of VLSI Symposium*, 1985, p. 114.
43. T. Toyoshima et al., *Proc. of VLSI Symposium*, 1985, p. 362.
44. T. Noguchi, *Tech. Dig. IEDM*, 1986, p. 730.
45. S. Jain et al., *IEEE Electron Dev. Lett.*, Oct 1988, p. 539.
46. T. Y. Chan et al., *Electron Dev. Lett.*, Jan. 1986, p. 16.
47. J. R. Pfiester and F.K. Baker, "Asymetrical High Field Effects in Submicron MOSFETs", *EDM Tech. Dig.*, 1987, p. 51.
48. P. K. Ko et al., *Tech Dig. IEDM*, 1986, p. 292.
49. E. F. Da Silva, *Tech. Dig. IEDM*, 1987, p. 848.
50. M. -L. Chen, "Hot Carrier Aging in Two-Level Metal Processing," *Tech. Dig. IEDM*, 1987, p. 55.
51. F. C. Hsu, *Proc. of VLSI Symp.*, 1985, p. 108.
52. S. Fujita et al., *Tech. Dig. IEDM*, 1985, p. 64.
53. J. Y. Chen, "CMOS-The Emerging VLSI Technology," *IEEE Circuits and Devices Magazine*, March, 1986, p. 16.
54. M. Koyanagi et al., *IEEE Trans. Electron Dev.*, **ED-34**, 1987, p. 871.
55. C. Hu et al., *IEEE J. Solid-State Circuits*, **SC-20**, 1985, p. 295.
56. Y. Hiruta et al., *Tech. Dig. IEDM*, 1986, p. 718.
57. T. Mizuno et al., *Tech. Dig. IEDM*, p.726.
58. T. Y. Chan, J. Chen, P.K. Ko. and C. Hu, *IEDM Tech. Dig.* 1987, p. 718.
59. C. Chang and J. Lien, *IEDM Tech. Dig.* 1987, p. 714.
60. C. T. Sah, "The Evolution of the MOS Transistor," *Proceedings of the IEEE*, Oct. 1988, p. 1280.
61. P. Balk, "Effects of Hydrogen Annealing on Silicon Surfaces," presented at the *Electrochem. Soc. Spring Meeting*, May, 1965. Ext Abs. No. 109, p. 237.
62. M.-C. Jeng et al., "Design Guidelines for Deep-Submicrometer MOSFETs," *Tech. Dig. IEDM*, 1988, p. 386.
63. M. L. Chen et al., *Tech. Dig. IEDM*, 1988, p. 390.
64. R. W. Gregor, "Consequences of Ion Beam Shadowing in CMOS Source/Drain Formation," *IEEE Electron Dev. Letts.*, Dec. 1986, p. 677.

65. I. Ahmed, H. Naguib, and C. Gomez, *Ext. Abs. ECS Meeting,* Fall 1988, Abs. No. 282, p. 405.

66. S. H. Dhong and E. J. Petrillo, *J. Electrochem. Soc.*, Feb. 1986, p. 389.

67. F. Hsu and K. -Y. Chiu, *IEEE Electron Dev. Letts.*, **EDL-5**, p. 162, 1984.

68. K. - Y. Toh, P.-K. Ko, and R. G. Meyer, *IEEE J. Solid-State Ckts,* Aug. 1988, p. 950.

69. J. W. Brews et al., "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Lett.*, **EDL-1**, p. 2, 1980.

70. R. H. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits,* **SC-9**, Oct. 1974, p. 256.

71. S. G. Chamberlain and S. Ramanan, *IEEE Trans. Electron Dev.*, Nov. 1986, p. 1745.

72. K. Marayam, J. C. Lee, and C. Hu, *IEEE Trans. Electron Dev.*, July 1987, p. 1509.

73. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Soild-State Electronics,* vol. 17, p. 1059, 1974.

74. P. Wright et al., *IEEE Electron Device Letts.,* August 1989, p. 347.

75. V. L. Rideout, F. H. Gaensslen, and A. LeBlanc, "Device Design Consideration for Ion-Implanted *n*-Channel MOSFETs," IBM J. Res. Dev., p. 50, (1975).

76. R. Siqusch et al., *Tech. Dig. IEDM,* 1980, p. 429.

77. F. H. Hsu and K. Y. Chiu, *Tech. Dig. IEDM,* p. 96, 1984.

78. P. Wright et al., "Hot-Electron Immunity of SiO_2 Dielectrics with Fluorine Incorporation," *IEEE Electron Device Letters,* August 1989, p. 347.

79. Y. Nishioka et al., "Hot-Electron Hardened Si-Gate MOSFET Utilizing F Implantation," *IEEE Electron Device Letters,* April 1989, p. 141.

80. W. M. Werner, *Solid-State Electronics,* **17**, 769, (1974).

81. R. R. Troutman, *IEEE J. Solid-State Circuits,* **SC-9**, 55 (April 1974).

82. J. R. Pfiester et al., *Tech. Dig. IEDM,* 1989, p. 781.

83. C.-Y. Lu et al., *IEEE Trans. Electorn Dev.,* November, 1989, p. 2530.

PROBLEMS

5.1 (a) A *p*-channel MOS transistor is fabricated on an *n* substrate doped with 10^{15} phosphorus atoms per cm^3, and a gate oxide which is 1000 Å thick. Calculate the threshold voltage if $\phi_{ms} = -0.6$ eV and $Q_{tot} = 5x10^{11}$ cm^{-2}. (b) The threshold voltage of the MOSFET in part (a) is reduced by using ion implantation of boron. What is the required boron dose in order to obtain a threshold voltage of -1.5 V?

5.2 In the E-D NMOS process, depletion-mode NMOSFETs are used for load devices. This requires a negative threshold, which can be obtained by implanting a shallow arsenic or phosphorus dose into the channel region. Calculate the arsenic dose needed to achieve a -3 V threshold in a n^+ polysilicon-gate NMOS transistor which has a substrate doping of $3x10^{16}/cm^3$ and a gate oxide thickness of 50 nm?

5.3 (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) How is a self-aligned gate obtained and what are its advantages? (d) List three functions of the PSG overglass layer?

5.4 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

80 keV: the channel is implanted with 8×10^{11} boron ions/cm^2 at 30 keV through a gate oxide of 250 Å thickness. (a) Estimate the threshold voltage change caused by the boron ion implantation step. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through (b1) the channel region, and (b2) the source region.

5.5 Describe punchthrough current and subthreshold current and explain the differences between them.

5.6 Design a submicron MOSFET with a gate length of 0.75 μm. (The gate length is the channel length plus twice the junction depth.) If the junction depth is 0.2 μm, the gate oxide is 200 Å, and the maximum drain voltage is limited to 2.5 V, find the required channel doping so that the MOSFET can maintain it long-channel characteristics.

5.7 An n-channel MOSFET has a gate oxide 200 Å thick, an $L_{eff} = 1.0$ μm, source/drain junction depths of 0.2 μm, and a threshold voltage of 0.6 V. If the device is biased at $V_{GS} = 3$ V and $V_{DS} = 5$ V, calculate the saturation voltage V_{Dsat}, and the maximum electric field, E_{max}.

5.8 Explain the difference between the gate-threshold voltage, V_T, and the field-threshold voltage, V_{TF}, of a MOS device.

5.9 (a) Sketch a set of masks for fabricating an MOS transistor using Si-gate technology. (b) Repeat this exercise for an E-D NMOS inverter circuit with LOCOS isolation, buried contacts, and a depletion-mode load transistor.

5.10 An NMOS structure consists of an n-type substrate with $N_D = 5\times10^{15}$ cm^{-3}, a gate oxide of 100 nm thickness, and an Al contact. The measured threshold voltage is -2.5 V. Calculate the surface charge density.

5.11 If a circuit designed wants to keep the off-state leakage current to less than 1 pA and also wants the threshold voltage value to be 0.9 V (where V_T is defined as the gate voltage that provides $I_{DS} = 1$ μA), what value of subthreshold swing (S.S.) will this device exhibit.

CHAPTER 6

CMOS PROCESS INTEGRATION

Complementary MOS (CMOS) is so-named because it uses both *p*- and *n*-type (complementary) MOS transistors in its circuits. (Figure 6-1 depicts a CMOS inverter.) Since CMOS technology is significantly more complex than NMOS with respect to the device physics and fabrication issues, the discussion here will include a thorough introduction to these subjects. (Note that an excellent comprehensive text on CMOS, dealing with both circuit and design issues, has recently been published.)[1]

6.1 INTRODUCTION TO CMOS TECHNOLOGY

6.1.1 The Power-Dissipation Crisis of VLSI, and How CMOS Came to the Rescue

NMOS remained the dominant MOS technology as long as the integration level of devices on a chip was sufficiently low. It was inexpensive to fabricate, very functionally dense, and faster than PMOS. The earliest NMOS technologies required only 5 masking steps (including the pad mask).* On the other hand, NMOS logic-gates (e.g., inverters) draw dc power during one of the inverter states. Therefore, an NMOS integrated circuit will draw a steady current even when being operated in the standby mode (i.e., even when no signal is being propagated through the circuit). Consequently, as the number of logic gates on the chip grows, the current being drawn (and hence the power being dissipated) also increases. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a drawback for most other applications when the number of devices on a chip was rela-

* Early NMOS logic gates used butted contacts as well as enhancement-mode transistors for both the driver and the load. Depletion-mode loads and buried contacts came later. Hence, of the seven masks of the E-D NMOS process described in chapter 5, only five were required in early NMOS technology. This resulted not only in lower manufacturing costs per wafer, but also in higher yields. Thus, the cost of manufacturing NMOS circuits was brought even lower.

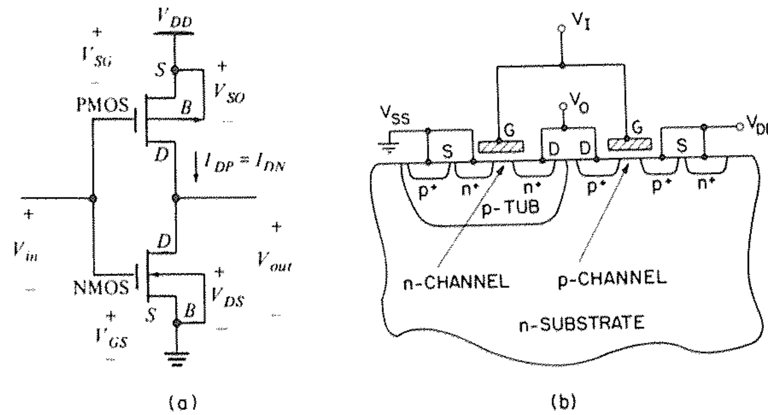


Fig. 6-1 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

tively small. Such was the situation at the level of device integration that existed up to the mid-1970s.

With the dawning of the VLSI era, however, power consumption in NMOS circuits began to exceed tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density, and the twenty-fold increase in speed of integrated circuits between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12-μm design rules to create the six-transistor SRAM cells; each cell occupied 20,600 μm^2. In 1987, however, SRAM cells using 1.2-μm design rules occupied only 150 μm^2. (The memory access-time in these respective memory circuits decreased from 1 μs to 46 ns.) By 1987 the decreased size and attendant increase in chip size allowed 256-kbit SRAMs to be built.

To take another example: The Intel 4004 4-bit microprocessor, introduced in 1971, had 2300 devices and was built in PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips can dissipate a maximum of about 5 W of power and still be used in conventional, but expensive, IC ceramic packages. In order for the much less expensive plastic packages to be used, however, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power when operated at 8 MHz. Thus, by the late 1970s it was already possible to manufacture NMOS chips whose power dissipation approached unacceptably high values. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 250 mW.)

In a CMOS inverter (unlike in an NMOS inverter) only one of the two transistors is driven at any one time. This means that when a CMOS inverter is not switching from one state to the other, a high-impedance path exists from the supply voltage to ground, regardless of the state the inverter is in. Hence, virtually no current flows, and almost no dc power is dissipated. CMOS thus allows the manufacture of circuits that need only several microwatts of standby power (Fig. 6-2).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, if microprocessors of the 32-bit generation were built in NMOS, they would dissipate 5 to 6 W of power. This would lead to severe heating and reliability concerns. In addition, expensive ceramic packages would be needed to house such chips. When such microprocessors are designed in CMOS, the power dissipation decreases to about 1 W.

From the system perspective, let's consider memory chips. Although a 1-Mbit DRAM may consume only 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be thousands of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation are significant. Smaller power-supplies and smaller cooling fans are but two of these ramifications.

6.1.2 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, for much of its 25-year history it was considered to be only a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power ICs was originally proposed by Sah and Wanlass in 1963.[2] The first CMOS ICs were fabricated

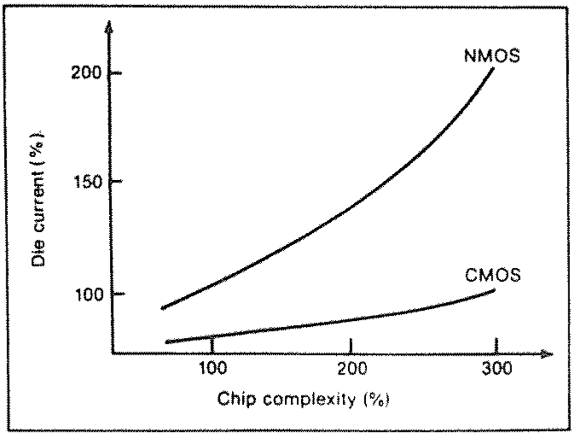


Fig. 6-2 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.

in 1966, and subsequent development of the technology was spearheaded by the RCA Corporation. The earliest volume commercial application was the use of the CMOS logic-inverter in the frequency divider circuits of digital watches.

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

At the time CMOS circuits were first fabricated, the processes of ion implantation and local oxide isolation (LOCOS) had not yet been developed. In addition, metal gates were being used for MOS devices and control had not yet been established over the large and quite variable positive oxide charges in the gate oxide. As a result, *p*-well technology was the only means of fabricating CMOS.

While PMOS enhancement-mode transistors could be successfully fabricated in a lightly doped (e.g., 10^{15} cm^{-3}) *n*-substrate with a V_T of about -2 V, NMOS enhancement-mode transistors could not be fabricated on lightly doped *p*-substrates because the V_T of NMOS metal-gate transistors on such substrates is negative. In addition, the problems of oxide charge and the segregation of boron at the field-oxide/silicon-substrate interface, when combined with necessity of having to use relatively-thin field oxides (in order to be able to achieve adequate step coverage over nonrecessed field-oxide steps; see chap. 2), made it likely that parasitic channels would be established in the field regions between NMOS devices built on lightly doped substrates. Therefore, the only reliable way to manufacture enhancement-mode NMOS transistors for CMOS inverters was on regions with boron surface concentrations high enough to overcome these problems..

The *p*-well approach to building CMOS provided such regions, since the well had to be doped about ten times as heavily as the substrate for adequate control of doping in the well to be achieved. As a result, *p*-well technology became established in the companies that pioneered CMOS technology. Long after the problems of fabricating NMOS on lightly doped substrates had been solved (i.e., through control of the gate-oxide charge and of V_T by ion implantation) most of these companies continued to use *p*-well technology to design new circuits. While this was primarily due to historical inertia, *p*-well technology did have a few advantages over *n*-well technology (as will be discussed in section 6.2.2)

The packing-density limits of the early *p*-well CMOS technology, however, were responsible for its poor performance. The packing density was much worse than that of NMOS, primarily because the *n*- and *p*-devices had to be surrounded by guard rings (n^+ or p^+ diffusions that surround the device, as shown in Fig. 6-3)[114] to prevent the inversion of the field regions.

Before LOCOS isolation and ion implantation became available, guard rings were needed to provide adequately large values of V_T in the field regions. However, their use

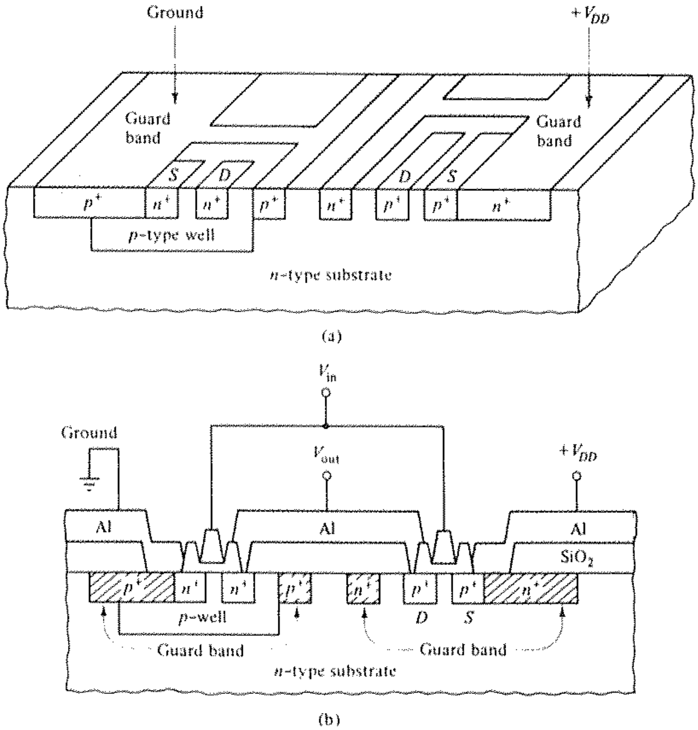


Fig. 6-3 CMOS structure with guard bands. A CMOS inverter circuit is depicted.[114] From W. C. Till and J. T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication*. Copyright 1982 Prentice-Hall. Reprinted with permission.

results in very large-area devices.* As a result, until oxide isolation and channel-stop implants were developed, interconnect capacitance and resistance severely degraded the speed performance.† Once guard rings were no longer needed, however, the devices could be brought closer together, and the speed of CMOS circuits improved dramatically.

When it became apparent in the late 1970s that the increases in power density and dissipation would make it impossible to design future generations of MOS circuits in

* The drain regions must also be isolated from the guard rings by a lightly doped region to prevent avalanche or Zener breakdown, as seen in Fig. 6-3. This requirement exacerbates the area penalty when guard rings are used.

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

NMOS, the companies that had been stubbornly continuing to use NMOS for design and fabrication finally began to consider CMOS. It was natural for them to seek a technology that was compatible with the modern high-production-volume Si-gate-NMOS processes that they had successfully perfected. Since *n*-well CMOS offers near compatibility with such processes, and since it allows *n*-channel transistor performance to be optimized (through fabrication in the lightly doped *p*-substrate regions), it became the technology of choice for many companies that had formerly been manufacturing NMOS integrated circuits.[8]

It became evident, however, that neither *p*-well nor *n*-well would be the optimum choice for submicron CMOS. Instead, it appeared that *twin-well CMOS* would be more effective. As a result, many such processes were subsequently developed.

We will examine the various well technologies with regard to their advantages and disadvantages in modern CMOS processes. Circuit designers and process engineers should be aware of the trade-offs. The advances developed through the refinement of NMOS were incorporated into CMOS technology, and the performance was thus dramatically improved over that exhibited by primitive CMOS circuits.

6.1.3 Operation of CMOS Inverters

The CMOS inverter (for which the circuit schematic and a sample layout are shown in Fig. 6-1a and 6-4, respectively) uses enhancement-mode transistors for both the NMOS driver and the PMOS load transistors. The gates of the two transistors are connected and serve as the input to the inverter. The common drains of each device are connected to the output of the inverter, and we assume that the inverter is driving some load capacitance, C_L (e.g., the input to another CMOS logic gate). Note that both the source and the body of the NMOS transistor are connected to ground, while those of the PMOS transistor are connected to V_{DD} (e.g., 5 V).

The threshold voltage of the NMOS driver transistor, V_{Tn}, is positive (e.g., V_{Tn} = 0.8 V), while that of the PMOS load transistor, V_{Tp}, is negative (e.g., V_{Tp} = -0.8 V). Figure 6-5a shows the inverter's output voltage, V_o, as a function of the input voltage V_{in}. This curve is known as the *static voltage input/output characteristic*, or the *transfer characteristic* of the gate.

When $V_{in} = 0$, $V_{GSn} = 0$ and the NMOS transistor is *OFF* (since V_{GSn} is <0.8 V). The PMOS transistor, however, is *ON*, since V_{GSp} = -5 V, which is much more negative than -0.8 V. Thus, when $V_{in} = 0$, C_L is charged to V_{DD} through the turned-on PMOS load transistor, and V_o = 5 V.

When V_{in} = 5 V, the NMOS transistor is turned *ON*, since V_{GSn} = 5 V (which is >0.8 V). The PMOS transistor is turned *OFF*, since V_{GSp} = 0 V (which is more positive than -0.8 V). Consequently, when V_{in} = 5 V (V_{DD}), the output is connected to ground through the turned-on NMOS driver, allowing C_L to be discharged. Since the PMOS device is off, C_L will be completely discharged, and V_o will be 0 V.

The most important property of the CMOS inverter is that when the gate is sitting quiescently in either logic state ($V_o = V_{DD}$ or 0), one of the transistors is *OFF* and the

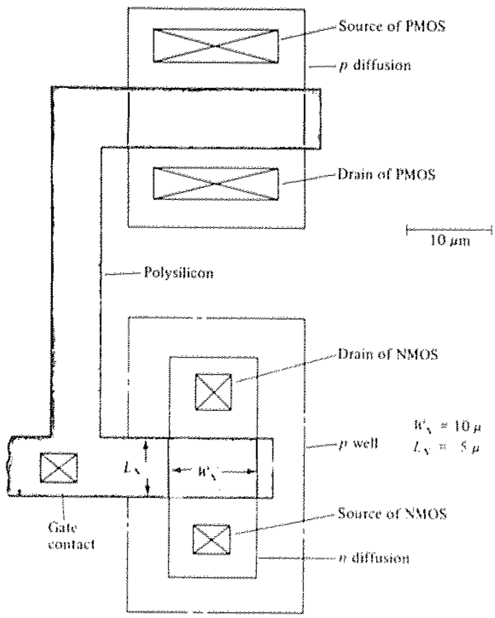


Fig. 6-4 CMOS inverter layout. From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright 1983, McGraw-Hill Book Co. Reprinted with permission.

current conducted between V_{DD} and ground is negligible (i.e., it is equal to the leakage current of the *OFF* device). This feature can be seen in Fig. 6-5b, which plots the current through the inverter, I_{DD}, as a function of V_{in} (solid curve). The power dissipated in the static (or standby) mode is then determined by the product of the leakage current and the supply voltage. Since the leakage current of an MOS transistor in cutoff (subthreshold current) is so small, very little power is consumed in the static mode. Another important feature is that the V_o swings all the way from V_{DD} to 0 as the inverter changes state; this characteristic of the output-voltage swing is referred to as swinging from *rail to rail.*

Two other operational features of the CMOS inverter must also be mentioned. First, as can be seen in Fig. 6-5b, significant current is conducted through the inverter only when both transistors are *ON* at the same time (i.e., during switching). Therefore, most of the power dissipation is due to the charging and discharging of C_L. In fact, it can be shown that the power dissipation is essentially given by $f C_L V_{DD}^2$, where f is the switching frequency.

Second, because the CMOS inverter output voltage can swing from rail to rail, it can inherently provide excellent noise margins. Noise margins are usually defined in terms of the *logic-gate threshold voltages*, V_{OH}, V_{OL}, V_{IH}, and V_{IL} (which are *not* the same parameters as *device threshold-voltages*, see Fig. 6-5a). The noise margins NM_L and NM_H (Fig. 6-5b) are defined according to the following equations:

$$NM_L = V_{IL} - V_{OL} \qquad (6\text{-}1a)$$
$$NM_H = V_{OH} - V_{OL} \qquad (6\text{-}1b)$$

Briefly, the argument for Eq. 6-1a is that the logic gate (e.g., *Inverter 1* in Fig. 6-6a) should provide a maximum "low output" that is less than the maximum "low input" that the subsequent logic gate (e.g., *Inverter 2*) can accept without causing the output voltage of *Inverter 2* to be driven into the ambiguous portion of the transfer characteristic. The difference between V_{IL} and V_{OL} (the noise margin) then specifies how much noise voltage can be tolerated at the input node of Inverter 2 before its output will be driven to a voltage value that is logically undefined.

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

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

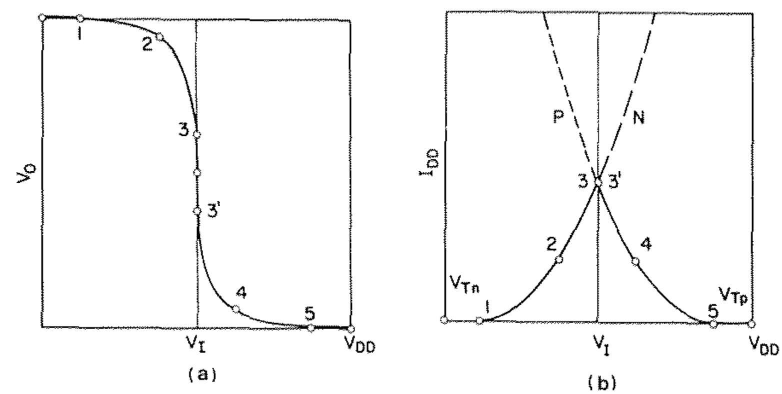


Fig. 6-5 (a) Output (V_o) versus input (V_I) voltage of CMOS inverter. (b) Current through inverter as a function of input voltage (solid curve); I-V characteristics of *n*- and *p*-channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.[115]

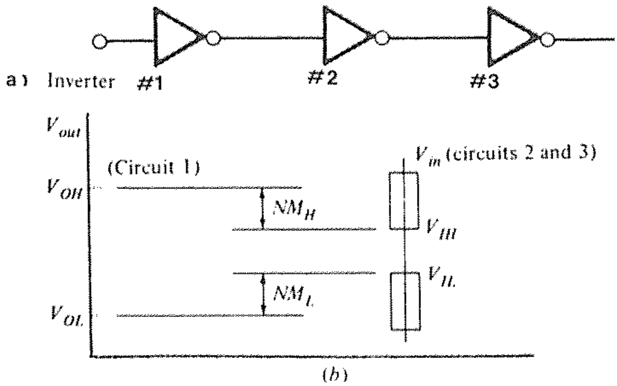


Fig. 6-6 (a) String of inverters connected in series. (b) Definition of noise margins.

For example, if the guideline described in the section on the subthreshold current in MOS devices is used (section 5.2.3 in chap. 5), a minimum value of V_{Tn} would be 0.5 V, to keep V_{OL} at least 0.5 V below V_{Tn} and to maintain sufficiently small I_{Dst}.

6.1.4 Advantages of CMOS

As noted earlier, the most important advantage of CMOS is its significantly reduced power density and dissipation.[3,4,5,6] There are other advantages as well, which fall into the following main categories:

- device/chip performance
- reliability
- circuit design
- cost issues

These advantages, as well as the disadvantages of CMOS, will be discussed in this section.

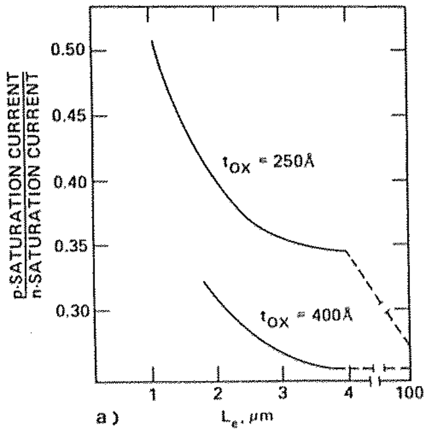
6.1.4.1 Device/Chip Performance Advantages.

- Although logic gates designed in CMOS are larger than those of NMOS (primarily because of the increased spacing needed to isolate *n*-channel from *p*-channel devices), this packing density penalty is becoming less important, for several reasons.

First, the CMOS gate uses a PMOS device rather than a depletion-mode NMOS device as the load. The PMOS device is usually made about twice as wide as the NMOS device in order for symmetrical driving capability to be achieved. The NMOS depletion-mode load, however, is four times as large as the NMOS driver. Hence, the area of the two CMOS devices is actually smaller than the area of the two NMOS devices in an inverter.

Second, since the devices are built with submicron dimensions, the difference in the output drive capabilities (I_D) of identically sized PMOS and NMOS devices is decreased due to velocity saturation effects (Fig. 6-7a). The area of the CMOS inverter devices will thus grow proportionately even smaller.

Third, as devices become smaller, the fraction of the chip area required for interconnections becomes larger. Hence, the fact that the gate density is lower in CMOS in NMOS becomes less important. This is especially significant in random logic designs.



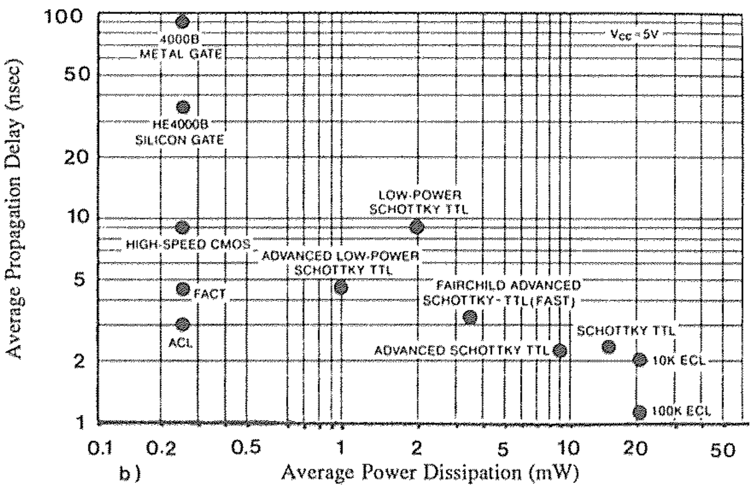


Fig. 6-7 (a) Ratio of p-channel saturation current to n-channel saturation current increases as the effective channel length of the devices shrinks, due to velocity saturation effects. (b) Comparative speed and power characteristics of various CMOS and bipolar logic families.

• Although the input capacitance of CMOS is larger than that of NMOS (since there are two MOS gates connected in parallel at the input of CMOS logic elements versus only one in NMOS), the interconnect capacitance is becoming more significant than the gate capacitance as gate sizes shrink and interconnect lengths grow with increasing chip size. NMOS gates are thus no longer significantly faster than CMOS gates. Furthermore, as the chip heats up from excessive power dissipation, the circuits slow down (due to degradation of carrier mobility). The lower power advantage might actually make a CMOS circuit operate faster than a comparable, but hotter-running, NMOS part.

• Since more devices can be placed onto a single chip in CMOS than in NMOS, less chip-to-chip driving is necessary, and the overall system speed is improved.

• The fan-out capacitance for MOS devices is much smaller than that of bipolar transistors. Therefore, while the transconductance is also much smaller, if the load capacitance (other than fan-out) is small, much less current is needed to charge the next MOS gate compared to a bipolar logic gate. As the channel length of MOS device shrinks, it may be possible to decrease the delay time required to charge other gates on the same chip. On the other hand, because the performance of bipolar devices is relatively insensitive to shrinking device dimensions, the delay times in CMOS are approaching those of bipolar ECL, but at much lower levels of power dissipation (Fig. 6-7b).

• CMOS can be operated over a wider range of V_{DD} values (e.g., 2-7 V) than NMOS.

• Threshold body-biasing sensitivity is less important in CMOS, and bootstrapping is not needed for transference of a signal through a string of inverter.

• The radiation hardness of CMOS circuits is much higher than that of NMOS.

6.1.4.2 Reliability Advantages.

• The heat generated during operation can raise the chip temperature to the point at which it becomes more prone to failure. Since CMOS circuits dissipate much less power, in most cases they should be inherently more reliable.

• Hot-carrier degradation of MOS devices should be decreased in CMOS for several reasons. First, since hot-electron effects are much less severe in PMOS devices, the load devices in CMOS will suffer less degradation than the depletion-mode load devices used in NMOS. Second, the CMOS gates do not draw static current, so long-term, cumulative hot–electron induced degradation will be smaller. Finally, unlike NMOS, CMOS generally does not use bootstrapping (which increases the electric field in the device and thus aggravates hot-electron degradation).

• Electromigration failures in the metal lines of the circuit are reduced, since no static current flows in the metal lines.

• The soft error rates (SERs) of DRAMs and SRAMs can be reduced by one to two orders of magnitude when the memory arrays are fabricated inside a well region where doping type is the opposite of the substrate's. The added SER protection arises because the reverse-biased well-substrate junction of CMOS creates a potential barrier against carriers generated in the substrate.

6.1.4.3 Circuit-Design Advantages.

• CMOS can achieve "static ratioless" logic design. Circuits that contain a *p*-channel transistor for every *n*-channel transistor are said to be "static," because such gates are triggered by the data path signal and do not require the use of an external clock. The design is said to be "ratioless" because the inverter voltage transfer characteristic does not depend strongly on the relative geometric sizes of the *p* and *n* transistors. By contrast, commonly used "ratioed" NMOS circuitry must have transistor widths and lengths that are chosen both to balance the currents between transistors and to ensure that this balance is maintained, given changes in operating temperatures, power-supply voltage variation, and day-to-day differences in the manufacturing process.

• As noted earlier, since the output of CMOS logic gates swings from rail to rail, excellent noise margins are inherently provided.

• The flexibility in selecting transistor geometry provided by the ratioless nature of CMOS makes it much easier for "uncommitted circuit" designs, such as gate arrays and standard cells, to be implemented. In gate arrays and standard cells, the number of input and output signals (fan-ins and fan-outs) are normally not known when the individual transistors are laid out. As a result, it is very difficult with PMOS or NMOS to create gate-array and standard-cell configurations that possess sufficient design flexibility. Instead, most standard cells and virtually all MOS gate arrays are implemented in CMOS.

• In NMOS, gating (or pass) transistors reduce the transmitted signal by the so-called *threshold loss*, whereas in CMOS such transmission gates leave the signal unchanged. As a result, signal regeneration can be less frequent in clocked CMOS circuits.

• CMOS allows both analog and digital functions with high circuit densities to be implemented on the same chip. This has stimulated the development of switched-capacitor techniques for analog-to-digital conversion and allowed their integration on a chip with a digital-signal algorithm.

• The circuit-design benefits of CMOS for analog applications are that the switches have no offset voltage, and that the area required for operational amplifiers is much smaller than that needed for NMOS op amps. That is, while an NMOS op amp might take 30 transistors, the same op amp in CMOS might take one-third the number of transistors, as well as one-third the area. Furthermore, CMOS op amps are three to five times smaller than bipolar op amps.

6.1.4.4 Cost Analysis.

• When CMOS and NMOS ICs were first being manufactured, CMOS required almost twice as many masking steps as NMOS. As NMOS processing grew more complex, however, the addition of depletion-mode loads, buried contacts, punchthrough-prevention implants, and lightly doped drains significantly increased processing costs. The costs of fabricating CMOS circuits, however, did not increase proportionately. Hence, the cost differential between the two technologies has steadily been reduced; at present the cost of manufacturing CMOS may be only 20 percent higher than that of manufacturing advanced NMOS circuits. The slightly increased cost of CMOS manufacturing is more than offset by the savings in design, packaging, system heat management, and reliability.

• The complexity of the design task is reduced in CMOS, lowering costs and allowing designs to be created more rapidly (which in turn, decreases costs further). Furthermore, since the time it takes to bring a product to market often has significant impact on market share and profitability, a shortened design time may represent a large increase in profit.

• Packaging costs can represent 25-75% of the total chip-manufacturing cost. Since the reduced power dissipation of CMOS allows the use of cheaper IC package technology, there is a significant cost savings with CMOS packaging compared to NMOS. In addition, the elimination of cooling measures and the reduced failure rates of CMOS result in lower costs. These savings translate into larger profit margins for both chip manufacturers and suppliers of electronic systems that use CMOS.

• Because grounding of the substrate can be done on the front side of the wafer, CMOS may not need back grinding or gold on the backside of the wafer, leading to further savings.

6.1.5 Disadvantages of CMOS

As is to be expected, CMOS also possesses disadvantages. Some of those listed here are inherent to all MOS circuits; these have been considered in in more detail in earlier sections on NMOS technology. Other disadvantages, however, are unique to CMOS and will be given more in-depth treatment in this chapter:

• Like NMOS, CMOS is susceptible to short-channel and hot-carrier effects when device channel lengths drop below about 2 μm (although as mentioned, the hot-carrier problem is reduced to some degree in CMOS). In addition, hot-electron effects in *p*-channel devices apparently do not become severe until channel lengths of below 1 μm are reached.

• CMOS has a somewhat lower packing density than NMOS.

• Static CMOS logic gates exhibit larger input capacitance than NMOS logic gates due to the additional input capacitance of the *p*-channel transistors, which are in parallel with the *n*-channel transistors.

- The need to simultaneously manufacture high-quality PMOS and NMOS devices on the same chip can give rise to processing difficulties.
- There are constraints on the scaling of PMOS devices manufactured with n^+ polysilicon gates.
- Well contacts must be provided, which takes up more chip area than required in NMOS.
- The well drive-in step requires a long process time (e.g., four hours or more at 1100°C).
- CMOS is susceptible to latchup, and hence needs guard bands or epi (see section 6.4.8). In addition, it is often very difficult to identify the exact location of the latchup site (i.e., special liquid-crystal or infra-red techniques must be used).
- Most current CMOS technologies use n^+-doped polysilicon as the gate material. An interconnect routing problem arises because the metal layer must be used when contact is made between this n^+ polysilicon and the p^+ source/drain region of PMOS devices.
- Like all other MOS technologies, CMOS is vulnerable to electrostatic-discharge damage.

6.2 THE WELL CONTROVERSY IN CMOS

There are many trade-offs involved in the optimization of a CMOS process. The choices revolve around the highly interrelated parameters of circuit performance, layout density, fabrication cost, and tolerance to latchup. As described in chapter 5, obtaining the best circuit performance from an MOS device involves maximizing the drive current and minimizing junction capacitances and body effect – all of which favor lower doping concentrations in the device body. Optimizing density, however, favors raising these same doping concentrations (to avoid punchthrough and to achieve high field thresholds). Higher density is thus achieved by allowing closer packing of adjacent *n*- and *p*-channel transistors. (Issues relating to CMOS isolation will be described in more detail in a later section.) Latchup tolerance can also be improved by spacing *n*- and *p*-channel transistors farther apart (see section 6.4), but this in turn lowers circuit density.

These complex interacting tradeoffs converge on several processing configurations that are determined at the outset of the processing sequence through the selection of the type of *well doping*. The ramifications of this choice must therefore be considered in more detail.[1,4,6]

6.2.1 The Need for Wells in CMOS

Both *n*- and *p*-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting

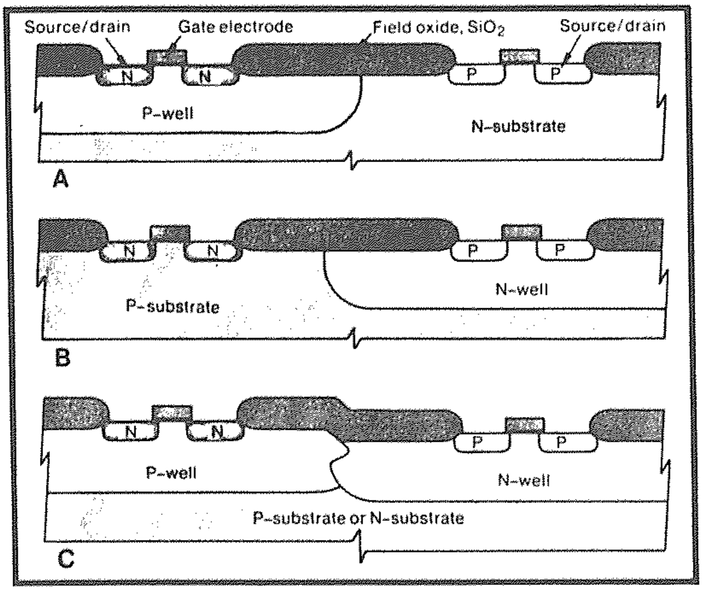


Fig. 6-8 CMOS circuits can be designed using (a) p-well, (b) n-well, or (c) twin-well technology.[3] (© 1983 IEEE).

substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite that present in the starting material must be formed, as shown in Fig. 6-8. These regions of opposite doping, called *wells* (or sometimes *tubs*) are the first features to be defined on a starting wafer. This is done by implanting and diffusing an appropriate dopant to attain the proper well depth and doping profile. The well's doping type becomes the identifying characteristic of the CMOS technology.

Normally, a p-type substrate is connected to the most *negative* circuit voltage, and an n-type substrate to the most *positive*, to ensure that pn junctions are not forward-biased during circuit operation. Similarly, the respective well regions must also be connected to the appropriate circuit voltages to prevent forward-biasing of the junctions within the well (and between the well and substrate). Because the wells are totally junction isolated from the rest of the wafer, it is especially important that such *well contacts* be made. That is, it may still be possible to contact the substrate from the backside of the wafer even if no provision is made for a *substrate* contact from the top surface of the wafer. Such a backside connection, however, cannot be established to the *well* regions.

Because one of the device types must be located in the well, there has been some controversy as to which type of well should be used in CMOS-circuit fabrication. The performance of devices in the well will suffer as a consequence of the higher doping, exhibiting higher junction capacitance, increased sensitivity to body effect, and decreased transconductance (due to reduced carrier mobility). Furthermore, substrate currents (e.g., due to hot-carrier effects) will be harder to collect from the well regions. The issue thus becomes that of deciding which device type should be subjected to such performance degradation.

Some argue that the NMOS devices should be built in the substrate, where their better performance can be optimized (*n*-well CMOS). This argument is persuasive for companies that have had long experience in producing high performance NMOS, since they can merely transfer this technology to the building of NMOS devices in the *p*-substrate starting material. Furthermore, if a circuit-design technique rich in NMOS devices is used (e.g., Domino logic),[97] most of the devices on the chip will be NMOS, which will again favor the building of NMOS devices in the *p*-substrate (and PMOS devices in the *n*-well).

On the other hand, the hot-electron-induced substrate current is much higher in NMOS than in PMOS and, as noted, is harder to collect from the well regions than from the substrate. In addition, device technologists might argue that the better-performing NMOS devices can afford to have their circuit behavior somewhat degraded, as this will balance the performance between these and PMOS devices.

The next sections will outline the pros and cons of both *p*-well and *n*-well configurations, as well as those of the more complex well configurations that have been developed (e.g., twin-well and retrograde-well CMOS).

6.2.2 *p*-Well CMOS

The *p-well process*, illustrated in Fig. 6-8a, involves the creation of *p*-regions in an *n*-type substrate for the fabrication of NMOS devices. The *p*-wells are formed by implanting a *p*-type dopant into an *n*-substrate, at a high enough concentration to over-compensate for the substrate doping and to give adequate control over the *p*-type doping in the well. The starting *n*-type substrate, however, must also have sufficient doping to ensure that the characteristics of devices fabricated in the substrate regions are adequate (a minimum doping concentration of of $3x10^{14}$-$1x10^{15}/cm^3$ is required). The *p*-well doping must therefore be about five to ten times higher than the doping in the *n*-substrate. If the *p*-well is doped *too* highly, however, the performance of the *n*-channel devices will be degraded through lower carrier mobility, increased source/drain to *p*-well capacitance, and increased sensitivity to body-biasing effects.

As noted in section 6.1.2, *p*-well CMOS was the first type of CMOS that could be practically manufactured. The first companies to commercially offer CMOS components produced many designs in *p*-well technology. As this experience was spread throughout the industry, *p*-well CMOS became widely established.

There are several advantages of *p*-well over *n*-well CMOS. First, *p*-well technology may be the better choice for pure-static logic, in which a good balance between the performance of both MOS device types is beneficial. Second, it is attractive for applications that require an isolated *p*-region (such as those using an *npn* bipolar transistor as an on-chip driver or *n*-channel FETs for an analog input). Third, it is less susceptible to field-inversion problems than *n*-well CMOS, and can thus be slightly easier to fabricate. (We will describe later how *p*-well CMOS can use the well itself as a channel stop, whereas *n*-well CMOS must use a separate channel-stop process.) Fourth, if the so-called *retrograde-well* process (rather than a diffusion of a shallow implant) is used to form the wells, *p*-well technology is more feasible. It is easier to form a *p*-retrograde than an *n*-retrograde well, since boron ions penetrate deeper than arsenic or phosphorus ions at a given implant energy.

Finally, *p*-well CMOS may be better for fabricating SRAMs. Since the alpha-particle–induced soft-error rate (SER) becomes significant even in SRAMs if feature sizes are scaled to submicron dimensions, the cells should be made inside a well. The sensing of the state of an SRAM cell depends on the current provided by the cell. As a result, high-gain NMOS devices are more desirable for the pass gates and drivers in the cell, and must be built in a *p*-well.

The organization that provides university communities with IC fabrication services, the MOS Implementation Service (MOSIS), offers a standard *p*-well CMOS process (as well as standard NMOS and advanced CMOS [twin-well] processes). MOSIS cooperates with various IC-manufacturing vendors that fabricate the designs submitted to it. Circuits designed to the specifications of the standard *p*-well process can be executed by vendors, which serve as *silicon foundries* for MOSIS.

6.2.3 *n*-Well CMOS

In the *n*-well process, shown in Fig. 6-8b, the *p*-channel devices are formed in the more heavily doped *n*-well. As noted earlier, *n*-well technology became the choice of companies with extensive experience in building NMOS ICs.[8] Because the NMOS device could be fabricated in a lightly doped substrate, virtually all of the experience that had been amassed in fabricating high-performance NMOS could be transferred to an *n*-well CMOS process. As a result, virtually all EPROMs, microprocessors, and dynamic RAM designs in the generations of technology built with 1.25-2.0 μm dimensions were implemented with *n*-well CMOS.

This technology does have some disadvantages. First, as mentioned earlier, it is more sensitive to field-inversion problems than *p*-well CMOS. Second, it may be more difficult to build pure-static, high-performance logic circuits with *n*-well CMOS.

EXAMPLE 6-1: An *n*-well CMOS process is to be developed for operation with a power-supply voltage of $V_{DD} = 5$ V. The substrate doping of the *p*-type wafers is $1 \times 10^{15}/cm^3$. The *n*-wells are to have an average dopant concentration of $1 \times 10^{16}/cm^3$. The *p*-channel MOSFET sources and drains are to have junction depths of 0.4 μm and an average dopant density of $10^{18}/cm^3$. What is the minimum *n*-well depth needed to avoid vertical punchthrough to the substrate?

SOLUTION: Vertical punchthrough will occur if the depletion region of the source/drain-to-well junction were to contact the depletion region of the well-substrate junction when $V_{DD} = 5$ V (see section 5.5.2).

The source-to-well junction is essentially a one-sided *pn* junction with a built-in voltage of V_{bi} = 0. 82 V. From Fig. 5-20, we estimate that the depletion-region width extends into the *n*-well ~0.35 μm, since there is no applied voltage across the junction (i.e., both the source and the well are connected to V_{DD}). The *np*-junction to the substrate has a built-in voltage of 0.63 V, and the total depletion width of this junction at a 5 V bias is ~1.9 μm. We calculate that about 0.19 μm of this depletion region is in the *n*-well. The *n*-well must therefore be deep enough to accommodate the depth of the source junction (0.4 μm) as well as the sum of the depletion region widths in the well in order for vertical punchthrough to be avoided. While the total of these dimensions is 0.94 μm, it is good engineering practice to increase the depth of the well by about 50% to account for process variations. A reasonable well depth might therefore be 1.5 μm.

6.2.4 CMOS on Epitaxial Substrates

As will be described in the section dealing with latchup prevention (section 6.4.8.2), heavily doped substrates with a more lightly doped surface epitaxial layer have been utilized to suppress latchup in CMOS.[11] When such starting material is used with single-well CMOS, the epitaxial layer is doped to a concentration equal to that of the substrate in a nonepitaxial wafer used for that process. If a twin-well CMOS process is used, the epitaxial layer is doped to a level significantly lower than that required for building either the *p*- or *n*-channel MOSFETs (see section 6.2.5).

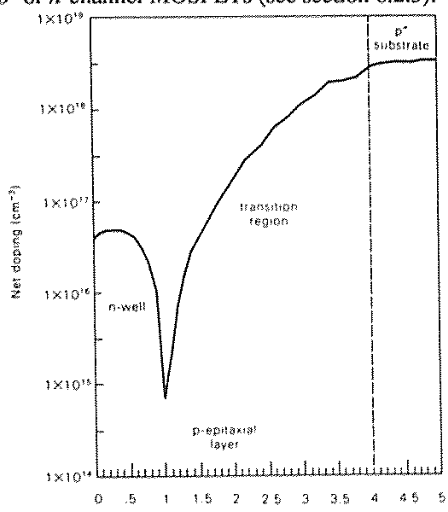


Fig. 6-9 A *p*-on-*p*-epi doping profile and diffused *n*-well profile measured after the entire CMOS process has been completed. The as grown epi thickness is 4 μm and the final lightly doped epi thickness is about 2 μm.[116] (© 1987 IEEE).

The epitaxial layer is made thicker than the well depth, since the dopants in the heavily doped substrate under the epi layer diffuse toward the surface as the well dopants are diffused toward the bulk (Fig. 6-9). Thus, some of the epitaxial layer becomes more heavily doped during the CMOS process flow. The process is designed so that the bottom of the well is eventually adjacent to the heavily doped substrate region.

Either n-epi on n^+ substrates or p-epi on p^+ substrates can be used, with each method having advantages and drawbacks. Because the problems with n-epi on n^+ are more serious, p-epi on p^+ is more widely used. The major limitation of the latter approach is that the outdiffusion of boron from the p^+ is much more severe than it is in n-epi on n^+. (The reason for this is that boron diffuses much more rapidly than antimony, which is the most widely-used n^+ dopant material.) Thus, a thicker p^-- epitaxial layer must be used.

In addition, the transition region between the p^+ substrate and the p^-- epi layer is thicker, producing a larger series resistance (R_{sub}), which in turn reduces latchup immunity. On the other hand, the p-on-p^+ material is less sensitive to process-induced defects, and the p-type substrate provides higher conductivity under NMOS devices. Such additional conductivity is desirable, since it reduces the voltage drops caused by the substrate currents (generated as a result of the hot-carrier effects in short-channel devices). It is especially important in the regions containing NMOS devices, since the hot-electron substrate current is much higher in such devices than in PMOS devices.

The n-on-n^+ epitaxial substrates also offer some advantages. First, SRAMs are often built on n-type substrates, because the p-well to n-substrate junction provides protection from radiation-induced discharging of the memory's n-type storage nodes in the p-well.[12] Second, retrograde p-wells are easier to implement because their implantation energy requirements are much lower. Third, the n-to-n^+ transition region is smaller than that in p-on-p^+ substrates, and the smaller value of R_{sub} provides improved latchup protection.

The limitations of n-epi on n^+ involve the process by which the heavily doped substrate is grown. Antimony (Sb) is the n-type dopant used, both because its diffusion coefficient is so low and because it exhibits much less lateral autodoping than arsenic (the other slow-diffusing n-type dopant). The problem with Sb is that its segregation coefficient, k_o, is very small (i.e., $k_{o,Sb} = 0.023$; see, Vol. 1, chap. 1), and thus a large quantity of Sb must therefore be put into the Si melt to ensure that a sufficiently heavily doped ingot will be produced. In even the most highly refined Sb there are high concentrations of unwanted heavy metals, which become incorporated into the growing silicon crystal. In addition, the oxygen content of the Sb-doped crystal is relatively small, due in part to the special growing conditions used when the Sb-doped ingot is pulled.[7,9,10] As a result, intrinsic gettering techniques that would getter the metals in the substrate are not as effective as they are in p-on-p^+ epi. For these reasons, n-on-n^+ substrates are less frequently chosen when epi-CMOS is implemented.[12]

A problem that exists with both types of epitaxial substrates for CMOS is that the wells cannot be made too deep, since the lateral diffusion would then take up too much area. On the other hand, if the wells are too shallow, vertical punchthrough will ensue.[13] A second problem is that of leakage current. Appreciable leakage current can

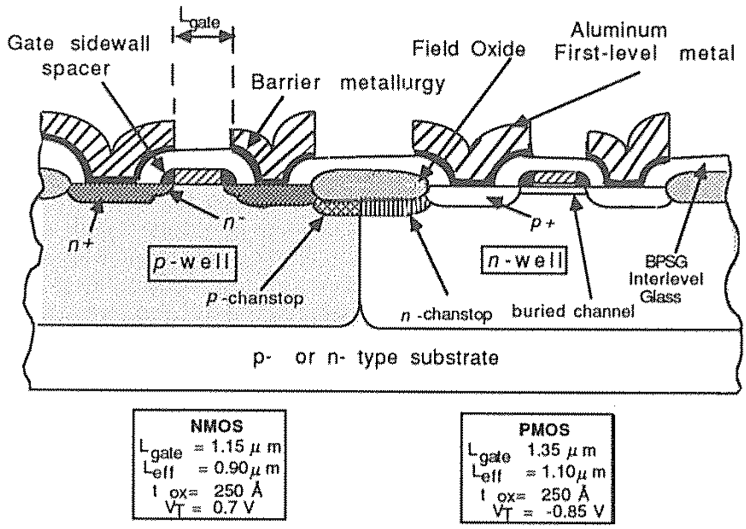


Fig. 6-10 Cross section of a twin-well 1.2-μm CMOS structure.[12] Reprinted with permission of Semiconductor International.

flow in the vertical path in two different situations. First, punchthrough can occur if the depletion regions of the p^+/n-well junction and the n-well/p^+-substrate junction touch each other. The problem is even more severe in the case of the heavily doped p^+ substrate, due to the high degree of boron outdiffusion. A major limitation is thus imposed on the minimum epitaxial-layer thickness. Second, it may be possible for the source regions to become biased to a potential below V_{SS}.

Another problem of epi-CMOS is back surface autodoping. For example p-type 20 Ω-cm (7×10^{14} boron/cm^3) epi on a 0.005 Ω-cm (2×10^{19} boron/cm^3) substrate is representative of epi for CMOS devices. If the substrate back is not sealed (e.g., using a sealing layer such as undoped silicon dioxide or silicon nitride), boron evaporation can contribute to autodoping on the front surface during the entire epi deposition cycle. This widens the epi/substrate interface and may even prevent the epi from reaching the 20 Ω-cm specified resistivity. Note that if silicon nitride is used as a sealing layer it should only be used in thin layers (e.g., less than 100 nm thick) since its high intrinsic stress causes it to bow the silicon.[113]

6.2.5 Twin-Well CMOS

With the twin-tub approach, two separate wells are formed for n- and p-channel transistors in a lightly doped substrate (Figs. 6-8c and 6-10). The substrate may be either a lightly doped wafer of n or p material, or a thin, lightly doped epitaxial layer on a heavily doped substrate. In either case, the level of surface doping is significantly lower than that required for building either the *p*- or *n*-channel MOSFETs. Each of the well dopants is implanted separately into the lightly doped surface region and is then driven in to the desired depth.

The doping profiles of each of the device types can be set independently, since the constraint of single-well CMOS does not exist (i.e., that the well doping must always be higher than the doping of the substrate in which one type of device is made). This was originally cited as an advantage of twin-well CMOS over single-well CMOS, with the argument made that both device types could thus be optimized.[14] This claim for 1-2 μm CMOS has been questioned by Chen,[4] who points out that in modern single-well CMOS processes, an additional implant is used to prevent punchthrough without the need to raise the entire substrate-doping concentration. By incorporating this additional implant, it is possible to build higher-performance devices than can be achieved with the twin-well approach, in terms of junction capacitance and sensitivity to body effect.

Twin-well CMOS does offer some significant benefits for devices with submicron-channel lengths (although these advantages are gained at the cost of greater process complexity).

The first, and most important advantage arises when devices with submicron channel lengths are fabricated. Since it is recognized that the two device types perform similarly as channel lengths approach 0.5 μm, it is useful to provide symmetrical *n*- and *p*-channel devices. Furthermore, at submicron dimensions the body doping of both transistor types must be raised significantly to prevent punchthrough and to maintain adequate threshold-voltage levels. Thus, the advantage of having one type of MOS transistor in a lightly doped region (to optimize its performance at the expense of the other) disappears. It is instead more beneficial to produce two types of active device wells, each with its own optimized doping profile (i.e., formed by separate implants into a lightly doped substrate).*

The second advantage of the twin-well process is that it is compatible with the technologies of either isolation by selective epitaxial growth (SEG) or trench isolation.[4] Both approaches restrict the lateral diffusion of the dopants in each of the wells. In addition, sidewall inversion along the trench is less likely when both device sidewalls are butted against a trench that has been formed in a more highly doped well.[15] Finally, when deep trenches are used with a thin epitaxial layer on a heavily doped substrate, latchup can be eliminated. The combined use of the twin-well process and advanced isolation techniques allows n^+ to p^+ spacing to be dramatically reduced in comparison to single-well technologies.

* It is not useful to start with a substrate doped to the optimum level needed for just *one* of the submicron devices. If this were done, a single well of much higher doping would have to be established for the other type of submicron device, which would unnecessarily degrade the device performance. It is possible, with the twin-well approach, to have both types placed in a well of optimum doping profile.

A third benefit is that the combination of epitaxial substrates and the twin-well process allows either substrate type to be chosen with no effects on transistor performance and essentially no change in process flow. Such flexibility is useful since some applications are best met with n^+ substrates and others with p^+ substrates. This advantage may be important if a single process is needed for implementing designs with different applications.

Finally, self-aligned channel stops can be easily implemented in the twin-well approach, allowing the spacing between *n*- and *p*-channel devices to be reduced. Although this spacing reduction is not as great as it would be if trenches or SEG were used, the process is much less complex with the twin-well method.

6.2.6 Retrograde-Well CMOS

Conventional wells are formed in single- and twin-well CMOS technology by implanting dopants and then diffusing them to the desired depth. However, the diffusion occurs laterally as well as vertically, which has the effect of reducing packing density. If a high-energy implant is used to place the dopants at the desired depth without further diffusion, much less lateral spread occurs (see Vol. 1, chap. 9). Such high-energy implants also cause the peak of the implant to be buried at a certain depth within the silicon substrate (depending on the implantation energy), and the impurity concentration decreases as it approaches the wafer surface. Since the well profile in this case is different from that of conventionally formed wells (in which the doping concentration is highest at the surface, and decreases monotonically with depth), such deeply implanted wells are known as *retrograde wells*. The retrograde-doping profile is retained by minimizing the temperature cycles of later process steps. Retrograde wells can be implemented on both bulk wafers and on epitaxial wafers.[5]

Besides the potential benefit of increased packing density, such wells offer the following advantages:[4]

- A retarded electric field is created in the parasitic vertical bipolar transistor (thereby providing some protection against latchup; see section 6.4).
- Susceptibility to vertical punchthrough is reduced.
- The conductivity in the bottom of the well is increased, which also provides some further latchup protection (as will be explained in section 6.4.5).
- A higher threshold voltage can be achieved in the field regions of the *p*-wells since the boron implant is done following field oxidation (i.e., the boron does not segregate out into the field oxide as it is grown).[16]
- Lateral diffusion of the boron is also eliminated, thereby reducing encroachment of the boron into the active regions.

A disadvantage of the retrograde-well approach is that both the junction capacitance and body factor are significantly increased. For example, in a simulation of a 32-bit CMOS arithmetic logic unit (ALU), it was found that the circuit delay increases by

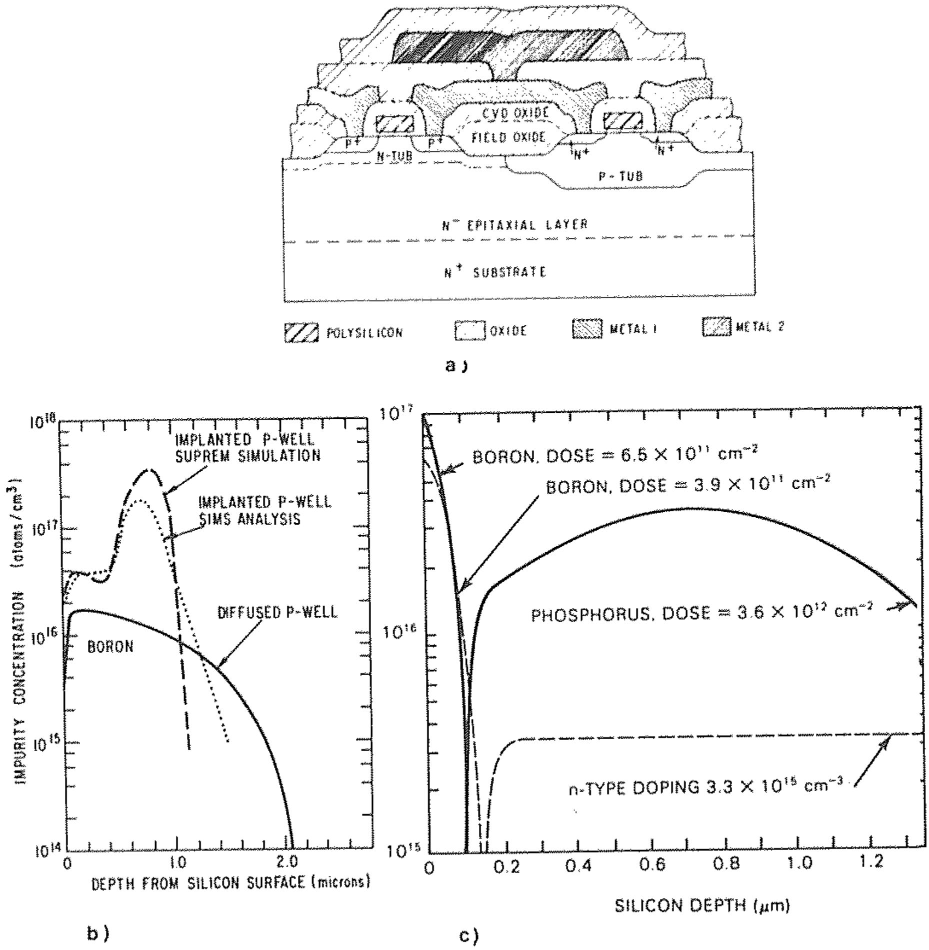


Fig. 6-11 (a) Twin-tub device cross section in which a retrograde p-well is used. General Electric AVLSI Process. (b) Retrograded p-well implanted impurity concentration profile. General Electric AVLSI process. Also shown is a conventionally thermally diffused well.[5] (© 1986 IEEE). (c) Simulated doping profiles in PMOS channel regions: dashed lines, lightly doped substrate; solid lines, retrograde n-well.[23] (© 1982 IEEE).

about 7% as the p-channel junction capacitance increases by 30 percent.[4] When the retrograde well is formed by means of a very high-energy implant, however, the doping concentration under the bottom of the source and drain regions is reduced, which reduces the junction capacitance. This implies that if a very-high-energy ion implanter (~1 MeV) were available as a manufacturing tool, the disadvantage of retrograde wells could be overcome.[17] In addition, the higher doping that will be required in the wells for fabrication of submicron CMOS devices may mean a less severe junction capacitance penalty.

Although both *p*-type[18] and *n*-type[19] retrograde wells have been demonstrated (see Figs. 6-11a and b,[5] and Fig. 6-11c,[23] respectively), the *p*-type technology has been more widely implemented. The reason for this is that a 700-keV (or greater) ion implanter is required for the formation of *n*-type retrograde wells. As of 1989 such machines were commercially available (e.g. NV 1003), but not yet found in ordinary production environments because of their relatively high cost. The *p*-type retrograde well, on the other hand, can be formed either by implanting singly ionized boron at 400 keV or doubly ionized boron at 200 keV (because of the larger projected range of boron compared with arsenic or phosphorus; see Vol. 1, chap. 9). Although the doubly ionized boron approach is achievable with conventional production implanters, it is not a trivial process to implement. In addition, a quadruple-well technology has been developed. This approach uses deep retrograde *p*- and *n*-wells, as well as shallow *p*- and *n*-wells (Fig. 6-12).[112]

A twin-retrograde-well 0.7-μm-CMOS process for fabricating 1-Mbit SRAMs has been reported.[74] The high energy implants also allow a restricted thermal budget to be used, thus reducing the up-diffusion from the p^+ substrate. This permits the use of a thinner epi layer (which also helps prevent latch-up), and also allows implantation of

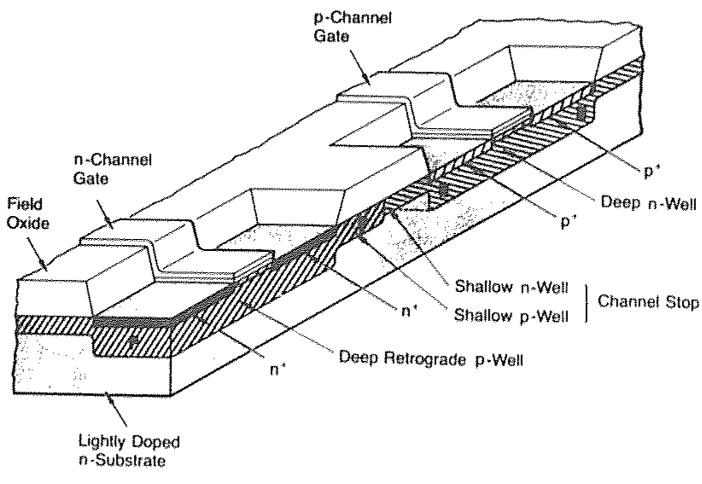


Fig. 6-12 Cross section of a quad-well CMOS device.[117] Reprinted with permission of VLSI Design.

the channel-stop dopants after the field oxide has been grown and planarized (thus minimizing lateral diffusion of the boron channel-stop dopants).

6.2.7 Summary of CMOS Well-Technology Issues

The selection of a particular CMOS process depends largely upon circuit applications, and to a lesser degree on technology evolution. For VLSI circuits with 1-2 μm design rules, *n*-well CMOS has been the more widely used single-well technology, both because *n*-channel devices can be optimized in the *p*-substrate and because most circuits use more NMOS than PMOS devices. In circuits that use transistor channel lengths smaller than 1 μm, twin-well and retrograde-well technology will become more attractive. As noted, the price will be increased process complexity.

The optimum process-integration design strategy is to first select the well technology (and to decide whether or not to use epitaxial substrates), and to then select the isolation method. Once these decisions have been made, the well depth and doping profile can be determined. The well depth will impact the lateral-diffusion distance of the well and the vertical-punchthrough voltage. The doping profile will affect transconductance, threshold voltage, source/drain punchthrough, junction capacitance, carrier mobility, source/drain-to-substrate breakdown, sensitivity to body effect, and hot-electron effects.

6.3 *p*-CHANNEL DEVICES IN CMOS

The fabrication of *p*-channel devices in CMOS presents some unique problems which arise from the need to build both NMOS and PMOS devices on the same chip. The problems revolve around the choice of a doping type for the polysilicon gate electrode and the impact of this choice on the threshold voltage and transistor action of PMOS devices.

6.3.1 PMOS Devices with n^+-Polysilicon Gates

As mentioned earlier, the threshold voltages of the *n*- and *p*-channel devices in a CMOS circuit should have comparable magnitudes for optimal logic-gate performance. To allow for maximum current-driving capability, the threshold voltages should also be as small as possible, with the minimum value dictated by the need to prevent excessive subthreshold currents under normal circuit operating conditions. For 5-V CMOS technologies, typical threshold voltages are 0.8 V for V_{Tn} and -0.8 V for V_{Tp}. Furthermore, the most common choice for the gate material is heavily doped *n*-type polysilicon. The work function of n^+ polysilicon is ideal for *n*-channel devices since these will yield threshold voltages of less than 0.7 V for reasonable values of channel doping and oxide thicknesses.*

* Note that the polysilicon layer may be combined with a layer of silicide for sheet resistance reduction; since the polysilicon remains as the underlayer of the polycide, the work function of the gate electrode will not be changed.

Figure 5-4, chapter 5 shows the value of V_T in devices manufactured with n^+ polysilicon gates (left scale)[18] as a function of doping and various gate-oxide thicknesses (Q_{tot} is assumed to be small enough that its effect on V_T can be ignored). The value of V_{Tn} is less than 1.0 V for gate-oxide thicknesses of 25 nm or less and a substrate doping of less than 10^{17} cm^{-3}. Thus, the threshold voltage of NMOS devices can easily be adjusted to the desired value of 0.7 V by means of ion implantation.

When n^+ polysilicon is used for the gate electrode of PMOS devices, however, it is not as easy to adjust V_{Tp} to -0.7 V. Figure 5-4 shows V_{Tp} (on the left scale for n^+ poly) as a function of substrate doping and gate-oxide thickness. In the doping range of 10^{15}-10^{17} cm^{-3}, V_{Tp} is already more negative than -0.7 V. Thus, implanting the n-doped body with more n-type dopant would only raise the *magnitude* of V_{Tp}, rather than bringing its value closer to the desired -0.7 V. To reduce the magnitude of V_T in PMOS devices using an n^+ polysilicon gate, it is necessary to implant the channel with a shallow layer of boron. The dose must be heavy enough to overcompensate the n-surface so that a p-region depleted of holes is formed. This shifts V_{Tp} toward more positive values by forming a compensating layer.

The fact that boron is implanted to adjust both V_{Tn} and V_{Tp} in CMOS circuits with n^+ polysilicon gates suggests that a single implant could be used instead of two separate implants. Figure 6-13 shows that this can be accomplished if the appropriate background dopings are chosen for the substrate and the well.[21] On the other hand, it

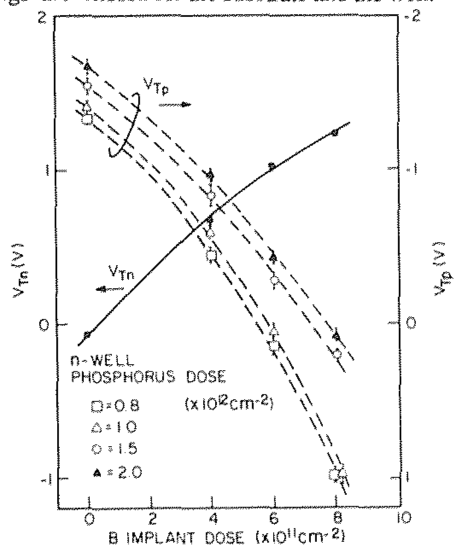


Fig. 6-13 Threshold voltages of n-channel (V_{Tn}) and p-channel (V_{Tp}) transistors as a function of boron threshold-adjustment dose. The CMOS structure uses an n-well implanted into a p-substrate (whose doping level is $6x10^{14}$ atoms/cm^3). V_{Tp} results are shown for various implant doses of the n-well.[21] (© 1980 IEEE).

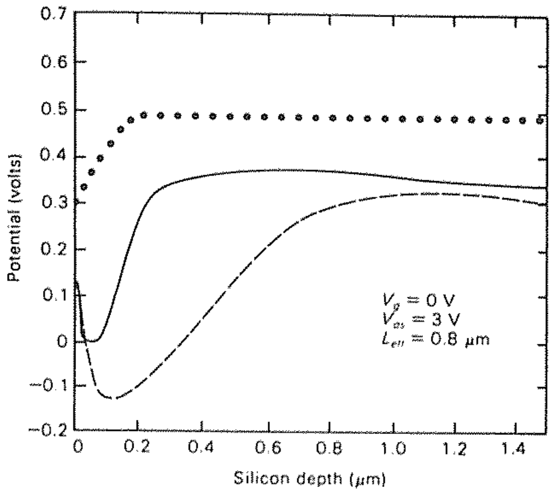


Fig. 6-14 Channel potential profiles in PMOS device in CMOS. *Dashed lines*: buried channel PMOS in a lightly doped substrate. *Solid lines*: buried-channel PMOS in a retrograde *n*-well. Dotted lines: surface channel NMOS drawn for comparison.[23] (© 1982 IEEE).

may be decided to use two separate implants in order to achieve better short channel behavior through individual optimization of the *n*- and *p*-channel devices.

6.3.1.1 Punchthrough Susceptibility. PMOS devices in which boron is used to adjust V_T exhibit a high susceptibility to punchthrough effects, since the boron implant produces a *p*-layer with a finite thickness. The potential minimum in the channel is thus moved away from the Si-SiO_2 interface (Fig. 6-14), causing more current to flow below the surface of the device.[23] Such PMOS devices are referred to as *buried-channel transistors*. As seen in Fig. 6-14 (which gives the *calculated* variation of the potential as a function of distance below the surface), the potential minimum moves further into the substrate as the thickness of the implanted *p*-layer is increased.

As the potential minimum moves deeper below the surface, the punchthrough susceptibility also becomes more pronounced (see section 5.5.2). This is illustrated in Fig. 6-15, which shows the results of a simulation[22] in which the lines of equipotential in the channel region were plotted for various depths of the channel junction Y_j (for a constant source/drain junction depth of 0.15 μm). As Y_j is increased from 0.1 to 0.2 μm, the drain electric field extends closer to the source for a constant gate and drain bias. Hence, this simulation predicts that more barrier lowering will occur as Yj is increased, leading to an increase in the punchthrough current.

The calculated predictions are supported by experimental data, as shown in Fig. 6-16, which plots the subthreshold I_D-V_{GS} characteristics of the structures described in Fig. 6-15. The subthreshold swing (S.S.) has the smallest slope when Y_j is 0.2 μm (indicating that the largest punchthrough current flows in this case). In fact, when

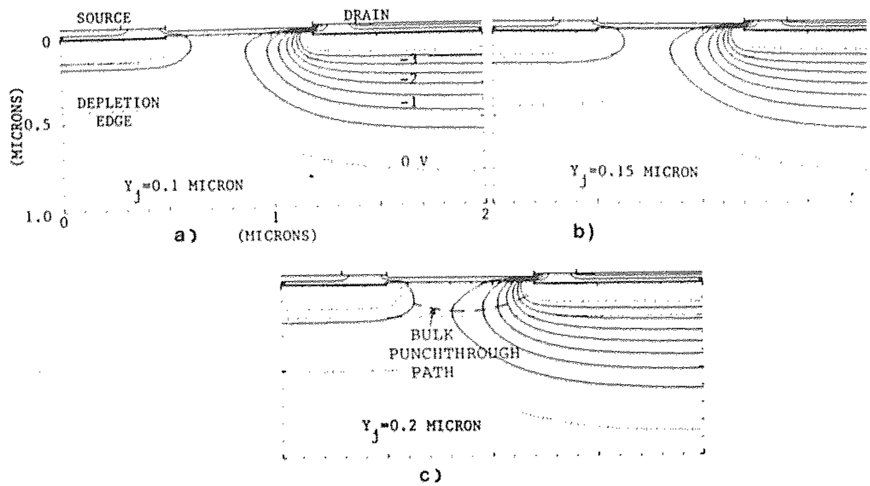


Fig. 6-15 Simulated 2-D potential profile for *p*-channel MOSFETs with Y_j = 0.1, 0.15, and 0.2 μm. The drain and gate bias are -3 V and 0 V, respectively.[22] From K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, 2nd Ed., Copyright Klewer Academic, 1989. Reprinted with permission.

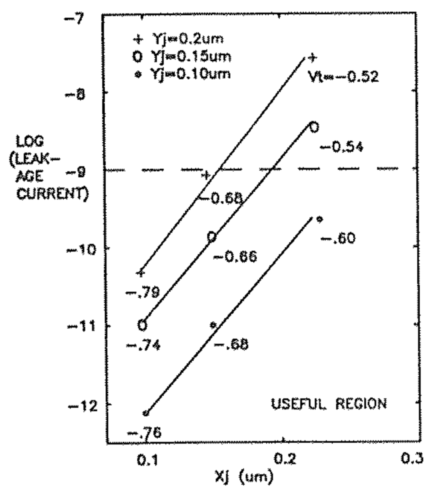


Fig. 6-16 Simulated subthreshold leakage current versus Y_j and X_j. L_{eff} = 0.5 μm, W = 50 μm, t_{ox} = 25 nm, V_{GS} = 0 V, and V_{DS} = -3 V.[22] From K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, 2nd Ed., Copyright Klewer Academic, 1989. Reprinted with permission.

$V_{GS} = 0$, I_{Dst} is increased by two orders of magnitude as Y_j is increased from 0.1 to 0.2 μm.

Leakage currents due to punchthrough in PMOS devices can be a significant problem in some CMOS IC applications. For ULSI devices, even a small value of leakage current per device may not be tolerable. For example, it has been shown that in submicron PMOS transistors ($L_{eff} = 0.8\ \mu$m), the punchthrough current will increase by two orders of magnitude if V_{DS} is increased from -1 V to -10 V.[24] Thus, if the punchthrough current were 1 nA at -1 V, it would be increased to 0.1 μA if V_{DS} were increased to -10 V. Such leakage would cause the dissipation of a few-tenths of a watt of power in a chip containing 1 million PMOS transistors.

The most obvious solution is to increase the PMOS device channel length. This is done in many CMOS technologies, and it is the reason that the minimum channel lengths of PMOS devices are often larger than those of the NMOS devices on the same chip. Another obvious technique is to make the *p*-buried layer as thin as possible. One of the reported approaches for achieving this involves implantation with BF_2^+ (which produces shallower boron layers than implantation with boron, see Vol. 1, chap. 9). Another approach is to use a high-energy *n*-implant (e.g., As at 400 keV) in order to place more *n*-type dopant atoms below the *pn* junctions (the more heavily doped regions absorb the drain voltage in a shorter distance, while simultaneously squeezing the channel *pn* junction toward the surface).[25]

To prevent a shallow implanted-boron layer from growing thicker, it is necessary to use a reduced thermal budget in order to restrict the process sequence following the implant in order to restrict boron diffusion. Specific steps for restricting boron redistribution include the following:[12]

1. Implant the boron through the gate oxide. This avoids the oxidation-enhanced diffusion of boron that would occur during the growth of the gate oxide if the implant were performed first. In this case, it is necessary to prevent the gate oxide from becoming contaminated during implant, either by material sputtered by the beam line or by vaporized resist material used as a mask against the implant. To prevent such contamination, a thin layer of polysilicon may be deposited on the gate oxide prior to the implant (in fact, immediately after the oxide is grown).[26]

2. After the implant has been performed, the remainder of the polysilicon film is deposited. This polysilicon is doped during the deposition step (at ~600°C) in situ with phosphorus, in order to avoid the 900°C phosphorus doping thermal cycle that would have to be used if the poly were doped following deposition.

3. A BPSG glass layer is used as the dielectric between the gate and the first level of metal. A significantly lower temperature can be used to flow BPSG than PSG (e.g., 850°C versus 1000°C). A lower temperature cycle can thus be used to smooth the surface topography (flow step) and gently taper the contact holes after etch (reflow step).

The use of a boron implant to adjust V_{Tp} becomes less feasible as devices use even thinner gate oxides, since larger doses of boron are needed. Y_j thus becomes deeper, and the punchthrough problem worsens. Solutions involving the use of gate electrodes other than n^+ polysilicon must therefore be explored. One alternative is to use p^+ polysilicon for PMOS devices; another is to use a material whose work function is very close to the mid-gap of silicon, thereby allowing symmetrical threshold voltages for n- and p-channel devices.[5]

6.3.2 PMOS Devices with p^+-Polysilicon Gates

When p^+ polysilicon is used for the gate material, V_{Tp} is shifted from the values that occur when n^+ polysilicon is used. Figure 5-4, chap. 5 plots V_{Tp} as a function of substrate doping when p^+ polysilicon is employed (using the scale on the right side of the figure). V_{Tp} can be seen to be less negative than -0.7 V over the substrate doping range of 10^{15}-10^{17} cm^{-3}. Thus, it can easily be made more negative through the implantation of phosphorus or arsenic into the channel. If p^+ polysilicon is used throughout the circuit, however, the NMOS device then has to be overcompensated in order for V_{Tn} to be reduced to sufficiently small values. (Note that some reports have been published on the use of p^+ polysilicon alone.[27,28]) This implies that it would be ideal to use both n^+ and p^+ poly gates on the same chip (with n^+ poly for NMOS devices, and p^+ poly for PMOS devices).[100]

Such a dual-doped poly approach, however, would add process complexity, and would also introduce other problems arising from the need to connect the two types of poly (e.g., at the input of an inverter). Such problems occur when a silicide overlayer (or *strap*) is used to make the connection between n^+ and p^+ poly (as a method to avoid a separate, space-consuming metal contact). Because the silicide strap provides a very rapid diffusion path for boron and arsenic, one type of poly can be counterdoped by the other when the device is subjected to high-temperature excursions. This counterdoping can occur to the degree that a region of poly can change doping types (i.e., from n^+ to p^+). In such cases, the threshold voltage of devices with counterdoped poly will be shifted from their designed value.

If the processing temperature is limited to 800°C after the two types of poly have been connected by the silicide, such lateral diffusion does not produce significant shifts in V_T.[29] On the other hand, temperatures of 900°C will produce sufficient counterdoping to significantly shift V_T. Since one of the last high temperature steps in CMOS is activation of the source/drain implants, this would mean either using a lower-temperature activation step, or performing the step prior to formation of the silicide layer. Process considerations for 0.5-μm CMOS technologies using both n^+ and p^+-poly gates are described in reference 85.

Another approach is to form a polysilicon electrode with an overlying conductor layer that suppresses such counterdoping. One such structure is a W-TiN-poly electrode structure.[105] The thin (30-nm) TiN film acts as a diffusion barrier to the dopants in the poly and also prevents reaction between the poly and the W. Very little of the dopant present in the poly is found to diffuse into the W, even following an anneal at 900°C for 1 hour.

Another problem encountered with p^+ poly gates when a thin gate oxide is used is poor V_T process control, due to penetration of the boron into the oxide (or further, into the silicon).[78] It is reported that boron will penetrate gate oxides that are ≤12.5 nm thick during a 900°C 30-minute post-implant anneal in N_2.[81] This implies that if p^+ poly gates are used, a lower processing temperature may be needed.[85]

If too low a temperature is used, however, the boron implanted into the polysilicon will not be sufficiently redistributed. The polysilicon dopant concentration at the polysilicon–gate oxide interface could thus end up being less than the mid-$10^{19}/cm^3$. This would make the work function of the polysilicon different from the desired degenerately doped value, creating V_T control problems in the MOS devices.

It has also been found that the presence of fluorine in the gate oxide worsens the boron penetration problem (Fig. 6-17a).[123, 124] Such fluorine can be introduced into the gate oxide if the PMOS source drain regions are implanted using BF_2. Elemental boron is therefore considered inherently superior to BF_2 as the implant species for surface-channel PMOS devices in CMOS technologies that use *p*-doped polysilicon. A study of enhanced boron diffusion through thin SiO_2 layers in a wet oxygen atmosphere is also reported in reference 107.

6.3.4 Gate Materials with Symmetrical Work Functions (for Both NMOS and PMOS Devices)

Because the larger work functions of molybdenum (4.7 V), tungsten, or refractory silicides produce low and nearly symmetrical threshold voltages for both PMOS and NMOS devices on moderately doped Si substrates, work has been conducted to investigate their suitability as gate-electrode materials.[30] For example, $TaSi_2$ gates have been successfully implemented.[31] Some of the benefits that such gate materials provide (besides symmetrical threshold-voltage values) include a reduction in subthreshold leakage currents and a decreased sensitivity to body bias (Fig. 6-17b).

Molybdenum and tungsten films deposited by means of magnetron sputtering have also been evaluated in terms of adhesion to SiO_2, mechanical stress, and compatibility with silicon processing techniques.[32,33] It was found that both could be deposited with low compressive stress by adjusting the deposition conditions of the sputter process. Both films can also exhibit good adhesion to SiO_2. (An additional advantage of Mo and W is that their resistivities are about 100 times lower than that of doped polysilicon and about 10 times lower than that of polycide gates.)

Some novel techniques had to be developed in order for compatibility to be established with the conventional Si-gate MOS process sequences. For example, it is necessary to use a wet H_2 ambient to oxidize the silicon without oxidizing the Mo[33] or W gates.[30] (This procedure is useful when a screen oxide is to be formed prior to the source/drain implant, but after the formation of the gate sidewall spacers used in LDD structure.) In addition, Mo and W form a layer of columnar grains, which makes such films susceptible to ion implant channeling along the grain boundaries. When the

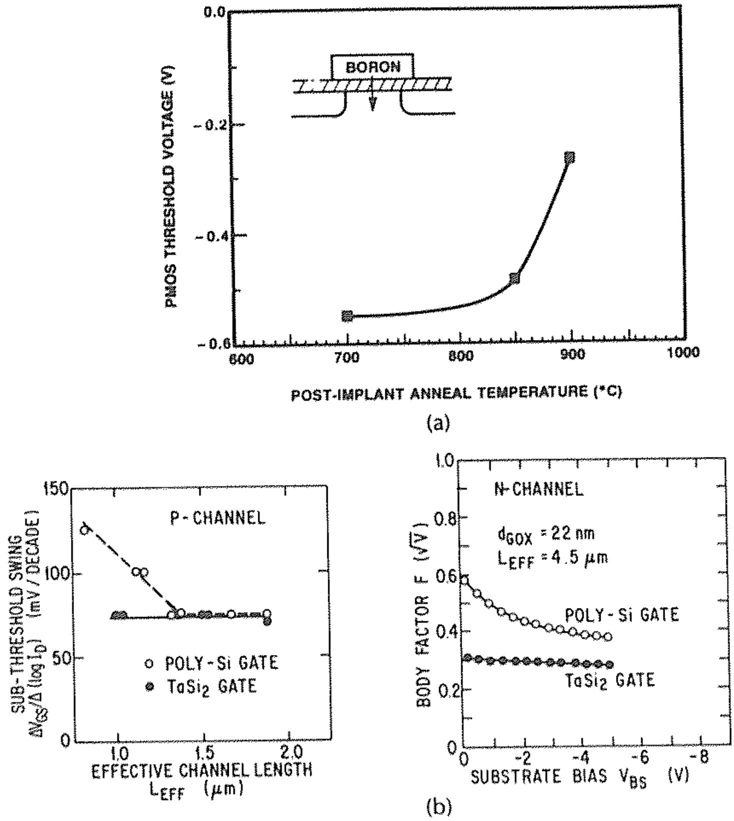


Fig. 6-17 (a) Boron penetration through the gate oxide causes the threshold voltage in BF_2- implanted PMOS devices to shift positive at anneal temperatures above ~800°C.[123] (© 1989 IEEE). (b) Sub-threshold characteristics versus effective channel length for poly-gate and $TaSi_2$-gate PMOS devices.[118] (© 1984 IEEE).

source/drain implantation is performed, some of the dopants may thus penetrate the gate film. This problem can be overcome by limiting the ion energy[33] or by depositing a thin layer of PSG (~60 nm thick) on the gate electrodes prior to implantation.[30]

Finally, V_T control with refractory metal gates was once a problem because of the relatively high concentrations of radioactive impurities (e.g., U and Th) in Mo and W sputtering targets. Such impurities can also produce soft errors in large memories. Recently developed chemical-purification procedures have significantly reduced the concentrations of U and Th in these sputtering target materials.[34] Methods for depositing Mo[35] or W[98] by CVD offer another option for forming of high-purity films. A deep-submicron CMOS process that uses a W gate has been reported.[77]

6.4 LATCHUP IN CMOS

A major problem in CMOS circuits is the inherent, self-destructive phenomenon known as *latchup*.[36] Latchup is a phenomenon that establishes a very low-resistance path between the V_{DD} and V_{SS} power lines, allowing large currents to flow through the circuit. This can cause the circuit to cease functioning or even to destroy itself (due to heat damage caused by high power dissipation).

The susceptibility to latchup arises from the presence of complementary parasitic bipolar transistors structures, which result from the fabrication of the complementary MOS devices in CMOS structures. Since they are in close proximity to one another, the complementary bipolar structures can interact electrically to form device structures which behave like *pnpn* diodes. In the absence of triggering currents, such diodes act as reverse-biased junctions and do not conduct. It is possible, however, for triggering currents to be established in a variety of ways during abnormal (but nevertheless frequently occurring) circuit-operation conditions (*e.g.*, terminal overvoltage stress, transient displacement currents, ionizing radiation, or impact ionization by hot electrons). Since there are many such parasitic *pnpn* structures on a VLSI CMOS chip, it is possible to trigger any one of them into latchup.

The phenomenon of latchup is well understood,[37] and many approaches have been implemented to control or even eliminate it. However, since the problem is increasing in severity as device dimensions continue to shrink, new latchup suppression techniques will be needed.

Latchup is fairly complex and can be a difficult concept to grasp for readers not well-versed in device physics. We will therefore briefly review the device concepts relevant to the problem before discussing the processing, layout, and circuit design techniques that have been developed to solve it.

6.4.1 Parasitic *pnpn* Structures in CMOS Circuits

Our example device configuration will be a *p*-well CMOS technology. As shown in Fig. 6-18a, lateral *pnp* and vertical *npn* transistor structures are inevitably created in *p*-well CMOS as a result of the multiple diffusions needed to fabricate *p*- and *n*-channel devices. The emitter of the lateral *pnp* transistor is the p^+ source and/or drain, while its base is the *n*-substrate and its collector the *p*-well. The n^+ source and/or drain comprises the emitter of the vertical *npn* device, while the *p*-well forms its base and the *n*-substrate its collector. The equivalent resistances of the substrate (R_{sub}) and the well (R_w) are also important elements of the latchup structure.

Several major aspects of these parasitic devices can be seen by studying the CMOS device cross-sections shown in Figs. 6-18a and 6-18b. First, the base of the *npn* transistor is connected to the collector of the *pnp* transistor (i.e., they are part of the same *n* region in the CMOS structure), and the base of the *pnp* is connected to the collector of the *npn*. We can therefore draw a simplified equivalent circuit diagram of this connection (Fig. 6-18c). It can be seen that the base of each transistor is driven by the collector current of the other, forming a positive feedback loop. Next, we note that

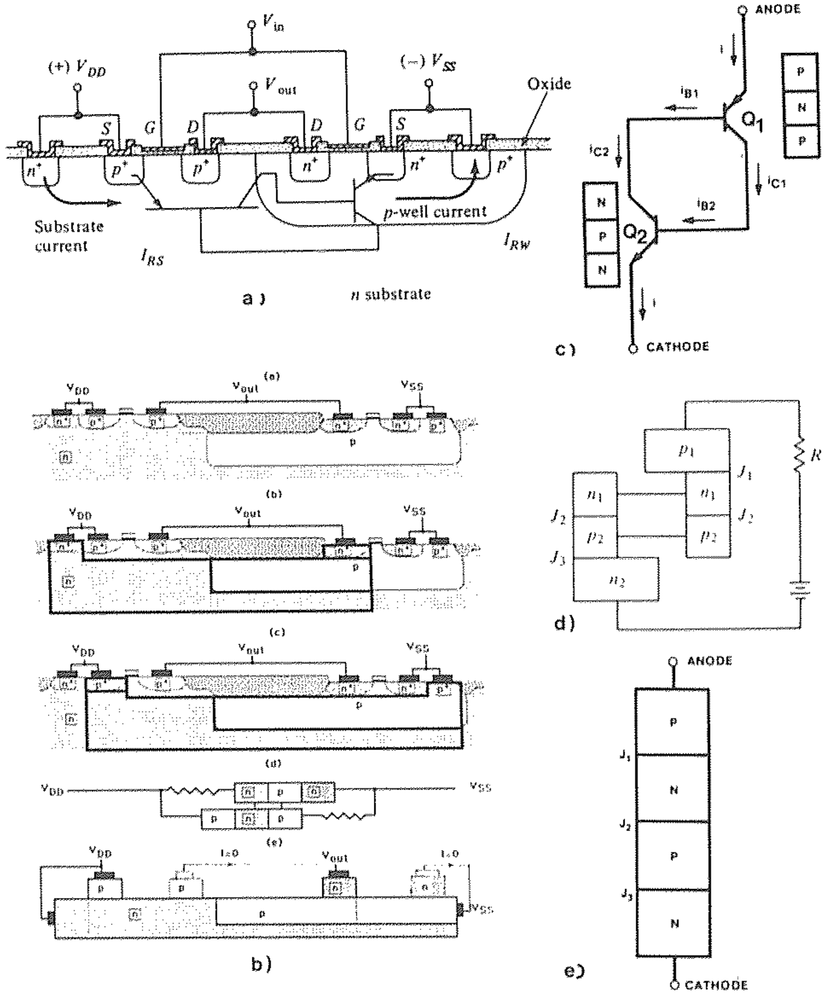


Fig. 6-18 (a) Cross-sectional view of a p-well CMOS inverter with parasitic bipolar transistors and lateral currents schematically shown.[4] (© 1986 IEEE). (b) Schematic showing the parasitic *npn* and *pnp* transistors in a p-well CMOS inverter. From W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission. (c) Simplified equivalent circuit diagram showing how these parasitic transistors are connected in the CMOS structure. (d) Schematic version of the circuit diagram of part (c). (e) *pnpn* diode obtained by merging the connected n_1 and p_2 regions shown in part (d).

the emitter of the *npn* is connected to V_{SS} (e.g., 0 V), and the emitter of the *pnp* is connected to V_{DD}. These external power-supply connections can also be added to the equivalent circuit diagram.

It should be noted that this simplified diagram ignores the parasitic resistances of the electrical paths through the bulk regions of the silicon, R_{sub} and R_w. These resistances, however, *are* very important in the understanding and control of latchup, and they will therefore be incorporated as soon as the simplified description of the device behavior is presented (Fig. 6-18c). The simple *pnpn* diode circuit to be used in the initial discussion turns out to be a "worst-case" possibility. That is, the addition of R_{sub} and R_w to the circuit model actually reduces the latchup susceptibility of CMOS. Since the two transistors are connected via their base and collector regions, by merging both the n_1 and the p_2 regions in Fig. 6-18d, we come up with a parasitic device structure (Fig. 6-18e).

6.4.2 Circuit Behavior of *pnpn* Diodes

Devices with structures like those shown in Fig. 6-18e and with external connections to the two end regions only, are known as *pnpn diodes*.* The terminal connected to the p_1 region is called the *anode*, and the terminal connected to the n_2 region is called the *cathode*. When an external voltage is applied with the anode voltage positive with respect to the cathode, and the resulting current I is measured, the I-V characteristic of this device is observed to have four distinct regions (Fig. 6-19); as follows:

Region 1. For voltages with values from *a* to *b*, very little current is observed to flow from anode to cathode, and the device is said to be in an *OFF*, *forward-blocking*, or *high-impedance* state. In this state, junctions J_1 and J_3 are forward-biased, and J_2 is reverse-biased. The externally applied voltage appears primarily across the reverse-bias junction.

Region 2. For voltages with values from *b* to *c*, the voltage across the reverse-bias junction, J_2, approaches the breakdown voltage of that junction. The current during this voltage excursion increases slowly up to the breakover voltage, V_{BO} (point *c*), at which point it suddenly increases abruptly.

Region 3. For voltages from *c* to *d*, the device exhibits a differential negative resistance (i.e., the current increases as the voltage sharply decreases). This is a transient state, which occurs as the device switches from the *OFF* state to *Region 4* operation (*ON* state).

Region 4. For voltages beyond *c*, the I-V characteristic exhibits *low-imped-*

* Note that the generic term *diode* simply means a device with two electrodes. Therefore, we should not expect the electrical behavior of all two terminal electronic devices to be alike. More specifically, we should not confuse *pnpn diodes* with *pn diodes*, with respect to electrical behavior. In fact, as we shall see, the I-V characteristic of the three junction *pnpn* diode is considerably different, and more complex, than that of the single-junction *pn* diode.

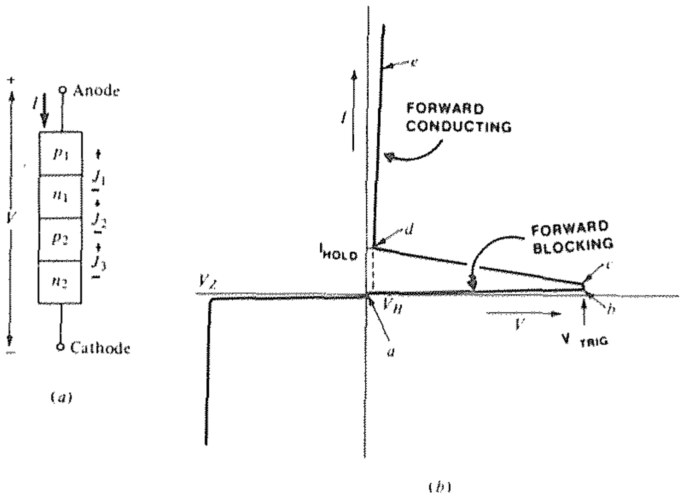


Fig. 6-19 (a) *pnpn* diode showing current reference direction and voltage polarities used in part b. (b) I-V characteristic of a *pnpn* diode.

> *ance* behavior. The I-V curve in this region behaves very much like that of a single forward-biased *pn* junction and is known as the *ON*, (or *forward conducting*) state. In this state the junction J_2 is forward-biased, and the voltage across the entire device is on the order of 1.0 V. If the current through the device is reduced while it is operated in this state, the device will remain *ON* only as long as I exceeds I_H (the so-called *holding current*). If the current is reduced below I_H, the diode switches back to the high-impedance state. The voltage across the device when I_H is flowing is called the *holding voltage*, V_H. Thus, a *pnpn* diode operated with a positive voltage between the anode and cathode is a bistable device that can switch from an *OFF* state to an *ON* state, or vice versa. If the external circuit can supply more current than I_H, the device will remain *latched* in the *ON* state as long as power is applied.

The parasitic *pnpn* structure in the CMOS circuits exhibits essentially the same I-V characteristic as the device just described. If the parasitic *pnpn* diode is triggered into operating in Region 4, and if the external circuit can provide the necessary holding current, the CMOS circuit will remain *latched up* in the *ON* state, even if the source of trigger current has been removed.

6.4.3 Device Physics Behavior of *pnpn* Diodes

Assume that a voltage source through a resistor once again applies a positive voltage to the *pnpn* diode, as shown in Fig. 6-19a, and that this produces a current I through the device. In *Region 1* the applied voltage forward-biases J_1 and J_3 and reverse-biases J_2.

If the emitter-base junction of a bipolar transistor is forward biased and the collector-base junction is reverse-biased, the transistor is in the *active* mode of operation. Thus, in the equivalent circuit of Fig. 6-18c, transistors Q_1 and Q_2 are both biased in the active mode when the diode is operated in *Region 1*. If a bipolar transistor is operated in the active region, the collector current of a transistor is given by $I_C = \alpha I_E + I_{CO}$. When this equation is applied to Q_1 and Q_2, then

$$I_{C1} = - \alpha_1 I + I_{CO1} \tag{6-2a}$$

and

$$I_{C2} = - \alpha_2 I + I_{CO2} \tag{6-2b}$$

where α_1 and α_2 are the common-base current gains of Q_1 and Q_2, respectively.

According to Kirchhoffs current law, the sum of the currents entering Q_2 must be zero. Using this law and the current components as shown in Fig. 6-18c, we get

$$I - I_{C1} - I_{C2} = 0 \tag{6-3}$$

Combining Eqs. 6-2 and 6-3, we obtain

$$I = \frac{\left(I_{CO_2} + I_{CO_1} \right)}{1 - (\alpha_1 + \alpha_2)} \tag{6-4}$$

It can be seen that as the sum of $\alpha_1 + \alpha_2$ approaches unity, the current I increases without limit. At this point the device is said to *break over*.

In bipolar transistors, the magnitude of α increases with collector current at low current levels (as shown in Fig. 6-20a). Since the value of I_C is increased as the avalanche breakdown voltage of the collector junction is approached, an increase in collector voltage can therefore lead to a significantly increase α.

When the collector currents are small, both α_1 and α_2 are much less than 1, and the current flowing through the diode is essentially the sum of the leakage currents, $I_{CO} = I_{CO2} + I_{CO1}$. As a result of the effects that lead to the increase of α, however, if the voltage across the *pnpn* diode is increased to a point near the collector-base junction breakdown voltage, the magnitude of the two alphas also increases. If the sum of the two alphas approaches unity, the current I begins to rise rapidly, which further increases the magnitudes of the alphas (this is an example of a *positive-feedback*, or *regenerative*, mechanism). When $\alpha_1 + \alpha_2 = 1$, breakover occurs, and *Region 3* operation sets in.

Beyond this point, device stability is provided by forward-biasing of the junction J_2. Since both junctions in each of the two transistors are now forward-biased, both Q_1 and Q_2 are in the *saturation* region of operation. In saturation, the current gain of a bipolar transistor α again becomes smaller. As a result, once the diode enters the *ON* region, the transistors enter saturation to the degree necessary to maintain the condition of $\alpha_1 + \alpha_2 = 1$. The current I then increases to a value that is essentially limited by the external circuitry.

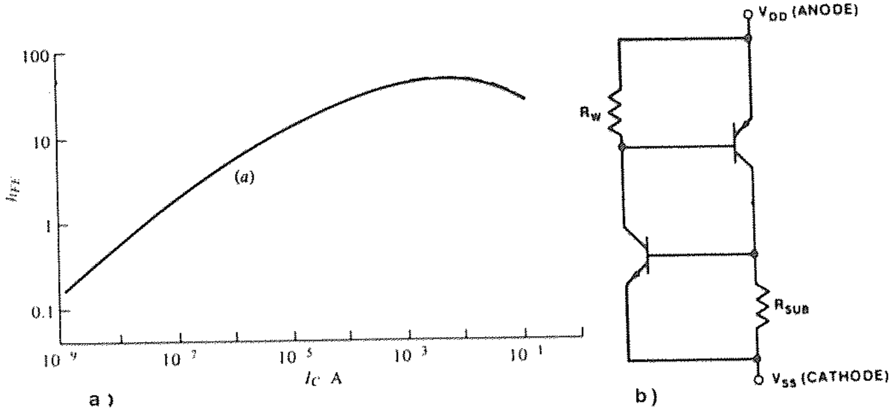


Fig. 6-20 (a) Dependence of current gain on the collector current. (b) Equivalent circuit model of the latchup structure in CMOS, including the parasitic resistances, R_W and R_{sub}.

In the *ON* region, the voltage across the device is the algebraic sum of the voltages across the three forward-biased junctions. Since the voltage drop across the center junction J_2 is in the opposite direction of the voltage drops across J_1 and J_3, the total drop across the *pnpn* diode in the *ON* region is about 1.0 V.

To maintain the diode in the *ON* region, the condition of $\alpha_1 + \alpha_2 = 1$ must continue to be satisfied. The holding current is the minimum current at which this condition is still met. If the current through the device is reduced to less than I_H (or if the applied voltage drops below V_H), the diode switches back to the *OFF* region of operation.

We can also express the condition $\alpha_1 + \alpha_2 = 1$ in terms of the β of the device by adding $(\alpha_1\alpha_2)$ to both sides of the equation and rearranging terms, to get

$$\alpha_1 \alpha_2 = \frac{(\alpha_1 \alpha_2)}{1 - \alpha_1 - \alpha_2 + \alpha_1\alpha_2} \qquad (6-5)$$

This can be further rearranged to give us

$$\frac{\alpha_1}{(1 - \alpha_1)} \frac{\alpha_2}{(1 - \alpha_2)} = 1 \qquad (6-6)$$

Since $(\alpha_1/1 - \alpha_1) = \beta_1$, and $(\alpha_2/1 - \alpha_2) = \beta_2$, we finally get

$$\alpha_1 + \alpha_1 = \beta_1 \beta_2 = 1 \qquad (6-7)$$

We will refer to the expression $\beta_1\beta_2 = 1$ as the *current-gain-product latchup condition for an ideal pnpn diode.*

6.4.4 Summary of Necessary Conditions for Latchup

Based on the principles of operation of *pnpn* diodes, the following four conditions must exist in a CMOS circuit in order for latchup to occur:

Latchup Condition #1. The emitter-base junctions of both parasitic bipolar transistors must be forward-biased. (It commonly happens, however, that initially only one transistor has its emitter-base junction forward-biased. This condition may then supply the current necessary to forward-bias the emitter-base junction of the second bipolar transistor.)

Latchup Condition #2. The product of the transistor gains must be sufficiently large to allow regenerative feedback. (We swill see that the minimum product of β_{npn} and β_{pnp} needed to induce latchup in real CMOS structures is usually much greater than 1. That is, the $\beta_{npn}\beta_{pnp}$ product needed to cause latchup = 1 in ideal *pnpn* devices, but >1 in actual CMOS structures.)

Latchup Condition #3. The external circuit must be able to supply a voltage equal to or greater than the holding voltage, V_H, of the *pnpn* structure.

Latchup Condition #4. There is a minimum latchup trigger time for which the stimulus must be present in order for the circuit to be latched.

6.4.5 Circuit Behavior of *pnpn* Structures in CMOS Circuits

In actual CMOS circuits, the parasitic *pnpn* device structures consist of two complementary parasitic bipolar transistors adjacent each other. The parasitic series resistances of the electrical path from the *n*-well contact to the *npn* collector, R_w, and of the path from the substrate contact to the *pnp* collector, R_{sub}, are also important circuit elements of the structure. Hence, the equivalent circuit that most accurately models the latchup structure in CMOS must include R_{sub} and R_w as well as the two parasitic bipolar transistors (Fig. 6-20b).

One of the effects caused by each of these two resistors is that a portion of the collector current of each transistor is siphoned away, reducing the base current. As a result of such current shunting, the effective current gains of the bipolar transistors are reduced. In fact, when finite R_{sub} and R_w are included in the circuit, Eq. 6-7 – in terms of the current gains β_{npn} and β_{pnp} – must be modified to more accurately express *Latchup Condition #2* for *pnpn* structures in CMOS circuits. For the latchup structures in actual CMOS, therefore, the inequality is expressed as

$$\beta_{npn}\beta_{pnp} > 1 + \frac{(\beta_{pnp}+1)\left(I_{R_{sub}} + \dfrac{I_{R_w}}{\beta_{pnp}}\right)}{I - I_{R_{sub}} - I_{R_w}\left(1 + \dfrac{1}{\beta_{pnp}}\right)} \qquad (6\text{-}8)$$

where I_{Rsub} and I_{Rw} are the currents that flow in R_{sub} and R_w, respectively. The minimum gain-product needed to induce latchup in CMOS circuits can thus be much greater than 1, depending on the values of the series resistances.

In general, the approaches to eliminating latchup can be divided into two categories: (1) those that reduce the bipolar transistor current gains, and (2) those that lower the value of the series resistances R_{sub} and R_w. If either of these approaches is successful, the latchup condition given by Eq. 6-8 will be harder to satisfy. In the extreme, for example, if R_{sub} and R_w could be made equal to zero, the two emitters would be short-circuited and would thus be prevented from ever turning on. Even if R_{sub} and R_w are only made smaller, the values of I_{Rsub} and I_{Rw} will increase, making the right side of the inequality larger. Therefore, various techniques have been developed to decrease the values of R_{sub} and R_w.

6.4.5.1 Value of β in CMOS Vertical Parasitic Bipolar Transistors. The current gain of the vertical parasitic bipolar transistor in an actual CMOS structure depends on the well depth, the well doping concentration, and the built-in field in the well (see chap. 7). In typical 1-μm CMOS structures, the well is 1-2 μm deep and is doped to ~1×10^{16} cm^{-3}. In *n*-well technology, the current gain of the vertical *pnp* transistor, β_{pnp}, will be ~100. In *p*-well or twin-well technology, the current gain of the vertical *npn* transistor, β_{npn}, will be two to three times larger than that of the *pnp* device in *n*-well technology (due to the higher minority-carrier mobility in the base region).

The transient response of the bipolar transistors is also important because of the minimum latchup trigger time (*Latchup Condition #4)*. In addition, latchup in a real circuit is normally induced by transient triggering. The transit time for minority carriers across the base region is a measure of the transient response of a bipolar transistor. The minimum latchup trigger time may be approximated by the sum of the vertical and lateral bipolar transit times.[40] A typical value of the transit time of a vertical transistor in a 1-μm *n*-well CMOS technology is several nanoseconds.

6.4.5.2 Value of β in CMOS Lateral Parasitic Bipolar Transistors. The current gain of a lateral bipolar transistor is determined primarily by the layout spacing between the diffusion outside the well to the well edge, since this is the dimension of the transistor's base width (although the gain is also impacted by the field doping outside of the well and the well depth). Since the base width of the lateral transistor is usually much larger than that of the vertical transistor, the β value is generally an order of magnitude lower (e.g., β ranges from ~2 to 4 in CMOS structures whose n^+ to *n*-well spacings vary from 5 μm down to 2 μm).

The larger base width also means that the base transit time is also much greater. The poor current gain of the lateral bipolar transistor reduces the tendency of the CMOS structure to undergo latchup when dc signals are applied, while the longer base-transit time increases the minimum triggering time for transient-induced latchup. As the layout spacing shrinks in high-density ULSI, however, the β is increased and the base transit time is reduced.

6.4.6 Circuit and Device Effects that Induce Latchup

For latchup to occur, the emitter-base junctions of both of the parasitic transistors must be forward-biased. This *circuit condition* can be produced in the latchup structure of CMOS circuits in three ways, each of which can be triggered by various physical stimuli. In addition, virtually all latchup failures occur as a result of transient stimuli. To fully describe the various latchup causing mechanisms, we will refer to Fig. 6-18d (the equivalent circuit model), and to Fig. 6-18a (the CMOS inverter cross-section). The three scenarios[36,38] that lead to latchup are described in the next paragraphs.

6.4.6.1 An external stimulus forward-biases the emitter-base of one transistor, and its collector current then turns on the second transistor. To explain latchup in such cases, let us assume that the externally applied stimulus is a *voltage overshoot at the output node of an* p-*well CMOS inverter driver circuit*. This is, in fact, the most common cause of latchup.[54]

The sequence of events is as follows: The *source* region of the PMOS device and the substrate contact are both connected to V_{DD} to ensure that the source-substrate junction is never forward-biased (Fig. 6-18a). The *drain* of the PMOS device, however, is connected to the output of the inverter. If this output is *low* (e.g., 0 V), the drain-substrate junction is reverse-biased, and no latchup can occur – that is, both of the structures that could be emitters for the parasitic *pnp* device (i.e., the source and the drain of the PMOS device) are at potentials lower than or equal to that of the base region. If the output state of the inverter is *high*, the drain-substrate bias should then equal 0 V; in this case there should still be no reason for latchup to occur (i.e., we assume that the output of the inverter is designed to reach V_{DD} in the high-output state).

If a voltage overshoot occurs at the output terminal, however, the output node experiences a condition in which V_{DD} is exceeded. (Overshoots and undershoots are both common, especially at input/output [I/O] device nodes, where signals tend to be noisy). If the overshoot causes the voltage at the drain region to exceed V_{DD} by more than about 0.6 V, the p^+-n drain-substrate junction becomes forward-biased. Holes are injected into the n-substrate (the base of the *pnp*; see Fig. 6-18a), from the p-type drain region of the output node. (This region behaves like a second emitter to transistor Q_1.) In essence, a triggering current flows through this emitter.

Some of the holes of this triggering current recombine in the base, while the remainder reach the p-well (the collector of the *pnp*). The latter represent the *pnp* transistor collector current, which now flows both into the base of Q_2 (I_{B2}) and through R_w to V_{SS}. The fraction of the current that flows through R_w causes a voltage drop across it; this voltage is also impressed across the emitter-base junction of the *npn* transistor (Fig. 6-18a).

If the voltage drop across R_w reaches 0.6 V, the *npn* device will be turned on. That is, the n^+ source (the emitter of the *npn*) will emit electrons into the *p*-well (the base of the *npn*). Some of these electrons will reach the *n*-substrate and will drift out of the V_{DD} terminal. If enough electron current exists in the *n*-substrate and if sufficient resistance, R_{sub}, exists between the V_{DD} contact and the p^+ source, an IR drop will develop, causing the p^+ source to inject holes into the *n*-substrate. This hole current adds to the initial hole current injected from the positively biased drain region. Thus, the positive-feedback scenario between the *pnp* and the *npn* transistors is triggered, and latchup is rapidly induced. A voltage undershoot when the output node of the inverter is in the low state will have the same net effect, except that the *npn* will be the first transistor to turn on.

Transient overshoot or undershoot voltages are a particular problem at the outputs of CMOS driver circuits. since impedance mismatches at the far ends of transmission lines or printed-circuit-board wiring traces result in reflections that return to the driver output node. It is therefore especially important to utilize stringent latchup-prevention measures at the input and output circuitry of CMOS chips (some combination of guard structures and multiple-well contacts should be used; see section 6.4.8.3).

6.4.6.2 An external stimulus causes current to flow through both bypass resistors, forward-biasing one or both bipolar transistors. Triggering mechanisms that can cause currents to flow in both bypass resistors (R_{sub} and R_w) include avalanche breakdown of the well-substrate junction (J_2, in Fig. 6-18e), photocurrents due to ionizing radiation, and *n*-well displacement currents (due to the charging or discharging of the large well-to-substrate junction capacitance). As shown in Fig. 6-21, the effects of these mechanisms can be modeled by adding a current source, I_o, and a capacitance, C_{ws}, to the equivalent circuit model of Fig. 6-20b. It is evident that leakage or displacement currents can still flow in both resistors, even if the bipolar parasitic transistors are in cutoff.

Examples of the external stimuli that produce these currents are: (a) voltages across the power-supply terminals that exceed the breakdown voltage of J_2 (and hence cause avalanche breakdown current, I_o); (b) ionizing radiation that causes photogeneration leakage current, I_o; and (c) external voltage transients (e.g., the voltage step function that occurs when a chip is powered-up) that produce displacement currents in the course of charging and discharging C_{ws}.

If the current in the bypass resistors is large enough, it can turn on both transistors. Typically, however, one of the resistances is larger than the other, and the voltage drop across the larger resistance causes the transistor to which it is connected to be turned on first. The positive-feedback mechanism that induces latchup is then set in motion, and when *Latchup Condition #2* is satisfied (as given by Eq. 6-8), the circuit latches up.

6.4.6.3 Current is shunted through one of the parasitic transistors by some degradation mechanism, and the resulting collector current flows through the bypass resistor of the second transistor, turning it on. The triggering mechanisms that effectively shunt

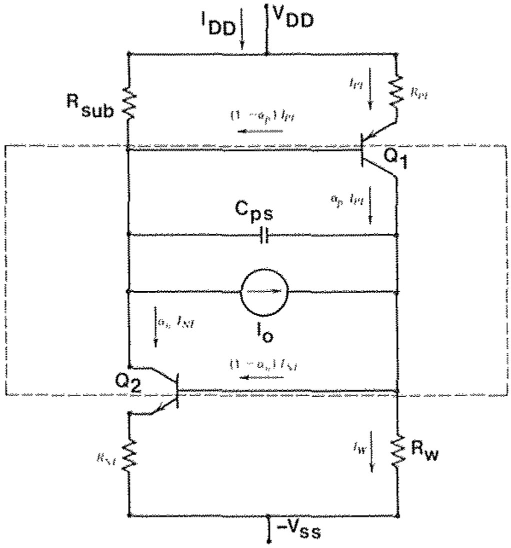


Fig. 6-21 Latch-up equivalent circuit including well-to-substrate capacitor, C_{ps}, and parasitic current source, I_o.

current through a low-impedance path from the emitter region to the collector region of one of the parasitic transistors include the following:

- Inversion of the field region between the source/drain regions of the MOS device in the substrate (emitter of lateral bipolar device) and the edge of the substrate (collector region).
- Punchthrough between the substrate region and the source/drain regions of the MOS device in the well.
- Avalanche ionization near the drain due to hot-electron effects. Note that in some cases this type of triggering mechanism can induce latchup even at voltages lower than the supply voltage, V_{DD}.

When one of these effects causes a large enough current to be shunted to the collector of one of the parasitic transistors, this current will flow through the bypass resistance of the second transistor, causing it to become turned on. Latchup will occur if the positive-feedback mechanism causes *Latchup Condition #2* to be satisfied.

6.4.7 Test Methods for Characterizing Latchup

The most common parameters used for characterizing latchup are trigger current (I_{trig}), holding current (I_H), and holding voltage (V_H). The trigger current is the current that must pass through the emitter-base junction of the *pnpn* device in order for latchup to

Table 6.1 Latchup Stimuli

I/O node voltage overshoot and undershoot	Type 1
Avalanche of well-substrate junction (extra-high voltage on some node)	Type 2
Photogeneration (ionizing radiation)	Type 2
Displacement current transient (voltage transient during power-up of chip)	Type 2
Inversion under field oxide	Type 3
Punchthrough between n^+ and n-well	Type 3
Hot-electron-induced current	Type 3

be induced (i.e., it is the current drawn from the power supply just before the test structure enters the latched state). Large values of I_{trig}, I_H, and V_H are desirable for reduced latch susceptibility. The efficacy of the processing and circuit-layout schemes designed to reduce the values of R_{sub} and R_w are therefore evaluated by measuring the values of I_{trig} and V_H. *If it is found that V_H is greater than the power-supply voltage, the circuit is said to be latchup-immune.* This assertion is based on *Latchup Condition #3*. Even if latchup is momentarily induced, it will not be sustained, because the power-supply voltage will be less than V_H.

The susceptibility of CMOS circuits to latchup is often determined experimentally by measuring the total current through the *pnpn* path while overstressing the anode voltage. With the source and n-well contact (Fig. 6-22) maintained at V_{DD}, the isolated voltage on the p^+ region in the n-well is raised above V_{DD}.[39] The value of I_{trig} is then measured for the circuit structure being evaluated. (Once a latchup has been triggered, the stressing voltage on the isolated region is returned to V_{DD}.) The value of V_H is measured by lowering V_{DD} after a latchup.

Techniques for studying the transient behavior of latchup have also been described. Such transient testing may provide better characterization of the latchup that occurs in

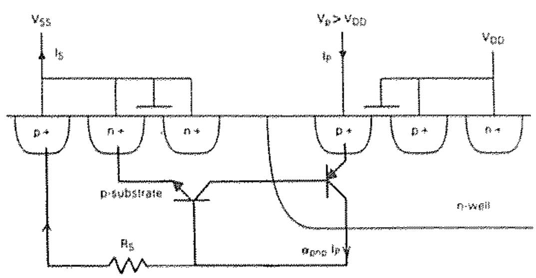


Fig. 6-22 Measurement technique for determining latchup triggering by p^+ overvoltage.

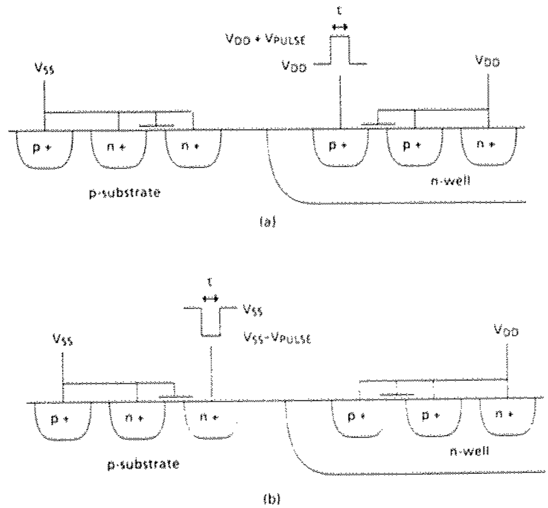


Fig. 6-23 Transient latchup measurement setups: (a) p^+ overvoltage triggering; and (b) n^+ overvoltage triggering.[116] (© 1987 IEEE).

actual circuit environments (*Latchup Condition #4*). One test involves the measurement of the *pnpn* diode current when different voltage ramp rates are impressed on the isolated p^+ region of Fig. 6-18a. This allows characterization of the displacement-current-triggering event, described as being one of the causes of Type 2 mechanisms. In another test setup, an overvoltage *pulse* is applied to the drain of the device described in Fig. 6-23. For a given pulse height, the pulse width is increased until latchup is triggered. As the pulse length approaches the bipolar transient response time, the pulse height needed to induce latchup increases. When the length is further reduced, latchup cannot be triggered, regardless of how large a pulse amplitude is applied.

A third transient test for latchup involves pulsing the base of one of the parasitic bipolar transistors.[40] Transient excitation with a pulse width shorter than the minimum regeneration time causes no latchup.

Another technique for evaluating the latchup hardness of a CMOS circuit design is given by Troutman,[36] who defines a differential latchup criterion. The following data need be gathered when this technique is used: the bypass resistance values; small-signal alphas of both parasitic transistors; base-emitter saturation currents for both transistors; and temperature.

6.4.7.1 Modeling Latchup in CMOS Technology. Modeling of latchup has been attempted by a number of different groups. It has turned out to be a difficult undertaking, since the phenomenon involves bipolar transistors with strong injection effects, with the structures also inherently distributed in nature. As a result, conventional one-dimensional device analysis is very difficult to apply.

6.4.8 Techniques for Reducing or Eliminating Latchup Susceptibility

Circuits are described as being either latchup-free or latchup-immune; the distinction between the two terms is important. *Latchup-free* refers to the ideal situation in which latchup will never occur under any circumstances. *Latchup-immune* refers to circuits that will not exhibit latchup under normal operating conditions, but that could be forced to do so through the application of sufficiently high voltages or the injection of high currents. While as of this writing there is no industrial standard for latchup hardness, a circuit is generally recognized as being latchup-immune if it can withstand an I_{trig} of more than 500 mA at the I/O pins.

Latchup immunity exists if V_H is less than the power-supply voltage per *Latchup Condition #3*. Many reports in the literature describe such a circuit as being *latchup-free*. However, Troutman, points out that even transient switching (unsustained latchup) can cause circuit problems and undesired power dissipation.[38] *Thus, he argues that the most effective way to prevent latchup problems is to ensure that the* pnpn *structures remain in the OFF state at all times.*

The approaches to reducing or eliminating susceptibility to latchup can be divided into three categories, as follows:

1. Processing techniques that reduce the current gains of the parasitic transistors to the degree that *Latchup Condition #2* cannot be satisfied.
2. Processing schemes that either reduce the values of R_{sub} and R_w or eliminate the parasitic *pnpn* diode structure.
3. Circuit-layout procedures that decouple the parasitic bipolar transistors.

6.4.8.1 Processing Techniques That Reduce Current Gains. Several techniques have been utilized in an attempt to reduce β of the parasitic bipolar transistors (bipolar spoiling). The first group involves methods that physically separate the emitter and collector regions of the lateral transistor (in which the n^+ regions are kept far from the *n*-well border). Obviously, for high-density circuitry, merely increasing the spacing of these regions on the wafer surface is not a good solution. Another approach for keeping these regions apart is to use recessed oxides and/or trench structures to increase the distance carriers must travel from the n^+ region to the well. While these techniques do help, by themselves they provide inadequate latchup protection. (These techniques will be described in more detail in the sections dealing with device isolation).

The techniques in the second group attempt to reduce the minority-carrier lifetimes in the base. These include gold doping to produce trapping sites,[41] neutron irradiation to cause structural damage and resultant recombination centers,[42] and harnessing of the oxygen precipitates formed in the substrate bulk by internal gettering processes for use as efficient recombination centers for minority carriers.[43] The first two of these lead to increased leakage currents, while the third has little effect on lateral transistors with small base widths, since the precipitates do not form in the lightly doped epi. As a result, the techniques in this group have not been widely implemented.* The third type of technique involves the use of a retrograde well, which reduces the vertical bipolar gain by providing a high Gummel number and a retarded E-field in the base.[4] However, care must be taken to remove injected carriers so that lateral transistor action to the well edge is not increased.[39] This approach also requires the availability of a high-energy implanter.

The fourth type involves the use of silicided source/drains[45,46] to reduce the emitter efficiency of the parasitic bipolar transistors. These techniques have the disadvantage of degrading the gain of the MOS devices.[47]

In summary, the methods described above can help to reduce latchup susceptibility, but provide an insufficient degree of latchup protection. The remaining two categories of latchup-suppression techniques, have been found to be much more effective, and as a consequence, they have been much more widely implemented.

6.4.8.2 Processing Techniques That Reduce R_{sub} and R_W or Eliminate the *pnpn* Structure. Techniques that reduce the series resistance values include the use of *epitaxial layers on heavily doped substrates* and the use of *retrograde implants*. Techniques that eliminate the *pnpn* path use either: (1) a combination of *deep trench isolation structures* and *epitaxial layers on heavily doped substrates*; or (2) *silicon-on-insulator* (SOI) substrates.

The use of lightly doped epitaxial layers on heavily doped substrates is effective in reducing latchup susceptibility for two reasons. First, the highly doped layer substantially reduces the value of R_{sub} by placing a low-resistance path for *majority carriers* in the substrate in parallel with the more lightly doped epi region in which the devices are formed. Hence, it very effectively shunts the lateral parasitic bipolar transistor. Second, any *minority carriers* injected into the epi layer that then diffuse into the highly doped substrate are more rapidly recombined there, so that fewer reach the collector of the lateral bipolar device.

Figure 6-24 shows the trigger current and the holding voltage as functions of n^+ to p^+ spacing, d, for various epi thicknesses.[48] Thin epi can be seen to dramatically improve latchup immunity from either standpoint. For a 12-μm epitaxial layer, the triggering current required for latchup is less than 1 mA when $d = 10$ μm. When the epitaxial thickness is reduced to 3 μm, d can be reduced to 5 μm, and I_{trig} is increased to 80 mA. The advantage diminishes as d approaches the epi thickness. However, the minimum epi thickness is limited by outdiffusion of impurity atoms from the heavily

* An approach to creating oxygen precipitates in a thin layer 2.5 μm below the surface by means of ion implantation of oxygen and epi growth has recently been reported.[44] Since these precipitates are positioned so that they reside in the base region of the lateral *pnp* device, they are effective in reducing its alpha by decreasing the minority-carrier lifetime in the base.

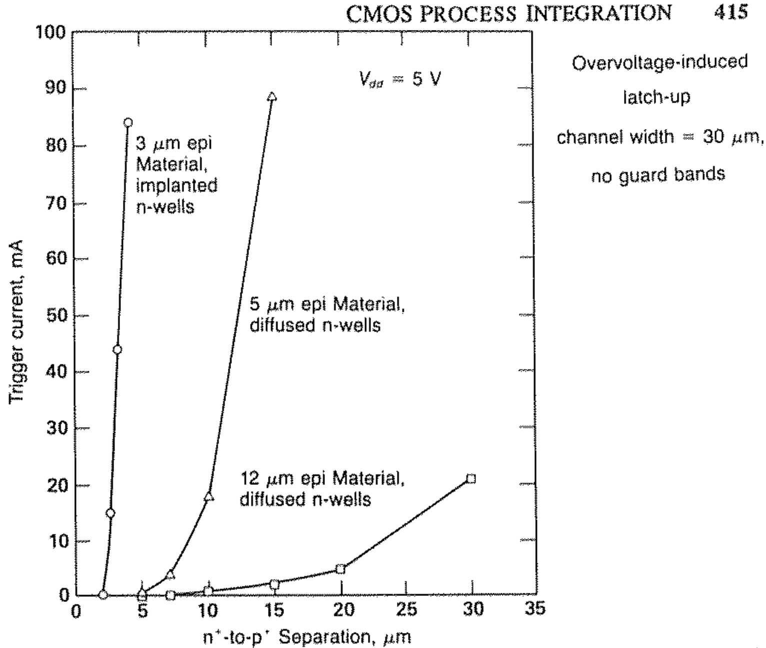


Fig. 6-24 Triggering current versus n^+-to-p^+ separation in an n-well CMOS structure.[48] (© 1985 IEEE).

doped substrate during epi deposition and subsequent wafer processing (see section on autodoping in Vol. 1, chap. 5). That is, due to this effect the *effective* epi thickness is thinner than the *initial* epi thickness. (The effective epi thickness is defined as the thickness of the epi layer when the doping concentration is less than some specific number – e.g., $1x10^{18}$ cm^{-3}.) The fact that n-epi on n^+ produces a sharper epi interface than p-epi on p^+ implies that n-epi should provide better latchup protection.[4] But as noted in section 6.2, on well technology in CMOS, this benefit is obtained only at the price of accepting the other drawbacks inherent to n-epi on n^+ wafers.

The disadvantages of epi CMOS include increased wafer cost, lower breakdown voltages, and increased leakage currents.[49] In addition, epi-layer growth may generate defects that will lead to reduced chip yields.[50]

With respect to *retrograde implants,* it has been found that retrograde wells can be effective against latchup for several reasons.[4] In p-well technology, the retrograde implant is used to form the p-well, with the following advantages being gained:

- The retrograde profile gives a high doping concentration deep in the well, which reduces the gain of the vertical bipolar device.
- The heavier doping near the bottom of the well reduces R_W, thereby helping to decouple the bipolar transistors.

- The retrograde well can minimize thermal processing following epitaxial-layer deposition. Epi layers with thinner *effective* thicknesses may therefore be possible if retrograde wells are combined with epitaxial layers.

- If the combination of the two techniques is used, a process that also reduces *both* R_{sub} and R_w is obtained.

In *n*-well technology, the retrograde implant is a *p*-type implant that is placed beneath the NMOS devices in the substrate.[51] This creates a p^+ layer in the substrate that has a much smaller transition region than a *p*-epi-on-p^+ substrate, thus providing even better latchup protection.

Trench isolation, as described in chapter 2, allows n^+ and p^+ regions to be placed close to one another in CMOS. If the trenches are deeper than the well regions, they provide physical separation among device types. However, when trenches are used without epitaxial layers on heavily doped substrates, their main contribution to latchup immunity is that they force carriers in the lateral transistor base region to travel greater distances to reach the collector. As a result, trench isolation alone does not significantly increase I_{trig} or V_H. The major advantage of using trenches on their own is the increase in latchup response time, which results in a substantial improvement in *transient upset* prevention. When deep trenches are combined with a lightly doped epitaxial layer on a heavily doped substrate, however, latchup-free structures are produced.[11,96] As minority carriers injected into the substrate attempt to diffuse toward the collector, the deep trench forces them into the heavily doped substrate. There they rapidly recombine, substrate and hence never arrive at the collector region.

Silicon-on-insulator technology is also receiving wide attention, since in addition to providing electrical isolation between the MOS devices, it results in latchup-free CMOS structures. Each MOS device is isolated from neighboring devices by the insulating layer; as a result, the *pnpn* path is no longer present.

6.4.8.3 Circuit-Layout Techniques for Decoupling Parasitic Bipolar Transistors. Two types of structures can be incorporated into the circuit layout that will provide latchup protection: *guard structures*, and *substrate and well contacts.*

Guard structures are heavily doped diffused regions that encircle the well. Troutman states that they are the most effective layout procedure available for providing latchup protection,[39,54] and this claim is supported by the simulated and experimental evidence of Menozzi et al.[52] Guard-ring structures can be fabricated either outside the well (in which case the guard surrounds the outer edge of the well region), or inside it (in which case the well is placed between the active device regions and the well border). The guard can be either a *minority-carrier* or a *majority-carrier* structure.

Minority-carrier guard rings have a doping type opposite to that of the region in which they are formed (Fig. 6-25). Normally, these guard rings are placed in the substrate outside of the well edge (e.g., an *n*-type diffusion in the *p*-substrate of an *n*-well CMOS technology). Since the guard is connected to the power supply in such a way that the *pn* junction it forms with the substrate is reverse-biased, this type of guard

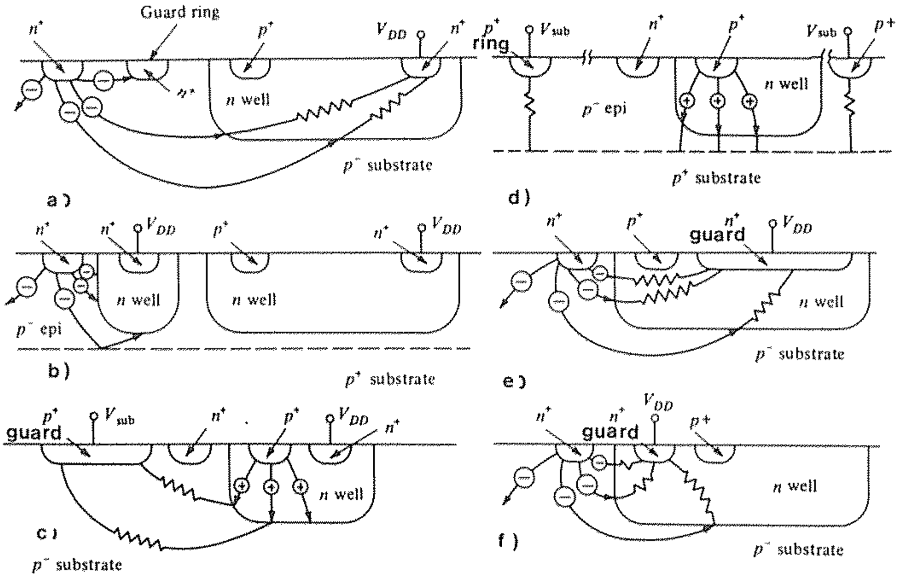


Fig. 6-25 [(a) and (b)] Minority-carrier guard in substrate: (a) n^+-diffusion guard; (b) Deeply diffused guard in epitaxial CMOS. [(c) and (d)] Majority-carrier guard in substrate: (c) p^+-diffusion guard for reducing substrate sheet resistance; (d) Contact ring is preferable to p^+-diffusion guard in epi-CMOS. (e) Majority-carrier guard in well. (e) n^+-diffusion guard to reduce n-well sheet resistance. (f) n^+-diffusion guard to steer current away from vertical *pnp* emitter.[36] From R. R. Troutman, *Latchup in CMOS Technology*, Copyright Klewer Academic, 1989. Reprinted with permission.

ring will collect any minority carriers in the substrate that diffuse into its depletion region. The carriers are thus prevented from reaching the well, and as a result, the collector current of the lateral bipolar transistor is reduced. The main role of minority-carrier guards in preventing latchup is to provide this reduction in effective current gain.

If a thin epitaxial layer is used together with a deeply diffused guard, the effectiveness of the guard is increased (Fig. 6-25b).[36] That is, the path of minority-carrier diffusion is narrowed (by the reflecting boundary where the high and low doped regions in the substrate meet); this forces the carriers closer to the guard ring, where they can be intercepted. In addition, any minority carriers that enter the heavily doped substrate recombine more rapidly there.

Majority-carrier guard rings, on the other hand, are doped with the same type of dopant as the regions in which they are formed. While they can be implemented outside the well, they are usually more effective when placed inside. The function of this type of guard is to provide a shorter (and, hence, a lower-resistance) path for the carriers that constitute the collector current in the well (or substrate). The majority carriers in the collector must drift to the power-supply node through R_w (or R_{sub}) once they enter the collector region, and the guard rings effectively reduce the value of R_w and R_{sub} by placing smaller resistances in parallel with these series resistances.

The reason that the majority guard is usually placed within the well is that the value of R_w is generally higher than that of R_{sub}. The effective lateral resistance can be quite high in shallow wells, since the source/drain diffusions in the well pinch the lateral current paths. Thus, a guard ring in the well that reduces the distance between the well edge and the well contact can be very effective in reducing the large value of R_w. A way to implement three guard structures with only two diffusions is to place one guard so that it overlaps the well and substrate boundary (Fig. 6-26).

Voltage drops along the power supply bus to which the guard ring is connected should never be allowed to become large enough to initiate a latchup condition. For example, if the bus to which the guard is attached has a sufficiently large resistance that the voltage of the n^+ guard in the n-well drops to 4 V, when the output rises to 5 V, the emitter tied to this higher voltage will turn *ON*, initiating a potential latchup event.

Troutman also emphasizes that a substrate-contact majority-carrier ring should be mandatory for all chips.[36,39] This would minimize lateral bypass resistance by distributing substrate majority carriers. When used with epitaxial CMOS, such contact rings can reduce lateral bypass resistance to below 1 Ω; in addition, they are as effective as backside substrate contacts in eliminating latchup. In fact, the need for multiple majority guard rings (which would surround each of the well regions) can be eliminated through the use of a single substrate contact ring (as long as it is used together with wafers that have a thin epi layer on a highly doped substrate).

The major limitation of guard rings is that they decrease overall circuit density. In the input/output circuits of CMOS, however, the MOS transistors must be made quite large in order for sufficient off-chip drive-current capability to be provided. Thus, the area penalty imposed by guard-ring use around these circuits is usually quite small. Because of the noisy signals encountered at the I/O nodes of a chip, the use of guard rings to suppress latchup in such circuitry is widespread.[54] Each of the output devices in the substrate is typically surrounded by a majority-carrier guard ring (tied to V_{DD} in p-well CMOS, and to V_{SS} in n-well).

Note that n-well CMOS generally provides better latchup immunity than p-well, for two reasons.[8] First, since the mobility of majority carriers is higher in an n-well than

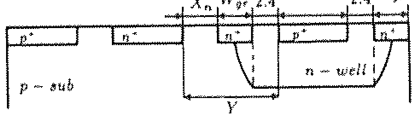


Fig. 6-26 Implementation of two guard structures with only one diffusion by placing one guard so that it overlaps the well–substrate boundary.[52] (© 1987 IEEE).

in a p-well, R_w in the n-well is lower than in a comparable p-well. Furthermore, R_w is usually much higher than R_{sub}, especially in thin wells that are pinched by source and drain diffusions. Thus, the lower value of R_w in the n-well can be of significant help. Second, because hot-electron-induced substrate currents are much higher in NMOS than PMOS devices, it is better to have the NMOS devices in the substrate. Such currents are more easily collected from here than from the well region, especially if an epitaxial layer on a heavily doped substrate is used.

The second type of circuit-layout technique to be discussed is that of *substrate and well contacts*. The use of guard rings in the well may represent too large an area penalty because in high density circuits multiple well contacts can serve as an alternative approach to improving latchup immunity. However, this approach is not as effective as the use of guard rings.[52]

The number and placement of power-supply contacts to both the well and the substrate also impact the effectiveness of this approach. Increasing the number of well contacts to V_{DD} reduces latchup susceptibility, since the resistor length of R_w for each FET in the well is reduced when a well contact is in its proximity. The contact spacings should therefore be no more than two squares apart, to ensure that no local high-resistance regions exist. The well contacts should be hard-wired to the power-supply connection (ground or V_{DD}) with a metal stripe. In this way, any injected charge will be shunted to the supply rail through a low-impedance path that will not contribute to the well's already relatively high lateral resistance.

The closer the contacts are to the source regions of the MOS devices, the smaller the potential drop during current flow, leading to increased latchup immunity. The use of *butted source-substrate* and *butted source-well* contacts, in which the n^+ and p^+ regions are contiguous and connected to the same potential has thus become popular.[101] While such butted contacts also conserve area, they are limited to FETs that operate with grounded sources.

6.5 CMOS ISOLATION TECHNOLOGY

In CMOS ICs, *like* kinds of devices within a given well must be isolated in the same manner as the devices in either NMOS or PMOS circuits (i.e., through a combination of a thick field oxide and channel-stop doping). However, the isolation requirements of CMOS technology extend beyond those of either PMOS or NMOS alone, in that in CMOS it is also necessary to isolate the p- and n-channel devices from one another. The isolation of p-channel from n-channel devices must satisfy two requirements: (1) any possible leakage currents that could flow between adjacent PMOS and NMOS devices must be suppressed; and (2) the susceptibility of CMOS to latchup must be minimized.

In CMOS structures, the isolation spacing between the n- and p-channel devices is defined as the total of the distances between the edge of the n^+ region of the n-channel device and the edge of the well, and the edge of the p^+ region of the p-channel device and the edge of the well (or, in other words, the n^+ to p^+ spacing – Fig. 6-27).

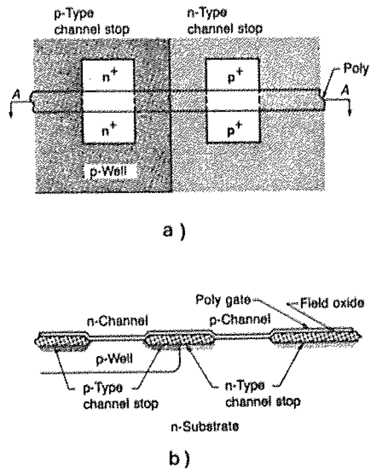


Fig. 6-27 (a) Layout top view of the isolation region between *n*-type and *p*-type transistors in CMOS. (b) Cross sectional view of a CMOS inverter showing the channel stops that are designed to prevent the surface under the field oxide from inverting.[53] (© 1986 IEEE).

Basic isolation between *p*- and *n*-channel devices in CMOS is established by both a reverse-biased well-substrate junction and by the regions under the field oxide (as long as they are not inverted). This isolation, however, may not be perfect. Unwanted current flow can arise due to junction breakdown, the formation of leakage paths between devices, or latchup. The key *leakage paths* involve the *parasitic MOS field devices* created in the field regions between the devices.* Leakage currents can arise if these well-border parasitic field devices conduct prematurely, as a result of surface inversion, or punchthrough below the surface.** As noted, the *latchup* effect in CMOS arises as a

* The parasitic NMOS field transistor at a *p*-well border consists of the n^+ source, the *p*-well body, the *n*-substrate drain, and the polysilicon gate runner over the field oxide, as shown in Fig. 6-27b. Similarly, the parasitic PMOS field transistor at this well border consists of the p^+ source, the *n*-substrate body, the *p*-well drain, and the poly runner over the field oxide.

** In Fig. 2-4 of chapter 2 we described how punchthrough can occur between the source and drain of the same device, or between the source/drain regions of neighboring devices of the same channel type. In this case, we refer to the punchthrough between the border of the substrate and source/drain regions in the well.

result of the *parasitic bipolar devices.* Various isolation techniques have therefore been developed to prevent both leakage and latchup. Device and process simulators used to model such CMOS isolation structures are described in chapter 9.

Interest in n- and p-channel isolation techniques is very keen, because such isolation requires much more area than between like types of devices. For example, in early single-well CMOS technologies, the isolation spacing required was typically about three times the diffusion depth of the well.[8] Thus, for a 4-μm well depth, a minimum n^+ to p^+ spacing of ~12 μm was needed. In early twin-tub CMOS technologies, which used dual-channel stops, a minimum of 3-9 μm of lateral space was necessary for effective isolation.* In chapter 2, however, we showed that isolation between *like* kinds of MOS devices can be accomplished with isolation spacings of only 1.0-1.5 μm. The large area penalty of p- to n-channel device isolation is one reason why CMOS technologies using conventional isolation methods cannot achieve as high a packing density as NMOS.

This isolation-spacing requirement is due in part to the processes used to fabricate CMOS devices. Wells are typically driven in quite deeply to ensure that enough charge exists below the transistor to prevent vertical punchthrough to the substrate. This results in both a lateral diffusion of the well dopant and a reduction in the surface concentration near the border of the well (Fig. 6-28). The channel stop doping in the

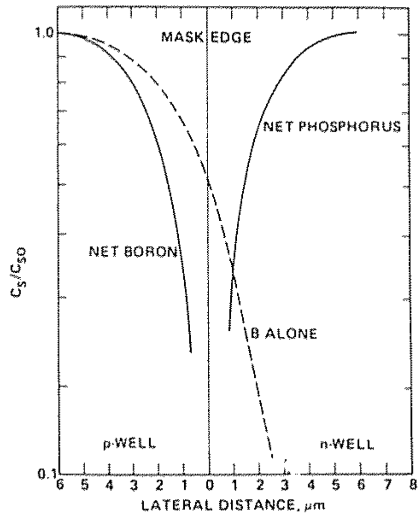


Fig. 6-28 Impurity surface concentration near the border of the two tubs in twin-tub CMOS.[14] (© 1980 IEEE).

* A 3-μm minimum space between n^+ and p^+ regions was reported for a non–trench isolated, twin-well CMOS process. To achieve such tight spacing, high-pressure oxidation was used to grow the field oxide. This was more effective in preventing interdiffusion of the impurities from the two wells than an atmospheric-pressure field-oxidation process would have been.[29]

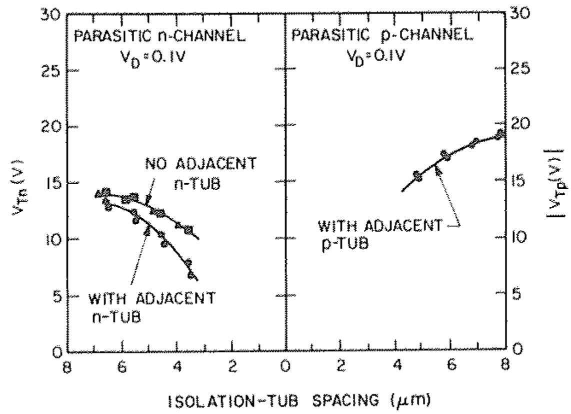


Fig. 6-29 Threshold voltage of n- and p-channel parasitic device as a function of the separation between the transistor edge and the tub border.[14] The upper curve in the left-hand graph shows the parasitic n-channel threshold-voltage reduction near the tub border when no adjacent n-tub is present. When an n-well is present, the threshold voltage is reduced (lower curve), since the interdiffusion of the two types of impurities reduces the net surface concentration of each dopant near the well border.[14] (© 1980 IEEE).

substrate at the edge of the well is also substantially reduced, due to the compensation between the acceptor and donor atoms; this causes a reduction in the threshold voltage of the parasitic MOS field transistors. We will next describe this phenomenon in more detail for the case of twin-tub CMOS.

Figure 6-29 shows the threshold voltage of each type of parasitic device as a function of the separation between the transistor edge and the tub border.[14] The upper curve in the left-hand graph shows the parasitic n-channel threshold-voltage reduction near the tub border when no adjacent n-tub is present. When an n-well is present, the threshold voltage is reduced (lower curve), since the interdiffusion of the two types of impurities (Fig. 6-28) reduces the net surface concentration of each dopant near the well border. Unless the net dopant-reduction effects are somehow counteracted, the spacing between adjacent n^+ and p^+ devices must be kept quite large (e.g., >10 μm) to prevent inversion beneath the field oxide.

If smaller spacings between n- and p-channel devices are to be possible, the channel-stop doping concentration must somehow be increased, particularly in the substrate regions in n-well CMOS. During field oxidation, boron segregates into the oxide, while phosphorus piles up at the silicon surface. As a result, in an n-well process a separate p-type channel-stop implant must be added to increase the surface concentration of the lightly doped p-substrate. Without such an implant, inversion between the n-channel devices is likely to occur. In *p*-well technologies the well itself is an adequate *n*-type channel stop, because heavier boron doping exists in the *p*-well, and the concentration of phosphorus is increased at the surface of the *n*-substrate during field oxide growth.

In twin-tub processes, two channel stops are needed to reduce the isolation distance as much as possible. Figure 6-30, which is based on a 2-D device model,[53] shows the impurity contours in the isolation region of a CMOS twin-tub process (with retrograde wells).

The AT&T Twin-Tub V technology,[55] is an example of an advanced CMOS process uses a LOCOS-based approach to isolate like devices. The starting material is a *p*-epi layer on a p^+ substrate. The technology is implemented with a single-mask, self-aligned twin-well process that uses two channel-stop implants (as well as two separate well implants).

First, phosphorus and arsenic are sequentially implanted into the *n*-well areas, forming a high/low doping profile (channel-stop and well; Fig. 6-31). Since arsenic diffuses much more slowly than phosphorus, it will remain near the surface of the well during the drive-in step. This high arsenic concentration provides an effective channel stop for *p*-channel devices and also protects against punchthrough.

Next, a relatively thick masking oxide is selectively grown over the *n*-well region, and the *p*-well implant (boron) is then performed. Because the implant is blocked by the oxide that covers the *n*-wells, it enters the silicon only in the *p*-well regions (a self-aligned implant step). Both well regions are then driven in (with the thick oxide over the *n*-wells being retained), and a second boron implant (which will serve as the *p*-well channel stop) is carried out. This implant is kept shallow because a high pressure LOCOS process is subsequently used to grow the field oxide, minimizing the lateral and vertical impurity-profile spreading. Once again, the implant is self-aligned to the *p*-well regions by the presence of the oxide on the *n*-wells (Fig. 6-31). The masking oxide is then removed, and a nitride masking layer is deposited and patterned to cover the active areas. Finally, the field oxide is formed.

This process makes it possible to achieve adequate isolation and a reduced tendency

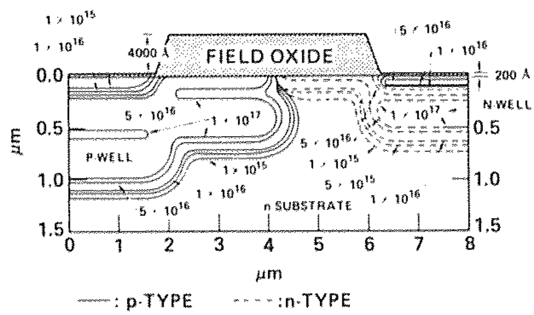


Fig. 6-30 2-D net-impurity contours in the isolation region of a retrograde-well CMOS structure.[53] (© 1986 IEEE).

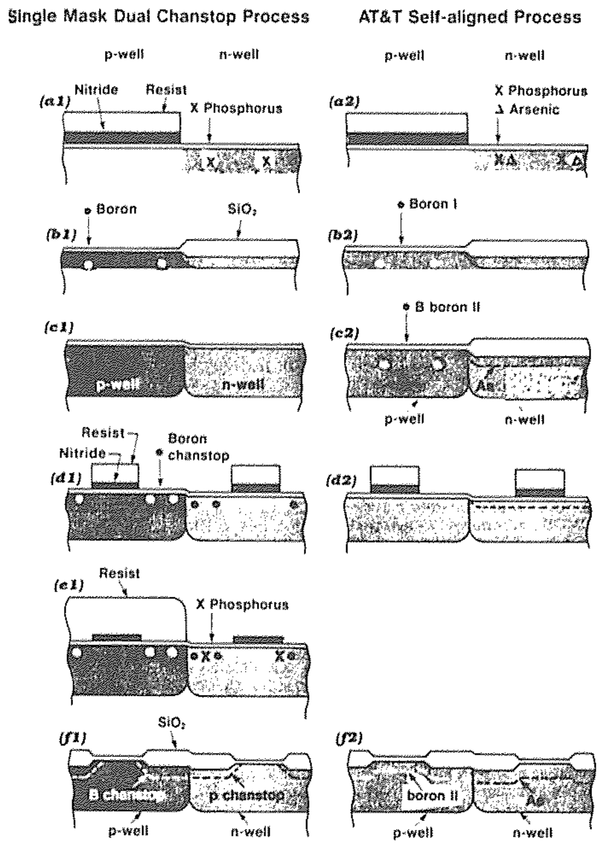


Fig. 6-31 (a1) - (f1) Twin-tub dual-field implant process.[119] (a2) - (f2) Self-aligned twin-tub and field-implant process using high-pressure field oxidations.[62]

toward latchup with smaller device separations than are possible with an atmospheric-pressure oxide-growth process. Nevertheless, a 7-μm spacing must be used between the n^+ and p^+ regions in the Twin-Tub V structure to provide a V_H greater than the power-supply voltage (5 V). Since this spacing still represents a significant layout-density limitation, alternatives to the LOCOS-only process are being vigorously pursued.

In another, more recent study using n-well technology, high-field channel-stop implant doses for the substrate (yielding a peak concentration of 5×10^{16} boron

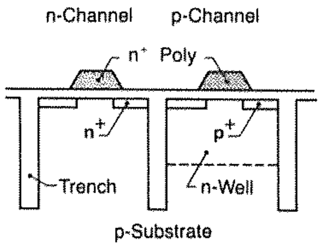


Fig. 6-32 Schematic of an *n*-well CMOS with trench isolation.[4] (© 1986 IEEE).

atoms/cm^3) allowed n^+ to p^+ spacings of 2.4 μm with conventional wells, and 1.8 μm spacings with retrograde wells.[106]

6.5.1 Trench Isolation for CMOS

The two major alternative CMOS-isolation approaches are *trench isolation*[58] and *selective epitaxial growth* (SEG) isolation. Trench-formation technology, discussed in chapter 2, is applied to CMOS to attack the problems of latchup and punchthrough. Its main advantage is that latchup can be completely eliminated if a process employing a thin epitaxial layer on a heavily doped substrate is used, and if the trench is allowed to penetrate to the heavily doped region.[11] Trenches of micron and submicron widths have been used to reduce p^+ to n^+ spacings to 2-2.5 μm.[57] In addition, when trenches deeper than the well depth are used, they replace the reverse-biased *pn* junction as the isolation structure at the well sidewalls (Fig. 6-32).

As of 1988, active devices can not yet be set against the trench sidewall, due to channel-inversion problems associated with the vertical trench sidewall (Fig. 6-33).[56] The sidewall inversion is caused by the horizontal parasitic MOS device, with the well acting as the gate electrode and the trench dielectric acting as the MOS gate oxide (with a thickness equal to the trench width). The voltage across this parasitic device is 5 V

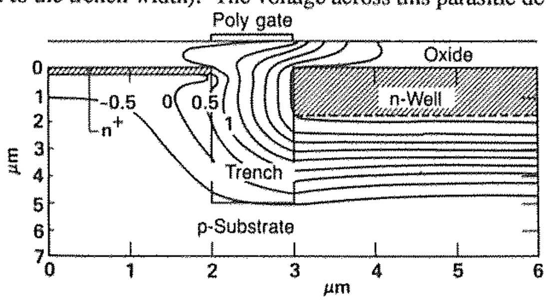


Fig. 6-33 The corresponding two-dimensional potential contours showing sidewall inversion; the simulation is done at $Q_b = 5\times10^{10}$ cm^{-2}, $N_A = 6\times10^{14}$ cm^{-3}, $V_{GD} = 3$ V, $V_{GS} = 0$ V, $V_{SB} = -1$ V.[120] (© 1983 IEEE).

under normal CMOS operating conditions. This gate voltage and the narrow (e.g., 1-μm) trench width can easily cause inversion along the sidewall, outside the well but facing it.

There are two reasons why *p*-wells exhibit more severe trench-sidewall inversion problems. First, some of the needed boron doping segregates into the trench oxide during the thermal oxidation step. The second reason is the presence of the fixed charge, which is normally positive.

Once sidewall inversion occurs, *n*-channel devices with regions butted to the same sidewall become electrically connected by a path along the sidewall. Two obvious solutions are to leave a separation between the n^+ and the sidewall or to increase the trench width, with the end result of either approach being an increase in the isolation spacing. Because of this limitation, the minimum n^+ to p^+ isolation distance in trench isolation is 2-2.5 μm (1988).

Trench isolation has several other disadvantages, as follows:

- Fabrication is much more complicated than in LOCOS, making this an expensive alternative approach.
- For adequate filling and planarization, only one trench width can be utilized (unless SEG refill of the trenches is used).
- Another type of isolation is also needed (usually LOCOS) for most inactive parts of a chip. If a trench deeper than the well depth were used to isolate each device within the well region, each device would be isolated from the substrate. To keep the device bodies from floating, it would be necessary to provide each one with its own well contact, rather than using a single contact for the entire well. Chip area would thus be wasted, defeating the purpose for which the technology was adopted.

6.5.2 Isolation by Selective Epitaxial Growth for CMOS

Various schemes using SEG have been explored as CMOS isolation alternatives (see chap. 2 for more details on SEG technology). One of the first, *selective-etch-and-refill-with epi* (SEREPI), uses an epi refill of recesses in silicon, with the sidewalls passivated prior to refilling. In one approach,[59] the well regions were anisotropically etched to a 5-μm depth, and a 200-nm thermal oxide was grown on the recess surfaces. This oxide was removed from the bottom of the recess with an anisotropic dry etch, leaving it on the sidewalls. Buried layers were formed by arsenic implantation into the bottoms of the wells, and then epitaxial silicon was selectively grown in the wells. By controlling the growth rate, it was possible to refill the wells so that the surfaces were level with the outside substrate.

The SEREPI structure offers the advantages of minimum isolation area and also the flexibility of dopant-profile control in the epitaxial silicon in the wells. A similar process, reported by Kasai et al.,[60] uses a sidewall layer consisting of a 250-nm-thick composite film of oxide and nitride (see Fig. 2-34b, chap. 2). After these layers have been removed anisotropically through dry etching, a sacrificial oxide is grown on the bottom of the trenches and is then wet-etched away to remove any dry etching damage. Finally, an SEG step is performed.

While it is possible to make the n^+ to p^+ isolation distance as small as 0.25 μm, with either approach, some problems exist. First, LOCOS must still be used to isolate devices from one another within the same well. Second, unless a deep boron implant is added to the n-channel side of the isolation, sidewall leakage or inversion can occur. Finally, the 0.25-μm distance between p^+ and n^+ regions can become a limitation for masking of opposite-type implants if n^+ and p^+ junctions are to be placed on opposite sides of the 0.25-μm-thick dielectric.

Another SEG-based approach for isolating CMOS has been described by Manoliu and Borland,[15] and by Stivers.[61] In this approach, windows are anisotropically etched into a 1-2-μm-thick surface-layer oxide (see Fig. 2-43, chap. 2). Separate implants are used to form p^+ layers in those recesses designated to be p-wells, and n^+ layers in those that are to be n-wells. The oxide recesses are then refilled, using an SEG step in which sidewall inversion in the p-well is suppressed by tailoring the doping profile of the boron doping (i.e., this region is doped with $1\text{x}10^{17}/\text{cm}^3$ in the active area of the n-channel devices). As a result, no subthreshold "kink" is observed in the turn-off current (such kinks in the subthreshold current curve are due to leakage from source to drain along the sidewall of the isolation structure).

The two buried layers increase the latch-up immunity of the CMOS devices considerably. Furthermore, this approach eliminates the need for an additional LOCOS isolation. Although this technique was only tested for minimum spacings of 2 μm, it is estimated that it will provide adequate isolation for spacings down to 1 μm. Finally, good planarity and gate oxide integrity were observed.

The technique of *retrograde wells* has also been investigated for implementing reduced isolation spacing and increased latchup immunity in CMOS (as well as for enhancing other device characteristics).[16,19] In this approach, the wells are formed by means of a high-energy implant, which is performed following formation of the field oxide. Because the peak of the implant is well below the silicon surface, the impurity concentration in the well decreases as it approaches the surface (hence the name "retrograde well").

It is easier to implement p-well CMOS with retrograde wells, since boron has a higher implant range in silicon than does phosphorus, and boron implants thus require a lower ion-implantation energy in order to achieve a similar depth. Because the well is formed after the field oxidation, retrograded p-wells can have higher NMOS field-threshold voltages compared to those exhibited in conventionally-formed wells. (In the latter, some of the boron segregates into the oxide during field oxidation, leaving less boron present at the silicon surface. This reduces the field-region threshold voltage.) In addition, the lateral diffusion of boron that occurs when the channel-stop implant is done prior to field oxidation is eliminated.

There are several drawbacks to the retrograde well approach. First, it produces devices with a high junction capacitance (C_j) and a high body-effect coefficient (γ) due to the relatively high doping concentration below the surface in the wells. These effects decrease the speed of circuits manufactured with such devices. Second, this approach may require the use of implanters with accelerating voltages greater than 400 keV. While such machines are available, their low beam current makes them incompatible with high-volume manufacturing needs. Note that doubly ionized boron can be implanted with an accelerating voltage of 200 keV to obtain the same result, and that such a voltage is within the range of ordinary implanters. However, the beam current in this case remains low (see Vol. 1, chap. 9.)

6.6 CMOS PROCESS SEQUENCES

Since there are so many options available for designing a CMOS process flow, no "standard" approach has been adopted. The process flows presented here are merely illustrations of the general sequence of process steps and of the types (and number) of masking layers used. We will first present examples of simple *p*-well and *n*-well single-level-metal CMOS process flows. Such process sequences might be used in the fabrication of CMOS circuits with a minimum feature size of 1.25-2.0 μm. We will then describe a twin-well, double-level-metal process that uses LDD structures. With additional enhancements such as trench or SEG isolation, the filling of contact holes and vias with CVD metal, and novel interlevel dielectric planarization methods, such a process would likely be used in the manufacture of CMOS circuits with submicron dimensions.

6.6.1 Basic *p*- and *n*-Well CMOS Process Sequences

A basic single-well CMOS process can be implemented in either *p*-well or *n*-well technology using eight masking levels. Figure 6-34 shows seven of the eight masking levels of a basic *p*-well process. Figure 6-35 illustrates the *front end* masking levels of an *n*-well process. (The *back end* steps of a process sequence are those that begin with the contact masking step. Thus, the term *front end processing*, refers to all of the steps up to that point.) We will describe an example *n*-well CMOS process in more detail.

The NMOS devices in the *n*-well technology are formed in the lightly doped *p*-substrate ($\leq 1\text{x}10^{15}/\text{cm}^3$), while the PMOS devices are formed in the more heavily doped *n*-well ($\sim 1\text{x}10^{16}/\text{cm}^3$). The starting material is either a lightly doped <100> *p*-type wafer or a heavily doped <100> p^+ wafer with a thin (5-10-μm thick), lightly doped *p*-type epitaxial layer at the surface. A process for enhancing the gettering capabilities of the wafer may be employed before feature formation on the wafer surface is begun (see Vol. 1, chap. 2).

The *n*-well regions are the initial features formed on the starting material. First, a thermal oxide is grown and a CVD nitride film deposited. Then *Mask #1* is used to pattern windows in these layers, through which phosphorus for the *n*-well is implanted. Since the implantation process is unable to place the phosphorus ions deeply enough into the silicon, these impurities must be driven in to the appropriate depth during

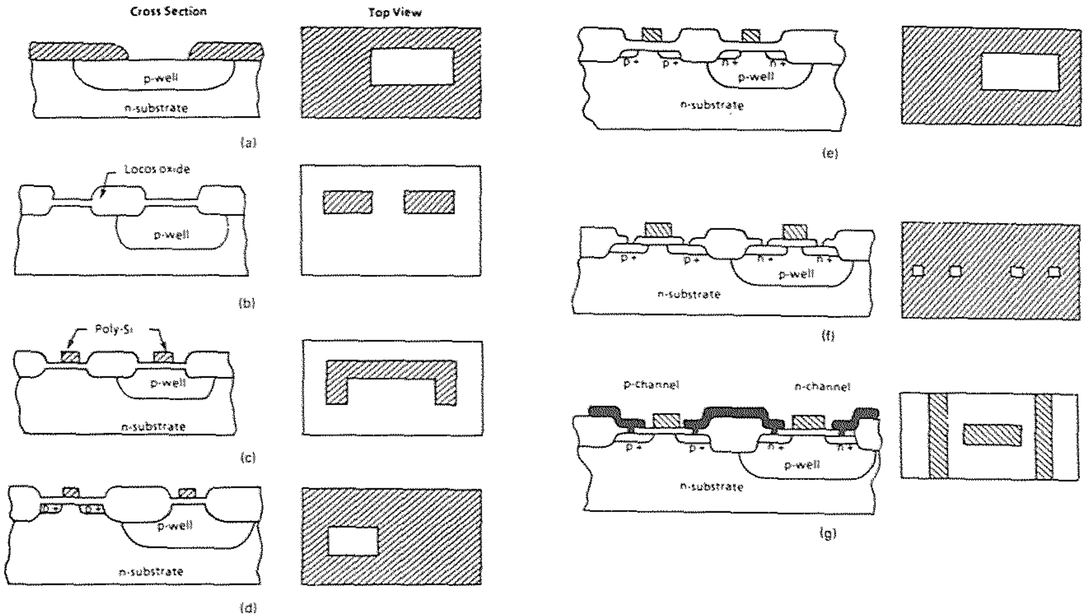


Fig. 6-34 Cross sections and top views of a typical p-well CMOS process at all mask levels: (a) p-well mask; (b) active-area mask; (c) poly-gate mask; (d) p^+ mask; (e) n^+ mask; (f) contact mask; (g) metal mask.[1] From J. Y. Chen, *CMOS Devices and Technology for VLSI*. Copyright Prentice- Hall, 1989. Reprinted with permission.

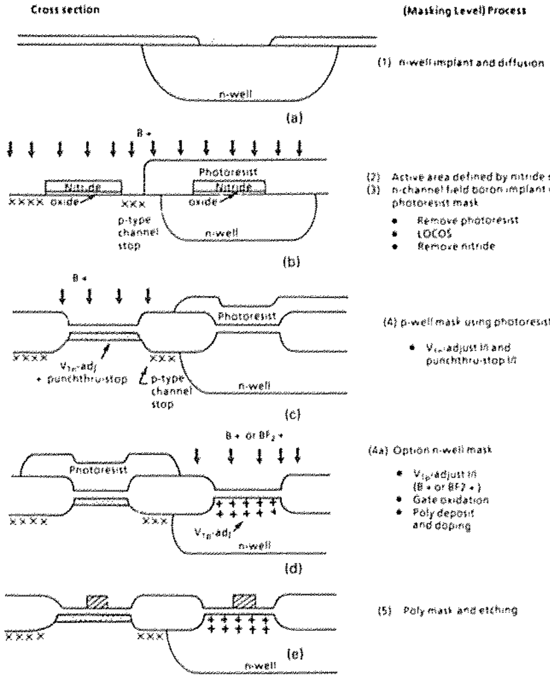


Fig. 6-35 Cross sections and masking levels and associated processes of an n-well CMOS front-end process.[1] From J. Y. Chen, *CMOS Devices and Technology for VLSI*. Copyright Prentice- Hall, 1989. Reprinted with permission.

subsequent high temperature cycles. An oxide is also grown on the n-well regions during the drive-in step. At the conclusion of the drive-in processes the surface concentration in the well is ~$1 \times 10^{16}/cm^3$, and the impurity concentration gradient within the well is also rather small. Note that the redistribution of the well dopant occurs laterally as well as vertically.

Next, a boron threshold-adjust implant is carried out. There is no resist mask for this step, as the thin oxide or oxide/nitride layer covering the silicon wafer surface protects it from contamination. As was described in section 6.3.1, this single implant can provide a correct V_T adjustment for both the NMOS and the PMOS devices.

The surface is then stripped of its oxide and nitride/oxide layers, and a new pad-oxide/nitride layer (needed for LOCOS) is formed. *Mask #2* is then used to pattern this layer to define the active device and field regions. A boron channel-stop implant is performed for the p-substrate field regions. Although no separate mask is used, the boron implanted into the *well field regions* is not of sufficient concentration to significantly alter the n-concentration there. In addition, when the field oxide is grown the phosphorus piles up beneath it in the well regions (see Vol. 1, chap. 7), while some of the boron implanted during the channel-stop implant segregates out into the field oxide. Hence, the surface concentration of phosphorus in the well remains high enough that a separate channel-stop implant is usually not needed for the well region.

The field oxide is then grown, after which the nitride/oxide layer is removed from the active device regions (see chaps. 2 and 5). Next, a gate oxide is grown, in the same manner as that described in chapter 5 for NMOS devices. (Note that a *sacrificial pre-gate oxide* is frequently grown and stripped prior to the growth of the actual gate oxide.)

The deposition of polysilicon by CVD for the gate layer is then carried out. This layer is subsequently doped with phosphorus to form an n^+ polysilicon gate material. The resistivity of the polysilicon should be as small as possible, since this layer also serves as an interconnect structure. (In more advanced processes, even lower resistivities are achieved by forming a silicide layer on top of the polysilicon layer.) *Mask #3* is used to pattern the polysilicon.

Masks #4 and *#5* are then used to selectively implant the source/drain regions of the PMOS and NMOS devices, respectively. The polysilicon protects the channel region under the gate from being implanted. Arsenic is preferable for the n^+ regions so that shallow junctions and minimum lateral diffusion under the gate can be obtained. (It is typically implanted to a dose of 3-6x10^{15} cm^{-2}, and with an energy of 40-60 keV.) Boron is often implanted as BF_2^+ (for shallow junction formation) at doses of 1-5x10^{15} cm^{-2} and energies of 30-50 keV. Typical values of source/drain sheet resistance should be below 30 Ω/sq. Ohmic contacts to the n-well and p-substrate are formed simultaneously with the implants that create the NMOS and PMOS source/drain regions, respectively. These implants are then annealed with a short thermal process at a moderate temperature (e.g., 900-1000°C). The gate oxide that covers the source and drain regions during the implant is usually later stripped and regrown, since it has been damaged by the heavy implants and may have been contaminated by the RIE step used to pattern the poly.

A CVD doped oxide (with the dopants being either phosphorus, or boron + phosphorus) is deposited to a thickness of 50-100 nm, to serve both as a dielectric between the polysilicon and metal layers, and as a gettering layer. This layer is flowed (see Vol. 1, chap. 6) to improve the wafer-surface topography with respect to metal step coverage.

Mask #6 is used to open contact windows in the CVD oxide so that connections can be made between the metal layer and the silicon and polysilicon. Following a *reflow step*, an aluminum (or aluminum-silicon) layer is deposited and patterned (using *Mask #7*). The wafers are subjected to a final anneal step, which is followed by the deposition of the passivation layer. *Mask #8* is then used to open windows in this layer.

An important aspect of this simple CMOS process is that it uses only one more mask than the E-D NMOS process described in chapter 5. The three extra masks in the CMOS sequence (two masks for the source/drain implants, and one for the well mask) replace the depletion-mode implant mask and the buried-contact mask of the E-D NMOS process. Thus, the process complexity of a modern simple CMOS process is not much greater than that of advanced NMOS.

6.6.2 Twin-Well CMOS Process Sequence

This section describes an advanced ten-mask, twin-well CMOS process (Fig. 6-36). Again the process sequence options are many, so there is no one standard process. What is described here is an academic, hypothetical baseline process sequence that *almost certainly does not exist exactly as presented,* and that is useful primarily as a vehicle for illustrating a general sequence of steps in advanced CMOS-circuit fabrication. Where the process steps are not significantly different from those of the single-well process, details will not be repeated; instead readers will be referred to the appropriate sections of the book.

6.6.2.1 Starting Material. The starting material in an advanced twin-well process is typically a <100>-orientation, heavily doped substrate on which a thin (5-10-μm thick), lightly doped epitaxial layer has been grown. Although n-epi-on-n^+ substrates have some advantages for a few special applications (e.g., for building CMOS SRAMs; see chap. 8), p-epi-on-p^+ substrates are the more common choice, because they are less sensitive to process-induced defects (see section 6.2).

6.6.2.2 Formation of Wells and Channel Stops. The twin wells are the first features to be formed, assuming that neither trench isolation nor SEG isolation is being used. (Note that, as described earlier, trench isolation would require an additional masking step.) The well-formation procedure can be carried out in a number of ways. The most obvious method is to use two masking steps, each of which blocks one of the well implants. A single masking-step procedure, however, has been developed,[63] and that is probably the most commonly used approach (see Fig. 6-31).

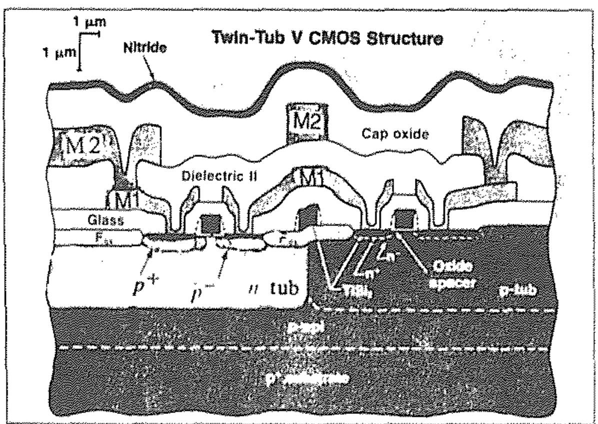


Fig. 6-36 Cross section of AT&T's Twin-Tub V CMOS structure.[121] Reprinted with permission of Semiconductor International.

In this method, a single mask is used to pattern a nitride/oxide film that has been formed on the bare silicon surface (*Mask #1*). The openings in the film become the *n*-well regions, and phosphorus is then implanted into them (e.g., at 80 keV). Next, a thermal oxide is grown on these regions, to a thickness that is sufficient to block the boron implant used to form the *p*-wells. After stripping of the oxide/nitride layer, the silicon in the *p*-well regions is exposed, while the *n*-well regions are covered with the implant-blocking oxide. Thus, during the *p*-well implant (e.g., at 50 keV), the boron penetrates the silicon only in the desired areas.

Next the wells are driven in (e.g., at 1100°C for 500 min). At the conclusion of the drive-in steps, the concentration in the wells for a 0.8-μm-CMOS process could be ~$1\times10^{16}/cm^3$ for the *p*-well, and ~$3\times10^{16}/cm^3$ for the *n*-well. The *n*-well might have a higher doping concentration, to improve the punchthrough performance of the PMOS devices and to eliminate the need for a separate channel-stop step for the *n*-well. A higher concentration in both wells would still produce devices with relatively low capacitances at the bottoms of the source/drain-to-well junctions.

The channel-stop implant procedure is also usually included in the part of the processing sequence in which the wells are formed. In one reported twin-well process,[26] only a *p*-well channel-stop implant (boron) is used, because the doping in the *n*-well is high enough (3×10^{16} cm^{-3}) that a separate channel stop implant is not required. In this case, an unmasked boron implant is done following both well implants, but prior to field-oxide growth.

In a second approach,[55] an additional mask can be used to provide both *n*-well and *p*-well channel stops (Fig. 6-31a). Note that in this procedure, the boron channel stop is implanted into both the *n*-well and *p*-well field regions; the phosphorus channel stop must be increased to compensate for the presence of the boron in the *n*-well field regions. The disadvantages of this method include the additional alignment step between the channel stops and the well masks, the interdiffusion that occurs during the oxidation step, and the asymmetry of the doping profiles. (The latter is due to the increased phosphorus concentration that must overcompensate the nonselective boron channel stop).

A maskless variation of this method has been developed to overcome the above drawbacks (Fig. 6-31b).[55,64] In this sequence, both arsenic and phosphorus are implanted into the *n*-well regions. This places both the dopants that form the well and the dopants that form the channel stop into the *n*-well regions prior to implanting of the *p*-wells. An oxide is then selectively grown on the *n*-well regions, and the boron dopant for the *p*-wells is implanted. After the *n*-wells have been driven in, the oxide over the *n*-wells is retained, and a second boron implant is carried out. This implant serves as both a channel stop in the *p*-well *field* regions and a punchthrough-prevention implant in the *active* regions of the *p*-well.

Various approaches have been proposed to prevent excessive redistribution of the boron channel stop implant during field-oxide growth. These include:

1. Co-implanting of Ge with the boron, which has been found to reduce the boron diffusion constant.[65]

2. Implanting of the field regions with Cl to increase the field-oxide growth rate

(thus reducing the amount of time the wafers must be exposed to the high temperature).[66]

3. The use of high-pressure oxidation to reduce the thermal cycle that must be employed to grow the field oxide.[12,26,55] This also reduces the boron channel-stop dopant loss to the growing field oxide, as well as the interdiffusion of the wells (As a result, smaller isolation spacings between *n*- and *p*-channel devices can be used).

6.6.2.3 Definition of Active and Field Regions. In a conventional advanced CMOS process, a standard LOCOS process (described in detail in chaps. 2 and 5) is used to define the active and field regions (*Mask #2*) and to grow a thick thermal oxide over the field regions. Enhanced LOCOS approaches, which overcome some of the limitations of conventional LOCOS, can be used as alternative processes to form isolation structures in the field regions. Some that have been reported as having been integrated into submicron CMOS processes include the SILO process[67] and a poly-buffered LOCOS process[26] (see chap. 2 for details on both of these). A high-pressure oxidation process for growing the field oxide, was reported to have been used in both of these processes.

6.6.2.4 Gate-Oxide Growth and Threshold-Voltage Adjustment. The gate-oxide growth process is essentially the same as that used in the NMOS process described in chapter 5. The threshold-voltage implant process can be accomplished with either a single implant step, using boron (see section 6.3.1), or two separate implants (in which case additional masking steps are required). A separate selective punchthrough-prevention implant for the NMOS devices is usually needed, especially for submicron devices; this step also requires the use of an extra mask. The V_T-adjust and punchthrough-prevention implants can also be done either before or after the gate-oxidation step. To keep the boron V_T implant shallow (i.e., to improve the PMOS punchthrough characteristics, as described in section 6.3.2), implanting is sometimes done through the gate oxide. To protect the gate oxide from contamination (as well as to keep the implant shallow), part of the polysilicon can be deposited prior to the implant step,[26,67] with the remainder deposited afterward.

Alternative gate-oxide materials are being studied as replacements for SiO_2. Gate oxides of less than 10 nm will not be practical unless the gate voltage is also reduced, for two reasons. First, in order to prevent tunneling, the minimum gate oxide thickness must be $2V_{DD}/E_{ox}(max)$, where $E_{ox}(max)$ is 6 MV/cm. For V_{DD} = 3 V, the minimum thickness is thus 10 nm. The second reason has to do with the reliability and burn-in techniques used in accelerated lifetime testing. It becomes very difficult to screen out weak devices from good ones when the breakdown-voltage criterion defining a weak oxide falls within the range of the distribution exhibited by the good devices. This situation is encountered at oxide thicknesses below 10 nm.[68]

If a material with a larger dielectric constant than that of SiO_2 (3.9) could be used, the gate dielectric could be thicker, while the same capacitance per unit area could be maintained. Extensive work has been conducted on such materials, including nitrided oxides, thermal nitrides (dielectric constant = 8), and tantalum pentoxide, Ta_2O_5 (22). Rapid thermal processing has been explored as a means for forming the nitrided oxides, since the high temperatures needed for their formation can be produced by rapidly heating and cooling the wafers.

The Ta_2O_5 films have been prepared through thermal oxidation of a tantalum layer,[69] by reactive sputtering,[70] and CVD deposition.[71] Several reports indicate that these films exhibit excellent characteristics, and thus show promise for potential applications in advanced integrated circuits.[72,75]

In addition, a novel technique for forming ultrathin (15 nm), low-defect SiO_2 gate layers has been been reported.[76] In this approach, a 5-nm thermal oxide is first grown. A 5-nm CVD oxide is then deposited, and finally another 5-nm oxide is thermally grown, *under* the CVD layer (Fig. 6-37). The layer is able to grow beneath the CVD oxide because the oxidizing species diffuse through the first two layers to the Si-SiO_2 interface (see Vol. 1, chap. 7). During the second thermal oxidation, the top CVD oxide is also densified. This combination layer exhibits a low defect density because the growth-induced micropores (which are the major factor contributing to the defect density in gate dielectrics), are misaligned. If they were to extend completely through the oxide film, these micropores would be potential paths for rapid diffusional mass transport, as well as for current leakage.

6.6.2.5 Polysilicon Deposition and Patterning. The process of forming the n^+ polysilicon gate structure involves the deposition of a polysilicon layer and patterning with *Mask #3*. The steps followed are essentially the same as those in the basic CMOS process. In most advanced CMOS processes the polysilicon film is overlaid with a refractory metal silicide to form a *polycide* structure (see Vol. 1, chap. 11). In some cases, a *salicide* (self-aligned silicided gate and source/drain regions – see Vol. 1, chap. 11) process is employed to reduce the parasitic resistance of the source and drain regions, as well as that of the gate material. In most such cases, the salicide

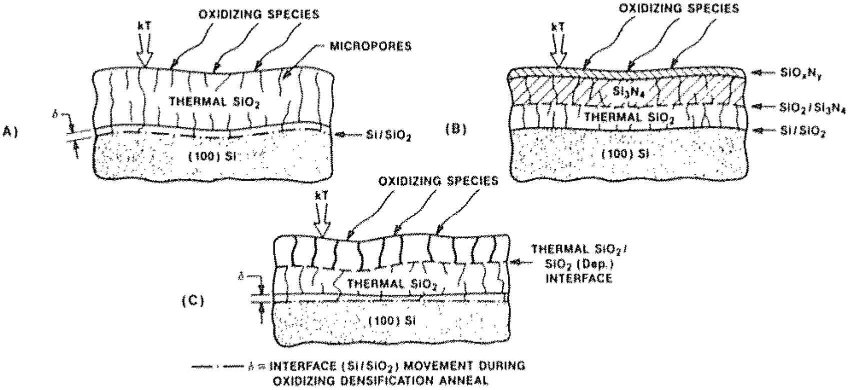


Fig. 6-37 Effects of oxidizing densification anneal: (a) SiO_2 films; (b) SiO_2- Si_3N_4 dual dielectric; (c) thermal SiO_2 - deposited SiO_2 multilayered stacked films.[76] (© 1988 IEEE).

formation is carried out after the source and drain have been implanted.

The polysilicon can be doped either after deposition by diffusion or ion implantation, or *in situ* during deposition. The former approaches were used in most MOS processing prior to submicron device generations. For submicron devices, in which a reduced thermal budget becomes more imperative, *in situ* doping becomes more attractive, (since the long, high-temperature drive-in step can be eliminated).[103] The redistribution of channel and source/drain implants is thus minimized, allowing improved control of threshold voltage, punchthrough, and L_{eff}.

As was described in sections 6.3.3 and 6.3.4, alternative materials to n^+ polysilicon (or to polycides with n^+ polysilicon as the underlayer) have been investigated as gate materials (e.g., p^+ *and* n^+ polysilicon, or refractory metal gates). Each alternative process introduces some changes into the process flow; nevertheless it is predicted that as device dimensions shrink, such measures will need to be used to maintain adequate device performance.

Etching of the gate structure can be a difficult step, especially when gate lengths are in the submicron range. First, the dimensions must be accurately and uniformly produced across the wafer, as I_D is strongly dependent on the gate dimensions. Second, the sidewalls must be vertical in order to reduce asymmetric I_D characteristics in devices built with LDD structures (see section 5.6.5.2). Third, the formation of vertical sidewall gate structures implies the need for a completely anisotropic etch step (see Vol. 1, chaps. 15 and 16). In turn, such a step requires that the etch process be highly selective against etching the thin underlying gate oxide layer.

6.6.2.6 Formation of Source/Drain Regions. In advanced-CMOS processes, the gate lengths are short enough that LDD structures must be used to minimize hot-electron effects, especially in NMOS devices. Therefore, the procedures outlined in section 5.6.5 are used to fabricate such LDD structures. If these structures are used only for the NMOS transistors, *Masks #4* and *#5* are used to allow the sources and drains of the two transistor types to be selectively implanted.* Note that if LDD structures are also used for the PMOS devices, two additional masking layers may be needed (6-38a).

A removable-spacer LDD process for both NMOS and PMOS devices has been reported (Fig. 6-38b) that does not require the use of any other masks than the two needed to selectively form the sources and drains of the two transistor types.[67,73] In the removable spacer process, the heavily doped source/drain implant is performed first, with the spacers in place. After the spacers have been removed, the implant that forms the lightly doped drain regions is carried out (Fig. 6-38b). This process can be used to provide LDD structures for one or both transistor types, as desired. Polysilicon is used as the material of the removable spacers in one approach,[67,99] while a low-temperature

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

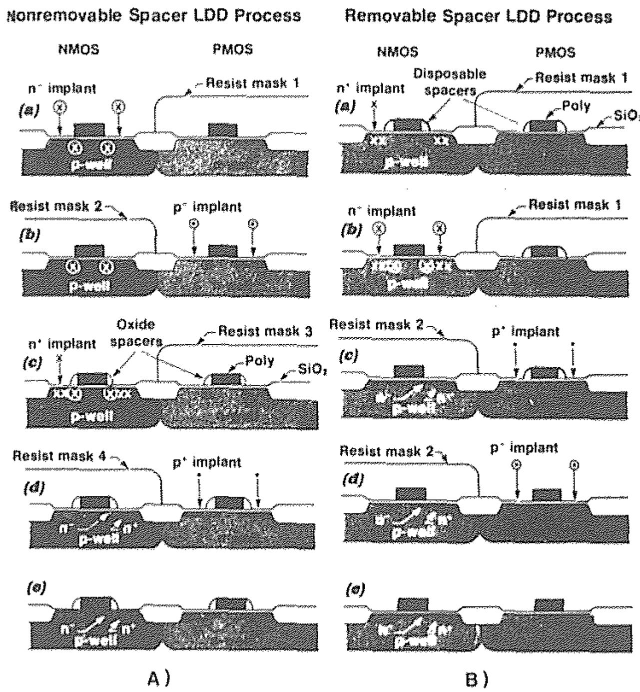


Fig. 6-38 (A) Non-removable oxide spacer LDD process: (a) n^- implant, (b) p implant, (c) after LDD spacer etch and n^+ implant, (d) after n^+ and p^+ implant, (e) after p^+ anneal. **(B)** Removable spacer LDD process: (a) after spacer etch and n^+ implant, (b) after NMOS spacer removal and n^- implant, (c) p^+ implant, (d) after PMOS spacer removal and p^- implant, (e) after p^+ anneal.[67] (© 1986 IEEE).

CVD oxide (over a thin polysilicon film) is used in the other.[73]

Various techniques have been developed for forming the shallow source/drain junctions needed for submicron CMOS devices. First, As is implanted for the n-channel devices and BF_2^+ for the p-channel devices, since both species have very shallow projected ranges at implant energies of 30-50 keV (see Vol. 1, chap. 9). These implants are usually performed through a screen oxide to protect the source/drain regions from contamination during the implant procedure. Second, preamorphization of the silicon by implantation with Si or Ge reduces channeling and helps produce shallow junctions. It is necessary, however, to diffuse the implanted species past the layer of implant damage that cannot be annealed out, in order to prevent junction leakage (see Vol. 1, chap. 9). RTP techniques have been explored as a means of carrying out these anneal and diffusion thermal cycles. Shallow p^+n junction formation through the use of diffusion (i.e., by utilizing RTP and either a solid diffusion source[82] or a spin-on diffusion source[83]) has also been reported. The use of an antimony implant to obtain shallow n^+p junctions with n^+ layers of lower resistivity than is possible with arsenic has also been described.[81]

Note that the screen oxide is usually stripped off following the source/drain implant, since it is damaged (and may be contaminated) by the heavy implants. A new oxide is grown over the source/drain regions and on the sidewall of the etched polysilicon electrode (and this step is referred to as *poly reoxidation*). During this thermal cycle, some (but not all) of the implantation damage is annealed out. The oxide formed on the poly sidewall shifts the remaining damage away from the gate edge to maintain the integrity of the thin gate oxide (Fig. 6-39).[104,110] At the same time, the sidewall oxidation produces a *gate bird's beak*, or GBB, under the polysilicon edge. This reduces the the gate-to-drain overlap capacitance and relieves the electric-field intensity at the corner of the gate structure. However, because the GBB encroachment can degrade transconductance of submicron MOSFETs and impact their subthreshold swing and threshold voltage,[109] this poly reoxidation process must be optimized for fabrication of submicron MOSFETs.[111]

An approach to forming shallow junctions that uses $CoSi_2$ source/drain junctions has been reported.[29] In this approach, the $CoSi_2$ is formed before the source and drain junctions are created by means of a heavy ion implantation step. The implant is then performed so that the damaged layer, which would have occurred in the silicon crystal, is kept within the $CoSi_2$ layer. As a result, it is necessary to drive the implanted impurities just far enough so that they enter the silicon region to form the required *pn* junction. If a salicide process is used, the device has a cross-section, as shown in Fig. 6-40.

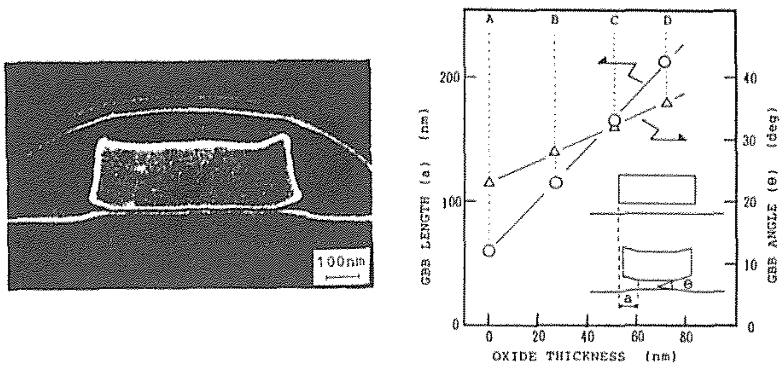
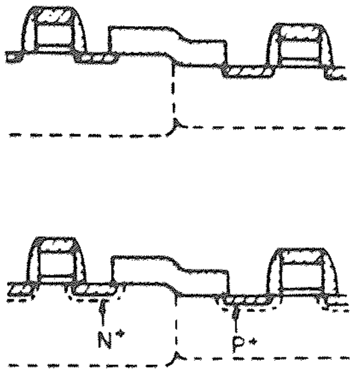


Fig. 6-39 (a) SEM micrograph of a gate structure with a GBB. (b) Profiles of GBB versus oxide thickness grown during re-oxidation.[104] This paper was originally presented at the Spring 1989 Meeting of The Electrochemical Society, Inc. held in Los Angeles, CA.



- Form $CoSi_2$ S/D regions
- Selectively implant As and BF_2
- Drive dopant from $CoSi_2$ to form n^+ and p^+ regions

Fig. 6-40 Cross section of CMOS structure in which $CoSi_2$ is first formed on the source and drain regions, and then As and BF_2^+ is selectively implanted into the $CoSi_2$. These dopants are then driven out of the $CoSi_2$ and into the Si substrate to form shallow p^+ and n^+ source/drain regions.[29] (© 1986 IEEE).

Another approach is to create so-called *elevated source-drains*. A thin (e.g., 200-nm) epitaxial layer of silicon can be selectively deposited onto the exposed source-drain areas of the MOS transistor, following the implantation of the lightly doped region of the LDD structure and formation of the spacers (Fig. 6-41a). This method has been used for the NMOS devices of a 1 Mbit SRAM.[79] In this case, a heavy BF_2^+ implant is done such that the SEG film becomes heavily doped; the boron penetrates to the substrate during an RTP anneal that follows the implant. In this way, elevated, heavily doped, shallow source/drain regions are formed. The source-drain junction depths in the substrate are less than 0.2 μm deep (Fig. 6-41b). As noted earlier, the gate oxide that covers the source and drain regions is usually etched away and regrown following the implant step.

6.6.2.7 CVD Oxide Deposition and Contact Formation. Following the formation of the source and drain regions (and the salicide layers), a doped oxide is deposited by CVD, with procedures very much like those of the basic NMOS and CMOS processes (see section 5.4.1.6).

Contact windows are opened in this CVD layer to allow electrical connections to be made between the Metal 1 layer and the source/drain, gate, and substrate and well contact regions (*Mask #6*). Again, the details are similar to those of the basic NMOS and CMOS processes.

Advances in contact technology for CMOS include the use of barrier layers to prevent spiking through the shallow junctions, and the filling of contact holes by such materials as CVD W, polysilicon, or even SEG. In addition, as device dimensions decrease, the parasitic resistances of the source and drain regions become more significant. More details on all of these topics are provided in chapter 3.

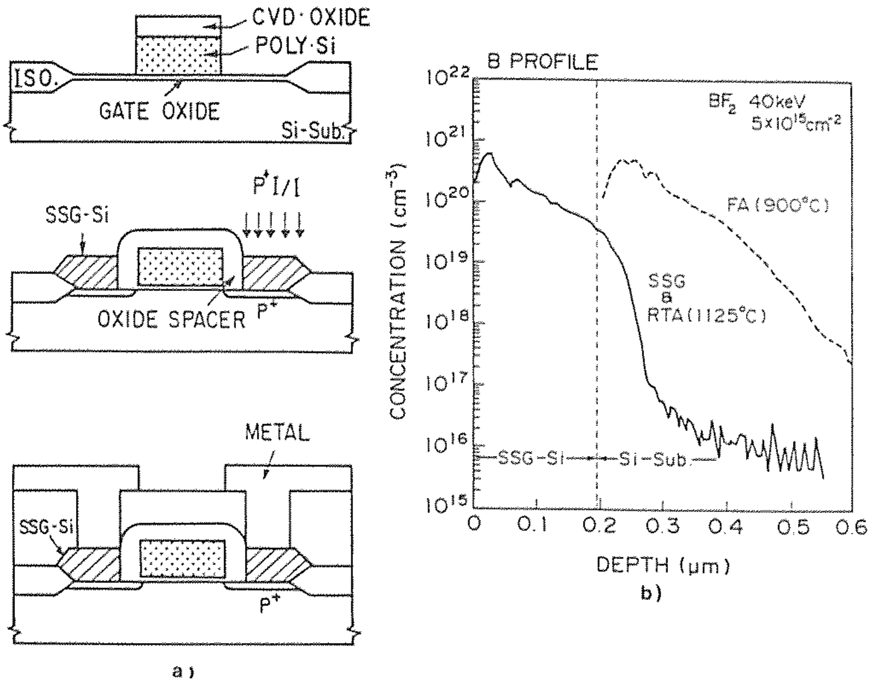


Fig. 6-41 Elevated source/drains. (a) Key process steps in forming elevated source/drains using SEG. (b) Comparison of the doping profile in source/drain regions of a PMOS device after a BF_2^+ implant directly into the regions (and a conventional furnace anneal) versus a BF_2^+ implant into the SEG regions and an RTP anneal.[79]

6.6.2.8 Metal 1 Deposition and Patterning. The issues involved in the deposition and patterning of the Metal 1 layer are not significantly different from those described in chapter 5 for the NMOS process. *Mask #7* is used to pattern the Metal 1 layer. (For more information on the metal-layer deposition and patterning processes, see chapter 4 of this volume, and Vol. 1, chap. 9.)

Despite its higher resistivity compared to Al, tungsten's superior electromigration properties make it more appropriate for certain applications. As a result, CVD W has been selected as the Metal 1 material for a variety of circuit designs.

6.6.2.9 Intermetal Dielectric Deposition/Planarization and Via Patterning. Following the Metal 1 patterning, an intermetal dielectric must be deposited to electrically isolate Metal 1 from the Metal 2 layer. A variety of materials and deposition processes, have been utilized (see chap. 4).

However, deposition of this layer may make the wafer topography too severe to allow the Metal 2 film to be deposited with adequate step coverage. A number of *planarization* processes have been developed to overcome this problem (see chap. 4). If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical.

It is necessary to open vias in the intermetal dielectric layer so that an electrical connection can be established between Metal 2 and Metal 1 at desired locations (*Mask #8*). Metal 2 must be deposited with adequate step coverage into the vias.

A number of techniques have been studied to improve coverage of Metal 2 in the vias, including the following (see chap. 4):

- Sloping the via walls by means of the via-etch process.
- Filling the vias with a blanket W or polysilicon deposition, and then etching back to provide a planar surface.
- Filling the vias with a selective deposition of W or Al.
- Increasing the step coverage into the vias through bias sputtering and heating of the substrate during deposition.
- Laser melting the Metal 2 film to increase step coverage.

6.6.2.10 Metal 2 Deposition and Patterning. The processing issues of depositing and patterning (*Mask #9*) the Metal 2 layer are discussed in chapter 4.

6.6.2.11 Passivation Layer Deposition and Patterning. The passivation-layer deposition and patterning (*Mask #10*) issues are the same those of the basic CMOS or NMOS processes.

6.7 MISCELLANEOUS CMOS TOPICS

6.7.1 Electrostatic-Discharge Protection

The input signals to an MOS IC are fed to the gates of MOS transistors. If the voltage applied to the gate insulator becomes excessive, the gate oxide can break down. The dielectric breakdown strength of SiO_2 is approximately $8x10^6$ V/cm; thus, a 15-nm gate oxide will not not tolerate voltages greater than 12 V without breaking down. Although this is well in excess of the normal operating voltages of 5-V integrated circuits, voltages higher than this may be impressed upon the inputs to the circuits during either human-operator or mechanical handling operations.

The main source of such voltages is triboelectricity (electricity caused when two materials are rubbed together). A person can develop very high static voltage (i.e., a few hundred to a few thousand volts) simply by walking across a room or by removing an integrated circuit from its plastic package, even when careful handling procedures are followed. If such a high voltage is accidentally applied to the pins of an IC package, its discharge (referred to as *electrostatic discharge*, or ESD) can cause breakdown of the gate oxide of the devices to which it is applied. The breakdown event may cause sufficient damage to produce immediate destruction of the device, or it may weaken the oxide enough that it will fail early in the operating life of the device (and thereby cause device failure). A more detailed description of ESD failures in semiconductor devices can be found in reference 84.

All pins of MOS ICs must be provided with protective circuits to prevent such voltages from damaging the MOS gates. The need for such circuits is also mandated by the increasing use of VLSI devices in such high-noise environments as personal computers, automobiles, and manufacturing control systems. These protective circuits, normally placed between the input and output pads on a chip and the transistor gates to which the pads are connected, are designed to begin conducting or to undergo breakdown, thereby providing an electrical path to ground (or to the power-supply rail). Since the breakdown mechanism is designed to be nondestructive, the circuits provide a normally open path that closes only when a high voltage appears at the input or output terminals, harmlessly discharging the node to which it is connected.

Four types of circuits are used to provide protection against ESD damage, as follows:

1. diode breakdown
2. node-to-node punchthrough
3. gate-field-induced breakdown
4. parasitic *pnpn* diode latchup.

Often, a combination of protection methods is used, with a breakdown diode and one of the other protection devices connected in parallel with the gate being protected.

6.7.1.1 Diode Protection. Protection is obtained by using the diode-breakdown phenomenon to provide an electrical path in the silicon substrate that consists of a diffused resistor region (of a doping type opposite to that of the substrate). This diffused region is connected between the input pad and the gate (Fig. 6-42a). If a reverse-bias voltage greater than the breakdown voltage of the resultant *pn* junction is applied, the diffusion region (which otherwise works as a resistor), operates as a diode and undergoes breakdown. Furthermore, this diffused region will also clamp a negative-going transition at the chip input to one diode drop below the substrate voltage. In CMOS technologies, an additional protection diode can be added by utilizing the *pn* junction that exists between a p^+ node and the body region of the PMOS device (an n-type region, that is connected to V_{DD}). This diode is utilized as a protection device when a connection is made between the pad and a p^+ region. (Note that this second diode will clamp positive-going transitions to one diode drop *above* V_{DD}.)

6.7.1.2 Node-to-Node Punchthrough. As defined in section 5.5.2, *punchthrough* is the phenomenon by which the depletion region surrounding the drain of an MOS device extends along the channel of the device and contacts the depletion region of the source of the device. While the current that results from punchthrough is ordinar-

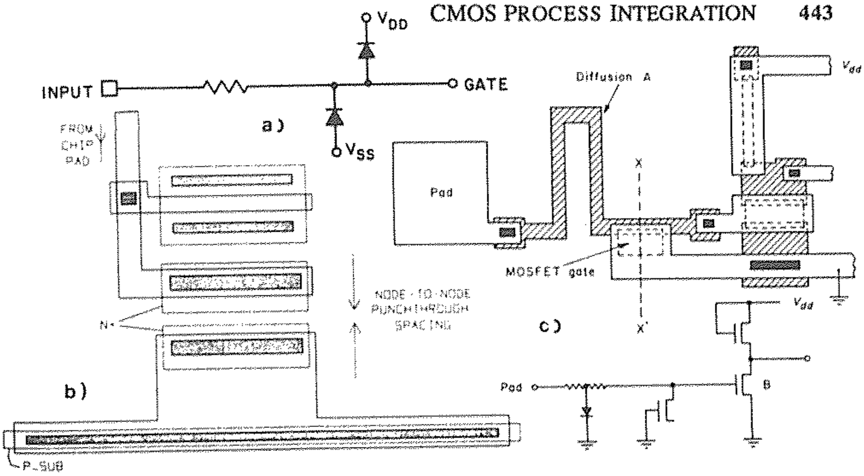


Fig. 6-42 MOS protective circuits: (a) Diode; (b) Punchthrough device; (c) Layout of an input protection structure containing both devices.[122] From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

arily an unwanted effect, it is harnessed in ESD protection circuits for beneficial purposes. The source and drain regions of an ESD punchthrough protection structure are placed close enough together that punchthrough occurs at a voltage lower than that of the gate-oxide breakdown (Fig. 6-42b). Such a structure is layed out and tested to meet a specific ESD transient reliability measure. Once it has been verified as functioning according to the design specifications, it can be connected to each chip signal I/O pad.

6.7.1.3 Gate-Controlled Breakdown Structure. The electric field near the corner of an MOS device's drain node is strongest at the surface, since the depletion region is narrowest at this point[85] and the entire voltage across the depletion region must be dropped over a very short distance. The fact that the gate voltage can increase the strength of the electric field at the corner of the drain is the principle used in designing the gate-controlled breakdown structure (Fig. 6-43).

As the reverse-bias voltage applied to the drain is increased, the gate-to-drain voltage also increases (note that the gate is tied to ground). Because the presence of the gate reduces the breakdown voltage of the junction near the corner of the drain, a relatively small voltage can cause the junction to break down at this point. At breakdown, the junction conducts a large current; in this device, the current flows from the drain to the substrate. The actual input voltage at which breakdown occurs can be controlled by altering the oxide thickness. This produces a large voltage drop across the relatively high resistance of the diffusion area, reducing the voltage applied to the gate of the MOSFET.

Such structures were the workhorses of ESD protection devices. It has been found, however, that processing enhancements that increase the device performance of small-dimension devices sometimes have a negative impact on the failure resistance of the

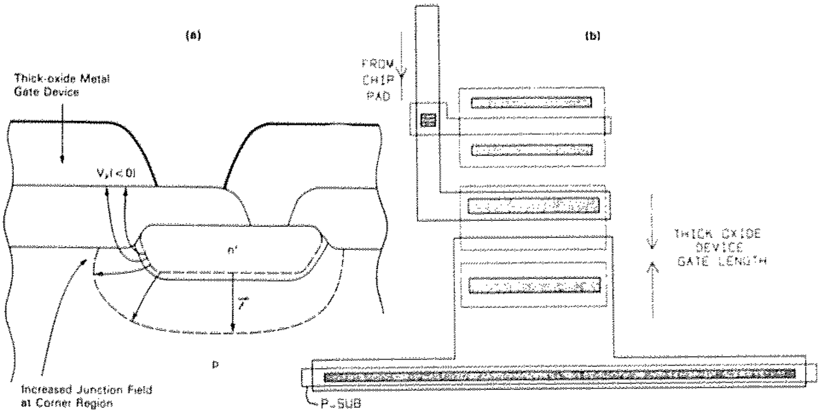


Fig. 6-43 (a) Cross-section, and (b) layout of a thick-oxide metal-gate device for node-to-substrate breakdown and input-pad protection.[122] From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

ESD protective devices. Specifically, silicided source/drain regions and LDD structures can degrade the ESD performance of the gate-controlled breakdown structures so that the voltage at which *these devices themselves fail* is reduced from 4 kV to 1 kV.[86,87] As a result, ESD protection structures can be rendered largely ineffective by the silicided regions, or else the circuits can be caused to fail by ESD in new failure modes as a result of the addition of performance-enhancement features.[86,88]

6.7.1.4 *pnpn*-Diode ESD Protection for Advanced CMOS Circuits.

Even as devices are made smaller, a fixed minimum volume of silicon must always be used to dissipate the power associated with the current flowing through the protective devices to ground. If the power is dissipated in too small a volume of silicon, the silicon can be heated beyond its melting point; this can destroy the device, even if the circuit-protection device is operating properly. As a result, a minimum unscalable area (Fig. 6-44) will be required in order for the same degree of protection to be maintained through each generation of scaled technology.[68,89]

Effective protection of VLSI must be provided for both automated and human handling. ESD from a machine is typically modeled as a high-voltage (4-kV), short-duration pulse (i.e., the *EIAJ machine model [MM] test method*), while ESD from a human operator is usually modeled as a longer, lower-voltage (300-V) but higher-current pulse (e.g., the *Mil. Std. human body model [HBM] test method*). Thus, if the gate-controlled breakdown structure itself fails at 1 kV, it ceases to provide protection against HBM ESD.[87]

To solve this problem on the inputs of a CMOS chip, it has been proposed that the parasitic lateral *pnpn* structure inherent in all CMOS circuits be exploited (Fig. 6-45a)

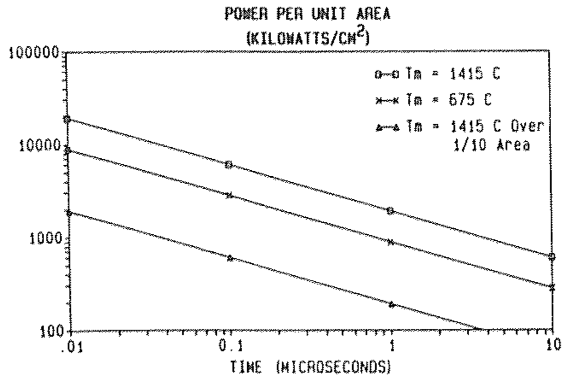


Fig. 6-44 Wunsch and Bell power curves for silicon melt, aluminum melt, and silicon melt for nonuniform current density.[89] (© 1968 IEEE).

as an ESD structure (see section 6.4). Such a structure – which is deliberately configured to latch up at a voltage lower than that required to damage the input MOS gate oxide – is built into each input circuit on the chip (D1 in Fig. 6-45b). Thus, this parasitic *pnpn* structure serves as the primary ESD structure of the input circuit by

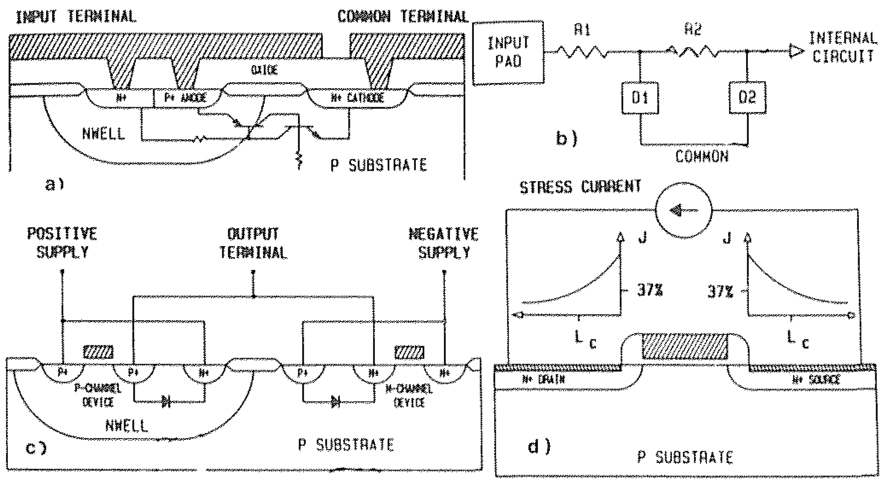


Fig. 6-45 (a) Lateral *pnpn* input protection structure showing parasitic and increased area at *n*-well and *p*-well junction resulting in lower power density. (b) Input-protection circuit showing primary protection elements R1 and D1 and secondary protection elements R2 and D2. (c) Representation of dual-diode output circuit cross section. Effectiveness depends on parasitic diode impedance as well as capacitance between supplies. (d) Cross section of MOS thin-oxide device showing current density at silicide and silicon interface.[87] (© 1979 IEEE).

providing a low-impedance path when it is turned on. It also offers increased area for power dissipation. Secondary ESD structures, however, must also be used on each input circuit (e.g., D2 and R2 in Fig. 6-45b) to provide effective protection until D1 achieves its low-impedance conducting state.

On the outputs of the chip, the parasitic diodes that exist in the CMOS output circuit are used to provide ESD protection, much as described in section 6.7.1.1 (Fig. 6-45c).[87] Such dual-diode CMOS outputs have proved to be effective when the resistance of the parasitic diode in the *p*-channel device is minimized and the failure threshold of the *n*-channel device is maximized. As shown in Fig. 6-45d, the stress current density in the diode of the silicided *n*-channel device is very nonuniform due to the silicided source/drain regions. As a result, the diode can fail if too high a current density passes through only a small fraction of the area of the silicided contact.

6.7.2 Power-Supply Voltage Levels for Future CMOS

As MOS has evolved, device scaling has been practiced to increase the speed and packing density of MOS ICs. Although scaling of the electric field to maintain a constant electric field (CE) would have promoted both high performance and high reliability, in practice a constant voltage (CV) has been maintained in order to maintain TTL compatibility and produce higher circuit performance. Such nonscaling of the voltage leads to reliability problems, caused by hot-electron effects and oxide breakdown. Scaling of the power supply voltage is thus desirable. M. Kakamu has calculated the optimum power-supply voltage as a function of shrinking line width, and has shown that this voltage can be scaled down in proportion to the square root of the design-rule shrinkage without sacrificing circuit performance.[90] The relationships given between the supply voltage and the design rule are

$$V_{DD} = 6.1 \times (L/2)^{1/2} \text{ (for conventional drain structures)} \qquad (6-9)$$

and

$$V_{DD} = 8.2 \times (L/2)^{1/2} \text{ (for LDD structures)} \qquad (6-10)$$

where L (μm) is the design rule and V_{DD} is the power-supply voltage (volts). These equations predict that at dimensions of 0.6 μm, the power supply voltage will have to be reduced to 3.3 V in order for reliable device operation and circuit performance to be maintained if conventional drain structures are used. On the other hand, if LDD structures are used, it is predicted that V_{DD} = 5 V will still be feasible for 0.7-μm dimensions. A recent report indicates that these equations may be slightly pessimistic, in that MOSFETs with gate lengths of 0.6 μm that operate from a 5-V supply have been successfully fabricated.[102] Through optimization of the LDD structures of these devices, a lifetime of more than 10 years in the face of hot-carrier degradation effects, is predicted.

6.7.3 Low-Temperature CMOS

Since it is possible to build MOS devices with minimum dimensions of less than 0.5 μm, low-temperature operation of CMOS (e.g., at 77°K) has been actively investigated. At low temperatures, the MOS devices exhibit lower subthreshold leakage, higher carrier mobility (which yields improved speed performance), and a steeper logarithmic current-voltage slope. It has been found that the normalized propagation delay in CMOS logic gates is reduced by a factor of nearly 2 or 3 when the devices are operated at 77°K.[91,92,93,94] The major reason for the slow commercial introduction of such systems has been the difficulty associated with liquid-nitrogen refrigeration.

6.7.4 Three-Dimensional CMOS

Since the limits of two-dimensional CMOS may soon be approached, a logical direction for continued advances in device scaling is three-dimensional devices. CMOS is well-suited for 3-D integration, since low-power circuits will be mandatory in order for heat dissipation problems to be avoided. The 3-D approach will offer the benefits of increased packing density and higher speeds (due to shorter interconnects).

The inability to fabricate single-crystal silicon over insulating layers, however, has historically been an obstacle to the implementation of 3-D ICs. Recent advances in SOI technologies show promise of being able to surmount this obstacle and much work is currently being done to develop 3-D CMOS.[95]

REFERENCES

1. J. Y. Chen, *CMOS Devices and Technology for VLSI*, Englewood Cliffs N.J., Prentice-Hall, 1989.
2. F. M. Wanlass and C. T. Sah, *IEEE Int. Solid-State Circ. Conf.*, February 1963.
3. R. D. Davies, "The Case for CMOS", *IEEE Spectrum*, October 1983, p. 26.
4. J. Y. Chen, "CMOS-The Emerging Technology," *IEEE Circuits and Systems Magazine*, March 1986, p. 16.
5. D. M. Brown, M. Ghezzo, and J. M. Pimbley, "Trends in Advanced CMOS Process Technology," *Proceedings of the IEEE*, December 1986, p. 1678.
6. W. C. Holton and R. K. Calvin, "A Perspective on CMOS Technology Trends," *Proceedings of the IEEE*, December 1986, p. 1646.
7. S. Prussin, private communication.
8. R. Chwang and K. Yu, "CHMOS-An *n*-Well Bulk CMOS Technology for VLSI," *VLSI Design*, Fourth Quarter, 1981, p. 42.
9. W. Wijaranakula et al., *J. Electrocchem. Soc.*, December, 1988, p. 3113.
10. H. Tsuya et al., *Japanese J. Appl. Phys.*, **22**, L16, (1983).
11. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, **ED-31**, 205 (1984).
12. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. Internat.*, April 1988, p. 64.

13. A. G. Lewis et al., "Vertical Isolation in Shallow *n*-Well CMOS Circuits," *IEEE Electron Dev. Letts.*, March 1987, p. 197.
14. L. C. Parrillo et al., "Twin-Tub CMOS," *IEDM Tech. Dig.*, 1980, p. 752.
15. J. Manoliu and J. O. Borland, "A Submicron Buried-Layer Twin-Well CMOS SEG Process," *IEDM Tech. Dig.*, 1987, p. 20.
16. R. D. Rung, C. J. Dell' Oca, and L. G. Walker, "A Retrograde *p*-Well for High-Density CMOS," *IEEE Trans. Electron Dev.*, **ED-28**, p. 1115, 1981.
17. R. A. Martin and J. Y. Chen, "Optimized Retrograde *n*-Well for 1 μm CMOS," *Proc. Custom Integ. Circ. Conf.*, 1985, p. 199.
18. S. R. Combs "Scalable Retrograde *p*-Well CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-28**, p. 346, 1981.
19. Y. Taur et al., *J. Solid-State Circuits*, **SC-20**, p. 123, 1985.
20. S. M. Sze, Ed., *VLSI Technology*, 2nd. Ed., New York, McGraw-Hill, 1988, p. 491.
21. T. Ohzone et al., *IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
22. K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, Boston, Klewer Academic Publishers, 1986 p. 182.
23. G. J. Hu et al., *IEDM Tech. Dig.*, 1982, p. 710.
24. J. Zhu et al., *IEEE Trans. Electron Dev.*, February 1988, p. 145.
25. A. Schmitz and J. Y. Chen, "Design, Modelling, and Fabrication of Submicron CMOS Transistors," *IEEE Trans. Electron Dev.*, **ED-33**, p. 148, (1986).
26. R. A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.
27. L. C. Parrillo et al., *Tech Dig. IEDM*, 1984, p. 418.
28. K. M. Cham et al., *IEEE Electron Dev. Lett.*, January 1986, p. 49.
29. S. J. Hillenius et al., "A Symmetric Submicron CMOS Technology," *Tech. Dig. IEDM*, 1986, p. 252.
30. Y. Pauleau, "Interconnect Materials for VLSI Circuits, Part-1," *Solid State Technology*, February 1987, p. 61.
31. V. Schwabe, F. Neppl, and E. P. Jacobs, *IEEE Trans. Electron Dev.*, **ED-31**, 1984, p. 988.
32. S. Iwata et al., *IEEE Trans Electron Dev.*, **ED-31**, 1984, p. 1174.
33. R. F. Kwasnick et al., *IEEE Trans. Electron. Dev.*, September 1988, p. 1432.
34. H. Oikawa and T. Amazawa, *Proc. 3rd. Intl. VLSI Symp., ECS Meeting*, May, 1985, p. 131.
35. D. M. Brown et al., *IEEE Trans. Electron Dev.*, **ED-18**, 1971, p. 931.
36. R. R. Troutman, *Latchup in CMOS Technology*, Klewer, Boston, MA., 1986.
37. G. J. Hu, "A Better Understanding of CMOS Latchup," *IEEE Trans. Electron Dev.*, **ED-31**, p. 62, 1984.
38. R. R. Troutman, "Latchup in CMOS Technologies," *IEEE Circuits & Sys. Mag.*, May 1987, p. 15.
39. A. G. Lewis et al., *Tech. Dig. IEDM*, 1986, p. 248.
40. R. D. Rung and H. Momose, *IEEE Trans. Electron Dev.*, **ED-30**, 1983, p. 1647.
41. W. R. Dawes and G.F. Derbenwick, "Prevention of CMOS Latchup by Gold-Doping," *IEEE Trans. Nucl. Sci.*, **NS-23**, 1976, p. 2027.

42. J. R. Adams and R. J. Sokel, Presented at the1979 Nuclear and Space Radiation Effects Conf., Santa Cruz, CA., July, 19, 1979.
43. J. O Borland and T. Deacon, "Advanced CMOS Epitaxial Processing for Latchup Hardening," *Solid-State Technology*, p. 123, August 1984.
44. S. Ratanphanyarat et al., *Tech. Dig. IEDM*, 1987, p. 744.
45. F. S. Lai et al., *IEDM Tech. Dig. 1985*, p. 513.
46. M. L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
47. S. Swirhun et al., "Latchup-Free CMOS Using Guarded Schottky Barrier PMOS," *Tech. Dig. IEDM*, 1984, p. 402.
48. R. A. Martin et al., *IEDM Tech. Dig.*, 1985, p. 403.
49. A. G. Lewis, *IEEE Trans. Electron Dev.*, Oct. 1984, p. 1472.
50. D. Takacs et al., *Tech. Dig. IEDM*, 1983, p. 159.
51. K.W. Terrill et al., *Tech. Dig. IEDM*, 1984, p. 406.
52. R. Menozzi et al., "Layout Dependence of CMOS Latchup," *IEEE Trans. Electron Dev.*, Nov. 1988, p. 1892.
53. J. Y. Chen and D. E. Snyder, "Modeling Device Isolation in High-Density CMOS," *IEEE Electron Dev. Letts.*, February 1986, p. 64.
54. L. Herman, "Controlling CMOS Latchup," *VLSI Design*, April, 1985, p. 100.
55. M. -L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
56. K. M. Cham and S.-Y. Chiang, "A Study of Trench Surface Inversion Problem for the Trench CMOS Technology," *IEEE Electron Dev. Lett.*, Sept., 1983, p. 303.
57. Y. Nitsu et al., "Latchup Free CMOS Structure Using Shallow Trench Isolation," *Tech. Dig. IEDM*, 1985, p. 509.
58. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *IEDM Tech. Dig.*, 1982, p. 6.
59. T. I. Kamins and S. Shiang, *Electron Dev. Letts.*, December 1985, p. 617.
60. N. Kasai et al., *IEEE Trans. Electron Dev.*, June 1985, p. 1331.
61. A. Stivers et al., *10th Internat. Conf. on CVD, ECS*, **87-8**, p. 389.
62. R. D. Rung et al., *IEEE Trans. Electron Dev.*, **ED-28**, 1981, p. 1115.
63. L. C. Parrillo et al., "Twin-Tub CMOS II," *IEDM Tech. Dig.*, 1982, p. 706.
64. S. J. Hillenius and L.C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.
65. J. Pfiester and J. Alvis, "Improved CMOS Field Isolation Using Ge/B Implantation," *IEEE Elect. Dev. Lett.*, August 1988, p. 391.
66. J. Pfiester, *Electron Dev. Letts.*, November 1988, p. 561.
67. L. C. Parrillo et al., *Tech. Dig. IEDM*, 1986, p. 244.
68. M. H. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings of IEEE*, Dec. 1986, p. 1715.
69. G. S. Ohrlein, *J. Appl. Physics*, **59**, 1587 (1986).
70. B. W. Shen et al., *Tech. Dig. IEDM*, 1987, p. 582.
71. M. Saitoh, T. Mori, and H. Tamura, *Tech. Dig. IEDM*, 1986, p. 680.
72. K. Ogiue et al., "Technology Improvement for High Speed ECL RAMs," *Tech. Dig. IEDM*, 1986, p. 468.
73. J. Pfiester, *IEEE Electron Dev. Letters*, April, 1988, p. 189.
74. R. de Werdt et al., *Tech. Dig. IEDM*, 1987, p. 532.

75. S. G. Byeon and Y. Tzeng, *Tech. Dig. IEDM*, 1988. p. 722.
76. P. K. Roy et al., *Tech. Dig. IEDM*, 1988, p. 714.
77. N. Kasai, N. Endo, and A. Ishitani, *Tech. Dig. IEDM*, 1988, p. 242.
78. L. Manchanda, *Intl. Reliability Symp.*, 1986, p. 183.
79. K. Hashimoto, Presented at the Third Annual Applied Materials Inc.,"*Innovations in Epitaxial Technology for Advanced Device Structures*" Seminar, December 15, 1988, Palo Alto, CA.
80. G. Sai-Halasz and H. B. Harrison, *Electron Dev. Letts.*, September 1986, p. 534.
81. M. L. Chen et al., "Constraints in P-Channel Device Engineering for Submicron CMOS Technologies," *Tech. Dig. IEDM*, 1988, p. 390.
82. K. -T. Kim and C. -K. Kim, *IEEE Electron Dev. Letts.*, December 1987, p. 569.
83. E. Ling et al., *IEEE Electron Dev. Letts.*, March, 1987, p. 96.
84. B. A. Unger, *Intl. Reliability Physics Symp.*, 1981, p. 193.
85. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, Section 10.5.
86. K.- L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 484.
87. R. N. Rountree, *Tech. Dig. IEDM*, 1988, p. 580.
88. C. Duvvury et al., "ESD Protection in 1 μm CMOS," *Intl. Reliabilty Physics Symp.* 1986.
89. D. C. Wunsch and R. R. Bell, "Determination of Threshold Failure Levels of Semiconductor Diodes and Transistors due to Pulse Voltages," *IEEE Trans. Nucl. Sci.*, **NS-15**, p. 244, Dec. 1968.
90. M. Kakumu et al., *Tech. Dig. IEDM*, 1986, p. 399.
91. J. D. Plummer, "Low Temperature CMOS Devices and Technology," *Tech. Dig. IEDM*, 1986, p. 378.
92. J. Watt and J. D. Plummer, *Tech. Dig. IEDM*, 1987, p. 393.
93. J. Y-C. Sun et al., *Tech. Dig. IEDM*, 1986, p. 236.
94. J. S. T. Huang and J. W. Schrankler, *IEEE Electron Dev.*, January 1987, p. 101.
95. Y. Akasaka, "Three-Dimensional IC Trends," *Proc. IEEE*, December 1986, p. 1703.
96. M. H. El-Diwany et al., *Tech. Dig. IEDM*, 1987, p. 917.
97. R. H. Krambeck, C. M. Lee, H. F. S. Law, *IEEE J. Solid-State Circuits*, **SC-17**, (1982), p. 614.
98. M. Wong and K. Saraswat, *IEEE Electron Dev. Letts.*, November 1988, p. 579.
99. C. -Y. Yang et al., *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1988, Abs. No. 138, p. 207.
100. C. Y. Wong et al., *Tech. Dig. IEDM*, 1988, p. 238.
101. G. J. Hu and R. H. Bruce, *IEEE Electron Dev. Letts.*, **EDL-5**, p. 211, 1984.
102. H.-M. Mulhoff, K. H. Kusters, and H. Melzner, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 133, p. 186.
103. S. M. Kugelmass and J. P. Krusius, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 135, p. 188.
104. M. Kishimoto et al., *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 137, p. 191.
105. P.-H. Pan, J. G. Ryan, and M. A. Lavoie, *Ext. Abs. of the Electrochemical Soc. Meeting*, Spring, 1989, Abs. No. 138, p. 193.
106. A. G. Lewis et al., *IEEE Trans. Electron Dev.*, June 1987, p. 1337.
107. Y. Sato, K. Ehara, and K. Saito, *J. Electrochem. Soc.*, June, 1989, p. 1777.

108. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, MA, 1980.
109. J. R. Pfeister et al., *IEEE Electron Dev. Letts.*, August 1989, p. 367.
110. C. Y. Wong et al., "Process Induced Degradation of Thin Oxides," in *Proc. 1stInt. Symp.Ultra Large Scale Integration Sci. Technol.*, S. Broydo and C. M Osburn, Eds., Electrochem. Soc., 1987, p. 155
111. C. Y. Wong et al., *IEEE Electron Dev. Letts.*, September, 1989, p. 420.
112. J. Y. Chen, *IEEE Trans. Electons Devices*, ED-31, p. 910 (1984).
113. K. Tanno, F. Shimura, and T. Kawamura, *J. Electrochem. Soc.*, 128, 395, (1981).
114. W. C. Till and J. T. Luxton, *Integrated Circuits: materials, devices, and fabrication*, Prentice-Hall, Englewood Cliffs, N. J., 1982.
115. B. Hoefflinger and G. Zimmer, in "J. Carrol, ed., *Solid-State Devices 1980, Institute of Physics Conf. Ser. 57*, from the 10th European Solid-State Device Research Conf., Sept. 1980.
116. A. G. Lewis et al., *IEEE Trans. Electron Dev.*, **ED-34**, 2156, (1987).
117. J. Y. Chen, *VLSI Design*, July 1984, p. 78.
118. H. Momose et al., *Tech. Dig. IEDM*, 1984, p. 706.
119. L. C. Parrillo, G. W. Reutlinger, and L. K. Wang, U.S. Patent No. 4,435,895, March 31, 1984.
120. K. M. Cham et al., *Tech. Dig. IEDM*, 1986, p. 296.
121. M. L. Chen, *Semiconductor International*, April 1988, p. 78.
122. T. E. Dillinger, *VLSI Engineering*. Englewood Cliffs, N.J., Prentice-Hall, 1988.
123. F. K. Baker et al., *Tech. Dig. IEDM*, 1989, p. 443.
124. J. M. Sung et al., *Tech. Dig. IEDM*, 1989, p. 447.

PROBLEMS

6.1 An *n*-well CMOS process starts with a substrate doping of 3×10^{15} cm^{-3}. The well doping g near the surface is approximately constant at a level of 3×10^{16} cm^{-3}. The gate oxide thicknesses are both 40 nm. (a) Calculate the threshold voltages of the *n*- and *p*-channel transistors. (b) Calculate the boron doses needed to shift the NMOS threshold to +1 V and the PMOS threshold to -1 V. Assume that the threshold shifts are achieved through shallow ion implantations. Neglect oxide charge.

6.2 In a *p*-well CMOS technology, the *n*-substrate has a doping concentration of 1×10^{15} cm^{-3}, and the *p*-well concentration is 5×10^{16} cm^{-3}. Calculate the source/drain junction capacitances and body-effect coefficients for both *n*- and *p*-channel MOSFETs assuming the gate oxide is 20 nm thick in both devices.

6.3 In an *n*-well CMOSS process, specify the necessary masking sequence in correct order and also state the number of masks needed a process that uses both a double-poly and double-level metal process.

6.4 A CMOS inverter is driven with a square wave voltage source with a period of 1 μs. The power supply voltage, VDD = +12 V and the load capacitance CL is 20 pF. Find the power dissipated by the inverter.

6.5 If an LDD structure is needed for both *n*- and *p*-channel MOSFETs in a submicron CMOS process, list the CMOS masking sequence and associated major processing steps. If the doping concentration of the p^- LDD is an order of magnitude larger than that of the n^- LDD, can any of these steps be omitted?

6.6 Comparing a retrograde well CMOS process to a conventional well CMOS process, what is the most important difference in the two process sequences and in the completed device structures? How do these differences impact the device isolation?

6.7 In a (p^+ diffusion)- (*n*-well)-(*p*-on-p^+ substrate) structure, the *n*-well is uniformly doped at 1×10^{16} cm^{-3} and is 1 μm deep, the *p*-epi thickness is 4 μm and is uniformly doped at 1×10^{15} cm^{-3}, and the p^+ diffusion is 0.4 μm deep. If a substrate bias generator is needed and the p^+ diffusion and *n*-well are both biased at 5 V, what is the maximum (in absolute value) substrate voltage that can be applied before punchthrough occurs?

6.8 Explain why using a lightly doped epitaxial layer on a heavily doped substrate is one of the most effective ways to increase latchup resistance. Why is a thinner epi even more helpful and what limits the continuous scaling of epi thickness?

6.9 For an *n*-well CMOS structure, draw a lumped equivalent circuit model of the parasitic device structures that give rise to latchup and derive the conditions for latchup to occur.

6.10 Explain why latchup cannot be sustained if the holding voltage is greater than the supply voltage. What is the most effective technique for increasing the holding voltage?

6.11 If the n^+-to-*n*-well spacing is 3 μm, the *n*-well depth is 1.3 μm, and the p^+ diffusion depth is 0.3 μm, estimate the minimum pulse width needed to trigger latchup.

6.12 In an *n*-well CMOS structure the well depth is 1 μm and the well concentration 1×10^{16} cm^{-3}. The p^+ diffusion is 0.3 μm deep and it is doped at 1×10^{19} cm^{-3}. Estimate the vertical *pnp* transistor common-emitter current gain (β) using one-dimensional analysis. *(Hint:* You will need to calculate the minority-carrier diffusion length in the base of the transistor, and compare this with the base width in order to estimate α, which can then be used to find β.)

CHAPTER 8

SEMICONDUCTOR MEMORY PROCESS INTEGRATION

Memories store digital information (or data) in terms of *bits*, or **b**inary dig**its** (ones or zeros). Modern digital systems use *memory devices* to store and retrieve large quantities of digital data at electronic speeds. Early digital computers used magnetic-cores as the devices in fast-access memories. With the introduction of semiconductor memory chips in late 1960s, however, magnetic cores began to be replaced by integrated circuits (which implemented a much higher-density digital-memory function). This not only increased the performance capabilities of the memory, but also drastically decreased its cost. By the end of the 1970s, magnetic-core memories had been completely displaced as high-speed memory devices.

8.1 TERMINOLOGY OF SEMICONDUCTOR MEMORIES

Memory capacities in digital systems are usually expressed in terms of bits, since a separate storage device or circuit is used to store each bit of data. Each storage element is referred to as a *cell*. Memory capacities are also sometimes stated in terms of *bytes* (8 or 9 bits) or *words* (32 – 80 bits). Each byte typically represents an alphanumeric character. Every bit, byte or word is stored in a particular location, identified by a unique numeric address, and only a single bit, byte, or word is stored or retrieved during each cycle of memory operation.

Memory-storage capability is expressed in units of kilobits and megabits (or kilobytes and megabytes). Since memory addressing is based on binary codes, capacities that are integral powers of 2 are typically used. As a result, a memory device with a 1-kbit capacity can actually store 1,024 bits, and a 64-kbit device can store 65,536 bits.

In digital computers, the number of memory bits is usually 100 to 1000 times greater than the number of logic gates, which implies that the memory cost per bit must be kept very low. In addition, it is desirable for the memory devices to be as small as possible (since this will allow the highest density of cells on a chip), to operate at a high speed, to have a small power consumption, and to operate reliably.

Memory cells could be designed to possess a set of characteristics close to those of an ideal digital-logic-element. Such an ideal cell would be able to

1. perform the desired logic function;
2. robustly quantize the signal levels of the stored data;
3. exhibit a high degree of input-output isolation and fan-out; and
4. regenerate the stored logic-levels.

However, to enable each memory cell to possess all of these attributes would require the use of a complex circuit to implement each cell. Memory-cell design therefore involves trading off most of the desired properties of digital-logic devices in order to achieve a cell that is as simple and compact as possible. Consequently, the cell itself is not capable of outputting digital data in an electrical form compatible with the requirements of the remainder of the system. To restore the electrical characteristics of the cell's outputted data to adequate values, properly designed peripheral circuits (e.g., sense amplifiers, memory registers, and output drivers) are necessary. These circuits are designed to be shared by many memory cells. The trade-off thus made is that of a less-robust output signal from the cell, in exchange for a simple, compact memory cell design (consisting of only 1 to 6 transistors).

8.1.1 Random-Access and Read-Only Memories (RAMs and ROMs)

The most flexible digital memories are those that allow for data storage (or *writing*) as well as data retrieval (or *reading*). Memories in which both of these functions can be rapidly and easily performed, and whose cells can be accessed in random order (independent of their physical locations), are referred to as *random-access memories* (RAMs). *Read-only memories* (ROMs) are those in which only the read operation can be performed rapidly (although ROMs are generally configured so that their cells are also randomly accessible, and data can be entered into them). Entering data into a ROM, however, is referred to as *programming* the ROM, to emphasize that this operation is much slower than the writing operation used in RAMs.

8.1.2 Semiconductor-Memory Architecture

The organization of large semiconductor memories is shown in simplified form in Fig. 8-1.[1] The storage cells of the memory are arranged in an array consisting of horizontal rows and vertical columns. Each cell shares electrical connections with all the other cells in its row, and column. The horizontal lines connected to all the cells in the row are called *word lines*, and the vertical lines (along which data flows into and out of the cells) are called *bit* lines. Each cell therefore has a unique memory location, or address, which can be accessed at random through selection of the appropriate word and bit line. (Some memories are designed so that four or eight cells are accessed simultaneously.) Thus, in semiconductor memories such as that shown in Fig. 8-1, any cell can be accessed in random order, at a fixed rate, for the purpose of either reading

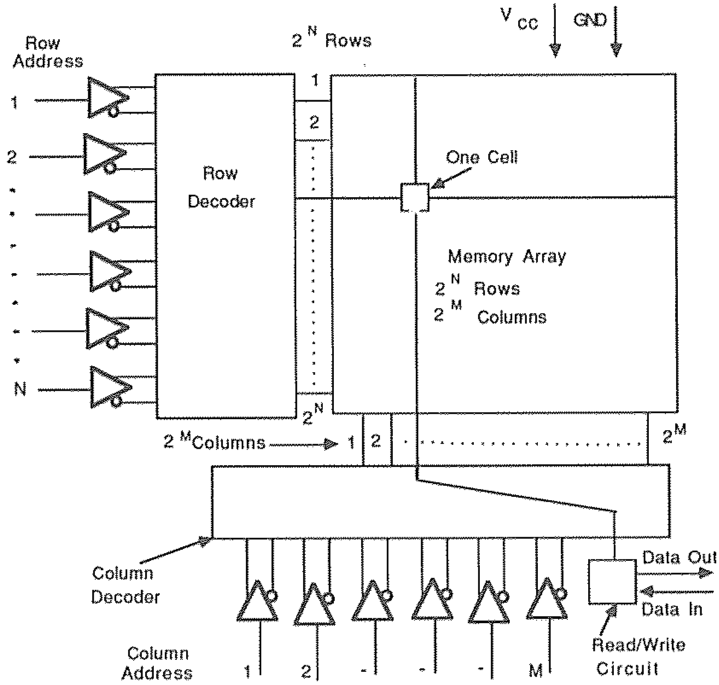


Fig. 8-1 Organization of a random access memory (RAM).

or writing data. The array configuration of semiconductor memories lends itself well to the regular, structured designs favored in VLSI.

There are also a number of important circuits at the periphery of the array. The first such peripheral circuit is the *address decoder*. Two of these are used on each chip – one for the word lines, the other for the bit lines. These circuits allow a large number of word and bit lines to be accessed with the fewest number of address lines. Address decoders for this purpose have 2^n output lines, with a different one selected for each different n-bit input code. In later generations of memory circuits, address multiplexing was integrated on some memory chips to reduce the number of address pins by half. (Note that the address decoder was the first peripheral logic-circuit to be built on the memory chip).

The read/write control circuitry shown in Fig. 8-1 determines whether data is to be written into or read from the memory. Because such circuits also amplify and buffer the data signals retrieved from the cells, one of the important circuits in this subsystem is the *sense amplifier*. In dynamic memories that need periodic data refreshing, refresh circuitry may also be provided.

Note that most RAMs have only one *input-data lead* and one *output-data lead* (or only one combined *input/output lead*). Writing into and reading from such RAMs is done one bit at a time. Other RAMs have a number of input- and output-data leads, with the number determined by the word length of the system's data bus. ROMs, on the other hand, are typically organized so that the number of output-data leads (usually eight) is the same as the number of lines on the data bus. ROMs are programmed word by word (i.e., eight bits, or one byte, at a time) and are read from in the same manner.

8.1.3 Semiconductor-Memory Types

In semiconductor RAMs, information is stored on each cell either through the charging of a capacitor or the setting of the state of a bistable flip-flop circuit. With either method, the information on the cell is destroyed if the power is interrupted. Such memories are therefore referred to as *volatile memories*. When charge on a capacitor is used to store data in a semiconductor-RAM cell, the charge needs to be periodically refreshed, since leakage currents will remove it in a few milliseconds. Hence, volatile memories based on this storage mechanism are known as *dynamic RAMs*, or *DRAMs*.

If the data is stored (i.e., written into the memory) by setting the state of a flip-flop, it will be retained as long as power is connected to the cell (and no other write signals are received). RAMs fabricated with such cells are known as *static RAMs*, or *SRAMs*. Volatile RAMs can be treated as nonvolatile if they are provided with battery backup. Some DRAM and SRAM chips are even packaged together with a battery to facilitate implementation of this approach.

It is often desirable to use memory devices that will retain information even when the power is temporarily interrupted (or when the device is left without applied power for indefinite periods). Magnetic media offer such nonvolatile-memory storage. In addition, a variety of semiconductor memories have been developed with this characteristic. At present, virtually all such nonvolatile memories are ROMs. While data *can* be entered into these memories, the programming procedure varies from one type of ROM to the other (and none of them can be considered to be RAMs).

The first group of nonvolatile memories consists of those ROMs in which data is entered during manufacturing, and cannot subsequently be altered by the user. These devices are known as *masked ROMs* (or simply *ROMs*). The next category consists of memories whose data can be entered by the user (*user-programmable* ROMs). In the first example of this type, known as a *programmable ROM*, or *PROM*, data can be entered into the device only *once*.

In the remaining ROM types, data can be erased as well as entered. In one class of erasable ROMs, the cells must be exposed to a strong ultraviolet light in order for stored data to be erased. These ROMs are called *erasable-programmable ROMs*, or *EPROMs*. In the final type, data can be *electrically erased* as well as entered into the device; these are referred to as *EEPROMs*. The time needed to enter data into both EPROMs and EEPROMs is much longer than the time required for the *write* operation in a RAM. As a result, none of the ROM types can at present be classified as fully functional RAM devices.

A few nonvolatile RAMs have been developed, but these have not yet reached the state of development at which they can be considered an important class of semiconductor memory.

8.1.4 Read-Access and Cycle Times in Memories

The two principal time-dependent performance characteristics of a memory are the *read-access time* and the *cycle time*. The first is the propagation delay from the time when the address is presented to the memory chip until data stored at that address is available at the memory output. The cycle time is the minimum time that must be allowed after the initiation of a read operation (or a write operation, in a RAM) before another read operation can be initiated. The minimum cycle times for reading and writing in a RAM are not necessarily equal, but for simplicity of design most systems employ a single minimum cycle time. For semiconductor RAMs, the read-access time is typically 50 – 90% of the read-cycle time.

8.1.5 Recently Introduced On-Chip Peripheral Circuits

Additional peripheral circuits have recently been added to the basic memory-organization structure shown in Fig. 8-1. These circuits serve mainly to improve the manufacturability and testability of the chips. Those designed to increase manufacturability include redundancy circuits and error-correction circuits. Redundancy circuits allow some defective chips to be salvaged, while self-testing circuits reduce testing time.

Redundancy allows a defective row or column of cells to be replaced with a spare. Replacement techniques include the use of electrically or laser-blown fuses, or of one-time-programmable memory cells (which control on-chip multiplexers that switch in spare rows or columns). Redundancy measures typically improve manufacturing yields by factors of between 1.5 and 5.

Error-detection and correction techniques involve the addition of parity bits to allow the system to detect bad data, as well as circuitry to accomplish parity checking and error correction. This imposes an area penalty: for example, if one parity bit is added to each byte, the size of a 32k x 8k chip would be increased to 32k x 9k. If a failure is detected during programming, the parity bits can be sacrificed to act as a spare bit field for byte-wide applications. In DRAM designs, the use of error-correction coding (ECC) requires another 27% of memory cells. Because the ECC approach corrects soft errors as well as hard errors, the problem of soft errors can be reduced for the life of the product.

8.1.6 Logic-Memory Circuits

Special-purpose circuits that combine logic and memory on the same chip have recently been introduced. These fall into two categories: (1) memories that have some additional logic capabilities (*logic-in-memory circuits*, or *special-application RAMs)*; and (2) logic circuits that contain some memory capability (*memory-in-logic circuits*).

A number of different special-application RAMs have been developed, including *video RAMs* (VRAMs) and *multiport SRAMs*.[45] Video RAMs are DRAMs designed to support the high-capacity requirements of frame buffers and display memories found in graphics terminals and systems. They have two input/output ports – one for random access (as in conventional RAMs), and one for serial-access. The serial port accesses the memory sequentially and performs the various serialization tasks necessary to drive cathode-ray tubes or other serial-data devices. The fast serial-readout-rate also enables quick refreshing of the graphics screen. The random-access port is driven by a graphics processor and is used to build the screens of displayed data (i.e., the random-access capability allows real-time updating of pictures).

Multiport SRAMs are being offered for use in multiprocessor systems. For example, dual-port SRAMs are becoming widely used to allow two independent logic circuits to simultaneously access one memory in a read and write mode. The two circuits can thus communicate with each other by passing data through the common memory. (Two processor-containing components of a digital system might be a CPU and a disk controller, or two processors working on two related but different tasks.) The use of the dual-port memory would eliminate the need for any special data-communication hardware.

Many applications for memory-in-logic circuits have also been envisioned, but these have only begun to be implemented. Some manufacturers of *application-specific circuits* (ASICs) do offer large memory blocks as part of prefabricated gate arrays, as well as increasingly larger memory blocks in their standard-cell libraries. Other, more advanced circuits are still being developed. For instance, more extensive use of memory will be needed in *image processing* and *coder-decoder* (CODEC) ULSI circuits. Image processing systems will be greatly enhanced when a processor is available with a memory of at least 2 Mbits that can store a frame of a picture. Similarly, CODECs used to move pictures could use such an image processor with an on-chip memory for data compaction. Finally, microprocessor chips are being built with on-board memories. (For example, the Intel 486 μP contains *cache memory* on the chip, which consumes ~40% of the chip area.) Even more memory integration might allow for an on-chip memory hierarchy. Products that could use such circuits include *point-of-sale terminals*, *smart cards*, and *telecommunications circuits*.

8.2 STATIC RANDOM-ACCESS MEMORIES (SRAMS)

SRAMs, the first type of semiconductor memory to be implemented, are referred to as *static memories* because they do not require periodic refresh signals in order to retain their stored data. The bit state in an SRAM is stored in a pair of cross-coupled inverters, which form a circuit known as a flip-flop. The voltage on each of the two outputs of a flip-flop circuit is stable at only one of two possible voltage levels, because the operation of the circuit forces one output to a high potential, and the other to a low potential. The memory logic state of the cell is determined by whichever of the two inverter outputs is high. Flip-flops maintain a given state for as long as the circuit receives power, but they can be made to undergo a change in state (i.e., to flip), through the application of a trigger voltage of sufficient magnitude and duration to the appropriate input. Once the circuit has settled into its new stable state, the trigger voltage can be removed. SRAM cells can be implemented in NMOS (Fig. 8-2a), CMOS (Fig. 8-2b), bipolar (Fig. 8-2c), or BiCMOS technologies.

The chief disadvantage of an SRAM cell is that it consists of at least six devices, as compared to only two for the dynamic-memory cell. Thus, even when the same set of design rules is used, an SRAM chip cannot be built with as many cells as a DRAM chip (as is illustrated in Figs. 8-3a and 8-3b).

On the other hand, SRAMs are the fastest semiconductor memories. Their speed is derived from the self-restoring nature of the flip-flop and the static peripheral circuits of the memory chip. Bipolar SRAMs are the fastest of all, and MOS SRAMs are the fastest among MOS memories. Bipolar SRAMs, however, dissipate much more power than CMOS SRAMs (e.g., 0.1-to-1.0 mW/bit versus ~25 μW/bit).

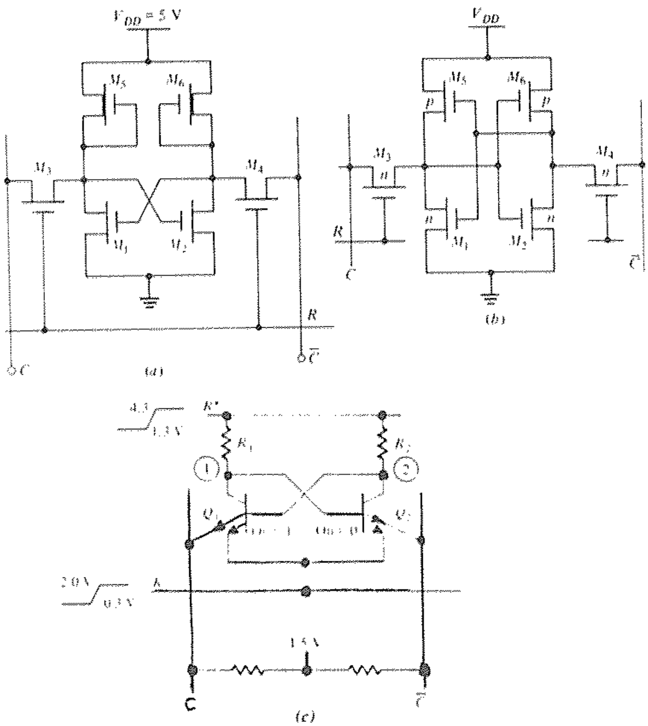


Fig. 8-2 Circuit schematic of (a) NMOS SRAM cell, (b) CMOS SRAM cell, (c) Bipolar SRAM cell.[1] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

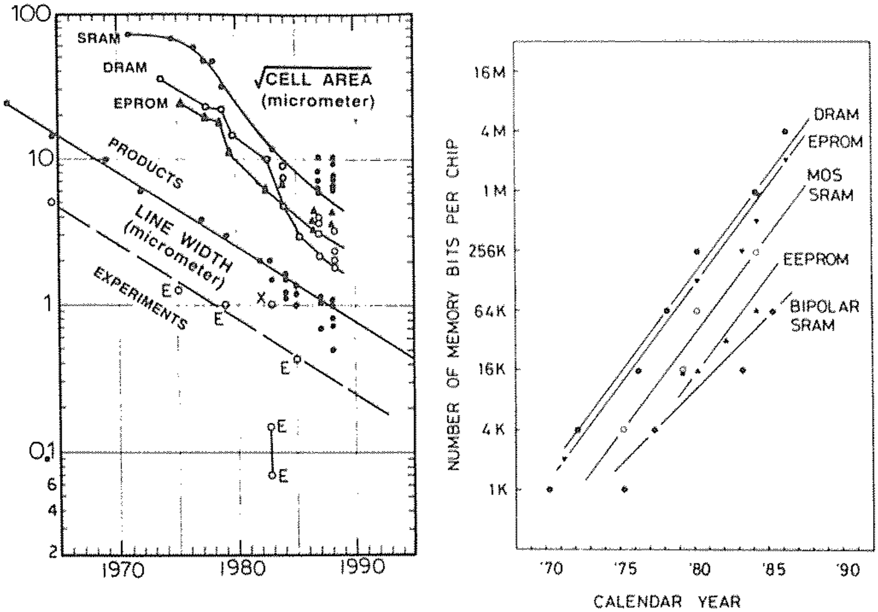


Fig. 8-3 (a) Trends of the laboratory and the production lithographic line width (two lower straight lines) and of the cell size (the square root of cell area) of production SRAM, DRAM, and EPROM cells.[3] (© 1988 IEEE). (b) Level of integration of semiconductor memories which have been presented at the *IEEE International Solid-State Circuits Conference* versus calendar year.[2] (© 1986 IEEE).

SRAMs are also characterized by their input/output (I/O) capability: one group has a TTL I/O capability, while the other has an ECL I/O capability (see chap. 7, section 7.2.2 for a comparison of TTL and ECL I/O signals). For the past several generations, high-speed CMOS technologies have been used to build TTL I/O SRAMs, while ECL I/O SRAMs have been implemented using bipolar technology. This is changing, since BiCMOS technology has begun to be used to fabricate both TTL and ECL I/O SRAMs. For example, 256-kbit ECL I/O SRAMs with 15-ns access times were commercially introduced in 1989, and 1-Mbit devices are expected shortly. TTL-I/O BiCMOS SRAMs are challenging incumbent CMOS and bipolar devices and already offer higher performance than all but the best of the CMOS SRAMs.

As SRAMs have evolved, they have undergone an increase in density. Most of this has been due to the use of smaller line widths (e.g., the 4-kbit MOS SRAM used 5-μm lines, while the 16-kbit, 64-kbit, 256-kbit, and 1-Mbit SRAMs were built with 3.0-μm, 2.0-μm, 1.2-μm, and 0.8-μm lines, respectively). The remainder of the density increase has been due to improvements in process technology, novel cell designs, and circuit innovations.[2]

Figure 8-2 indicates that one word line and two bit lines are connected to each SRAM cell: consequently two access transistors are also provided in each cell. In principle, it should be possible to achieve all memory functions using only one column line, one bit line, and one access transistor. In practice, however, normal variations in device parameters and operating conditions, make it difficult (or impossible) to obtain reliable operation at maximum speed using a single access-line to flip-flop cells. Therefore, the symmetrical bit lines (bit 0 and bit 1), are necessary. In a matrix of memory cells, each pair of bit-0 and bit-1 lines is shared by all memory cells in each column, and each word line is shared by all memory cells in each row.

8.2.1 MOS SRAMs

MOS SRAMs can be fabricated in either NMOS and CMOS, with early MOS SRAMs implemented in the former (Fig. 8-2a depicts a schematic diagram of an NMOS-SRAM cell). The load devices, M_5 and M_6, are depletion-mode NMOS transistors; the driver transistors, M_1 and M_2, and the access transistors, M_3 and M_4, are enhancement-mode NMOS transistors. The fully static CMOS SRAM *(full-CMOS)* cell shown in Fig. 8-2b was developed next. In this cell, the load devices are PMOS enhancement-mode transistors, while the other four transistors are enhancement-mode NMOS devices. Finally, a four-transistor SRAM cell was developed, with high-valued polysilicon resistors used as the load devices *(poly-load cell).* Since all the transistors in this cell are NMOS devices, the cells can be built in CMOS *p*-wells. When an array of such poly-load cells was fabricated in a *p*-well and combined with full-CMOS peripheral circuits, the resulting SRAMs demonstrated a significant decrease in power consumption compared to NMOS SRAMs (since most of the power in an SRAM is dissipated in the peripheral circuits, such as the off-chip drivers). These SRAMs also displayed a higher packing density than did SRAMs built with fully static CMOS cells.[4]

The full-CMOS cells dissipate less power than do the other types of MOS SRAM cells when in the standby mode of operation (i.e., when the cell is not being written into or read from). In the other types of cells, one inverter is always *ON*, and hence significant current is drawn from V_{DD} (much more, of course, in the six-transistor NMOS cell). In the full-CMOS cell, however, one transistor in each of the coupled inverters is *OFF*; thus, only junction-leakage current is drawn from V_{DD}. This current is approximately three orders of magnitude less than that drawn by the poly-load cell,[5] demonstrating that very little power is dissipated in the full-CMOS cell. In addition, the stability of the full-CMOS cell is high, since higher alpha-particle immunity and smaller junction leakage sensitivity is exhibited. The insensitivity to leakage allows operation at higher temperatures. Fully static CMOS SRAMs have been exploited as low-power, battery-backed memory devices in battery-operated consumer goods and portable office equipment.

The isolation of *n*-channel from *p*-channel devices means that the full-CMOS cell generally requires a larger area than does the poly-load cell. Furthermore, in order to establish contacts from the drain regions of the *p*-channel devices to those of the *n*-channel devices (as well as to the n^+ polysilicon-gate materials), the metal layer must be used. In four-transistor/poly cells, there are no *p*-channel devices; so no *n*-channel-to-*p*-channel isolation is needed in the memory array of the chip. Buried contacts which take up less space can also be used to connect the drains of the access transistors and the gate of driver-transistors. Since it is also more costly to manufacture SRAMs with full-CMOS cells, four-transistor/poly-load cells have been used in the design of most high-density CMOS SRAMs. By 1989, 1-Mbit CMOS SRAMs were being offered commercially.

Alternative cell structures have also been investigated as a way to increase the density of SRAMs. One such approach is to stack the transistors on top of one another. For example, the full-CMOS cell can be built with the active *p*-channel transistor load stacked above the *n*-channel devices. The second layer of transistors can be fabricated on recrystallized silicon[12] or built using a hydrogen-passivated polysilicon transistor.[13] Although the processing techniques needed to fabricate such three-dimensional stacked structures are complex and difficult to control, these devices will become more attractive as the need to form higher-density structures becomes greater.

A novel SRAM cell based on the reverse base current of a bipolar transistor and consisting of only one bipolar transistor and one MOS transistor was described by Sakui et al.[14] Such a compact cell would allow SRAMs to be built with the same densities as DRAMs.

8.2.1.1 Circuit Operation of MOS SRAM Cells. NMOS and CMOS SRAM cells all exhibit the same basic circuit behavior (Figs. 8-2a and b). When writing or reading data in such a cell is desired, the word line of the cell (which is held low in the standby state) is raised to V_{DD} (e.g., +5 V). This causes the enhancement-mode NMOS access transistors M_3 and M_4 to be turned *ON*. *Writing* is performed by forcing one of the bit lines low (e.g., close to 0 V), while maintaining the other at its standby value (about 3 V). For example, to write a *1*, the bit-*0* line must be forced low.*

When this occurs, M_1 turns *OFF* and its drain voltage rises due to the currents flowing through M_5 and M_3. When M_2 has been turned *ON*, the bit line can be returned to its standby level, leaving the cell in the state of storing a *1*. (The operation of writing a *0* is complementary to that just described).

For *reading* a *1*, the bit lines must both be biased at about 3 V. When the cell is selected, current flows through M_4 and M_2 to ground and through M_5 and M_3 to the bit *1* line. The gate voltage of M_2 does not fall below 3 V, so M_2 remains *ON*. The voltage of the bit-*0* line is thus reduced to less than 3 V, while the voltage of the bit-*1* line is pulled up above 3 V, since M_1 is *OFF* but M_5 is *ON*. As a result, a differential output signal exists between the bit-*0* and bit-*1* lines. This signal is fed to the sense amplifier, which in SRAMs is a differential amplifier capable of providing rapid sen-

* The cell must be designed so that the conductance of the access transistor is several times larger than that of the load transistor (i.e., comparing M_4 to M_6), in order for the drain of M_2 and the gate of M_1 to be brought below V_T.)

sing. Consequently, one of the bit lines needs to be only slightly discharged in order to generate a differential input signal large enough to drive the sense amplifier. To avoid a change in the state of the cell during reading, however, it is necessary for the conductance of M_2 to be around three times as large as that of M_4 so that the drain voltage does not rise above V_T. (The operation of reading a *0* is complementary to the one just described).

Among the most important factors limiting the maximum speed of MOS SRAMs are the delay associated with signal propagation through address buffers and decoders (which gets longer as the number of inputs and outputs increases), and the delay associated with the charging and discharging of the word and bit lines (which increases as the RC product of the word- and bit-line structures increases).

Bit-lines are typically formed in metal (Al), and hence their resistance is not a significant limitation. Word-lines, however, are normally implemented with polysilicon or polycide, so their higher resistance is considerably larger than that of the bit-lines. This then becomes one factor that limits SRAM speed. The parasitic capacitance of the word- and bit-line structures themselves, combined with the many paralleled access transistors (which are connected to each word and bit line), results in a large equivalent lumped capacitance on each of these lines.

Finally, there is a delay associated with the signal propagation through the sense-amplifier and data-output circuits. Considerable effort has been expended to develop high-speed sense amplifiers for SRAMs. In addition, a circuit technique known as *address-transition detection* (ATD) has also been used to speed up sensing in MOS SRAMs. In this technique, the bit lines are equilibrated upon detection of a change in the address input. (Sense-amplifier design, and the details of ATD are discussed in texts on VLSI circuit design, and so will not be further described here.

8.2.1.2 SRAM Cell Layout and Processing Issues. For maximum density to be achieved in a memory device, the cells must be be laid out in as small a size as possible. The size is determined by the cell's topology and by the design rules of the IC fabrication technology. A completed layout design represents the outcome of years of development, and a great deal of design experience. Figures 8-4a, 4b, and 4c are examples of SRAM-cell designs for a six-transistor NMOS SRAM cell,[6] a poly-load cell,[8] and an advanced full-CMOS cell,[7] respectively. These cell layouts also reflect some of the process enhancements that have made possible the improvements in SRAM performance, speed, and density.

Table 8-1 shows the evolution of the MOS SRAM; shrinking line widths and a variety of process enhancements can be seen to have been primarily responsible for the density and performance improvements. Figure 8-5 presents the same information graphically. The process enhancements are summarized in the following paragraphs describing the cells used in 1-Mbit SRAMs.

Several 1-Mbit CMOS SRAMs based on the poly-load cell were described in detail in the *IEEE Journal of Solid-State Circuits* (October, 1988), and a 4-Mbit SRAM (October, 1989).[109] The access time of these devices ranges from 7.5 to 18 ns (although the access times of commercially available 1-Mbit SRAMs in 1989 were

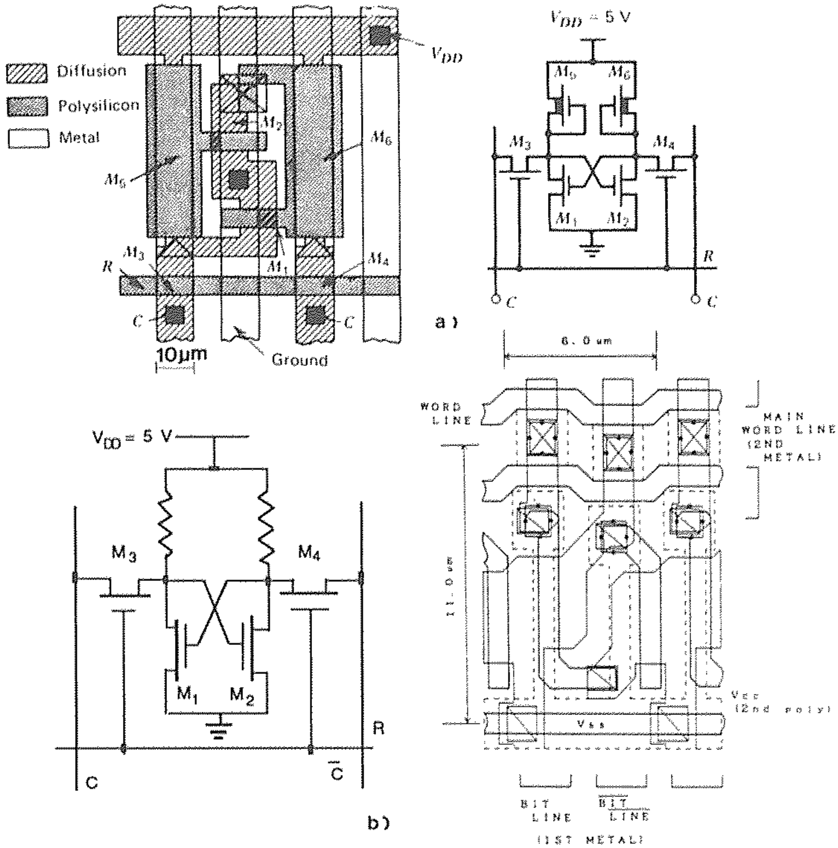


Fig. 8-4 (a) Schematic and layout of NMOS SRAM cell.[6] After R. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, eds., *Large Scale Integration*. Copyright 1981, John Wiley & Sons. Reprinted with permission. (b) Schematic and layout of poly load SRAM cell.[8] (© 1988 IEEE).

being given as 25-120 ns).[32] From the layout of the 7.5-ns CMOS SRAM cell (Fig. 8-4b), it can be seen that a double-polysilicon, double-level metal process is used, and that the cell area is 66 μm^2 (6.0 x 11 μm). The first level of polysilicon is a polycide structure, which is used for the V_{SS} power line in the memory array as well as for the gates of the MOS transistors. The second poly layer is used to form both the high-valued load resistors and the low-resistance V_{DD} lines. The bit lines are formed in Metal 1, and the word lines in Metal 2. A 0.8-μm twin well-CMOS process is used in which the channel lengths of the *n*- and *p*-channel transistors are 0.8 and 1.0 μm, respectively.

Other advanced poly-load-based CMOS SRAM designs include such process enhancements as trench-isolation structures, triple-level polysilicon, self-aligned con-

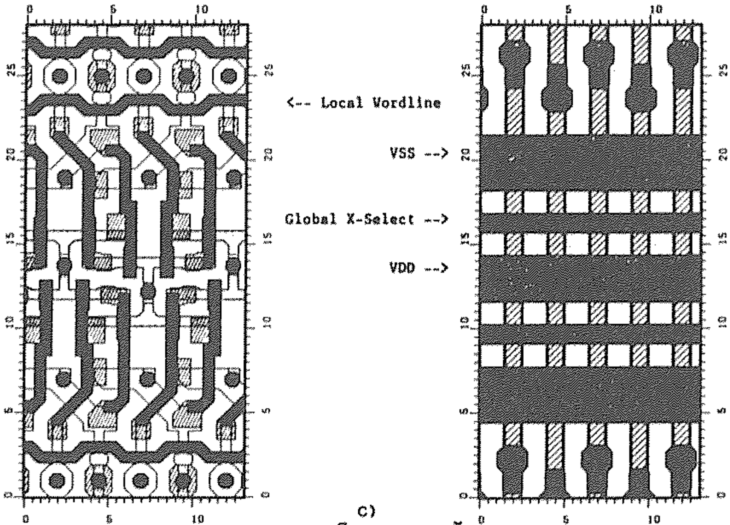


Fig. 8-4 (c) Layout of a full-CMOS SRAM cell.[7] (© 1988 IEEE).

tacts, buried contacts, and spare rows and columns for redundancy.

A 1-Mbit CMOS SRAM based on a full-CMOS cell was reported in the same issue of the *IEEE Journal* mentioned above (Fig. 8-4b).[7] It has a longer access time (25 ns) but consumes only 1 μW of power in the standby mode. Its memory-cell size is 5x12 μm, which implies that clever cell design and advanced processing techniques can produce fully static cells with sizes comparable to those of poly-load cells. (Note that the major reduction of cell area in such full-CMOS cells has been attributed to the use of a local interconnect layer[108] – chap. 3, section 3.11.2. The local interconnect layer allows the two inverters of the cell to be cross coupled with only two contacts per cell

Table 8-1. Evolution of MOS SRAM Technology

Introduction Date		Size (bits)	Access Time	Minimum Feature Size	Process Enhancements
1969	PMOS	256 bit			Silicon Gate, CVD Oxide
1972	NMOS	1k		8 μm	Depletion-Mode Load
1975	NMOS	4k	4 ns (1988)	5 μm	Ion-Implant V_T Adjust
1978	NMOS	16k		3 μm	Plasma Etching /Wafer Stepper
1982	CMOS/NMOS	64k	15 ns	2 μm	Double-Poly
1985	CMOS/NMOS	256k	25 ns (1988)	1.2 μm	Polycide/Poly, LDD Structures
1988	CMOS/NMOS	1M	25 ns (1988)	0.8 μm	(Polycide/Poly, Double-Metal,
	Full CMOS	1M	25 ns (1988)		Twin-Well, LDD Structures)
1989	CMOS/NMOS	1M	10 ns		
	CMOS/NMOS	4M	25 ns	0.5 μm	3.3 V, Retrograde p-Well,
	BiCMOS	1 M	8 ns	0.8 μm	25 Mask Levels, Twin-Well

– compared to nine contacts in a cell implemented with double-level metal, but with no local interconnect level – and thus the cell size can be made signigicantly smaller.)

This SRAM is fabricated using a 14-mask process, and it employs a single level of polysilicon and two layers of metal. The poly layer is selectively doped p-type when it acts as the gate for PMOS devices, and n-type when it is a gate for NMOS devices. A silicide strap is also used to connect poly lines to diffused regions. Spare rows and columns are included for redundancy.

Another full-CMOS SRAM (256 kbits in size) has also been described.[9] This 35-ns access-time part uses TiN as a local-interconnect structure between gates and diffused regions, and 0.8-μm MOS devices. The 100-nm-thick TiN layer has a sheet resistance slightly lower than that of a 500-nm doped-poly layer (14 Ω/sq vs. 20 Ω/sq). The TiN makes contact to the $TiSi_2$ layer that is formed on the surfaces of both the diffusion and gate regions. Since TiN is also an effective diffusion barrier, it prevents the phosphorus dopant in the n^+ polycide structure from diffusing and counterdoping the diffused drain regions of the PMOS devices when a connection is formed between them. (The formation and properties of TiN as an interconnect and barrier material is discussed in greater detail in chap. 3).

Several 4-Mbit CMOS SRAMs are described in the October, 1989 issue of the *IEEE Journal of Solid-State Circuits.* The decrease in SRAM cell size as a function of minimum feature size is shown in Fig. 8-5.

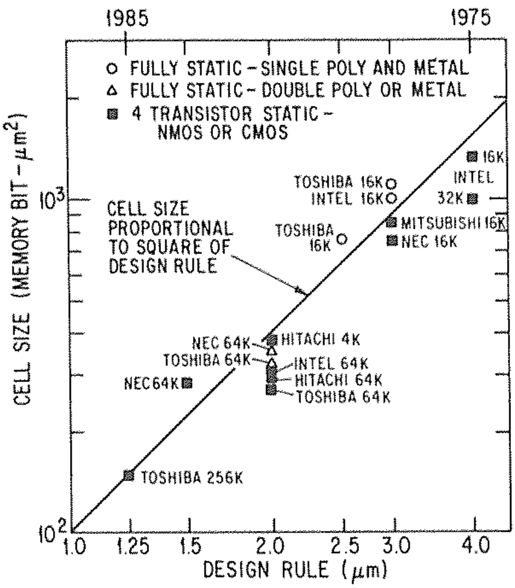


Fig. 8-5 MOS SRAM cell size versus design rule history.[116] (© 1986 IEEE).

8.2.1.3. High-Valued Polysilicon Load Resistors for MOS SRAMs

High-valued resistors are used as the load devices in the poly-load SRAM cell (Fig. 8-2b). In order to minimize power consumption and yet maintain an optimum soft-error rate, the load current of the cell is set to about 31 pA.[15] Very high-valued load resistors must be used to obtain such small load currents. For example, it has been calculated that 164 GΩ resistors must be used for 64-kbit and 256-kbit SRAMs, and 97 GΩ resistors are needed for 1-Mbit and 4-Mbit SRAMs (Fig. 8-6a).[15] Films made of materials with very high sheet resistances must be used to fabricate these load resistors to avoid the consumption of excessive area.

Undoped polysilicon films exhibit high sheet-resistivity values, making them good candidates for fabricating such structures (Fig. 8-6b). When undoped polysilicon films are implanted with arsenic in doses from $\sim 1\times10^{13}/cm^2$ to $1\times10^{15}/cm^2$, the sheet resistivity can be controlled from 10^4 Ω/sq up to about 10^{12} Ω/sq (Fig. 8-6c). Hence, to fabricate a high valued resistor (for example, a 97-GΩ resistor for a 1-Mbit SRAM cell), a polysilicon film with a sheet resistance of 26 GΩ/sq can be used. This sheet resistance can be obtained with an As implant dose of $\sim 3\times10^{13}/cm^2$. The length of a 97-GΩ resistor fabricated in such a 50-nm-thick, 1.2-μm-wide line of polysilicon would be 4.0 μm.

Undoped polysilicon exhibits such high resistivity because some of the impurities in the films segregate to the grain boundaries and do not effectively produce free carriers. In addition, the grain-boundary regions trap some of the free carriers that are produced (see Vol. 1, chap. 6).

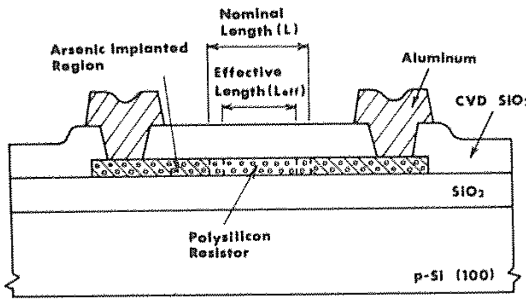
The I-V characteristics of the high-valued polysilicon resistors predicted by the trapping model of Lu et al. fit the experimental data fairly well if the resistor length is not too short.[16] On the other hand, lateral diffusion from adjacent higher-doped regions in the poly can significantly alter the resistance value if such diffusion takes place over a large enough fraction of the resistor length. Because the potential energy barrier to diffusion along the grain boundaries is lower than that in the bulk (see Vol. 1, chap. 8), the rapid diffusion of impurities along grain boundaries can bring impurities to the lightly doped poly regions, even at relatively low temperatures.

The effect just described can be important in the design of polysilicon-load SRAM cells. The resistors are normally formed in the second polysilicon layer, and the remainder of this layer is implanted with a much higher dose (so that it can serve as a low-resistance interconnect path). During this implant, the high-resistivity poly regions are covered with a mask to avoid impurity doping. The minimum size of the mask is limited by the effect of the lateral diffusion of impurities from the highly doped regions during the activation anneal of the polysilicon following ion implantation (e.g., 950°C for 30 min). Hence, a lower limit of about 3 μm was initially predicted for the length of such resistors.

A technique for reducing the extent of the lateral diffusion by implanting the polysilicon with a very heavy dose of oxygen ($\sim 1\times10^{22}/cm^3$) has been reported.[17] High-valued resistors can be fabricated with lengths as small as 0.8 μm (Fig. 8-6d). The oxygen apparently segregates to the grain boundaries, retarding the diffusion of the

	Memory size (bits)	Feature Size (µm)	Power Supply Voltage (V)	Load Current per Bit (pA)	Typical Memory Standby Current (µA)	Load Resistance (GΩ)	Memory Cell Size (µm²)	L/W of Load Resistor (µm)	Sheet Resistance (GΩ/□)	Thickness of Poly-silicon Resistor (nm)	Chip Size (mm²)
	[14] 64K	2.0	5.0	31	2	164	16×19 (304)	7.0/2.0	47	100	5.44×5.80
Estimated Value	256K	1.2	5.0	31	8	164	10×11 (110)	4.0/1.2	47	70	~6.5×7.5
	1M	0.8	3.0	31	33	97	7.0×7.5 (52.5)	3.0/0.8	26	50	~8.5×9.5
	4M	0.5	3.0	31	130	97	3.4×4.2 (14.3)	2.0/0.5	24	30	~8.5×10
	16M	0.25	1.5	31	520	48	1.7×2.1 (3.6)	1.0/0.25	12	30	~8.5×10

a)



b)

Fig. 8-6 (a) Comparison of parameters about the load resistors in SRAM cells. (b) Schematic cross section of a polysilicon resistor.

arsenic (and perhaps also increasing the potential barrier height by forming silicon-oxygen bonds).

The high sheet resistance of polysilicon-load resistors is also reduced by hydrogen diffusion into the polysilicon from plasma-deposited nitride passivation films. This was found to be controllable by sandwiching the polysilicon film with an LPCVD silicon-nitride film (which contains much less hydrogen than does plasma-deposited nitride; see Vol. 1, chap. 6).[19]

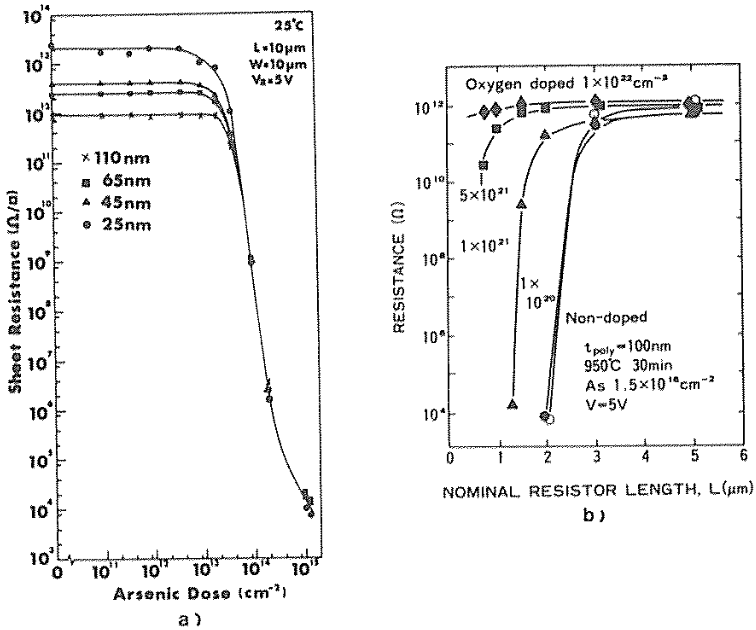


Fig. 8-6 (c) Sheet resistance versus arsenic dose for ion-implanted polysilicon resistors.[15] (© 1985 IEEE). (d) Resistance versus nominal resistor length for oxygen-doped and undoped polysilicon resistors. Resistance is normalized to 1-μm width.[17] (© 1987 IEEE).

8.2.1.4 Soft Errors in SRAMs. SRAMs offer better resistance than DRAMs to both transient and total-dose radiation, making them better suited for some military and space applications. Until recently, the soft-error rates of SRAMs (see section 8.3.4) were negligible compared to those of DRAMs. However, as geometries have been scaled down to produce circuits of greater density, alpha-particle–induced soft-error rates have also become a concern in SRAMs.[33] Although p-well CMOS in itself raises the threshold against soft-error failures, the use of extra buried p-layers has also been explored as a way to reduce such errors by an additional three orders of magnitude.[10] In addition, the full-CMOS cells exhibit less susceptibility than poly-load cells to single-event upsets and soft errors.

CMOS/SOS technology provides inherently harder parts than does bulk CMOS, and SRAMs have thus been built in CMOS/SOS for such applications. A report detailing the causes of the increase in soft-error rates in densely packed SRAM cells is given in reference 11. Another report on the modeling of alpha-particle sensitivities of SRAMs indicates that state-of-the-art CMOS SRAMs from 64 kbits to 1 Mbit can be made to exhibit sufficient insensitivity to alpha particles.[18]

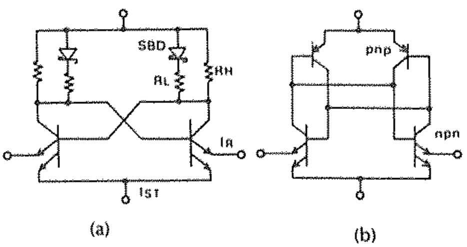
8.2.2 Bipolar and BiCMOS SRAMS

Although MOS SRAMs have achieved much higher device densities on a chip, as well as lower cost and lower power per bit, bipolar SRAMs using emitter-coupled logic (ECL) technology are still faster. Hence, they are chiefly used in applications where highest-speed operation is required (e.g., in the cache memory of high-speed computers). ECL SRAMs are classified into two groups: *high speed* (7-15 ns access times), and *ultra-high speed* (<7 ns access time).[24] The fastest 16-kbit bipolar SRAMs have access times of <4 ns,[20] and a subnanosecond 5-kbit bipolar SRAM has been reported.[21,31] As noted earlier, such bipolar SRAMs exhibit an ECL I/O capability.

The emitter-coupled cell shown in Fig. 8-2c is the most widely used bipolar SRAM cell. The load devices affect both the current in the standby mode and the saturation conditions of the driver transistors. In addition, since these devices also determine the read/write current, they have a significant impact on the access time. In early bipolar SRAMs the load for the flip-flop was simply a resistor, which caused the driver transistors to saturate when accessed. In more recent high-speed ECL SRAMs, another cell (as shown in Fig. 8-7a), has been used that utilizes a *pnp* transistor as the load. The advantage of this is that it allows for a smaller cell size, since the transistors are fabricated in parasitic elements. However, because the loads still allow the driver transistors to saturate, SRAMs which use such cells cannot achieve ultra-high speeds. Hence, such cells are used in medium-speed, high-density bipolar SRAMs.[23]

In ultra-high-speed ECL cells, saturation is avoided through the use of a Schottky diode in combination with resistors (the so-called *Schottky-barrier-diode, [SBD] switched-load-resistor cell;* Fig. 8-7b.[22] The load-resistance value in this cell is high during standby and low when active, making the current during standby about one-thousandth of that drawn by the cell during reading and writing.

Although this cell was invented quite some time ago, it did not gain wide acceptance because of its low density. Recent cell designs that incorporate trench isolation (to allow closer spacing of the transistors) and two levels of polysilicon (one for the interconnections and the other for the resistors) have allowed cells about half the size of conventional SBD cells to be realized.



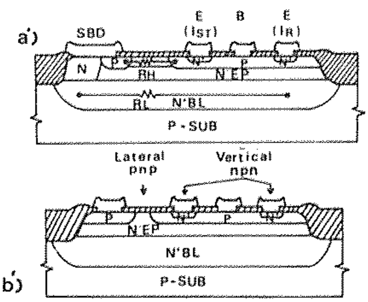


Fig. 8-7 Equivalent-circuit diagrams and cross-sectional views of (a) switched-load memory cell, and (b) cross-coupled *pnp* load cell.[2] (© 1986 IEEE).

In ultra-high-speed SRAMs, the Schottky diode speeds the switching response and increases the soft-error immunity by sustaining most of the stored charge on the SBD capacitance. A large SBD capacitance is therefore needed in order for high reliability to be maintained. The SBD capacitance, however, is quite small per unit area (2.8 fF/μm^2). An alternative cell that incorporates a separate Ta_2O_5 capacitor has been reported.[24,27] The use of this capacitor, which has a capacitance of 8.5 fF/μm^2, makes it possible to reduce the cell size by 30% compared to a conventional SBD cell.

Another ECL-SRAM cell using polysilicon diodes as the load elements has been developed.[113] The advantages of this approach include: compact cell size, very low standby current, very low parasitics, and small active junction area. Access times of 1.5 ns for a 1-kbit SRAM have been demonstrated.

Process innovations that have been incorporated to allow faster, higher-density SRAMs include the use of electromigration-resistant aluminum alloys (to permit higher current densities in metal interconnect stripes) and U-groove isolation (see chap. 2) which reduces the isolation width by a factor of three compared to fully recessed LOCOS isolation (see Fig. 2-35a).

Circuit techniques have also been used to enhance the performance of bipolar SRAMS. For example, read/write current has been concentrated in the active region of device operation, and the word delay has been reduced through the use of Darlington drivers. Table 8-2 summarizes the evolution of bipolar SRAM technology.

Table 8.2 Bipolar SRAM Evolution

Introduction Date	Size	Access Time	Load Device	Process Enhancements	Circuit Enhancements
1975	1 k	1.5 ns	Resistor	Al-Cu	Non-Saturated Read/Write Current
1978	4 k	2.2 ns	Schottky Diode		Darlington Drivers
1982	16 k	3.0 ns	*pnp* Transistor	U-Groove Isolation	
1986	64 k	5.0 ns			

8.2.2.1 BiCMOS SRAMs. Although high-speed ECL SRAMs up to 64 kbits in size have been fabricated, such large bipolar SRAMs have power dissipation and yield problems. Power dissipation increases because each cell draws a minimum standby current of about 2 μA to maintain sufficient noise margins and immunity from alpha particle soft errors. Defects in the narrow base region make it difficult for high-yielding circuits containing 262,144 narrow bases to be produced (i.e., each cell of a 64-kbit ECL SRAM designed with *pnp* loads contains four transistors).

On the other hand, CMOS alone cannot be used to build such high-performance, higher-density SRAMs because the driving capability of CMOS is inferior to bipolar, and it is practically impossible to design an input and output buffer circuit in CMOS that has an ECL I/O capability.

SRAMs have been developed which combine both bipolar and CMOS devices on the same chip. A comparison of 64-kbit SRAMs built using ECL and BiCMOS techno-

Table 8-3 Comparison of 64-kbit Bipolar and BiCMOS SRAMs[24]

	1.3-μm BiCMOS	2.0-μm BiCMOS	1.2-μm Bipolar	Unit
Organization	64 k x 1	16 k x 1	64 k x 1	word x bit
Address Access Time	7	13	10	ns
Write Pulse Width	4	7	11	ns
Operating Power	350	500	1300	mW
Memory-Cell Size	97	230	524	μm^2
Die Size	20	30	55.4	mm^2

logies is given in Table 8-3.[24] More recently, 256-kbit BiCMOS SRAMs have been announced, with access times of 8-12 ns.[25,26] As noted earlier, BiCMOS SRAMs can be designed with TTL as well as ECL I/O capabilities.

In early BiCMOS SRAMs, moderate-speed bipolar transistors were integrated into what was essentially a CMOS technology in order to provide faster output buffers and sense amplifiers. In more recent designs, the CMOS devices are being fabricated on a basically high-speed bipolar chip. In an early 16-kbit BiCMOS SRAM, the cells of the memory array were poly-load cells, the peripheral circuits were CMOS, and the I/O buffers and sense amplifiers were bipolar circuits. In a more recent 256-kbit BiCMOS SRAM design, the memory array was implemented with full-CMOS cells, and bipolar sense amplifiers and ECL output buffers were used. This memory had a reported access time of 8 ns and could be operated with battery backup (i.e., since it draws only 1 μA during standby, a battery with minimal power can provide long-term backup).

Figure 8-8 shows the speed-versus-density characteristics of bipolar, CMOS, and BiCMOS SRAMs.[34] These curves indicate that bipolar technologies no longer offer the fastest performance at such densities, and that ECL I/O BiCMOS offers higher

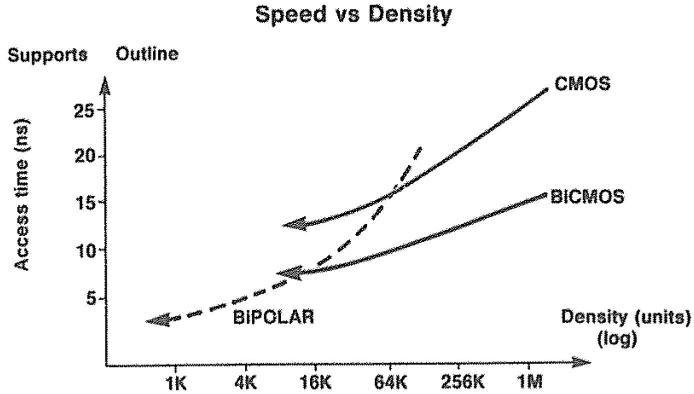


Fig. 8-8 Comparison of speed and density of CMOS, Bipolar and BiCMOS SRAMs.[34] Reprinted with permission of Semiconductor International.

speeds than bipolar ECL at middle and higher densities (because of the severe speed, power, and density trade-off in bipolar ECL). In addition, BiCMOS SRAMs are somewhat faster than CMOS SRAMs at equivalent power and density for the same geometry, and TTL I/O BiCMOS SRAMs are a little slower than ECL I/O SRAMs because of their larger output voltage swing. Finally, BiCMOS offers a range of speed and power trade-offs (from very fast, low-density BiCMOS at high power, to slightly slower, mid-density memory circuits at moderate power).

Recently, an 8 ns 1-Mbit BiCMOS SRAM[114] and a 3.5 ns 16-kbit ECL BiCMOS SRAM[115] have been reported.

8.3 DYNAMIC RANDOM ACCESS MEMORIES (DRAMS)

As noted earlier, *dynamic random access memories* (DRAMs) are so named because their cells can retain information only temporarily (on the order of milliseconds); even with power continuously applied. The cells must therefore be read and refreshed at periodic intervals. While the storage time may at first appear to be very short, it is actually long enough to allow for many memory operations between refresh cycles.

Despite of this apparently complex operating mode, the advantages of cost per bit, device density, and flexibility of use (i.e., both read and write operations are possible) have made DRAMs the most widely used form of semiconductor memory to date.

The earliest DRAMs used three-transistor cells and were fabricated using PMOS technology. Nevertheless, their introduction represented an immediate, dramatic decrease in the minimum semiconductor memory-cell size, since they could replace SRAMs based on a six-transistor cell. As a result, more cells per chip could be implemented. However, DRAM cells consisting of only one transistor and one capacitor were quickly implemented,[28] and such cells have been used in DRAMs ever since.

8.3.1 Evolution of DRAM Technology

The earliest MOS combinational logic networks, referred to as *static logic circuits,* operated without any need for periodic clock signals. However, it was recognized that clock signals could be used to advantage in combinatorial and sequential logic circuits. By introducing clock signals at arbitrary circuit nodes, it was possible to achieve faster operation, greater circuit density, and reduced power dissipation. Such logic circuits became known as *dynamic logic circuits.*

Data in these circuits was temporarily stored in dynamic registers (in the form of charge on the gate of an MOS transistor), rather than in static registers (which store data as the state of a flip-flop circuit). Thus, dynamic shift registers could be built with fewer transistors and, consequently, on much smaller areas of silicon than were needed for static shift registers. This allowed a dramatic increase in logic-circuit density.

A question arose, however, as to how long the gate of a MOS transistor could store charge before that charge would be lost through leakage currents. It turned out that at near room temperature, the charge could stored for more than 10 milliseconds. If a clock signal were to arrive at intervals significantly shorter than this, a large fraction of the initially stored charge would still remain on the MOS transistor gates. Therefore, if a clock signal were to be applied at least this frequently, dynamic registers could serve as efficient charge-storage nodes.

It was also quickly realized that if the dynamic stored-charge approach was practical in dynamic logic circuitry, it might also work for semiconductor-memory designs. In the case of a dynamic memory, however, a *refresh signal* had to be applied to each charged cell node at sufficiently frequent intervals (typically, every 4-8 ms), to allow the temporarily stored data on each cell to be retained indefinitely.

The concept of the DRAM was patented by Dennard of IBM in 1968, and the first commercial DRAM was introduced by Intel in 1970. The latter was built using a three-transistor cell (Fig. 8-9) in PMOS silicon-gate technology, while Dennard's patent used a one-transistor cell. In the three transistor cell, the charge is stored on the parasitic capacitance of the gate of transistor M_1. (Although the capacitance in this cell is a parasitic effect, it is drawn explicitly in the circuit schematic of Fig. 8-9 because it is essential for normal memory-cell operation.) The leakage current of the reverse-bias junction of the drain region of transistor M_3 discharges this capacitance over a period of several milliseconds or more. Hence, a periodic signal must arrive at the node in order for the charge stored on the capacitor to be maintained. Since one-transistor DRAM cells quickly replaced the three-transistor cells, the rest of our discussion will be restricted to them.

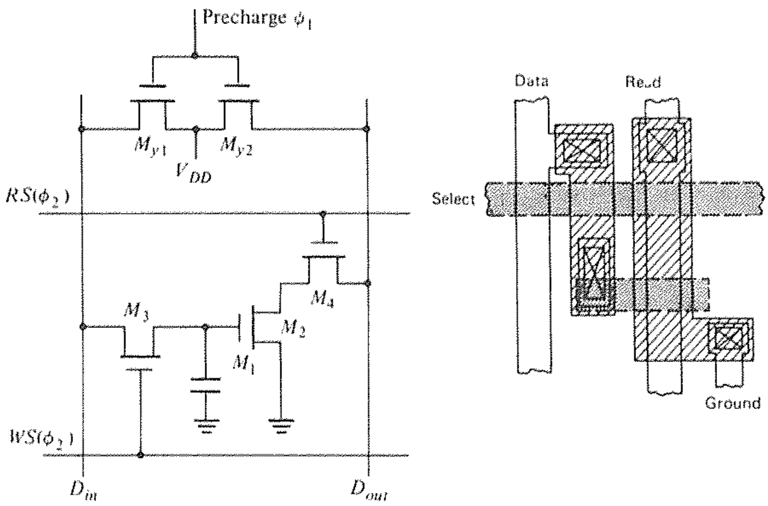


Fig. 8-9 Layout and circuit for a 3-transistor DRAM cell.

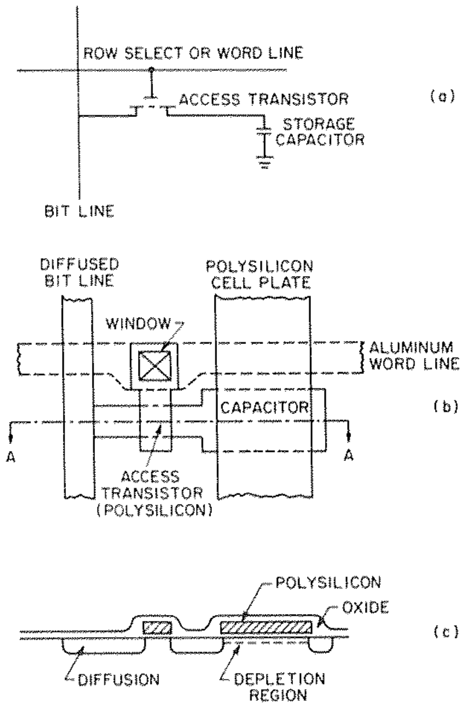


Fig. 8-10 Single-transistor DRAM cell and storage capacitor. (a) Circuit schematic. (b) Cell layout. (c) Cross section through A - A.[6] After R. Hunt, "Memory Design and Technology," in M.J. Howes and D.V. Morgan, eds., *Large Scale Integration.* Copyright 1981, John Wiley & Sons. Reprinted with permission.

8.3.1.1 One-Transistor DRAM Cell Design. The influence of Dennard's one-transistor cell[28] (which actually consists of one transistor and one capacitor) is considered to be comparable to that of the invention of the transistor itself.[3] The design of this cell has been rendered in many versions since its invention, and we will begin the description of its evolution (which, by the way, is far from over), with a description of a simple cell that utilizes a polysilicon layer as one plate of the cell capacitor (Fig. 8-10). The first such cell was fabricated with 8-μm features, used 1280 μm^2 of silicon area, and was employed in the design of the 4-kbit NMOS DRAM. There have since been many variations of even this simple cell, with single or double layers of polysilicon, different methods of capacitor formation, and different materials used for the word and bit lines. These are reviewed in detail in reference 30.

In the cell shown in Fig. 8-11a, the capacitor stores the charge on the cell (storage capacitor), and the NMOS transistor allows the bit line to access the charge-storage

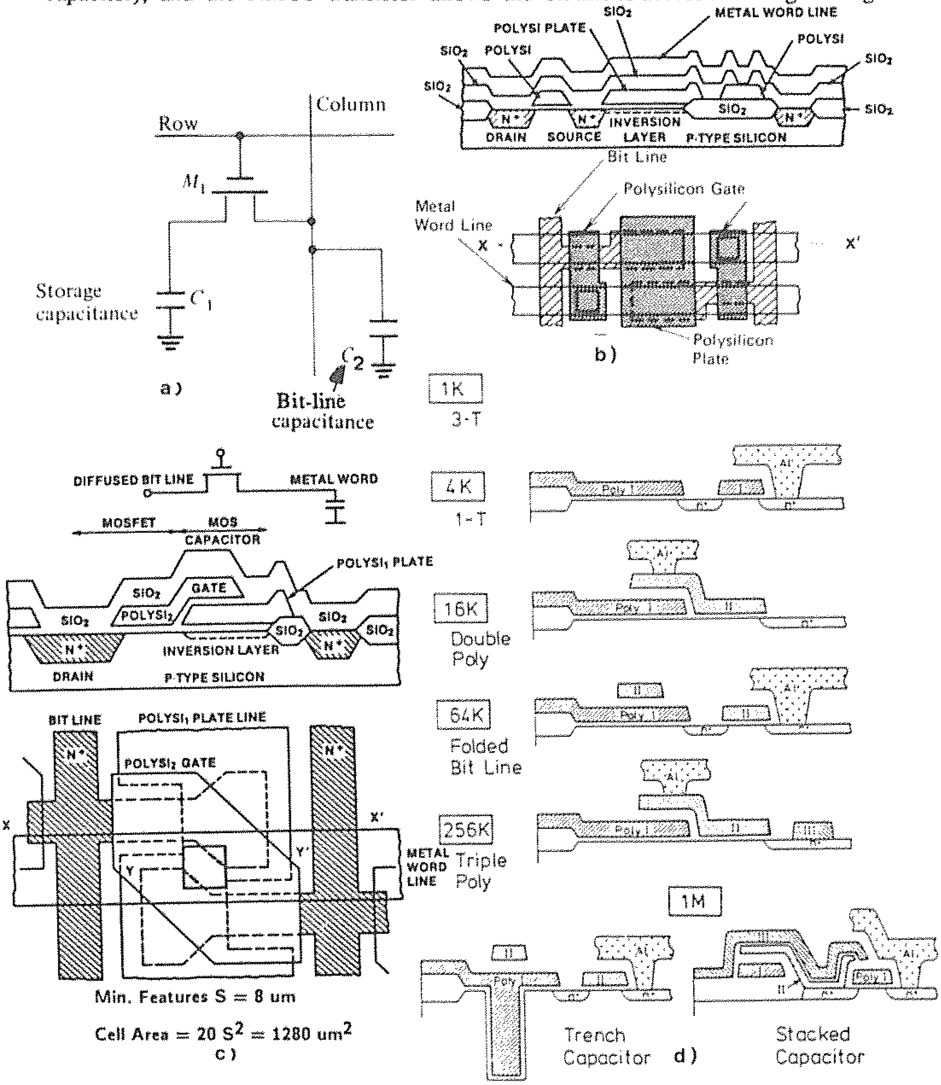


Fig. 8-11 (a) DRAM cell connections to word and bit lines. (b) and (c) Cross-sectional and layout views of (b) the single-poly-word-line/diffused-bit-line 4-kbit DRAM cell, and (c) the double-poly diffused bit line merged DRAM cell.[3] (After C. N. Berglund) (d) Structural innovations of DRAMs.[41] (© 1985 IEEE).

region of the capacitor. The storage capacitor consists of a polysilicon plate over a thin oxide film (which is the capacitor dielectric), with the semiconductor region under the oxide serving as the other capacitor plate. An n^+ diffused region in the semiconductor substrate serves as the bit line and an aluminum stripe as the word line. As can be seen in the cross-section of the cell (Fig. 8-11b), the bit-line diffused region makes contact with the n^+-diffused source region of the access transistor. A contact between the word line and the polysilicon gate of the access transistor is made, as also shown in this figure.

One widely implemented enhancement of this basic cell design is shown in Fig. 8-11c. In this modified cell, the floating drain region of the access transistor is eliminated and a second layer of polysilicon transfers the charge from the bit line to the storage capacitor. This not only allows the cell to be reduced in size, but also increases its storage capacity.[35] The disadvantage is that a double-polysilicon process must be used. As has often been the case in the evolution of VLSI, increased packing density and better performance are achieved at the price of somewhat greater process complexity. Figure 8-11d summarizes the structural innovations used as DRAMs have evolved.

8.3.1.2 Operation of the One-Transistor DRAM Cell. To study the operation of the cell in Fig. 8-11c, assume that the substrate is grounded and that 5 V are applied to the polysilicon top plate of the storage capacitor (which we'll refer to as the *plate electrode* of the capacitor). The semiconductor region under the polysilicon plate serves as the other capacitor electrode, and in an NMOS cell this *p*-type region is normally inverted by the 5-V bias. As a result, a layer of electrons is formed at the surface of the semiconductor, and a depleted region is created below the surface. (The electrode on which the charge is stored will be referred to as the *storage electrode.)*

To write a *one* into the cell, 5 V are applied to the bit line, and a 5-V pulse is simultaneously applied to the word line. The access transistor is turned *ON* by this pulse, since its V_T is about 1 V. The source of the access transistor is biased to 5 V, since it is connected to the bit line. However, the electrostatic potential of the channel beneath both the access-transistor gate and the polysilicon plate of the storage capacitor is less than 5 V, because some of the applied voltage is dropped across the gate oxide. As a result, any electrons present in the inversion layer of the storage capacitor will flow to the lower potential region of the source, causing the storage electrode to become a depletion region that is emptied of any inversion-layer charge. When the word-line pulse returns to 0 V, an *empty potential well* remains under the storage gate. This empty well represents a *binary one*, and it is shown as the deep-depletion space-charge region in Fig. 8-12.

For writing a *zero*, the bit-line voltage is returned to 0 V, and the word line is again pulsed to 5 V. With the access transistor turned *ON*, electrons from the n^+ source region (whose potential has been returned to 0 V) have access to the empty potential well (whose potential is now lower than that of the source region). Hence, the electrons from the source move to fill it, thus restoring the inversion layer beneath the poly plate. When the word-line voltage is returned to zero, the inversion-layer charge present

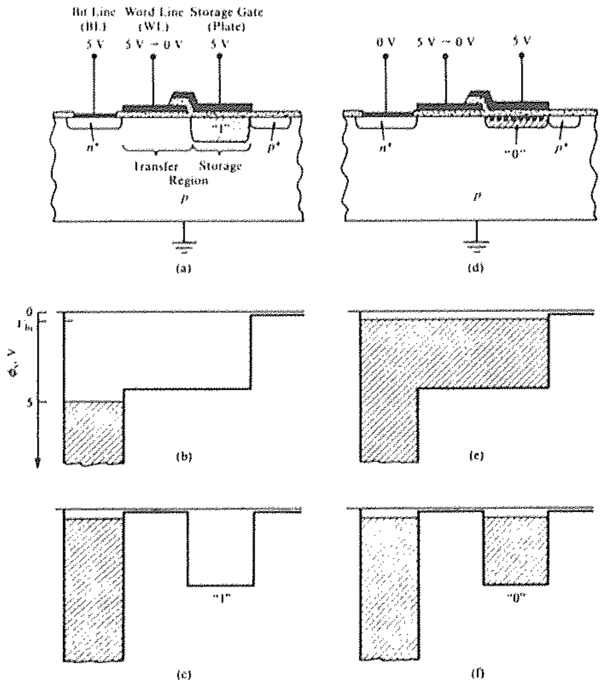


Fig. 8-12 A basic dynamic RAM cell, showing (a) a stored **one** and (d) a stored **zero.** The writing of a **one** is shown in (b) and (c), and the writing of a **zero** is shown in (e) and (f). From D. K. Schroder, *Advanced MOS Devices*. Copyright 1988, Addison-Wesley. Reprinted with permission.

on the storage capacitor is isolated beneath the storage gate. This condition represents a stored *binary zero*.

Note that when a one is stored, an empty well exists. This is not an equilibrium condition, since electrons that should be present in the inversion layer have been removed. As electrons are thermally generated within and nearby the depletion region surrounding the well, they will move to re-create the inversion layer. A stored one thus gradually becomes a stored zero as electrons refill the empty well. The nature of the planar one-transistor DRAM cells is that ones become zeros, and zeros remain zeros.

To prevent the ones from being lost, each cell must be periodically refreshed by the memory so that the correct data remains stored at each bit location. The time interval between refresh cycles is called the *refresh time*. The total leakage current of the cell must be low enough that the cell does not discharge and lose its memory state between refreshes. A typical guideline has allowed a 20% degradation of the charged state of a cell during the refresh time.[105] For example, if this interval is 8 ms and the charge stored on a cell is 10^6 electrons, the maximum allowed leakage current at maximum operating chip temperatures and worst-case voltages and timing conditions is ~4 pA. This specification can be achieved fairly easily at the average cell level. However, since the leakage currents in a large number of cells exhibit a distribution, the high side of the distribution must be monitored in the course of technology development – and manufacturing – to ensure that the maximum allowable current is not exceeded in any of the memory chip cells.

It is predicted that as memory sizes continue to increase, the maximum allowable leakage current will be reduced. Since alternative capacitor dielectric materials may be used in advanced DRAMs, the leakage currents across such dielectrics will be an important material characteristic. In addition, the access transistor of the cells will have to be turned off very "hard" when the memory cell is not being accessed, since any source-drain leakage can also discharge the storage capacitor.

8.3.1.3 Writing, Reading, and Refreshing DRAM Cells. It's useful to consider the reading and writing of DRAM cells from a circuit perspective as well as from the device physics perspective just discussed. Figure 8-13 provides a simplified schematic of this operation.

Switches S1 and S2 are used to select the appropriate bit and word lines that are connected to the cell in question. Such switch closure is the response to an address

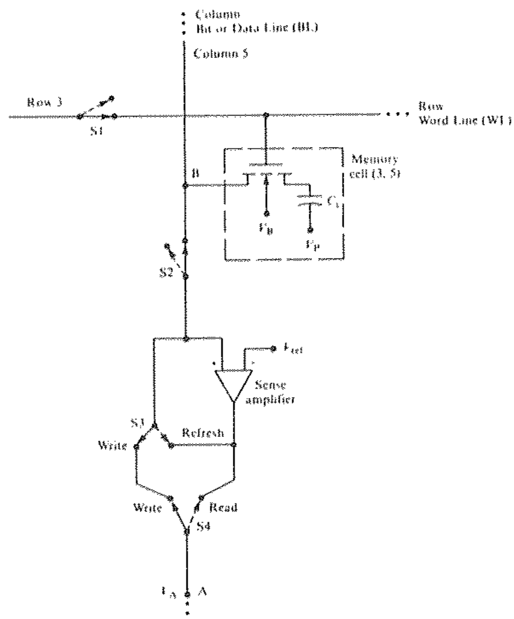


Fig. 8-13 Schematic diagram showing the writing, reading, and refreshing operations of a DRAM. After ref. [117], reprinted by permission of Scientific American, Inc.

signal fed to the bit (column) and word (row) decoders. The closure of S1 causes a "high" voltage to be applied to that particular word (row) line (in turn causing the access transistors controlled by this line to be turned *ON*).

If it is desired to write data into the cell at that point, an appropriate voltage must also exist on the correct bit line. In addition, switches S3 and S4 must be set to the write positions. This allows the voltage at point A on the bit line, V_A, to be applied to the source of the cell being addressed (point B). As long as Switch S1 is closed and the access transistor is turned on, the capacitor of the cell can be charged to $V_A - V_p$ (where V_p is the *storage-electrode potential*). When the write operation is completed, switches S1 and S2 are opened, and another cell can be written.

If it is desired to read data from a particular cell, the appropriate switches S1 and S2 are once again closed (it is assumed that all of the cells already contain the stored information). Switch S4, however, is set to the read position. The storage capacitor of the cell being read is now connected to one input of the sense amplifier. The sense amplifier is a *comparator circuit*, with its other input being connected to a reference voltage, V_{ref}. Therefore, if the cell-capacitor voltage is larger than V_{ref}, a logic *1* is read; if it is smaller, a logic *0* is read.

A logic *1* corresponds to the condition in which the storage electrode is depleted of its inversion layer charge. If a logic *1* is read immediately after the cell has been written, the signal will be strong. As time passes, thermal electron-hole pair generation will cause refilling of the empty potential well, thereby degrading the amplitude of the logic *1* signal. If too long a time elapses between the writing and reading, the inversion charge will be reestablished, and a logic *0* will be produced when the cell is read. Once a logic *0* is stored on a cell, however, it will continue to be read for as long as power remains applied to the capacitor plate electrode.

This indicates that the logic *1* must be periodically refreshed to allow it to be retained on the cell for indefinite time periods. Because it is not known what logic level is stored on each cell at any instant of time (especially since the cells are randomly read and written), it is mandatory that the entire memory be refreshed at periodic intervals (usually every few milliseconds). Furthermore, since reading a cell changes the charge on its capacitor, the cell must also be refreshed immediately following each read operation.

The refresh procedure is accomplished by switching S3 to the refresh position after the sense amplifier output has been set by the read operation. The output voltage of the sense amplifier will then write the appropriate information back onto the cell capacitor. If it is desired to refresh the entire memory, each cell can be read and refreshed. It is apparent, however, that data cannot be written while reading or refreshing is in progress.

One sense amplifier must be available for each bit line. Note that the sense amplifier is an extremely sensitive comparator (basically of the cross-coupled flip-flop type), and its design is critically important to the success of DRAM manufacture. Although the details of the design task are not our subject here, some aspects of sense-amplifier performance should be mentioned. When a cell is read, the charge stored on the cell capacitor is shared with the 10 to 20 times larger capacitance of the bit line (which, as we saw earlier, is a long conductor line connected to the sources of all of the cells in the column). After the time interval between refresh pulses has elapsed (e.g., 8 ms) the difference in stored voltage between a *1* and a *0* may be as small as 2 V. As a result, there may only be a 100-200 mV difference between the *1* and *0* signals applied to the sense-amplifier input.

8.3.1.4 DRAM-Cell Charge Storage and Capacitance. A *one* must be clearly distinguishable from a *zero* when the read operation is performed. The zero is represented by the inversion charge present when the potential well is full. This quantity in an MOS capacitor, Q_s, is given by

$$Q_s = (V_{G'} - 2\varphi_f)\, C_{ox} \approx V_{G'}\, C_{ox} \qquad (8-1)$$

where $V_{G'}$ is the voltage applied to the gate, φ_f is the difference in potential between the intrinsic Fermi level (E_i) and the Fermi level (E_F), and C_{ox} is the capacitance of the capacitor oxide.* In order to pack a great many cells onto a DRAM chip, the cell size is made as small as possible. This implies that it is also desirable to make the area of the storage capacitor as small as possible. On the other hand, Q_s of the storage capacitor must be large enough to send a sufficiently strong signal to the sense circuitry and to provide sufficient immunity from soft errors (see section 8.3.5). Novel cell designs have been developed in an attempt to satisfy these apparently contradictory requirements (such designs will be discussed later).

EXAMPLE 8-1: Calculate the charge stored in the inversion layer when a *zero* is stored (both in units of *coulombs*, and in terms of the number of electrons present) when 5 V are applied to an MOS capacitor whose dimensions are 4 x 4 μm, and which has an SiO_2 dielectric that is 15 nm thick. Also, find its capacitance, C_s.

SOLUTION: $Q_s\,(0) \approx V_{G'}\, C_s = (\varepsilon \times A \times V_{G'}) / t_{ox}$

$$= (3.9 \times 8.85\times10^{-14}\ \text{F/cm} \times 16\times10^{-8}\ \text{cm}^2 \times 5\ \text{V}) / 1.5\times10^{-6}\ \text{cm}$$

$$= 18.6 \times 10^{-14}\ \text{C} = 186\ \text{fC, or since } q = 1.6 \times 10^{-19}\ \text{C/electron}$$

$$Q_s = 1.15 \times 10^6 \text{ electrons; and}$$

$$C_s = Q_s / V_A = 37\ \text{fF}.$$

The above shows that the most important parameters involved in increasing the charge stored on the capacitor are the dielectric constant and thickness of the insulator, and the area of the capacitor.

The capacitance of the DRAM cell is also important. As described earlier, when the

* The approximation used in Eq. 8-1 is valid when $V_{G'} > 2\varphi_f$, which is the case for $V_{G'} = 5$V and $\varphi_f \approx 0.4$ V.

contents of the cell are sensed, the charge stored on it is "dumped" into the bit line connected to the sense amplifier. Because the bit-line capacitance (C_B) is typically 7 to 15 times larger than the cell capacitance, the capacitively-divided voltage difference applied to the sense amplifier (ΔV_{sa}) is substantially smaller than that existing in the cell alone. ΔV_{sa} is given approximately by

$$\Delta V_{sa} = (1/2)\, V_A\, (C_{ox}) / (C_{ox} + C_B) \qquad (8-2)$$

The minimum detectable voltage difference that the sense amplifiers of 1-Mbit DRAMs can detect (i.e., their *sensitivity)* is in the neighborhood of 150-200 mV. (Note that this sensitivity must be maintained under worst-case operating conditions of voltage, temperature, and noise, as well as worst-case variations of processing conditions.) It is predicted that the sensitivity of sense amplifiers will have to be significantly increased in order for higher-density DRAMs to be fabricated.

8.3.1.5 High-Capacity (Hi-C) DRAM Cells. Another technique for increasing the cell's charge-storage capacity without increasing its size was suggested independently by Sodini and Kamins,[37] and Tasch, et al.[106] This novel technique involves multiple ion implantations to increase the substrate doping in the local vicinity of the storage node.

A deep implantation of *p*-type impurities (boron) is first performed under the storage-plate area. This increases the substrate doping, which in turn increases the depletion-region capacitance of the storage capacitor. However, this single implant alone does not increase the charge-storage capacity of the cell, since the extra *p*-doping also reduces the difference in the surface potential between an empty and a full potential well.

To restore surface potential to its previous value it (without compromising the increased capacitance), a very shallow layer of *n*-type dopant (arsenic) is implanted under the storage plate area (Fig. 8-14). The implanted *donor* atoms, which are very close to the Si/SiO_2 interface, behave like a fixed positive oxide charge (and the presence of such oxide charge acts to increase the surface potential in NMOS structures). By increasing the depletion-region capacitance without simultaneously increasing the surface potential, the charge-storage capacity of these so-called *high capacity cells* (or *Hi-C cells*) is enhanced by about 50 percent compared to conventional cells.[35,104] Although, this technique increases the storage capacities of planar capacitor structures in these types of one-transistor cells, the cell size that must be used in order for adequate charge-storage capacity to be obtained eventually becomes too large for this type of cell to be used in advanced DRAMs (i.e., >1 Mbit). New cell structures have thus been developed for larger DRAMs.

8.3.1.6 CMOS DRAMs. With the introduction of the 256-kbit DRAM, the design of DRAM circuits began to change from NMOS to a mixed NMOS/CMOS technology. The cells of the mixed-technology memory array are all built in a common well of a CMOS wafer. The access transistor and storage capacitor of each cell are usually still fabricated using NMOS technology, while the peripheral circuits are

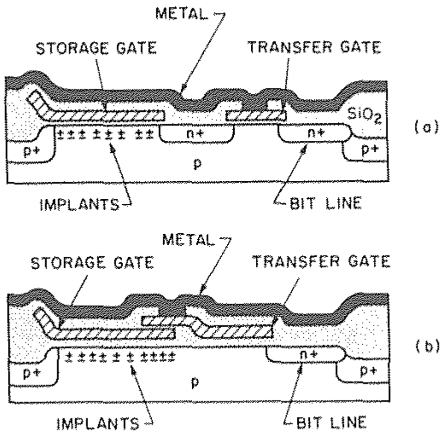


Fig. 8-14 High-capacity (Hi-C) dynamic RAM structure with arsenic (+) and deeper boron (-) implants. (a) One-transistor cell with single-level polysilicon. (b) Double-level polysilicon cell.[35] (© 1978 IEEE).

designed and fabricated in CMOS technology. The majority of 1-Mbit DRAM designs were executed in NMOS/CMOS technology, and this trend is expected to continue.

The advantages of NMOS/CMOS over NMOS DRAMs include lower power dissipation (i.e., by a factor of around 3) and smaller soft-error rates (see section 8.3.5). In addition, as power-supply voltages are reduced to allow smaller MOS transistors to be built, NMOS device design must become much more complex to allow the cells to function properly. CMOS on the other hand, can easily operate at such lower voltages. Finally, a circuit technique known as *static column decoding*, which significantly reduces the memory access time, can be successfully implemented in CMOS but not in NMOS (i.e., since to do so in NMOS would require a circuit dissipating excessively large standby power).[38]

In late 1989 1-Mbit BiCMOS DRAMs were introduced. Their access times of 40 ns placed them between the high-speed 1-Mbit SRAMs (access times of 20 ns), and the slower 1-Mbit CMOS DRAMs (access times of 60-80 ns).

8.3.2 Design and Economic Constraints on Advanced DRAM Cells

As the DRAM cell has been scaled down in size, the minimum amount of stored charge needed to maintain reliable memory operation has remained the same. This constant charge-storage value has had to be maintained within fabrication-cost constraints. From the system point of view, a new generation of DRAM will be embraced if it allows a density increase of about fourfold at the circuit board level, provided that this is accompanied by a cost reduction. To allow such an increase to be realized, the new

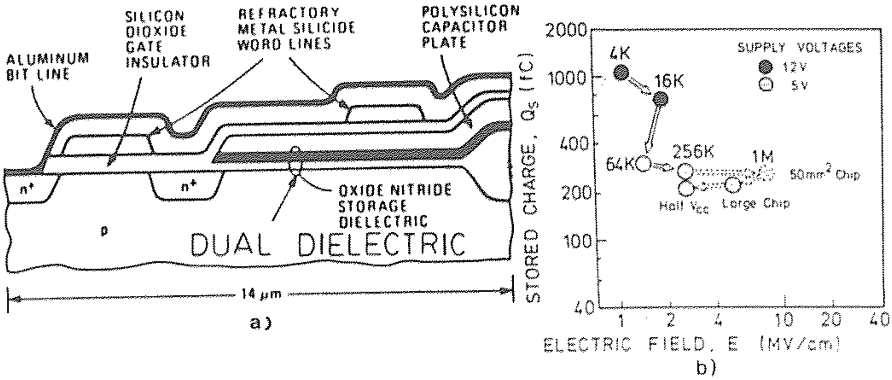
DRAM generation must be able to use the same size package as that used by the previous generation. This explains why the 300-mil package housed DRAMs for five generations. To squeeze enough cells onto a chip to allow this package to be retained implies that the cell size of the 1-Mbit DRAM cannot exceed 20 μm^2.

For future DRAM generations, however, increased memory size will be achieved by both increasing the chip size and shrinking the cell size. Since the chip size is predicted to increase by a factor of 1.5 from generation to generation, the cell area will therefore need to shrink to 40% of the size of the previous generation (which can be done by shrinking the minimum line width to 70% of that used previously). For 4-Mbit DRAMs, the cell size must therefore be no larger than 9 μm^2; for 16-Mbit DRAMs, no larger than 4 μm^2, and so on. To meet the signal-to-noise ratio constraints and the soft-error rate requirements, a minimum of ~200 fC of charge (~10^6 electrons) will have to be stored. For the fabrication process to be economically feasible, as few steps as possible should beyond those required to fabricate the transistors and interconnects should be added.

Recovery of the equipment and development costs of the 256-kbit, 1-Mbit, and 4-Mbit DRAMs has become the major factor dictating the three-year product-introduction and delivery cycle so that a profit can be generated from each DRAM generation. In a typical three-year cycle, fewer than 1 million chips would be shipped for sampling during the introduction year. In the first production year, 5 million chips would be shipped; production would increase to 50 million chips in the second year, and 500 million in the third. (Note that in the second year the market would be mainframe computers, and in the third year it would be personal computers.)[3] The three-year model predicts that 16-Mbit DRAMs will be introduced in 1992, and 64-Mbit DRAMs in 1995.

It was noted in section 8.3.1.3 that the cell's storage capacity could be increased by making the capacitor dielectric thinner, by using an insulator with a larger dielectric constant, or by increasing the area of the capacitor. The first two options are not currently viable, since capacitor dielectrics thinner than those now being used in DRAM cells (10 nm) will suffer leakage due to Fowler-Nordheim tunneling, and dielectrics with significantly larger dielectric constants than of SiO_2 have not yet been accepted for DRAM-cell application (although research work is under way to develop such higher-dielectric-constant materials).[39] One recent report described a plasma-CVD process for depositing high-quality Ta_2O_5 films.[119] (Ta_2O_5 exhibits a much higher dielectric constant than SiO_2 [22 vs. 3.9], but normally also suffers from a much higher leakage current.) However, by reacting $TaCl_5$ with N_2O under optimized plasma-CVD conditions, Ta_2O_5 films with a thickness that yielded capacitances equivalent to those of a 30-Å SiO_2 film, demonstrated very low leakage currents for up to 10-year operation at 3.3 V.

It should also be noted that since the 256-kbit DRAM generation bilayer films (consisting of both silicon nitride and SiO_2), have been used as the capacitor dielectric to increase cell capacitance (Fig. 8-15a). The higher dielectric constant of Si_3N_4 (twice as large as that of SiO_2) was responsible for this increase.



Factors	256K	1M	4M
Type	NMOS	CMOS	CMOS-{Trench or Stack
No. of Masks	9∼10	17∼19	20∼25
Total Process Steps	200	350	450
Test Time (Sec.)	60	120	240
Chip Size (Ratio)	1	2	3
Logical No. of Chips/ Wafer (Ratio)	3	2	1
Good Dies/Lot (Ratio)	6	2	1
Clean Room Class	100	10	1
Cost Ratio (on a Mature Production)	1	4	10

c)

Fig. 8-15 Cross-sectional view of a planar-capacitor DRAM cell with a two-layer capacitor dielectric.[3] (© 1988 IEEE). (b) Trends in storage charge and electric field across capacitor insulator.[41] (© 1985 IEEE). (c) Comparison of DRAM production factors.[44]

One technique that did allow thinner dielectric films to be used was the *half-V_{CC}* approach.[40] That is, the *plate electrode* is biased to $V_{CC}/2$ (i.e., to 2.5 V when V_{CC} = 5 V), and the *storage electrode* is allowed to swing between 0 V and 5 V. As a result, the same quantity of charge can be stored on the capacitor, but the value of the electric field acting on the dielectric is only half the value that exists when the voltage between the two plates equals V_{CC}. This technique has been implemented in the fabrication of 1-Mbit DRAMs (Fig. 8-15b).

The third option (increasing the capacitor area) can be effective if the area is increased by forming the storage capacitor in a trench etched in the substrate or by using a stacked capacitor structure. Both approaches have been implemented, and many variations of such three-dimensional capacitors have been reported.

The planar capacitor structure used in the one-transistor DRAM cell described in section 8.3.1.1 was predicted to be usable up to the 256-kbit DRAM generation. In this generation, the capacitor consumes 30 to 40% of the cell area. It was generally agreed that beyond this, a three-dimensional capacitor structure would be needed in order for sufficient charge storage to be obtained. It turned out, however, that virtually all of the DRAM manufacturers elected to squeeze everything they could from the planar capacitor, and continued to use it to manufacture 1-Mbit DRAMs. This decision was due largely to the difficulty in achieving a reliable capacitor dielectric in a trench cell at the time 1-Mbit DRAMs were introduced. The use of both larger chip sizes and the half-V_{CC} plate-electrode voltage technique permitted the planar capacitor to perform adequately for 1-Mbit DRAMs. Reference 42 presents the details of a 1-Mbit DRAM technology using a 38-fF planar-capacitor structure in which the cell size is 37 μm^2.

As DRAM size increases, process complexity is expected to increase markedly as well. For example, a 1-Mbit DRAMs is reported to require ~18 masks and 350 processing steps, all of which could be successfully carried out in a Class 10 cleanroom (Fig. 8-15c). In comparison, the 4-Mbit DRAM is expected to need 20-25 masks and in excess of 450 processing steps, and will thus require a Class 1 cleanroom processing facility.[43,44] A detailed report on the technology issues that will need to be addressed in the design and fabrication of 64- and 256-Mbit DRAMs has recently been published.[105]

In 1989 1-Mbit CMOS DRAMs with access times ranging from 6-100 ns were being commercially offered. (The fabrication of a high speed 22-ns CMOS DRAM was announced in late 1989, but it was not being offered for sale.)[110] At that time, 4-Mbit DRAMs with access times of 80-120 ns were also being offered, and 16-Mbit CMOS DRAMs with access times as small as 45 ns were being reported.[111] Finally, 1-Mbit BiCMOS DRAMs with access times of 30 ns were being introduced.[112]

8.3.3 Trench-Capacitor DRAM Cells

8.3.3.1 Trench Capacitor Processing for DRAMs. Trench-capacitor structures have been developed as a way to to achieve DRAM cells with larger capacitance values without increasing the area these cells occupy on the chip surface. (For example, the silicon-area reduction of a trench capacitor compared to a planar capacitor for the same specific capacitance is a factor of 18 or more. Specifically, a 4.0-μm-deep trench capacitor with surface dimensions of 0.87 x 2.4 μm will occupy less than 3 μm^2 of chip area but will have a capacitance of 40 fF.)[66] Many of the processing details involved in trench- capacitor fabrication are the same as those described in chapter 2, section 2.6.3, which deals with the process technology of trench-isolation structures. In this section we discuss those issues that are unique to the fabrication of trench capacitors used in DRAM cells.

There are several differences between the trench structures used for isolation and those used as DRAM capacitors. In the former, the dielectric film on the trench walls can be relatively thick, and the trench can be refilled with polysilicon or CVD SiO_2. In the latter, the insulator formed on the trench walls serves as the capacitor dielectric, and it must therefore be as thin as possible. Since the material that refills the trench serves as one plate of the capacitor, it must consist of highly doped polysilicon. Furthermore, in order for increased capacitance to be obtained through increases in trench depth (while all other parameters remain constant), the trench walls must be highly vertical. To allow for reliable refilling of the trenches, however, some trench sidewall slope must be allowed, and a compromise process that produces a nominal sidewall slope of 87° has been suggested.[46] Finally, to obtain such structures as Hi-C capacitors, the trench walls may need to be selectively doped.

Several techniques have been developed for achieving a dielectric capacitor film that is thin enough to provide both high capacitance and high reliability (that is, the dielectric must be able to provide the same equivalent breakdown voltage as the planar capacitor used in previous DRAM generations). First, composite dielectric films (e.g., thermally grown oxide and CVD nitride) are frequently used.[47,62,68] Since the nitride has a higher dielectric constant than SiO_2, a thicker composite film will yield the same capacitance as a thinner single SiO_2 layer. This thicker film prevents capacitor leakage due to dielectric breakdown or Fowler-Nordheim tunneling.

The growth of the thermal oxide film is also a key step. Unless preventative measures are taken, a thinner oxide will grow in the bottom corners (concave) and top corners (convex) of the trench. A higher electric field will exist across these regions, causing trench capacitors to exhibit higher leakage currents than planar capacitors.

This problem is avoided for the bottom corners by ensuring that the etch process produces a trench with rounded bottom corners (see chap. 2). In addition, an oxidation step for edge rounding and stripping is performed prior to the growing of the actual capacitor SiO_2 film. One report indicates that a 50-nm SiO_2 film is grown in this process and is then stripped in dilute HF (Fig. 8-16a).[48] In addition to smoothing out any sharp bottom corners, this step also removes any plasma damage from the trench walls.

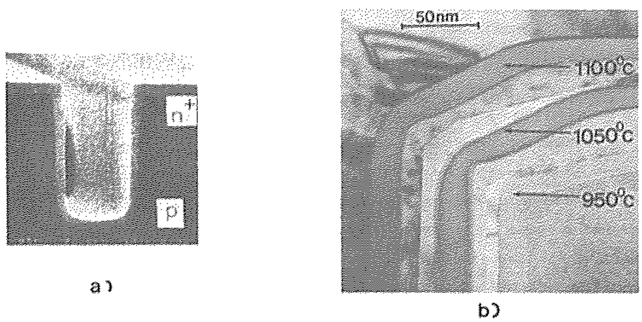


Fig. 8-16 (a) Rounding-off oxidation can produce trenches with smooth bottom corners.[48] (© 1985 IEEE). (b) Rounding-off oxidation can also reduce the severity of the sharp upper corner of the trench.[50] This paper was originally presented at the Spring 1989 Meeting of The Electrochemical Society, Inc. held in Los Angeles, CA.

The electric field is intensified at the top corners of the trench because they are normally quite sharp after etch, and this magnifies the effect of any oxide thinning that may occur. (The extent of the electric-field intensification is modeled in reference 49.) The edge-rounding oxidation step mentioned above also increases the curvature radius of the top corner of the trenched Si surface (Fig. 8-16b), thus helping to produce a trench capacitor with low leakage currents under high electric fields. However, because it is necessary to use a process that allows viscoelastic flow during oxide growth (to relieve the stresses that inhibit oxide growth at the convex corners), a higher temperature oxidation process (*e.g.* 1100°C) is usually involved.[49,50] The use of rapid thermal processing (RTP) to grow the trench oxide has also been reported.[51] Good leakage-current behavior is exhibited when RTP cycles of 1150°C for 25 sec in O_2 were used to grow the trench oxide.

The polysilicon that fills the trench must also be highly doped to prevent depletion effects. In situ doping of the poly is thus necessary. The conventional process for in situ doping of polysilicon employs gaseous phosphine as the dopant source. Unfortunately, this reduces the polysilicon deposition rate by a factor of about 25 (see Vol. 1, chap. 6). Specially designed LPCVD furnaces with caged boats are needed to improve the process.[117] However, these furnaces have particulate problems and cannot be automated, making them incompatible with a high-volume fabrication environment. A recent report described the use of t-butylphosphine (TBP) as an alternative doping source.[118] It can be used in standard automated 100-wafer LPCVD furnaces to produce in-situ doped polysilicon films. A higher deposition rate can be achieved (~20 Å/min), with adequate thickness uniformity. This material is also less toxic than phosphine.

8.3.3.2 First-Generation Trench-Capacitor–Based DRAM Cells. Trench structures for storage capacitor application in DRAMs were first reported in 1982-83 (Fig. 8-17a).[52] The processing technology that made these structures possible was anisotropic etching of Si by RIE. Earlier V-groove structures etched in Si by means of wet etching resulted in crystallographically produced sharp edges, which in turn degraded the gate-oxide integrity to the point where devices could not be reliably manufactured. One of the first tests that had to be met by RIE-etched trench capacitors was that of exhibiting breakdown characteristics equal to those of planar-type capacitors. As described in the previous section (and summarized in Fig. 8-17b), several reports showed that this could be achieved through the implementation of trench etching control measures, the use of edge-rounding procedures, or the use of combination films for the trench dielectric (e.g., thermal SiO_2 and CVD nitride).

In the first generation of trench-capacitor–based cells *the plate electrode of the storage capacitor is inside the trench, and the storage electrode is in the substrate.* The access transistor is a planar MOS transistor fabricated beside the trench capacitor, and the trenches are 3-4 μm deep. The cell size of the basic cells of this generation requires about 20 μm^2 of surface area, making the cells suitable for 1-Mbit DRAM designs. It was thought that with appropriate design-rule shrinkage, these cells would be appropriate for early 4-Mbit DRAM designs.[53]

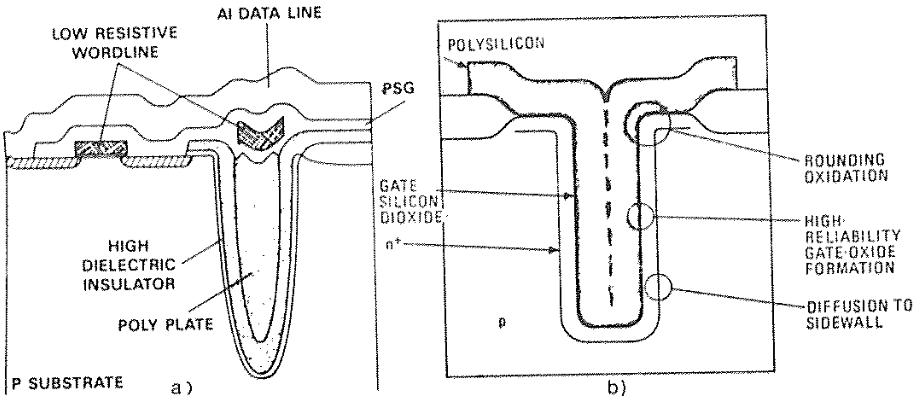


Fig. 8-17 (a) Basic DRAM trench capacitor structure.[52] (© 1982 IEEE). (b) Processing techniques used to insure fabrication of high-quality trench structures.

In one variation of this cell type, the plate electrode is grounded, and the substrate is biased between 0 and 5 V (which improves device isolation between adjacent cells).[54] The walls of the storage electrode (i.e., those in the *p*-type substrate) are doped *n*-type, creating a Hi-C type cell.

These first generation cells exhibit some disadvantages for smaller-sized DRAM cells. Since the charge is stored in a potential well in the substrate, if the cells are too close together, high leakage currents arise between adjacent cells (due to punchthrough or surface conduction). This problem can be alleviated through increased doping of the region between the cells or through the use of deeper, narrower trenches, but at the cost of creating other problems. First, the required doping in the substrate will lead to avalanche breakdown of the reverse-biased junction of the access transistors at spacings $\leq 0.8\ \mu m$. Second, deeper, narrower trenches are significantly more difficult to fabricate reliably and for practical trench dimensions the spacing limit is nearly reached for the cell sizes needed in 4-Mbit DRAMs. Further, since the storage node is in the substrate, there is no immunity to charge collection from alpha particles. Consequently, this type of trench capacitor is as vulnerable to alpha-particle–induced soft errors as cells made with planar storage capacitors. Several design modifications have been developed to increase capacitance without either making the trenches deeper or increasing cell size.

In the first modification, the plate electrode is folded around the sides of the storage electrode, creating a structure called the *folded-capacitor cell*, FCC (Figs. 8-18a and b).[55] A shallow trench is etched around most of the perimeter of the storage electrode. The plate electrode is deposited over this trench, much as a tablecloth is laid over a table top.[56] When both the sides and the planar area (tabletop) are covered, a capacitor with a larger area is obtained, and the capacitance is thereby increased (Fig. 8-18c).

Interestingly, the capacitor of this cell apparently utilizes *both* the planar- and trench-capacitor concepts. In addition, the cell's storage plate edges are electrically isolated

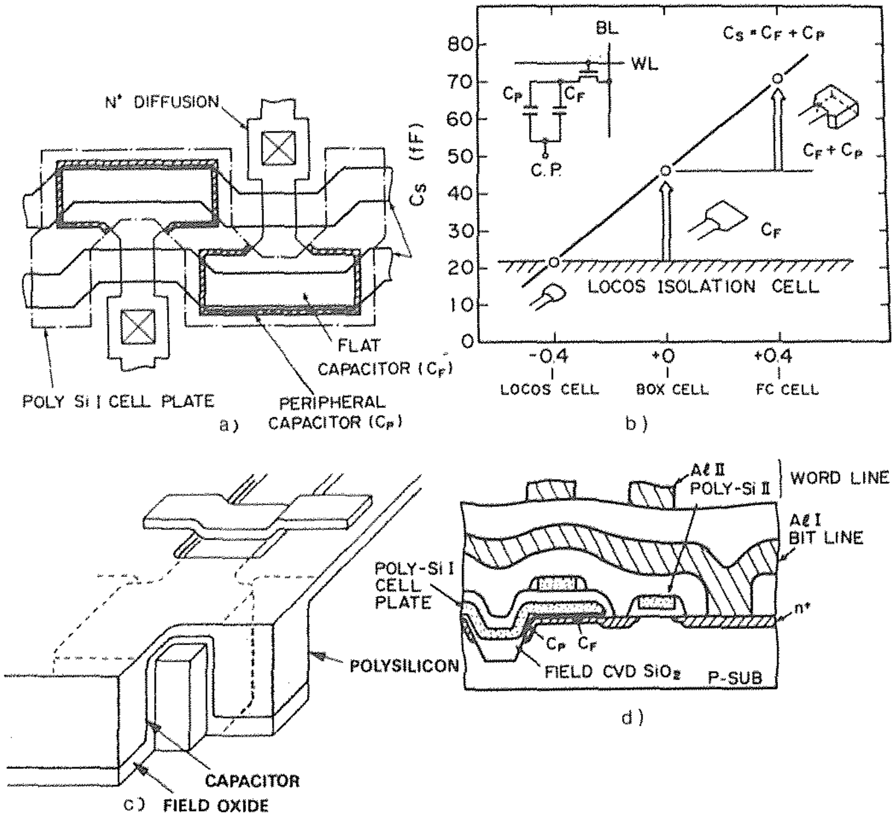


Fig. 8-18 (a) and (b) Top and perspective views of the folded capacitor cell (FCC).[55] (© 1984 IEEE). (c) Increase in capacitance by FCC. (d) Cross section of FCC showing CVD SiO_2 BOX-isolation structure.[56] (© 1986 IEEE).

from those of adjacent cells by means of a *BOX-type* isolation structure, rather than a LOCOS isolation structure (Fig. 8-18d). This increases the memory-array packing density (and in effect decreases the cell size), while also increasing the capacitance. An FCC cell size of 32 μm^2 with a 70-fF capacitor was reportedly used to fabricate 1-Mbit DRAMS. This cell appears to be scalable to 4-Mbit and 16-Mbit DRAM requirements.

In a second novel approach, the walls of the storage electrode were made to follow the outside edges of the cell perimeter, and the access transistor was placed inside (*Isolation VErtical Capacitor cell*, or IVEC, Fig. 8-19a).[57] A third invention folded the plate electrode around the storage electrode but used selective doping of certain trench walls to achieve isolation (i.e., the substrate trench walls that act as isolation structures were

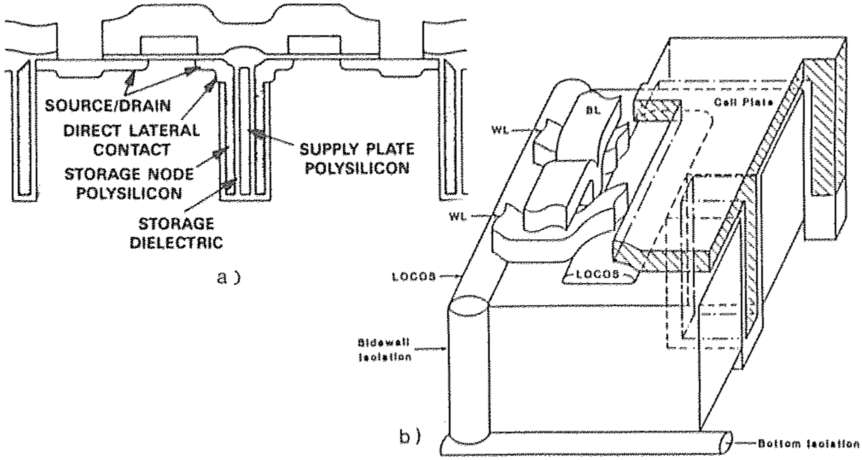


Fig. 8-19 (a) Isolation-merged VErtical Capacitor (IVEC) cell.[57] (© 1984 IEEE). (b) Perspective view of the FASIC cell.[58] (© 1987 IEEE).

selectively boron doped by means of oblique ion implantation) and storage (i.e., those walls used as the storage plate surfaces were arsenic doped by means of oblique ion implantation, creating a Hi-C storage capacitor). This latter cell was named the *folded bit-line adaptive sidewall isolated capacitor* (FASIC) cell (Fig. 8-19b).[58] FASIC cells can be made as small 10 μm^2 and with capacitances as large as 50 fF, making them suitable for use in 4-Mbit DRAMs. They require trenches of only 2 μm in depth.

8.3.3.3 Trench-Capacitor Structures with the Storage Electrode Inside the Trench (Inverted Trench Cell). One set of trench-capacitor designs sought to reduce punchthrough and soft-error problems by placing the plate electrode on the *outside* of the trench, and the storage electrode *inside* (Fig. 8-20a). Since the charge is stored inside the trench (which is therefore completely oxide isolated except in the region of lateral contact to the access transistor), it can leak only through the capacitor oxide or the lateral diffused contact.

Four examples of early approaches using such cell designs are the *buried-storage-electrode cell* (BSE) (Fig. 8-20b),[60] the *substrate-plate-trench cell*, (SPT) (Fig. 8-20c),[61] and the *stacked-transistor-capacitor cell,* (STT) (Fig. 8-20d).[62] In the first two, the plate electrode is heavily p-doped and is connected to the power supply, while the inside storage plate is heavily n-doped. Since the substrate is maintained at essentially an equipotential, the punchthrough problem exists only around the region through which the charge is introduced into the trench. Note that for heavily-doped storage electrodes (e.g., >2×10^{19} cm^{-3}), inversion will not occur at 5 V or less. Instead, the bias applied to the capacitor causes both plates of the capacitor to deplete; together with the oxide capacitor, these two depletion regions make this type of trench

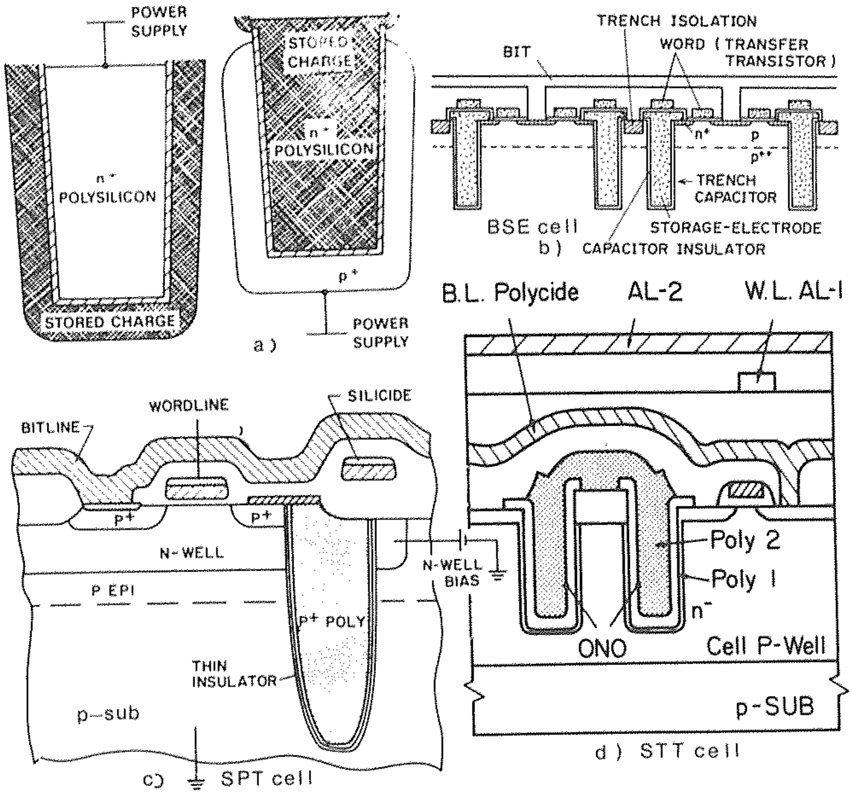


Fig. 8-20 (a) Mechanism of charge storage on outer and inner plates of the DRAM trench storage capacitor. (b) Cross section and process sequence of BSE cell.[60] (© 1985 IEEE). (c) Cross section of SPT cell.[61] (© 1985 IEEE). (d) Cross section of STT cell.[62] (© 1987 IEEE).

capacitor equivalent to three capacitor elements in series. Since the depletion regions grow with increasing voltage, the total trench capacitance decreases monotonically. The heavy doping of the plates therefore helps to maximize the cell capacitance. Finally, in such cells a *0* logic level is stored as 0 V and a *1* level as 5 V.

The problem with this type of cell is that the *gated-diode structure* shown in Fig. 8-21a can cause a significant leakage current to flow into the storage node, adversely affecting the cell's retention time. (The physics of the gated diode structure is treated in detail in reference 63.) An alternative cell (the IBM SPT cell) overcomes this problem by using PMOS access transistors and *p*-type doped inner-storage electrodes, and then creating the SPT cells in an *n*-well on a *p*-substrate (Fig. 8-21b).[64] As a result, the storage electrode gates the *n*-well-to-substrate junction, and the leakage current (as well as the *band-to-band tunneling-induced leakage current* generated in the bulk silicon)[65] is collected at the n-well contact instead of at the storage electrode. If such a cell is not built in a well (e.g., the BSE cell), the storage electrode will gate the junction formed by the storage electrode and the substrate, and the resulting leakage current will be collected by the storage electrode.

In the most advanced type of cell that does not use the substrate as the storage electrode, *both* the plate and storage electrodes are fabricated inside the trench opening, allowing both electrodes to be completely oxide-isolated. Lightly doped epitaxial layers on heavily doped substrates are not needed, and the cells will be free from punchthrough at arbitrarily small cell spacings. In addition, the soft-error rate will be reduced further than it is in the other inverted trench cells. However, these improvements are achieved through a substantial increase in process complexity.

Several such cells have been reported, including the *dielectrically encapsulated trench* (DIET) capacitor (Fig. 8-22a),[66] the *half-V_{CC} sheath-plate capacitor* (HPSC) (Fig. 8-22b),[67] and the *double-stacked capacitor* (DSP) (Fig. 8-22c).[68] The last has two polysilicon plates, one biased to V_{BB} and the other to $V_{CC}/2$. The capacitors formed by the lower poly plate and substrate (separated by the outer dielectric layer), and by the two poly layers (separated by the interpoly dielectric) act in parallel, almost doubling the cell's storage capacitance. A DSP cell of 6 μm^2 in size with trench depths of 4 μm is reported to exhibit a capacitance of 50 fF.

8.3.3.4 Trench-Capacitor Cells with the Access Transistor Stacked above the Trench Capacitor. The access transistor occupies a significant fraction of the cell area in trench-transistor cell designs. When this transistor is a planar transistor and is placed alongside the trench capacitor, surface area must be devoted to both structures. Attempts to use short-channel lengths for the access

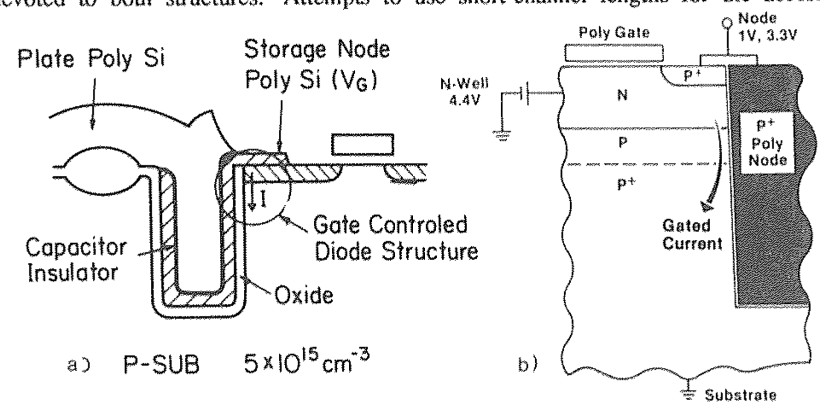


Fig. 8-21 (a) Cell structure with gate controlled diode. (b) Schematic representation of SPT cell bias conditions – p-substrate-to-n-well junction is gated by the polysilicon node.[64] (© 1987 IEEE).

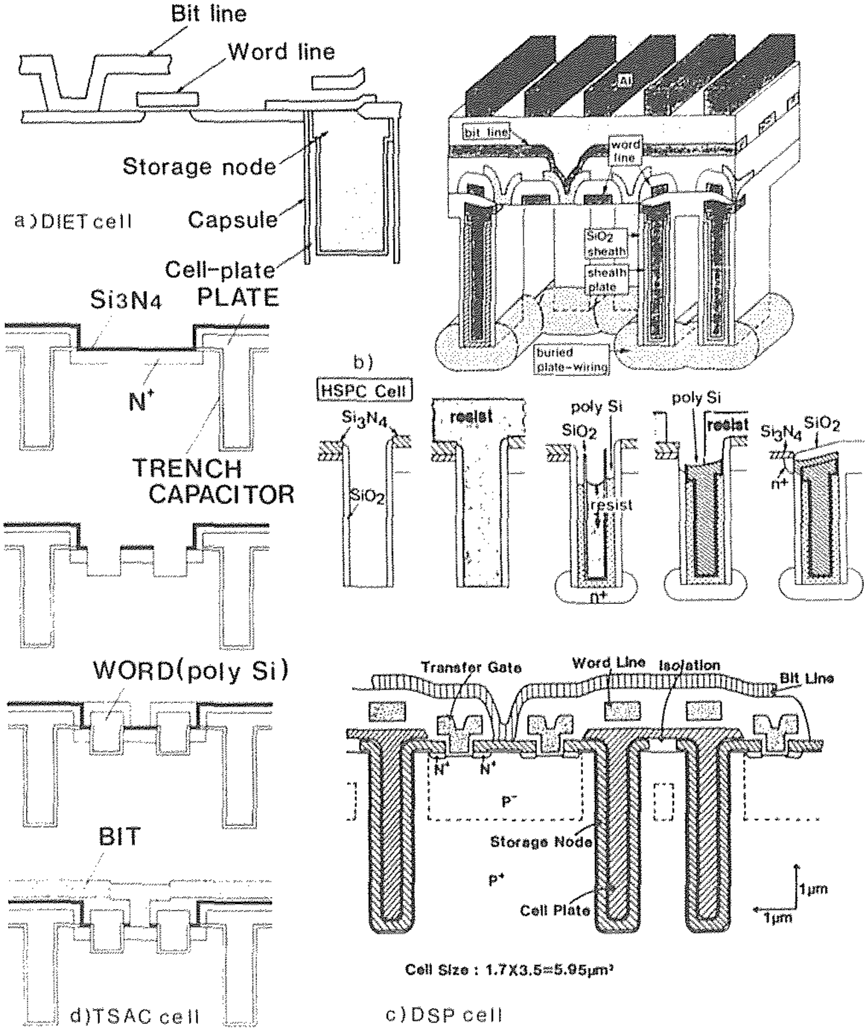


Fig. 8-22 (a) Cross section of DIET cell.[66] (© 1986 IEEE). (b) Perspective view and process sequence of the HPSCC cell.[67] (© 1987 IEEE). (c) Cross section of the DSP cell.[68] (© 1987 IEEE). (d) Fabrication process of the TSAC cell.[69] (© 1986 IEEE).

transistor have run up against the effects of drain-induced barrier lowering (see section 5.5.2).

One technique for overcoming this problem extends the gate length of the access transistor by forming a trench in the transistor channel (Fig. 8-22d). This reduces the area of the planar access transistor without decreasing its channel length.[69] Using this technique with a self-aligned contact structure, a cell size of 9 μm^2 was realized; such a cell is making it suitable for a 4-Mbit DRAM.*

A more efficient use of space would be to stack the transistor above the trench capacitor (and, if possible, to form a vertical-access transistor). Two examples of such cells are the *trench-transistor cell* (Fig. 8-23) and the *self-aligned epitaxy over trench cell* (SEOT) (Fig. 8-24).[71]

In the trench-transistor cell, the vertical-access (or trench) transistor is built in the top 2 μm of the trench. Its source is connected to the n^+ polysilicon storage electrode of the capacitor by a lateral contact, made by means of an oxide undercut etch and polysilicon refill. The drain, gate, and source of the trench transistor are formed by a diffused buried n^+ bit line, an n^+ polysilicon word line, and a lateral contact, respectively. The gate-oxide thickness is ~25 nm and the channel length is 1.5 μm. The transistor width is determined by the perimeter of the trench. The electrical behaviors of this trench transistor have been modeled, and the results are presented in reference 72. This cell has reportedly been used to build 4-Mbit DRAMs.

A *surrounding gate transistor* (SGT) cell that extends the trench-transistor cell approach has recently been reported (Fig. 7-23d).[123] This cell can be made smaller than the trench-transistor cell because it uses trench isolation for the bit-line isolation, rather than the LOCOS isolation used in the latter cell. The transistor and capacitor of this cell surround a silicon pillar. allowing the cell size to be shrunk to 1.2 μm^2 while still providing 30 fF storage capacitance. The SGT cell is being studied as a candidate for 64/256-Mbit DRAMs.

In the SEOT cell the storage electrode is first completely isolated from the substrate (Fig. 8-25a), and selective epitaxy is then grown. With the exposed Si area surrounding the trench acting as a seed, a single-crystal-silicon layer grows over the top of the trench (Fig. 8-25b). When the epitaxy growth is stopped before the lateral epitaxial film has grown completely over the trench, a self-aligned window is formed on top of the trench. The capping oxide on the top of trench surface is then etched, and a second epitaxial film is grown. A pyramidal window of polysilicon is formed on top of the exposed polysilicon in the trench; the material surrounding this pyramid is single-crystal silicon formed by means of lateral epitaxy. A planar surface is achieved after a specific minimum of epitaxial growth, and the isolation structure and MOS transistors are then fabricated. An 8-μm^2 cell size has been achieved using 0.85-μm design rules, making this cell suitable for 4-Mbit DRAMs. With some process improvements and design modifications, the cell appears to be scalable to 64-Mbit DRAM dimensions.

8.3.4 Stacked Capacitor DRAM Cells

Another approach that allows the cell to shrink in size without a loss of its storage capacity is that of stacking the storage capacitor on top of the access transistor, as

* A report that studied the design methodology and size limitations of submicron access transistors for DRAM applications is published in reference 70.

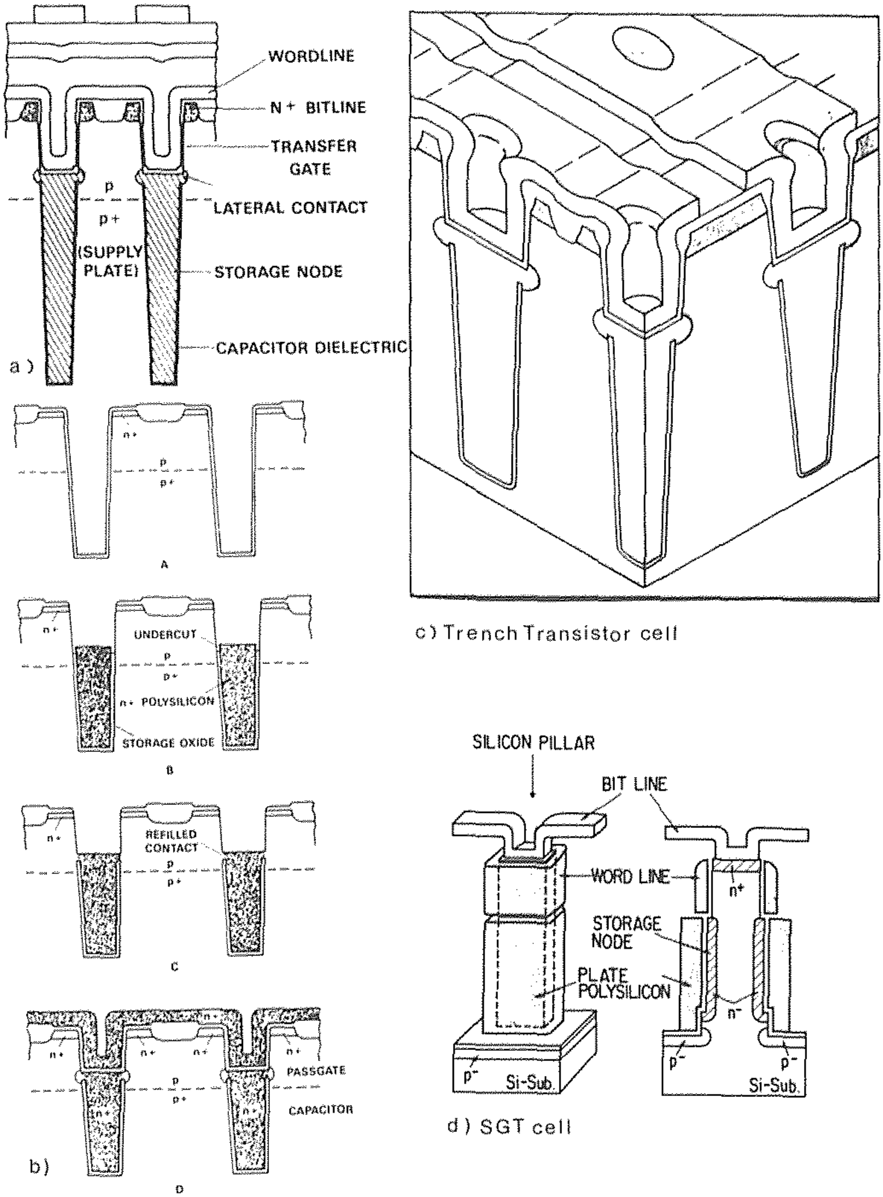


Fig. 8-23 (a) Cross section; (b) perspective view; and (c) fabrication sequence of the trench transistor cell;[59] (© 1985 IEEE). (d) Schematic view of SGT cell.[123] (© 1989 IEEE).

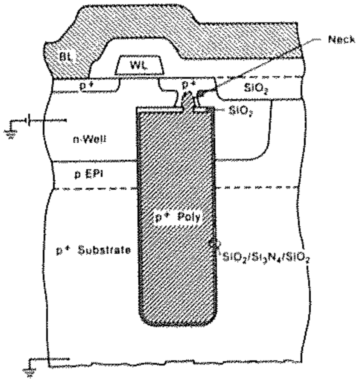


Fig. 8-24 Cross section of the SEOT cell.[71] (© 1988 IEEE).

shown in Fig. 8-26.[73] The lower electrode of the stacked capacitor is in contact with the drain of the access transistor, and the bit line runs over the top of the stacked capacitor. Although some stacked capacitor (STC) type cells have been used to fabricate

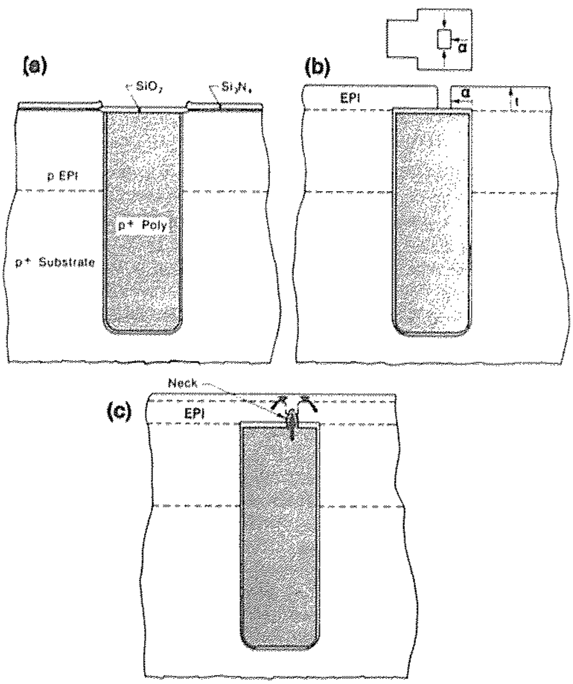


Fig. 8-25 Key processing steps of the SEOT technology.[71] (© 1988 IEEE).

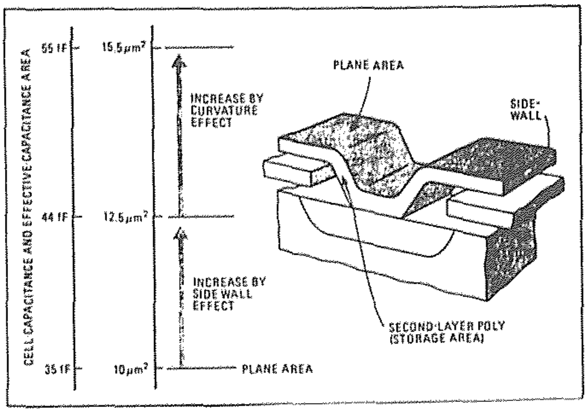


Fig. 8-26 Stacked capacitor (STC) cell structure.

256-kbit DRAMs, the minimum cell size required by conventional stacked-capacitor cells for adequate charge-storage capacity is too large for these cells to be used in larger DRAMs.

For STC cells to be made feasible for 1-Mbit DRAMs and beyond, an insulator with a larger dielectric constant than that of SiO_2 must be used, or novel cell structures must be developed. Although research is continuing into the use of such higher dielectric constant materials as tantalum pentoxide, currently acceptable insulators do not make *conventional* STC cells viable for 1-Mbit and larger DRAMs. However, several novel STC cell designs have been reported.

In the first of these, the contact hole used to connect the lower capacitor electrode to the drain of the access transistor is not filled up with polysilicon, as is the case in conventional STC cells (Figs. 8-27a and b).[74] Such filling up of the contact hole reduces the effective area of the capacitor, especially for holes with small dimensions. Instead, the contact hole is opened *after* the lower capacitor-electrode polysilicon film is deposited, and only a thin second polysilicon film is subsequently deposited into the hole (Fig. 8-27c). Good contact between the second film and the substrate is established by means of an ion-beam mixing implantation step following the thin-poly deposition (see section 3.4.2.5). This process produces a cell capacitance that is ~1.3 times as large as that obtained with a conventional STC. (Figure 8-27d shows an SEM photograph of the new cell.) A cell capacitance of ~35 fF with a cell size of 8.8 μm^2 is achieved with this design.

As shown in Fig. 8-27d, by trenching into the Si substrate it is possible to produce even more capacitance for a cell size of the same area (or a comparable capacitance for a smaller cell size). The trenched STC cell of Fig. 8-27d is predicted to be able to provide 30 fF of capacitance in a cell area of 1.3 μm^2, which would make it suitable for 64-Mbit DRAMs. A trench depth of ~1 μm is needed to achieve this capacitance value.

In the second novel STC cell, the bit lines are formed before the stacked capacitor is fabricated. In addition, the capacitor is laid out on a diagonal with respect to the bit and

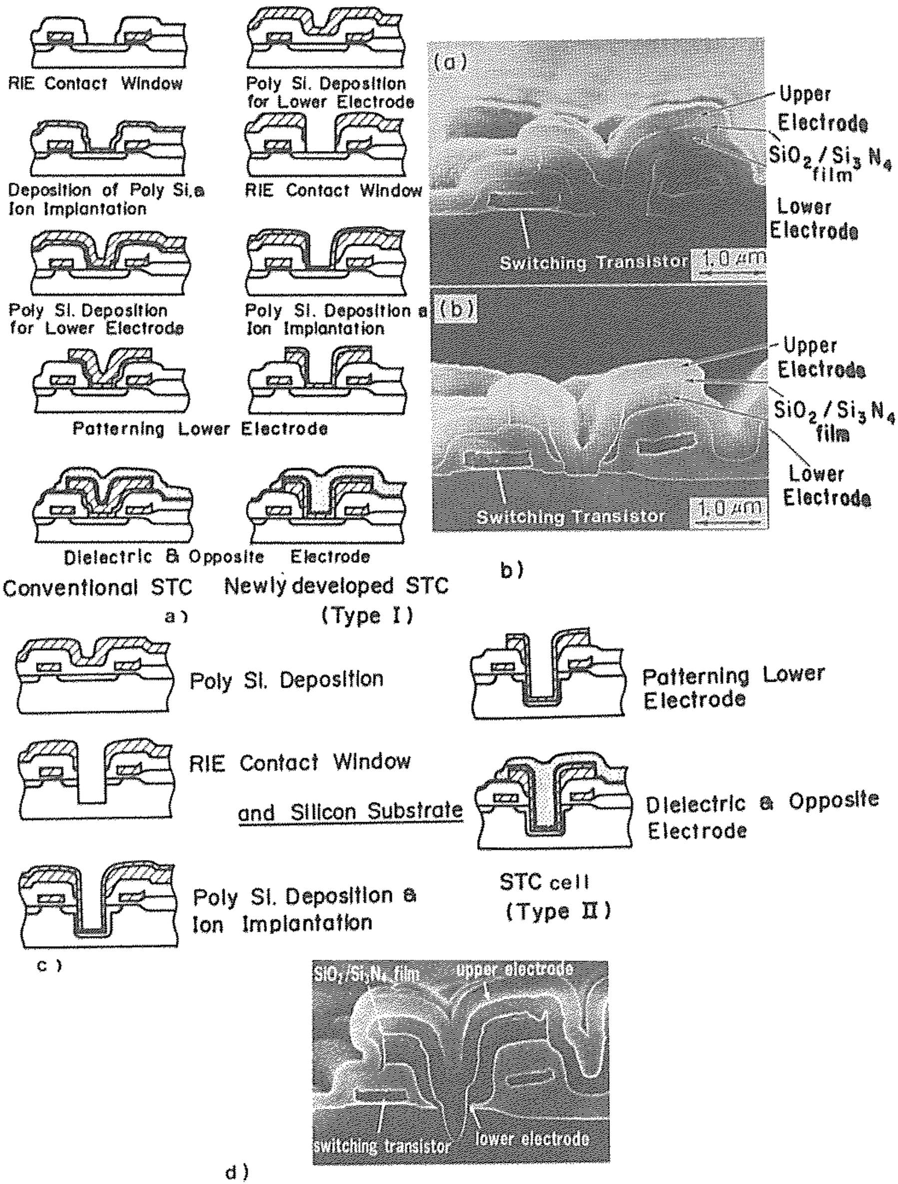


Fig. 8-27 (a) Process sequence of conventional and newly-developed type-I STC cells. (b) Cross-sectional SEM pictures of these two cells. (c) Process sequence of an advanced STC cell (type II). (d) SEM picture of this cell.[74] (© 1988 IEEE).

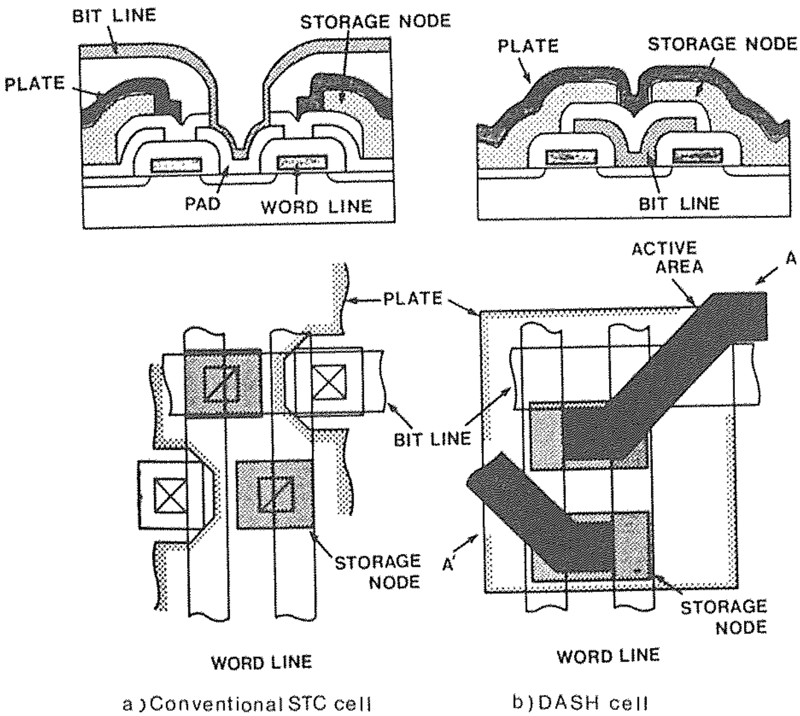


Fig. 8-28 Cross section and layout of: (a) conventional STC cell. (b) DASH cell.[75] (© 1988 IEEE).

word lines, which increases the cell area without increasing its size. This cell is thus called a *diagonal active-stacked-capacitor cell with a highly packed storage node (DASH)*.[75] Figure 8-28a shows the cross section of the DASH cell, and Fig. 8-28b compares the layouts of this type of cell with that of a conventional STC cell. The DASH cell yields a 35-fF capacitance for a cell size of 3.4 μm^2, and hence could be used in 16-Mbit DRAMs.

In the third novel STC cell, a unique fin structure is used to fabricate a capacitor with a high capacitance in a small area. (Figure 8-29a shows the fabrication process sequence.)[76] This structure would allow cells with two fins to be used in a 16-Mbit DRAM. When the bit lines are formed prior to formation of the fin structure (Fig. 8-29b), a cell structure can be obtained that has sufficiently high capacitance in a small enough area to allow fabrication of a 64-Mbit DRAM.

A fourth novel STC cell, called a *spread stacked capacitor* (SSC) cell,[122] the storage electrode is expanded into the neighboring 2nd memory cell area (Fig. 8-29c and d), and the storage electrode of the 2nd memory cell is expanded to the 1st memory-cell area.

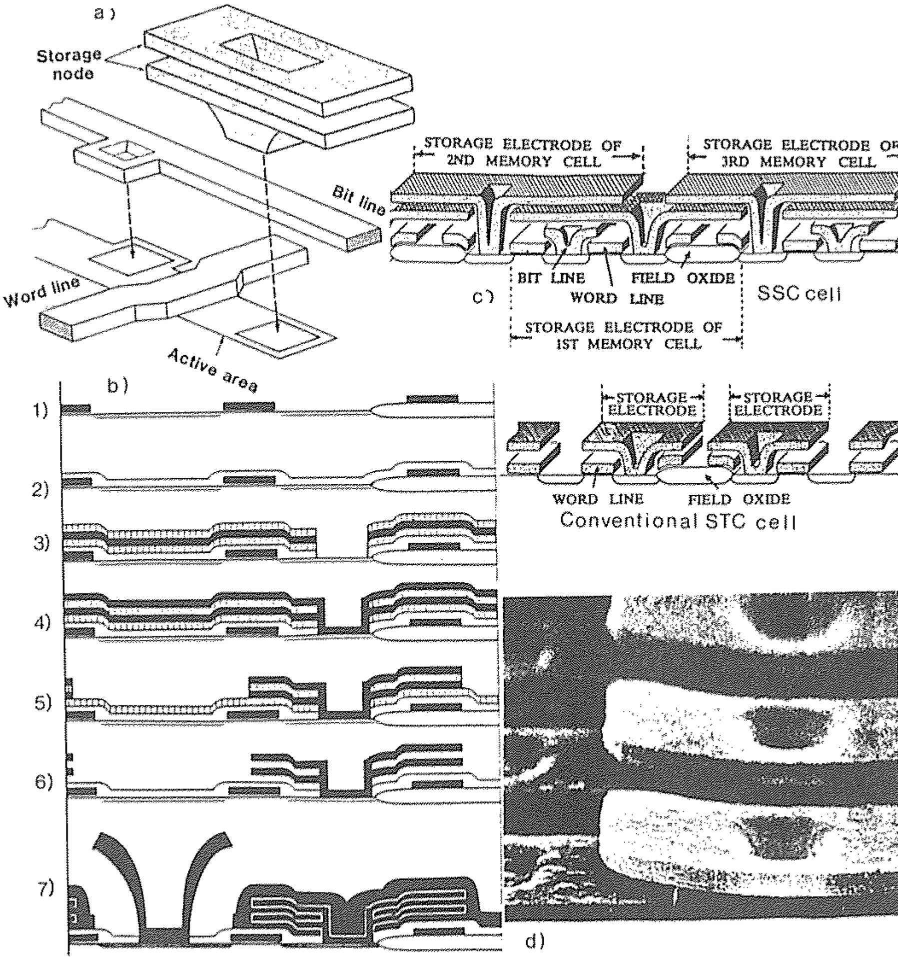


Fig. 8-29 (a) Schematic view of fin-STC structure. (b) Schematic view of fin structure and fabrication sequence.[76] (© 1988 IEEE). (c) Schematic cross sections of the SSC cell and a conventional-STC cell. (d) SEM photo of a spread-stacked capacitor.[122] (© 1989 IEEE).

This allows the storage capacitance of the SSC cell to be ~1.8x as great as that of the conventional STC cell, making a possible candidate for 64-Mbit DRAMs.

8.3.5 Soft-Error Failures in DRAMs

When a DRAM is tested, each cell is operated to verify that it functions correctly. A *hard fail*, or *hard error*, indicates that a particular array location repeatedly fails to output correct data values previously written into the location. Although there are a number of potential sources of hard fails, by and large they are caused by random physical defects. As noted earlier, various layout design approaches and on-chip error-detection and correction circuits (that replace failed bits with spares), have been implemented to eliminate faulty cells from a memory, allowing some defectively manufactured chips to be salvaged.

Soft errors, which are single-nonrecurring read errors on single bits of a memory array, are also a significant problem in DRAMs. A soft error is not a permanent error, in the sense that the cause is not a process defect. A write (then read) cycle in an array location that was previously in error carries no greater or lesser probability of error again than does a cycle in any other array location.

Although soft errors can be caused by such circuit related problems as supply-voltage noise, inadequate noise margins, and sense-amplifier imbalance, there is one specific *physical* failure mode that will cause soft errors, even when all circuit-related failure modes are eliminated. The cause of this failure mode of was identified in 1979 by May and Woods[77] as the alpha particles originating from the decay of uranium and thorium atoms. These radioactive atoms are naturally occurring trace impurities in the materials used to make IC packages, and the alpha particles they emit have energies in the 8-9 MeV range.

Since a bit of information is stored on a cell by the presence or absence of charge in the potential well of the storage capacitor, the number of electrons that distinguishes an empty well from a full one (and hence which differentiates between a logical *one* and a *zero*) is known as the *critical-upset charge*. The generally quoted value for this charge is 45-50 fC (2.5×10^5 electrons),[78] which is about 25% of Q_s (see section 8.3.1.8).

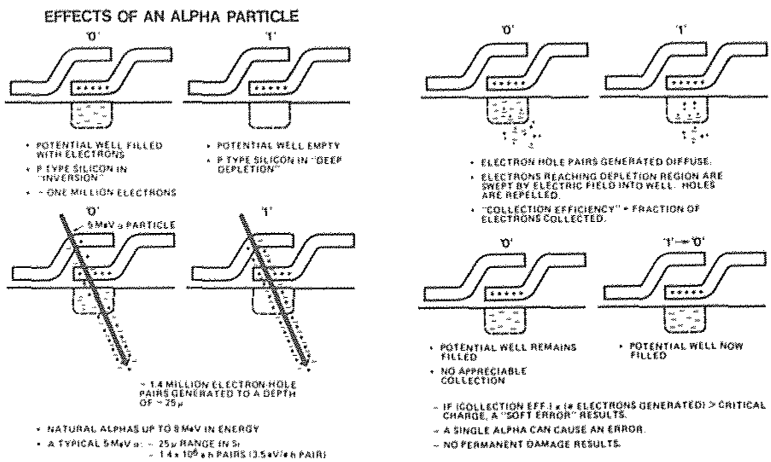


Fig. 8-30 Collection process of free carriers generated by an incident alpha particle. This shows that the cell is sensitive to the alpha particle hit when the storage node is depleted, which is when the cell is storing a one for the double-poly cell shown.[77] (© 1979 IEEE).

If the temperature is raised, or if light is incident on the memory, the generation rate of electron-hole pairs in the substrate can be significantly increased. If enough electrons are produced that an empty well fills up in the interval between refresh pulses, a soft error will occur. Electron-hole pairs are also produced by ionizing radiation incident on a memory chip – specifically, when energetic alpha particles strike a semiconductor substrate (Fig. 8-30).

The electron-hole pairs produced as an alpha particle first passes through the chip's passivation layers are not collected by the empty well, implying that some fraction of the initial alpha-particle energy is lost before it reaches the substrate. Even if only half of the initial energy is retained, there will still be enough energy (4 MeV) to produce roughly 10^6 electron-hole pairs along a trajectory ~25 μm in length.[79] Any of the electrons generated within the potential well will be swept into the storage node by the depletion region electric field. Those electrons generated in the bulk that diffuse to the edge of the well will likewise be collected by the storage node. The remainder will recombine in the bulk. If a large enough fraction of the electrons generated by an alpha-particle strike is collected by the empty potential well, a soft error can result. The entire collection process occurs in a matter of microseconds.

8.3.5.1 Techniques Used to Reduce Soft-Error Rates in DRAMs.

The *soft-error rate* (SER), expressed as the number of errors per hour, describes the degree of susceptibility to soft-error phenomena. When the storage nodes on silicon devices were on the order of 25 μm in length, it was possible (but not likely) for all of the electrons from one strike to be collected on one node. However, since Q_S for these structures was also larger (e.g., 2×10^6 electrons), the SER was so low that the problem was not noticed. As devices decreased in size,SERs grew significantly larger.

Soft errors in DRAMs were first reported in 16-kbit DRAMs (1 soft-failure per 1000 hours of operation was typical), and they became an appreciable concern in the design and packaging of 64-kbit DRAMs and larger. When new DRAMs are designed, an attempt is made to ensure that the SER is comparable to the hard-error rate.

It was found that as long as a DRAM cell can store more than 6×10^5 electrons, the SER can be made comparable to the hard-error rate. For storage devices in DRAMs of up to 16 kbits, this was accomplished by scaling the oxide thickness to maintain adequate capacitance in the cell. The reliability trade-off in this approach was an increased hard failure rate due to increased defects in the thinner gate oxide. The Hi-C cell was invented as a means of increasing the storage capacity without increasing the capacitor size. An added benefit of this cell is that the doping gradient of the deep *p*-implant produces a potential energy barrier to the diffusing electrons, preventing some of them from reaching the depleted well of the storage capacitor. Hence, the SER of memories built with such cells is decreased. An additional advance was the implementation of the insulator of the storage capacitor with a dual dielectric film (SiO_2 and Si_3N_4), which increased the cell's capacitance (see section 8.3.1.5).

The same SER-reducing phenomenon is exploited when DRAMs are built in CMOS technology. That is, the entire array of NMOS DRAM cells is built in a *p*-well. The well-substrate junction acts as a reflecting barrier for diffusing minority carriers created outside of the well, preventing most of them from reaching the unfilled storage nodes. Putting the memory on a heavily doped substrate with a lightly doped epitaxial layer on the surface is also beneficial. Since minority carriers generated in the heavily doped substrate have much shorter lifetimes than those in the lightly doped regions, they are prevented from reaching the epitaxial region in which the cells are built (see also section 6.8.4.2).

The use of trench-capacitor structures can also reduce the SER of DRAMs, increasing the number of stored electrons per cell without requiring the lateral device or chip to be made larger. A trench cell in which the storage plate is inside the trench provides better SER than one in which the storage plate is on the outer walls. (If there is little or no isolation between closely spaced trench cells and the track of an alpha particle intersects two of the cells, considerable charge can be transferred. The charge flow along the alpha track can cause transient forward-biasing of the cell junctions. This bipolar-like effect becomes more dominant as the cell-to-cell spacing decreases.)[105]

A final method used to reduce SER is to apply a thick coating of a radioactive-contaminant-free polymer on top of the IC. For example, alpha particles with energies of up to 8 MeV are completely absorbed by a 50-μm-thick layer of polyimide. In addition, packaging materials are now manufactured with much lower concentrations of radioactive impurities. Purification of the materials used in wafer fabrication (particularly metallization) is also being pursued, and trace levels of radioactive impurities are now low enough that they can be eliminated them as a significant source of alpha particles.

Even with thick polymer coatings, chips are still vulnerable to strikes by high-energy cosmic rays (in the form of high-atomic-number nuclei). A small fraction of such very energetic particles can penetrate the earth's atmosphere, the package, and the polymer coating, entering the silicon. Under some conditions, these nuclei can also produce soft-errors through the generation of electron-hole pairs. At present, it is estimated that cosmic-ray events produce an SER about an order of magnitude lower than that due to alpha particles from currently available "clean" packaging materials.

8.3.6 The DRAM as a Technology Driver

The DRAM has also been used as a *technology driver* over a large part of its life, since it makes a good test vehicle for advancing silicon integrated-circuit process technology. The regular, repetitive architecture of the DRAM chip requires the least amount of engineering design time to create a circuit with hundreds of thousands, or millions, of devices that can be produced in full-scale production in the factory. Hence, a new MOS technology can be fully tested in the shortest time using the DRAM as the production test vehicle. After the flaws in the process and manufacturing procedures have been ironed out using the technology driver, the process can be transferred to the manufacture of other, more design-intensive circuits.

In the mid-1980s, U.S. semiconductor manufacturers increasingly turned to high-density MOS logic arrays as a replacement technology-driver circuit, since most of them had been forced to abandon the the DRAM market after the price erosion of the 64-kbit

DRAM. The American preference for the logic test chip was also based on the fact that considerable engineering and manufacturing expertise from the production of such dense logic chips as the 80386 and 68000 MPU families. By the late 1980s, however, the world wide manufacturing trend had swung to CMOS SRAMs, since these circuits are easier to design than DRAM circuits (which require intricate clock circuitry). In addition, the 4-Mbit and larger DRAMs use three-dimensional capacitor structures that are unique to the DRAM, and their yield statistics may thus not be as valuable for learning how to manufacture other circuits that do not utilize such structures.

In 1989, however, Osamu of Toshiba presented a case for why DRAMs could still serve as excellent technology drivers.[80] First, the repetitive structure allows failure categories to be easily identified so that efforts can be undertaken to correct processes that lead to the failures. Second, the factors that impact the RAM yield can be analyzed more completely and in a much shorter period of time than those in logic devices (i.e., the tasks of test-vector generation and design for testability are much less onerous for RAMs than for logic circuits). Finally, the volume production of DRAMs is so great that memory-fabrication processing experience is acquired much more rapidly. These lessons can be transferred to other volume IC production lines. For example, DRAM production can reach 10^6 parts per year at a single manufacturer (with $2x10^6$ devices per chip in the 1-Mbit DRAM), versus ~50K microprocessor parts per year (with $1x10^6$ devices per typical 32-bit microprocessor).

8.4 MASKED READ-ONLY MEMORIES (ROMS)

Masked (or *mask-programmed*) ROMs are nonvolatile memories into which informa-ion is permanently stored through the use of custom masks during fabrication. Users thus must provide the ROM manufacturer with the desired bit pattern of the memory. Subsequent changes of stored data are impossible, and only *read* operations can be performed. Since only a single customized mask is required to personalize these ROMs for a specific application, however, many designs can be economically implemented.

Such memory circuits have been implemented in bipolar, NMOS, and CMOS technologies. As of 1989, 4-Mbit CMOS ROMs were the largest available, but parts containing as many as 32 Mbits are envisioned. The fastest high-density MOS-based ROMs have access times of about 80 ns,[81] while less dense (256-kbit) ROMs are faster (50 ns). The faster ROMs are most often used to simplify the interface to micro-processors by eliminating wait states, thereby permitting more rapid program execution. Bipolar ROMs are even faster (e.g., 10 ns access time for a 1-kbit ECL ROM), but they are also much less dense.

In addition to interfacing with microprocessors, ROMs have been used for a variety of applications, including these:

- *Look-Up Tables.* These are used for mathematical calculations in which evaluations of square roots and of trigonometric functions, logarithmic functions, and exponential functions are needed. The procedure would usually be much more time consuming if software subroutines were used to calculate the series expansions of particular functions. Other look-up applications include spell checking and dictionary servicing.

- *Character Generators.* All digital systems rely on the display of alphanumeric characters of such input and output devices as CRTs and dot-matrix printers. In most of these applications, the patterns of pixels, segments, or dots used to display the set of characters are stored in a ROM. In Japan, large (1-Mbit and larger) ROMs are used as character generators for the complex *kanji* characters.

- *Microcontrol Store.* In microprogrammed digital systems, the execution of an encoded macroinstruction involves the generation of a sequence of signal vectors for gating or control purposes. ROMs are often chosen for such applications because their higher packing density and simpler fabrication procedure make them less costly than static or dynamic RAMs. In addition, ROMs are more stable than SRAMs and DRAMs. Parts of computer programs that are completely debugged and that do not need to be rewritten during computation are thus often stored in ROMs.

8.4.1 Masked-ROM Implementation

Each bit of information in a ROM is stored by the presence or absence of a data path from the word (access) to a bit (sense) line. The data path is eliminated simply by ensuring that no circuit element joins a word and bit line. Thus, when the word line of a ROM is activated, the presence of a signal on the bit line will mean that a *1* is stored, whereas the absence of a signal will indicate that the bit location is storing a *0*. As shown in Fig. 8-31 (which uses an NMOS array as the example), there can be two basic forms of the ROM, with implementation by either the NOR function (Fig. 8-31a), or the NAND function (Fig. 8-31b). Note that programming of masked ROMs in bipolar technology is done by selectively omitting a contact, at the contact mask in the emitter-follower or Schottky-diode ROM arrays shown in Fig. 8-31d.

Although the speed of the ROM depends on the details of the MOS fabrication process, NOR arrays usually have faster access times; in addition, the stored bit pattern can be set by the metal-interconnection layer. Therefore, unprogrammed NOR ROMs can be manufactured up until the metal mask step and can then be stored in inventory. Such almost-completed wafers can be quickly completed (programmed) by using a custom mask that patterns the metal layer.

The NAND-type ROMs, on the other hand, have a longer access time, and they must be programmed through the implantation of dopants into the channel of selected transistors wherever stored zeros are desired (Fig. 8-31c). Such a step must be performed earlier in the manufacturing sequence, which increases the TAT. The advantage of NAND ROMs over the NOR type is that they have a considerably higher density when fabricated using the same process and design rules.

The access time of ROMs is limited by the resistance and capacitance of the word and bit lines, as well as by the currents available to drive these lines. Because of their higher density (i.e., 1-Mbit CMOS ROMs were introduced in 1983), ROMs were the

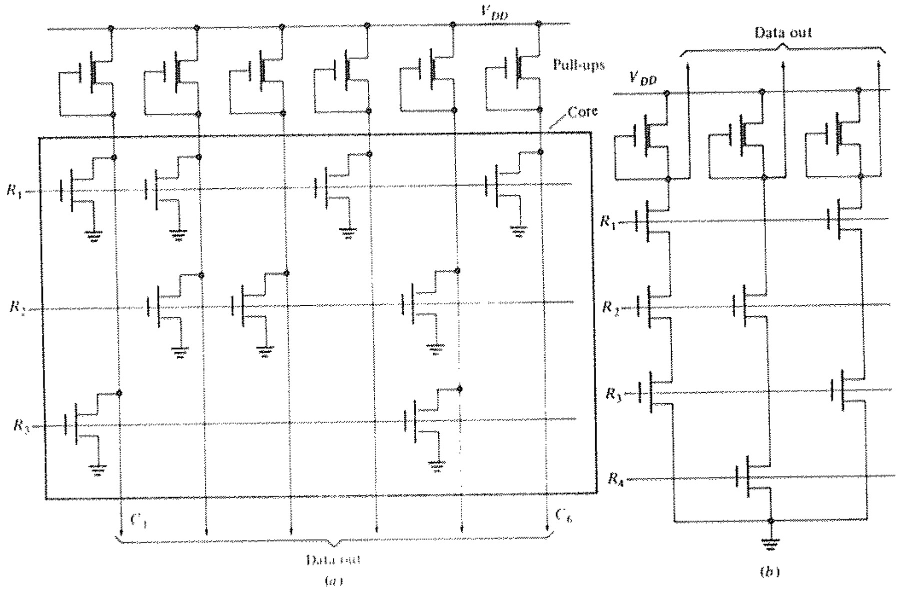


Fig. 8-31 (a) MOS ROM NOR array, (b) MOS ROM NAND array.[1]

first semiconductor memories in which on-chip error-correction circuitry (ECC) was implemented.

8.5 PROGRAMMABLE ROMS (PROMS)

If only a small quantity of ROM circuits is needed for a specific application, custom fabrication of even a single mask layer may be too expensive and/or time consuming. In such cases, it is faster and cheaper for users to program each ROM chip individually. ROMs with such capabilities are referred to as *user* or *field-programmable* memories. Many types of these have been developed.

In this section we describe *programmable read-only memories* (PROMs), a type of ROM into which information can *be programmed only once* and then cannot be erased. Subsequent sections describe ROMs that allow data to be erased after entry.

In PROMs, a data path exists between *every* word and bit line at the completion of chip manufacture (corresponding to a stored *1* in every data position). Storage cells are selectively altered to store a *0* following manufacture by electrically *blowing open* the appropriate word-to-bit connection paths. Since the write operation is destructive, once a *0* has been programmed into a bit location it cannot be erased back to a *1* in a PROM. PROMs were originally implemented in bipolar technology, although MOS PROMs have recently become available as well.

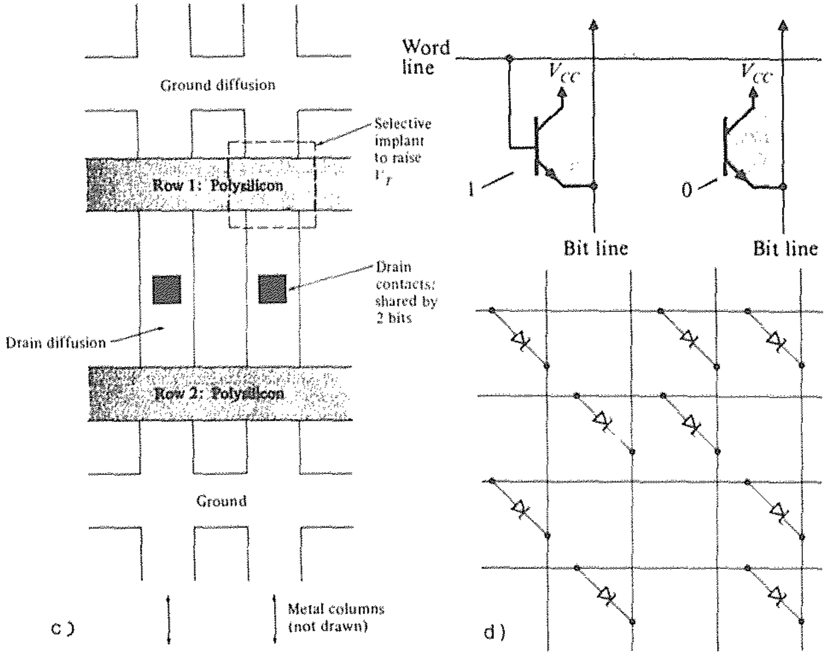


Fig. 8-31 (c) MOS NAND ROM - simplified layout, (d) BJT ROM cells.[1] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

In bipolar PROMs, two techniques are used to eradicate the word-to-bit-line paths in desired bit locations. The first involves the use of a bipolar transistor in series with a *fuse*, as shown in Fig. 8-32. In such *emitter-follower* bipolar PROMs, a small fusible link is placed in series with each emitter. Early links were implemented with nichrome thin films, but these exhibited a *growback* problem (in which disconnected fuses became reconnected after some time). PROM fuses are now generally made of polysilicon. Fuse-link PROMs are designed to operate with a 5-V supply for reading data, but higher voltages (10-15 V) are needed to produce the 10-30 mA required to blow open the fuses. Such large voltages may be supplied by off-chip programming devices or by special electronic circuits available on the chip.

The second technique makes use of a *pn* diode that is short-circuited by an avalanching pulse. An example of a 64-kbit bipolar PROM using *pn* diode cells with an access time of 50 ns is described in reference 82.

MOS PROMs have also been introduced, and in 1985 they became available in 1-Mbit size. Such components are actually MOS erasable and programmable read-only memories (EPROMs) housed in inexpensive plastic packages (rather than in the costly, quartz-windowed ceramic packages needed by *erasable* MOS PROMs). Without a quartz window these MOS PROMs can be field-programmed only once. As a result, they are commonly known as *one-time-programmable* (OTP) ROMs. The advantage of these over bipolar PROMs is that they can be fabricated in much higher densities. High-density CMOS OTP ROMs are now being built with access times close to those of bipolar PROMs, but with more bits per chip and much lower power dissipation. For example, a 256-kbit CMOS OTP ROM with an access times of 50 ns has recently been introduced; this approaches the access time (~40 ns) of large [64-kbit] bipolar PROMs.[87] In addition, OTP ROMs are no longer much more expensive than ROMs, and hence they are also expected to increasingly replace masked ROMs.

8.6 ERASABLE PROGRAMMABLE READ-ONLY MEMORIES (EPROMS)

Erasable PROMs depend on the long-term retention of electronic charge as the information-storage mechanism. The charge is stored on a *floating polysilicon gate* of an MOS device (the term *floating* refers to the fact that no electrical connection exists to this gate). The charge is transferred from the silicon substrate through an insulator.

Each of the various mechanisms implemented to transfer (and remove) charge from the floating gate has been the basis of a different erasable-PROM device type. This section describes the so-called *electrically programmable ROM* (EPROM), which also requires that the device be irradiated with ultraviolet (UV) light for removing (or *erasing*) the stored charge from the floating gate.

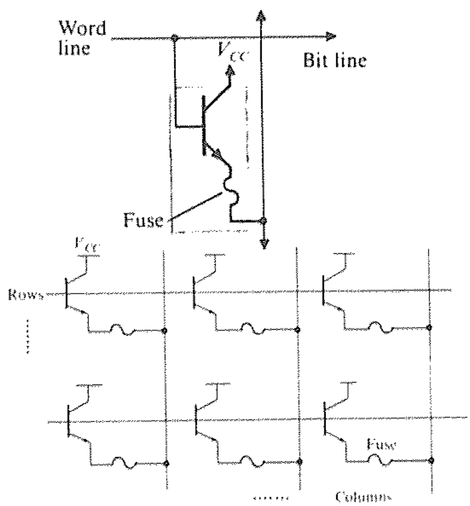


Fig. 8-32 (a) Emitter-follower bipolar PROM.[1] From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

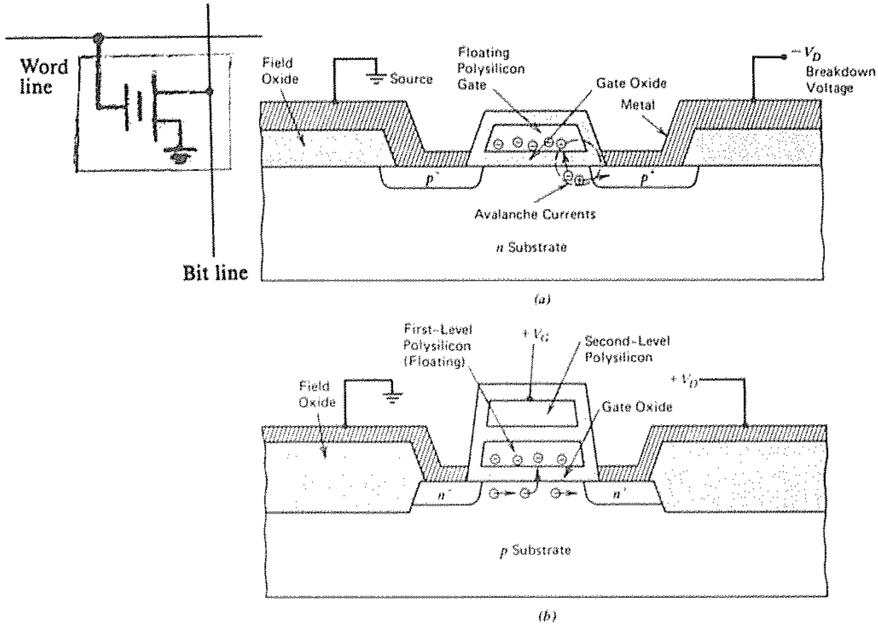


Fig. 8-33 (a) Circuit schematic and cross section showing the mechanism of charge injection into the gate by avalanche in a FAMOS memory element. (b) A FAMOS element made with two layers of polysilicon and suitable for n-channel MOS applications. From R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed. Copyright 1986, John Wiley & Sons. Reprinted with permission.

Traditionally, such EPROMs have been used as prototyping vehicles to ensure that no glitches remained in the code. Once the programs were finalized, the code was usually fixed into ROM components. However, the cost of EPROMs is shrinking with advances in technology, and as a result, their use is growing at the expense of ROMs. Another factor in favor of EPROMs is their faster turn around time (which also plays a role in the choice of technology used to implement masked ROMs).

The charge-transfer mechanism is based on the injection of hot electrons into the floating polysilicon gate, which is completely encapsulated by SiO_2. The original EPROM devices were fabricated in PMOS technology and consisted simply of a MOSFET with a floating gate (Fig. 8-33a). If a sufficiently high reverse-bias voltage is applied to the drain, the drain-substrate pn junction will experience avalanche breakdown, causing hot electrons to be generated. Some of these will have enough energy to pass over the oxide potential-energy barrier and charge the floating gate (see section 5.6.2). These EPROM devices were thus called ***Floating-gate, Avalanche-injection MOS transistors*** (FAMOS).

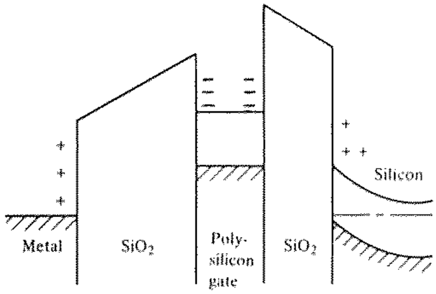


Fig. 8-34 Energy band diagram of a FAMOS device with charge stored in the silicon gate. From E. S. Yang, *Microelectronic Devices*, Copyright 1988, McGraw-Hill Book Company. Reprinted with permission.

Once electrons are transferred to the gate, they are trapped there, as illustrated by the energy-band diagram shown in Fig. 8-34. Since the potential-energy barrier at the oxide-silicon interface is greater than 3 eV, the rate of spontaneous emission of electrons from the oxide over this barrier is negligibly small. The electronic charge on the floating gate can thus be retained for many years.

If the floating gate is charged with a sufficient number of electrons, inversion of the channel under the gate will occur. A conducting channel then forms between the source and the drain, exactly as would occur if an external gate voltage were applied. The presence of a *1* or *0* in each bit location is therefore determined by the presence or absence of a conducting channel in a programmed device.

Subsequent advances in process technology (Fig. 8-33b) made it possible to implement EPROMs with 5 V, *n*-channel devices.[83,84] In such EPROMs the cells can also be laid out in NOR or NAND arrays; we will use the NOR array configuration to describe the operation of these newer cells.

Two layers of polysilicon are used to form a double gate in the transistor, as shown in Fig. 8-33b. Gate #1 is the floating gate and is placed under Gate #2. Cell selection is controlled by Gate #2, which therefore plays the role of the single gate in conventional MOS transistors. Initially, Gate #1 is uncharged; thus, if the drain, source, and Gate #2 of the transistor are grounded, Gate #1 will also be at 0 V. If a voltage (V_2) is subsequently applied to Gate #2, the voltage on Gate #1 (V_1) will be given by:

$$V_1 = [C_2 / (C_1 + C_2)]\ V_2 \tag{8-3}$$

because the two gates represent a capacitive divider as shown in Fig. 8-35. From the electrical perspective of Gate #2, the transistor appears to have a larger V_T. In order to turn on this transistor, a larger gate voltage must be applied to Gate #2 (typically somewhat more than twice the normal V_T). For example, if a conventional NMOS

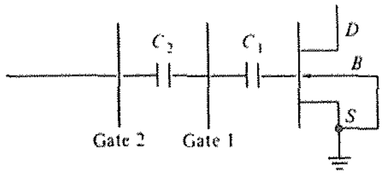


Fig. 8-35 Equivalent capacitive divider of an EPROM structure.

device with a $V_T = 1$ V was fabricated with the double gate of Fig. 8-33b, a voltage of 2 V would have to be applied to Gate #2 to turn it *ON*; a voltage of 5 V for reading the cell would also cause it to turn *ON* (Fig. 8-36). Such a turned-on device would cause a *positive logic stored zero* to appear at the output of the bit line if the device was used in a NOR array. As a result, the programming of the EPROM begins by discharging all of the floating gates through exposure to UV radiation, so that every cell initially stores a *0*. A *1* is then selectively written into the desired cells.

For a *1* to be written into a cell, both Gate #2 and the drain are raised to about 12 V (for a few hundred microseconds), while the source and substrate are kept grounded (early EPROMs required 30 V programming voltages for several milliseconds). Hot electrons are created near the drain and are attracted to the floating gate (which, due to capacitive coupling, has a more positive potential than the drain). Some fraction of the electrons will traverse the oxide and charge the floating gate. When the voltages on Gate #2 and the drain are returned to zero, these charges remain trapped on Gate #1. The electrons trapped on Gate #1 cause its potential to be at about -5 V. Therefore, if a signal of only 5 V is applied to Gate #2 when the EPROM is being read, no channel will form in the transistor. Under this circumstance, a *1* is stored in the cell. The electron- trapping process is self-limiting, because once electrons are stored on the floating gate they begin to inhibit further electron injection.

In order for the cells to be erased, the stored charge must be removed from the floating gate. This is accomplished by flood exposure of the EPROM with strong ultra-violet light for approximately 20 minutes. The UV light creates electron-hole pairs in the SiO_2, providing a discharge path for the charged floating gate.

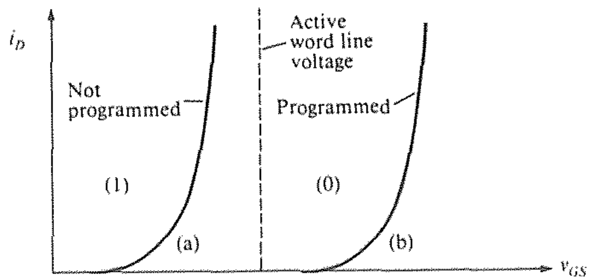


Fig. 8-36 Transfer characteristic of a floating-gate transistor.

One of the advantages of EPROMs is that the cells consist of only one transistor, allowing them to be fabricated with high densities (e.g., a 4-Mbit CMOS EPROM with an access time of 120 ns and 0.8-μm channel-length transistors has been reported).[85] In addition, they cost less to manufacture than electrically erasable PROMs (EEPROMs – see the next section).

A disadvantage of EPROMs is that they require UV light for erasing and must therefore be packaged in an expensive ceramic package with a UV-transparent quartz window. In addition, they must be removed from the circuit board and put into a special UV eraser. (Note that since sunlight and fluorescent lamps contain some UV, one week of sunlight or three years of room-level fluorescent lighting are likely to erase some of the cells. Therefore, except during erasure, the window should be covered at all times with an opaque label.) Another disadvantage is that the high voltage needed to program the EPROM is generally not available on the integrated circuit, so a special programming setup must also be provided. This limitation, combined with the fact that EPROM programming takes a relatively long time, means that these cells are used primarily for reading information and are only occasionally rewritten. (Note, however, that the programming time is decreasing dramatically. In the first 64-kbit EPROMs, it took about 50 ms to program each byte, adding up to almost seven minutes for the entire chip; in the 4-Mbit EPROM,[85] the program time has been reduced to 10 μs/byte, so that the entire chip can be programmed in only five seconds!)

OTP ROMs compete with high-density masked ROMs because they offer the benefit of a significantly shorter TAT (albeit at a somewhat higher cost). OTP ROMs are also less expensive than bipolar PROMs, and offer a PROM capability with much higher density. While bipolar PROMs are generally faster, a three-transistor EPROM cell was recently reported that would allow CMOS EPROMs to be built with the same speed and density as bipolar PROMs, but with much lower power dissipation and 100% testability.[86] Another one-transistor cell, split-gate 256-kbit CMOS EPROM with an access time of 50 ns has also been reported.[87]

Some of the relevant process and circuit-design enhancements used in fabricating the large CMOS EPROMs include the following:[88]

- Use of thin (20 nm) reliable interpoly dielectric materials, often consisting of composite films of $SiO_2/Si_3N_4/SiO_2$, for increased capacitive coupling between Gates #1 and #2.
- Self-aligned contacts to the control gate (as well as a self-aligned floating gate) to achieve the 3.1 x 2.9 μm^2 small cell size.
- Use of a low-resistance polycide gate for the word line to achieve high speed.
- Reduction of the programming voltage to 10.5 V, along with a reduction of the programming time to ~10 μs/byte.
- On-chip test circuits.

A novel self-aligned planar-array EPROM cell has also been proposed.[89,90] This cell appears to make possible the fabrication of 4-Mbit EPROMs with 1-μm design rules because it uses buried n^+ bit lines that are self-aligned to the FAMOS transistor.

8.7 ELECTRICALLY ERASABLE PROMS (EEPROMS)

In some applications it is desirable to erase the contents of a ROM electrically, rather than to use a UV light source. In other circumstances it is useful to be able to change one byte at a time, without having to erase the entire IC. A variety of *electrically erasable PROMs* have been developed to serve these applications. Such EEPROMs are the most sophisticated of the ROM families in terms of the physical operating principles and process complexity. For example, EEPROMs must be fabricated with unique tunnel oxides, as well as with high-voltage transistors (for programming and erasing the devices).

Three technologies have been developed for EEPROM fabrication: (1) MNOS transistors; (2) **Fl**oating-gate **T**unnel **Ox**ide (FLOTOX) MOS transistors; and (3) textured-polysilicon floating-gate MOS transistors. Although MNOS transistor-based devices were among the first EEPROMs to be commercially manufactured, their technology limitations have made them less widely adopted than the others. Therefore, we will devote most of our attention to FLOTOX and textured-poly EEPROMs.

8.7.1 MNOS-Based EEPROMs

The MNOS EEPROM cell consists of a single MOS-like transistor that employs a composite gate-dielectric layer (Si_3N_4, ~50 nm thick, on top of a very thin [~2 nm thick] SiO_2 layer), as shown in Fig. 8-37. (See ref. 118 for more details on MNOS devices.) Unlike in floating-gate MOS devices, the charge is stored in discrete traps in the nitride bulk. The charge transfers from the substrate to the nitride traps (and back, during erasure) by tunneling through the thin oxide layer. Programming is accomplished by applying a high voltage to the top gate; erasing is done by grounding the top gate and raising the well to a high potential. MNOS transistors are built within wells (akin

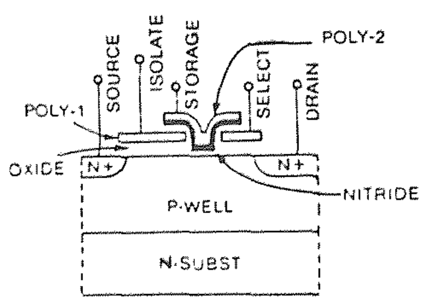


Fig. 8-37. Cross-sectional structure of an MNOS memory cell.[107] (© 1983 IEEE).

to those used in CMOS) so that their channel potentials can be controlled.

The manufacturing process of MNOS transistors involves the following modifications to the standard single polysilicon-gate MOS technology: thin oxide growth, nitride deposition, and post-nitride temperature cycles. Mastering the processes used to grow the ultra-thin oxide and deposit the nitride and to control their quality is a challenging task. Furthermore, while the basic transistor is very small and highly scalable, each cell of the memory array requires a select transistor. This requirement, coupled with the need to fabricate wells, produces a relatively large effective cell size. Finally, the charge stored on the nitride traps continually leaks away through the thin oxide by means of tunneling, even when no erase voltage is applied. The charge loss is thus time dependent, making charge retention the main reliability concern with MNOS devices. (With MNOS structures, as the switching speed is increased, the ability to retain stored charge is decreased. Thus, devices with a retention time of tens of years can be fabricated if a slow switching speed can be tolerated.)

Nevertheless, MNOS exhibits higher tolerance to ionizing radiation than do either of the other EEPROM technologies. Thus, MNOS EEPROMs currently find their main use in low-density military applications that need radiation-hardened EEPROMs; this appears to be the niche to which MNOS EEPROMs will be relegated in the future.[91]

8.7.2 FLOTOX EEPROMs

The *floating-gate tunneling oxide (FLOTOX)* transistor, shown in Fig. 8-38a, consists of an MOS transistor with two polysilicon gates. A thin (8-12 nm) gate oxide (or oxynitride) region is formed near the drain. The lower polysilicon layer is the floating-gate while the other is the control gate. The remainder of the floating-gate oxide is typically 50 nm thick, and the interpoly oxide is ~50 nm thick. Programming of this transistor is done by causing electrons to be transferred from the substrate to the floating

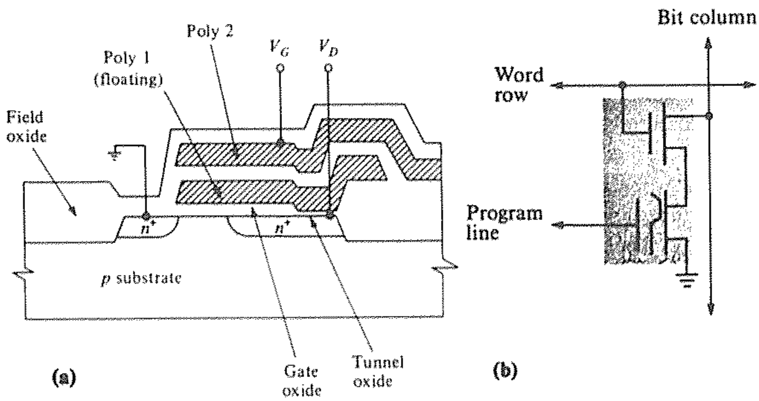


Fig. 8-38 (a) Cell structure of a Flotox transistor structure, (b) Connection in an EEPROM.[91] (© 1986 IEEE).

gate through the thin oxide layer by means of Fowler-Nordheim tunneling.[92]

The control-gate voltage is raised to a sufficiently high value so that tunneling ensues (e.g., 12 V in modern FLOTOX EEPROMs). As electronic charge builds up on the floating gate, the electric field is reduced, which decreases the electron flow. Since the tunneling process is reversible, the floating gate can be erased by grounding the control gate and raising the drain voltage, indicating that tunneling is used both to *program* and *erase* the FLOTOX transistor. Programming and erase times are on the order of 9 ms. Electron transfer by Fowler-Nordheim tunneling, however, requires a minimum electric- field strength of around 10 MV/cm. Thus, for oxides of 10 nm in thickness, such tunneling will be negligible when normal 5-V signals are applied. As a result, FLOTOX transistors can be expected to retain their charges for more than 10 years if the memory is subjected only to normal read cycles.

The FLOTOX transistor must be isolated by a select transistor. Otherwise, the high voltage applied to the drain of the selected cell during erasing would also appear on the drain of the other *unselected* cells in the same memory column. A FLOTOX EEPROM cell must therefore consist of two transistors (Fig. 8-38b). Although this limits the density of such EEPROMs in comparison to EPROMs and flash EEPROMs, it makes it possible to erase and re-program one byte of the memory without having to erase the entire IC. In addition, two cycles are needed to load the correct data into the memory. In the first, all the cells in a byte are programmed (i.e., the floating gates are charged); in the second, selected cells are erased, with the drain used for data control.

The fabrication of FLOTOX EEPROMs involves a modification of the polysilicon-gate MOS process. A double-polysilicon process is used, together with a thin tunnel-oxide growth process. The growth of a high-quality, thin tunneling oxide is, in fact, the critical manufacturing step in this technology. The tunneling dielectric reportedly can be successfully implemented with nitrided oxides, since the barrier between silicon nitride and silicon is lower than that between SiO_2 and Si. As a result, a higher tunneling current can be obtained for the same voltage.[93]

Despite the fact that a reliable process for growing thin tunneling oxides must be developed, FLOTOX-based devices have become the most widely manufactured of the EEPROM types. They are still the easiest to learn to manufacture for companies that have already successfully developed an EPROM process. Since it is desirable to be able to program and erase the EEPROM while it remains in place on a PC board, considerable effort has also been expended to make this memory type fully operational with a 5 V power supply. (This type of operational capability is referred to as *5 V only*.)

FLOTOX-based EEPROMs are best suited for applications in which low-cost, low-density, nonvolatile memories are required – for example, in microcontrollers and programmable logic devices. Another potential application is for smart credit cards; several Japanese companies have announced 64-kbit FLOTOX-based EEPROMs for this market.

On the other hand, scaling and reliability considerations appear to limit the maximum size of FLOTOX EEPROMs to 256 kbits. The need for two transistors, and the relatively large size of the select transistor (due to the large voltages needed for

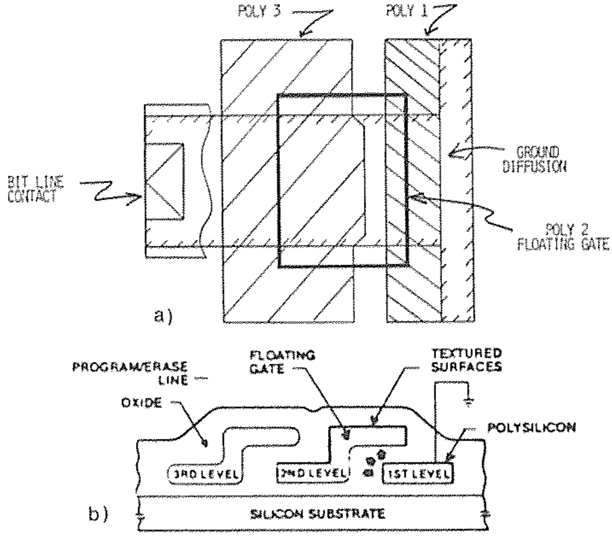


Fig. 8-39 Textured-polysilicon memory cell. (a) Top view. (b) Cross-sectional structure.[91] (© 1986 IEEE).

programming and erasure) are contributing factors. The poly-to-poly area of the sense transistor must also be large, due to the high oxide capacitance of the thin tunnel oxide.[40] Furthermore, the FLOTOX EEPROMs exhibit high failure rates, caused by defect-related oxide-breakdown problems as memory size is increased.[94] Error-detection and correction codes (EDCC) can be used to overcome this limitation, but such solutions impose a penalty of increased die cost. Reference 95 gives an example of a 50-ns access time, 256-kbit FLOTOX-based CMOS EEPROM using a single-bit EDCC.

The reliability of FLOTOX-based EEPROMs compares favorably with that of the other two EEPROM types. As with the others, there is a very low failure rate during 5-V operation; reliability problems occur as a result of the high voltages that must be used during programming and erasing. Random single-bit failures occur in FLOTOX devices due to oxide defects, resulting in leaky oxides that lose charge over time. The number of cycles that most FLOTOX EEPROMs are specified to be able to endure before the thin oxide becomes too leaky to retain data sufficiently, is 10^3–10^4 cycles. However, a process for increasing this endurance level to 10^6 cycles has been reported.[96]

8.7.3 Textured-Polysilicon EEPROMs

Textured-polysilicon EEPROMs, introduced in 1983 as an alternative to the tunneling oxide types of devices, are also based on the floating-gate MOS technology. The cell consists of three layers of polysilicon that partially overlap (Fig. 8-39) to create a cell that behaves like three MOS transistors in series. The floating-gate MOS transistor is formed by the middle polysilicon structure, which is encapsulated with SiO_2 to enable high charge retention. While charge is still transferred to the floating gate by means of Fowler-Nordheim tunneling, tunneling takes place from one polysilicon structure to another rather than from the substrate to the floating gate. The interpoly oxides through which the tunneling takes place can be made significantly thicker than the tunneling oxides in FLOTOX devices (60-100 nm in textured poly devices, versus <12 nm in FLOTOX devices), since the electric field that promotes the tunneling is enhanced by the geometrical effects of the fine texture at the surface of the polysilicon structures.

Textured-poly cells are programmed by causing electrons to tunnel from *poly 1* to the *floating poly*. Erasure is accomplished by causing electrons to tunnel from the floating poly structure to *poly 3*. The voltage of poly 3 is taken high in both the programming and erase operations. The drain voltage, however, determines whether tunneling occurs from *poly 1* to the *floating gate*, or from the *floating gate* to *poly 3*. As a result, the state of the drain voltage determines the final state of the memory cell. This provides an advantage, in that the cell represents a *direct write cell* – there is no need to charge all the cells and then remove the charge from selected cells, as with FLOTOX EEPROMs.

Textured-poly EEPROMs depend on a tunneling process whose physical mechanisms are not as well understood as those of tunneling through thin oxides, and which appears to require tighter control of empirically determined process parameters. In addition, the three poly layers require a more complex (and therefore more costly) fabrication sequence. Furthermore, textured-poly EEPROMs require a higher operating voltage than FLOTOX devices (>20 V). Finally, an intrinsic endurance problem is caused by the very high electron trapping that occurs as a result of tunneling in the poly oxides. This eventually leads to a condition in which the memory can no longer be programmed or erased.

For all of the above reasons, the textured-poly approach has been less widely pursued than the FLOTOX approach. Only one company, Xicor, is heavily involved in manufacturing these devices.[97] However, because the poly cells can be made about one-half the size of FLOTOX cells, it is possible to fabricate them in high-density configurations. In 1989, the largest textured-poly EEPROMs being offered had a 1-Mbit capacity. Although the cell-size advantage gives the textured-poly approach an edge over the FLOTOX EEPROMs for memories larger than 256 kbits, the flash-EEPROM technology described in the next section, provides a way to achieve equally high-density EEPROMs without the need to develop a textured-poly process.

8.8 FLASH EEPROMS

The *flash EEPROM* device is so named because the contents of all of the memory's array cells can be erased simultaneously as with a UV-EPROM, but through the use of an electrical erase signal. The term *flash* refers to the fact that the cells can be erased much more rapidly (1 or 2 seconds, compared to the 20 minutes required to erase a UV-EPROM). Although it was not possible to erase only a single byte in the first generation of flash EEPROMs, by 1989 parts had become available that offered a byte-by-byte erasable (and 64-byte erasable) feature in a 256-kbit memory.[98]

Flash EEPROMs are attractive for the middle of the programmable semiconductor spectrum, where neither EPROMs nor EEPROMs are particularly cost effective. The applications in this range typically require more memory capacity than EEPROMs can provide, but they also need faster and more frequent reprogramming than can be accomplished with EPROMs. Examples include automotive and automated factory equipment applications. As an example, the average EPROM cost about $7 in 1989, and a flash memory about $25. But the differential is wiped out by the expense of single reprogramming. The in-system reprogramming of a flash device may cost as little as $1, whereas pulling an EPROM out of a system to erase it by exposure to 20 minutes of UV light may cost over $80 when equipment, downtime and labor are factored in.

Meanwhile, EEPROMs are likely to remain popular wherever bytes will have to be erased selectively. But flash products, might do better for updating stored logic, when this must done more than once but less often than in main memory, cache memory, or registers. Reprogramming costs are similar, but flash memories are less than half the price of EEPROMs.

The erasing mechanism in flash EEPROMs is Fowler-Nordheim tunneling off the floating gate to the drain region. Programming of the floating gates, however, is carried out in most flash cells by *hot-electron injection into the gate.** Unlike floating-gate EEPROMs (which incorporate a separate select transistor in each cell to allow individual byte erasure), flash memories forego the select transistor to obtain bulk erasure. Thus, flash-EEPROM cells are roughly two to three times smaller than floating-gate EEPROM cells fabricated with the same design rules.[99] Figure 8-40 shows the cross-section of a CMOS flash-EEPROM cell implemented with triple polysilicon, and a SEM photo of a double-poly flash EEPROM cell.

Most flash-EEPROM cells use a double-poly structure, as shown in Fig. 8-41 (which also shows the Toshiba triple-poly cell, Fig. 8-41b). The upper poly forms the control gate and the word lines of the structure, while the lower poly is the floating gate. The gate oxide is ~10 nm thick,[100] and the interpoly dielectric is an oxide/nitride/oxide composite film ~45 nm thick.[99] In the structure shown in Fig. 8-40 and 8-41c the control-gate poly overlaps the channel region adjacent to the channel under the floating gate. This structure is needed because when the cell is erased, it leaves a positive charge on the floating gate. As a result, the channel under the floating gate becomes inverted. The series enhancement-mode transistor (formed by the control gate over the channel region), is needed in order to prevent current flow from source to drain. A more recently reported flash-EEPROM cell (Fig. 8-41a) does not require the control gate to form a series enhancement-mode transistor, because it uses a special software-controlled erase procedure that prevents the floating gate from being over erased.[100]

* The 5-V-only flash memories from Texas Instruments and Amtel depend on tunneling for both write and erase mechanisms.

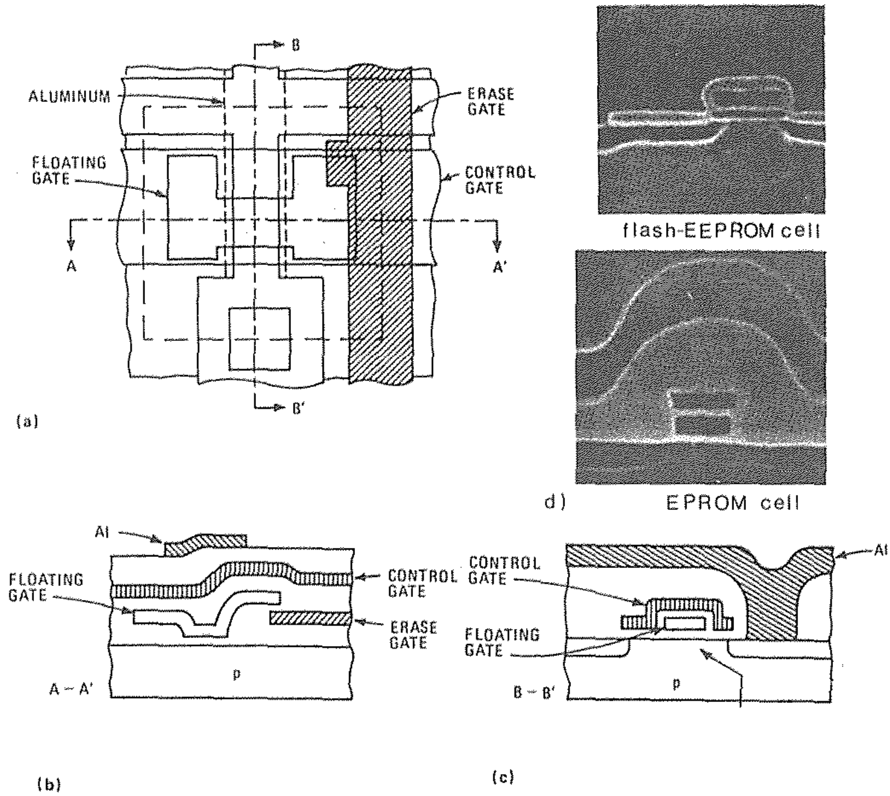


Fig. 8-40 Triple-poly flash-EEPROM cell from Toshiba: (a) Layout; (b) Cross section of the cell; (c) Section at right angles to the section shown in part (c);[101] (d) SEM pictures of double-poly flash-EEPROM and EPROM cells.[99] (© 1988 IEEE).

Flash EEPROMs can be seen to combine the advantages of UV-erasable EPROMs and floating-gate EEPROMs. They offer the high density (Fig. 8-42), small-die size, lower cost, and hot-electron writability of EPROMs, together with the easy erasability, on-board reprogrammability, and electron-tunneling erasure features of EEPROMs. High-density CMOS flash EEPROMs in 1-Mbit sizes are commercially available. It is projected that by the year 2000, 256-Mbit flash EEPROMs will be fabricated with 0.25 μm geometry.

With a memory-cell size of about one-quarter the size of current EEPROM cells, the flash EEPROMs also achieve EPROM die sizes. In addition, there are two types of flash EEPROMs: (1) those that are more akin to the EEPROM (and thus require a 12-V

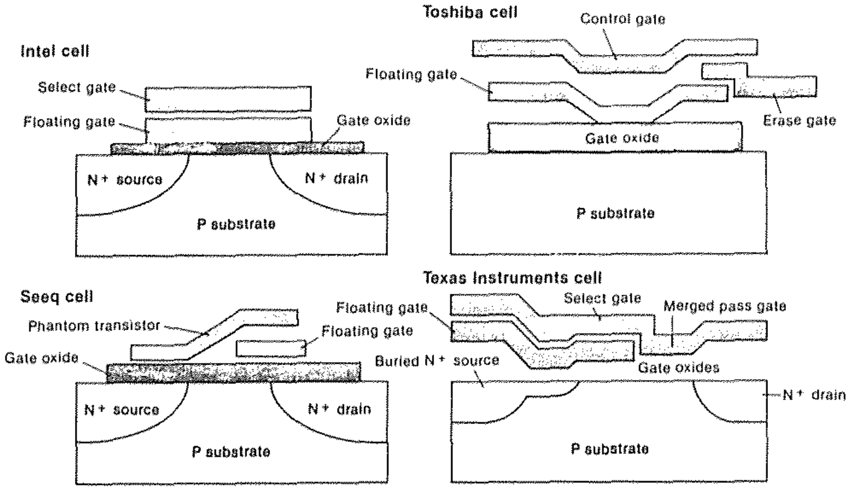


Fig. 8-41 Four approaches to flash memory technology: (a) Intel cell; (b) Toshiba triple-polysilicon cell; (c) SEEQ cell; (4) Texas Instruments 5-V-only cell.[125] (© 1989 IEEE).

external supply for programming and erasure); and (2) those that are closer to EPROMs, and hence need only a 5-V supply. Furthermore, the programming voltage can be applied during read operations, eliminating the need to switch it off when not erasing or programming. Byte-write times are 100 μs, and erasure times are 200 ms. Access times of 110 ns at 30-mA active-current consumption are provided by a 128-kBit CMOS flash EEPROM.[100,101] Endurance (i.e., the number of times a device can be erased and written) is a minimum of 100 cycles, and can be as high as 1000 cycles (note that this is lower than the endurance of EEPROMs, which is typically 1000 – 10,000 cycles).

8.9 NONVOLATILE FERROELECTRIC MOS RAMS

A novel type of nonvolatile MOSFET DRAM memory cell, introduced in late 1987, uses the electrical polarization of a ferroelectric capacitor to store information semipermanently.[102,103] Since ferroelectric polarization retention is nearly perpetual (just as in magnetic core memories), refresh is not needed. The reported write speed is 200 ns in one design,[102] and 60 ns in another,[103] which is much faster than that exhibited by an EEPROM (1 ms) or a UV-PROM (10 ms) without fatigue after 10^{12} write cycles. It is predicted that by 1991 products will be available with operating lifetimes of ~75 years at a cycle time of 100 ns, and 10^{12} read/write cycles. A recent review article has described the latest advances in such nonvolatile RAMs.[124]

The cell contains a ferroelectric capacitor as the charge storage element and an MOS transistor for sensing and writing (Fig. 8-43b). The ferroelectric insulator of the capacitor may be polarized by either a positive or negative voltage, and its polarization state is retained after the voltage is removed (this characteristic is called a *ferroelectric effect*). This effect occurs because in some materials dipoles will align in parallel under the influence of an externally applied electric field, and they will remain aligned (polarized) after the field is removed. Reversal of the field causes polarization in the opposite direction. Thus, a *ferroelectric* is a material which can be permanently polarized by the application of an electric field. (Contrary to the name, the ferroelectric effect has nothing to do with iron.)

A ferroelectric thin-film capacitor exhibits a characteristic hysteresis curve, which describes the amount of charge that the device can store as a function of the applied voltage (Fig. 8-42a). It has two stable polarization states and can be modeled as a bistable capacitor with two distinct polarization thresholds. The *coercive voltage* is the digital switching threshold of the capacitor. For memory applications, it is desirable for the two coercive voltage points to be symmetrical and less than 2.5 V, so that the memory may operate from standard memory power-supply voltages.

Memory arrays of such ferroelectric cells (*FRAMs*) have the potential to replace hard and floppy magnetic disks. Such memory arrays could provide increased reliability compared with the present magnetic disk drives, since the FRAM contains no moving parts. In addition, they could offer much shorter read, write, and access times than

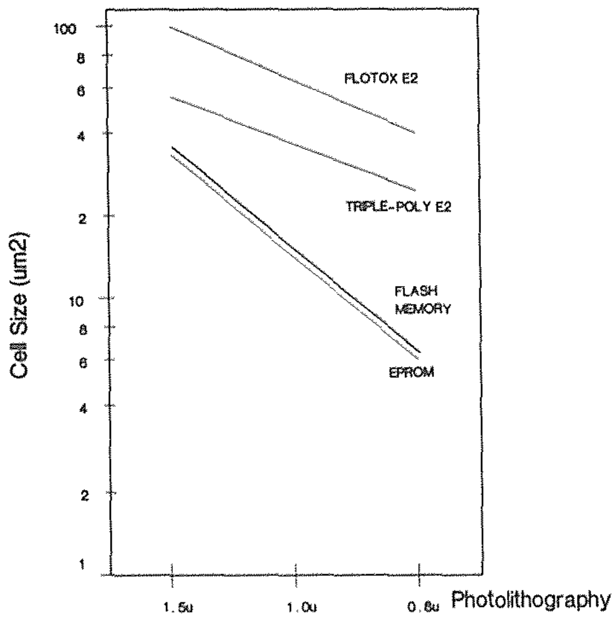


Fig. 8-42 Size of flash-memory cell versus lithographic feature size.[99] (© 1988 IEEE).

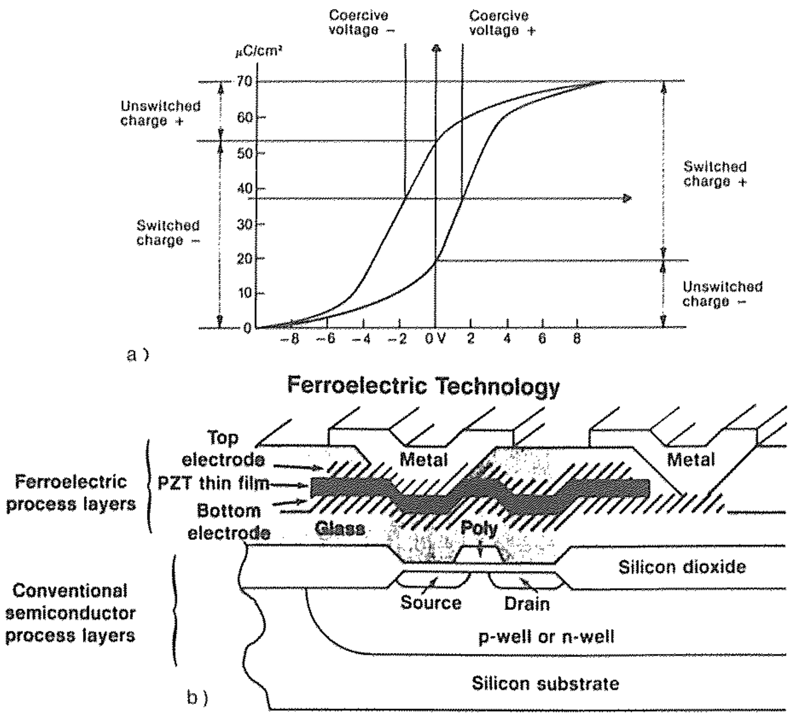


Fig. 8-43 (a) Thin-film ferroelectric capacitors exhibit a stable coercive voltage of less than 2.5 V and a switched charge of more than 20 μcoulombs cm^{-2}.[125] (© 1989 IEEE). (b) Ferroelectric-dielectric film (lead zirconate titanate, PZT) sandwiched between two metal electrodes to form a non-linear capacitor built above existing circuitry. Reprinted with permission of Semiconductor International.

current memory disks can.

Ferroelectrics are essentially compatible with conventional wafer processing and memory circuits. The manufacture of MOS transistors involves relatively mature technology, and only the fabrication of ferroelectric thin films on Si and SiO_2 substrates still needs to be further developed. In one reported design, a 256-bit demonstration chip uses capacitors fabricated with a thin film of lead zirconate titanate (PZT) ceramic as a dielectric sandwiched between two metal electrodes. This structure forms a "digital memory capacitor" built above existing semiconductor circuitry (Fig. 8-44). Such PZT films remain ferroelectric from -80°C to 350°C, well beyond the operating temperature range of existing silicon circuits. Other ferroelectric materials being studied are lanthanum-doped PZT, and lithium niobate.

FRAMs can be operated and programmed from a single 5 V power supply. In addition, because ferroelectric materials typically exhibit dielectric constants much larger than that of SiO_2 (e.g., 1000-1500 versus the 3.8 to 7.0 of current DRAM capacitors),

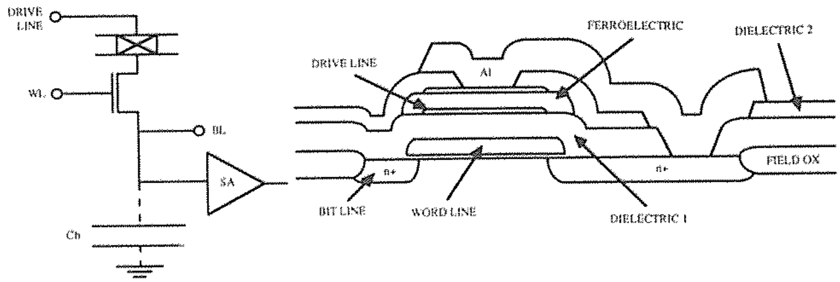


Fig. 8-44 Cross sectional view of a ferroelectric memory cell.[102] (© 1987 IEEE).

a larger charge can be stored on a capacitor of the same size (for example, if a capacitor using a PZT film could store 10 $\mu C/cm^2$ of charge, one of the same size using SiO_2 could store only 0.1 $\mu C/cm^2$). If the smaller ferroelectric capacitor could be used, DRAM manufacturing could thus revert to the simpler planar process.

REFERENCES

1. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill, New York, 1983.

2. S. Asai, "Semiconductor Memory Trends," *Proceedings of the IEEE*, Dec. 1986, p. 1623.

3. C. T. Sah, "The Evolution of the MOS Transistor," *Proceedings of the IEEE*, October 1988, 1280.

4. O. Minato, *IEEE J. Solid-State Circuits*, October 1982, p. 793.

5. W. C. Holton and R. K. Gavin, "A Perspective on CMOS technology Trends", *Proceedings of the IEEE*, December 1986, p. 1646.

6. R. Hunt, "Memory Design and Technology," in M.J. Howes and D.V. Morgan, Eds., *Large Scale Integration*, Wiley, New York, 1981.

7. S. T. Chu et al., *IEEE J. Solid-State Ckts.*, October 1988, p. 1078.

8. H. Okuyama et al., *IEEE J. Solid-State Ckts.*, October 1988, p. 1054.

9. R. A. Chapman, *Tech. Dig. IEDM*, 1987, p. 362.

10. H. Momose et al., *Tech. Dig. IEDM*, 1984, p. 706.

11. J. S. Fu et al., *Tech. Dig. IEDM*, 1987, p. 540.

12. C. E. Chen et al., "Stacked CMOS SRAM Cell," *IEEE Electron Dev. Letts.*, April, 1983, p. 272

13. L. R. Hite et al., *IEEE Electron Dev. Letts.*, October 1985, p. 548.

14. K. Sakui et al., *Tech. Dig. IEDM*, 1988, p. 44.

15. T. Ohzone et al., *IEEE Trans. Electron Devices*, September 1985, p. 1749.

16. N. C. C. Lu, L. Gerzberg, and J. D. Meindl, "Scaling Limitations of Polysilicon Resistors in VLSI SRAMs & Logic," *IEEE J. Solid-State Ckts*, April 1982, p. 312.

17. R. Saito, Y. Sawahata, and N. Momma, *IEEE Trans. on Electron Dev.*, March 1988, p. 299.
18. S. Voldman et al., *Tech. Dig. IEDM*, 1987, p. 518.
19. N. Hoshi et al., *Tech. Dig. IEDM*, 1986, p. 300.
20. N. Homma et al., *IEEE J. Solid-State Circuits*, October 1986, p. 675.
21. C.-T. Chuang et al., *IEEE J. Solid-State Circuits*, October 1986, p. 670.
22. M. Inadachi et al., *Tech. Dig. IEDM*, 1979, p. 108.
23. Y. Kato, M. Odaka, and K. Ogiuie, "A 16 ns, 16-K Bipolar SRAM," *IEEE Solid- State Circuits Conf.*, 1983, p. 106.
24. K. Ogiue et al., "Technology Improvement for High Speed ECL RAMs," *Tech. Dig. IEDM*, 1986, p. 468.
25. H. V. Tran et al., *IEEE J. Solid-State Circuits*, October 1988, p. 1041.
26. R. A. Kertis, D.D. Smith, and T.L. Bowman, *IEEE J. Solid-State Circuits*, Oct. 1988, p. 1048.
27. Y. Nishioka et al., *IEEE Trans. Electron Dev.*, September 1987, p. 1957.
28. R. H. Dennard, "Field Effect Transistor Memory," U.S. Patent 3,387,286, granted June 4, 1968.
29. R. H. Dennard, "Evolution of the MOSFET Dynamic RAM- a personal view," *IEEE Trans. Electron Dev.*, November 1984, p. 1549.
30. V. L. Rideout, "One-Device Cells for DRAMs: A Tutorial," *IEEE Trans. Electron Dev.*, June 1979, p. 839.
31. C.-T. Chuang et al., *IEEE J. Solid-State Circuits*, October 1988, p. 1265.
32. "Shipment of 1-Mbit SRAMs Ushers in Era of Submicron Linewidths," *Semicond. International*, November 1988, p.28.
33. J. S. Fu et al., "Scaling Studies of CMOS SRAM Soft-Error Tolerances - From 16K to 256K," *Tech. Dig. IEDM*, 1987, p. 540.
34. C. M. Hochstedtler, *Semiconductor International*, April 1989, p. 68.
35. A. F. Tasch, Jr., P. K. Chatterjee, H. S. Fu, and T. C. Holloway, *IEEE Trans. Electron Devices*, **ED-25**, p. 33, (1978).
36. D. K. Schroder, *Advanced MOS Devices*, Addison-Wesley, Redding, MA. 1988.
37. C. Sodini and T. Kamins, "Enhanced Capacitor for One-Transistor Memory Cell," *IEEE Trans. Electron Dev.*, **ED-23**, 1976, p. 1187.
38. S. Chou, *Electronic Design*, October 4, 1984, p. 138.
39. A. F. Tasch, Jr., *IEEE Proceedings*, March, 1989.
40. K. Shimtori et al., *ISSCC Tech. Dig.* 1983, p. 228.
41. H. Sunami, *Tech. Dig. IEDM*, 1985, p. 694.
42. D. S. Yaney et al., *Tech. Dig. IEDM*, 1985, p. 698.
43. L. Risch, W. Mueller, and R. Tielert, *Semiconductor International*, May 1988, p. 246.
44. *Electronics News*, March 6, 1989.
45. B. C. Cole, *Electronics*, February 5, 1987, p. 66.
46. K. V. Rao et al., *Tech. Dig. IEDM*, 1986, p. 140.
47. G. K. Herb, D. J. Riegler, and K. Shields, *Solid-State Technol.*, October 1987, p. 109.
48. K. Yamada et al., *Tech. Dig. IEDM*, 1985, p. 702.
49. K. Yamabe and K. Imai, *IEEE Trans. Electron Dev.*, 1987, p. 1681.

50. S. Rohl et al., *Ext. Abs. Electrochem. Soc. Meeting*, Spring, 1989, p. 194.
51. Y. Miyai et al., *J. Electrochem. Soc.*, January 1989, p. 150.
52. H. Sunami et al., *Tech. Dig. IEDM*, 1982, p. 806.
53. L. Risch, W. Mueller, and R. Tielert, "Four Megabit DRAM Processing," *Semicond. Internatl.*, May 1988, p. 246.
54. D. A. Baglee et al., *Tech. Dig. IEDM*, 1985, p. 384.
55. M. Wada, K. Heida, and S. Watanabe, *Tech. Dig. IEDM*, 1984, p. 244.
56. F. Horiguchi et al., *IEEE J. Solid-State Ckts.*, December 1986, p. 1076.
57. S. Nakajima et al., *Tech. Dig. IEDM*, 1984, p. 240.
58. K. Mashiko et al., *IEEE J. Solid-State Circuits*, October 1987, p. 643.
59. W. F. Richardson et al., *Tech. Dig. IEDM*, 1985, p. 714.
60. S. Sakamoto et al., *Tech. Dig. IEDM*, 1985, p. 710.
61. N. Lu et al., *Tech. Dig. IEDM*, 1985, p. 771.
62. F. Horiguchi et al., *Tech. Dig. IEDM*, 1987, p. 324.
63. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, Chapter 10, p. 296.
64. W. F. Noble, A. Bryant, and S. H. Voldman, *Tech. Dig. IEDM*, 1987, p. 340.
65. S. Banerjee et al, *IEEE Trans. Electron Dev.*, January 1988, p. 108.
66. M. Taguchi et al., *Tech. Dig. IEDM*, 1986, p. 136.
67. T. Kaga et al., *Tech. Dig. IEDM*, 1987, p. 332.
68. K. Tsukamoto et al., *Tech. Dig., IEDM*, 1987, p. 328.
69. M. Yanagisawa, K. Nakamura, and M. Kikuchi, *Tech. Dig. IEDM*, 1986, p. 132.
70. W.-H. Lee et al., *IEEE Trans. Electron Dev.*, November 1988, p. 1876.
71. N. C. C. Lu et al., *Tech. Dig. IEDM*, 1988, p. 588.
72. S. Banerjee and M. Bordelon, *IEEE Trans. Electron Dev.*, December 1987, p. 2485.
73. M. Koyanagi, *IEDM Tech. Dig.*, 1978, p. 348.
74. H. Watanabe, K. Kurosawa, and S. Sawada, *Tech. Dig. IEDM*, 1988 p. 602.
75. S. Kimura et al., *Tech. Dig. IEDM*, 1988, p. 596.
76. T. Ema et al., *Tech. Dig. IEDM*, 1988, p. 592.
77. T. May, and M. Woods, "Alpha-Particle-Induced Soft Errors in DRAMs," *IEEE Trans. Electron Dev.*, January 1979, p. 2.
78. G. Sai-Halasz et al., *IEEE Trans. Electron Dev.*, **ED-29**, 1983, p. 725,
79. C. M. Hsieh, P.C. Murley, and R.R. O'Brien, *Int. Reliab. Physics Symposium*, 1981, p. 38.
80. O. Osamu, *Ext. Abs. Electrochem. Soc. Meeting*, Spring, 1989, Abs. No. 130, p. 181.
81. F. Masuoka et al., *Dig. Tech. Papers, IEEE Int.Solid State Cirkts Conf.*, p. 146., 1984.
82. T. Fukushima et al., *IEEE Int. Solid State Circuits Conf.*, p. 14, 1983.
83. P. J. Salsbury et al., *Dig. Tech. Papers, 1977 Inter. Solid-State Circuits Conf.*, 1977, p. 186.
84. D. Frohman-Bentchkowsky, *Solid-State Electronics*, 1974, p. 517.
85. N. Ohtsuka et al., *IEEE J. of Solid-State Circuits*, October 1987, p. 669.
86. S.-S. Lee et al., *Tech. Dig. IEDM*, 1987, p. 588.
87. S. B. Ali et al., J. Solid-State Circuits, February 1988, p. 79.
88. S. Mori et al., Tech. Dig. IEDM, 1987, p. 556.

87. S. B. Ali et al., J. Solid-State Circuits, February 1988, p. 79.
88. S. Mori et al., Tech. Dig. IEDM, 1987, p. 556.
89. A. T. Mitchell, C. Huffman, and A.L. Esquivel, *Tech. Dig. IEDM*, 1987, p. 548.
90. A. Esquivel et al., *Tech. Dig. IEDM*, 1987, p. 859.
91. S. K. Lai, V. K. Dham, and D. Guterman, "Comparison of Trends in Today's Dominant EEPROM Technologies," *Tech. Dig. IEDM*, 1986, p. 580.
92. E.H. Snow, "Fowler-Nordheim Tunneling in SiO_2 Films," *Solid-State Communications*, 5 (1967), p. 813.
93. D. M. Brown et al., "Properties of $Si_xO_yN_z$ films on Si," *J. Electrochem. Soc.*, **15**, 1986, p. 311.
94. A. Bagles, "Characteristics and Reliability of 10 nm Oxides," *Internat. Reliability Physics Sympos.*, 1983, p. 152.
95. T.-K. J. Ting et al., *IEEE J. Solid-State Circuits*, October 1988, p. 1164.
96. D. Cioaca et al., *IEEE J. Solid-State Circuits*, October 1987, p. 684.
97. D. Guterman et al., *Tech. Dig. IEDM*, 1986, p. 826.
98. B. Santo, *IEEE Spectrum*, Dec. 1989. p. 47.
99. G. Samachisa et al., *J. Solid-State Circuits*, October 1988, p. 676.
100. V. N. Kynett et al., *IEEE J. Solid-State Circuits*, October 1988, p. 1157.
101. F. Masuoka et al., *IEEE J. Solid-state Circuits*, August 1987, p. 548.
102. W. I. Kinney et al., *Tech. Dig. IEDM*, 1987, p. 850.
103. S. Sheffield-Eaton et al., *IEEE Int. Solid State Circuits Conf.*, February 1988, p. 130.
104. A. F. Tasch et al., *Tech. Dig. IEDM*, 1977, p. 287.
105. A. F. Tasch and L. H. Parker, *IEEE Proceedings*, March 1989, p. 374.
106. A. F. Tasch et al., "The Hi-C RAM Cell Concept," *Tech. Dig. IEDM*, 1977, p. 287.
107. A. Lancaster et al., *Tech. Dig. ISSCC.*, 1983, p. 164.
108. C. G. Sodini, S. S. Wong, and P. -K. Ko, *IEEE J. Solid-State Ckts.*, Feb. 1989, p. 118.
109. F. Miyaji et al., "A 25-ns 4-Mbit CMOS SRAM," *IEEE J. Solid State Circuits*, October 1989, p. 1213.
110. N. C.-C. Liu et al., "A 22-ns High-Speed CMOS SRAM with Address Multiplexing," *IEEE J. Solid State Circuits*, October 1989, p. 1198.
111. S. Fujii et al, *IEEE J. Solid State Circuits*, October 1989, p. 1170.
112. K. Rogers, "Bipolar DRAM hits U.S.," *Electronic Engr. Times*, Sept. 11, 1989, p. 2.
113. B.-Y. Hwang et al., *IEEE J. Solid-State Circuits*, April 1989, p. 504.
114. M. Matsui et al., *IEEE J. Solid State Circuits*, October 1989, p. 1226.
115. M. Suzuki et al., *IEEE J. Solid State Circuits*, October 1989, p. 1233.
116. D. M. Brown, M. Ghezzo, and J. M. Pimbley, *Proceedings of the IEEE*, December 1986, p. 1678.
117. D. A. Hodges, "Microelectronic Memories," Scientific American, **237**, p. 130, September 1977.
118. E. S. Yang, *Microelectronic Devices*, McGraw-Hill, New York, 1988, p.342.
119. Y. Numasawa et al., *Tech. Dig. IEDM*, 1989, p. 43.
120. A. Learn et al., *J. Appl. Phys.*, **61**, p. 1898, (1987).
121. T. E. Tang, *Tech. Dig. IEDM*, 1989, p. 39.
122. S. Inoue et al., *Tech. Dig. IEDM*, *1989*, p. 31.

123. K. Sunouchi et al., *Tech. Dig. IEDM,* 1989, p. 23.
124. D. Bondurant and F. Gnadinger, *IEEE Spectrum,* July 1989, p. 30.
125. R. Pashley and S. K. Lai, "Flash memories: the best of two worlds," *IEEE Spectrum,* December, 1989, p. 30.

PROBLEMS

8.1 Determine the width to length ratios for the static NMOS static RAM cell layout shown in Fig. 8-4a. Assume that the source and drain diffuse 0.5 μm into the channel, so that the electrical channel length is 1 μm less than the channel length measured in Fig. 8.4a. Use the above device data to calculate the nominal power consumption at standby, the readout current when the column lines are biased at +3.0 V, and the column voltage needed to change the state of the selected cell.

8.2 For the cell of the previous problem, calculate the row and column capacitances and resistances per memory cell. Assuming that a read operation requires that the column voltage change from 3.0 to 2.5 V. Calculate the row and column delay times for a 4096 bit (64x64) array of these cells.

8.3 A DRAM is required to operate with a minimum refresh time of 4 ms. The storage capacitor in each cell is 10 μm square, has a capacitance of 50 pF (50×10^{-15} F), and is fully charged at 5 V. (a) Calculate the number of electrons stored in each cell. (b) Estimate the worst-case leakage current that the dynamic capacitor node can tolerate.

8.4 Calculate the area needed to make a trench-capacitor DRAM cell with the same capacitance as the capacitor in problem 8.3, assuming that the trench depth is (a) 1 μm deep, and (b) 5 μm deep, and the trench in both cases is 1 μm wide. Assume that the capacitance per silicon surface area is the same as in problem 8.3.

8.5 Draw the layout for a NOR ROM implemented in silicon-gate NMOS technology with a level of detail similar to that shown in Fig. 8-31c. Calculate row (gate) and column capacitances per bit, assuming the gate oxide thickness is 100 nm, substrate doping is 1×10^{15} cm^{-3}, source/drain doping is 1×10^{20} cm^{-3}, the S/D junction depths and lateral diffusions are 1 μm, and the minimum feature size (i.e., polysilicon gate width and length) = 10 μm. This ROM uses a load transistor formed with a depletion implant. (Do not neglect the capacitance from this implant to substrate.) In addition, calculate the nominal current available when reading a stored zero (low column voltage).

8.6 Given an EPROM cell as shown in Fig. 8-33b, assuming that the transistor characteristics if Gate 1 could be used as an input are those given above. (Of course, Gate 1 is not directly accessible in the EPROM device.) The oxide under Gate 1 is 100 nm thick, and that under Gate 2 is 120 nm thick. The drawn channel dimensions are W = L = 10 μm. (a) Calculate the saturated drain current for this device with source grounded and 5 V applied to drain and Gate 2, assuming that the potential on Gate 1 is zero with 0 V on all other electrodes. (b) Assume that during writing a current of 10^{-11} A flows from Gate 1 to drain, induced by the high field in the drain depletion region. How long will it take for the potential on Gate 1 to change by 5 V due to this current?

8.7 Calculate the threshold voltage seen from Gate 2 for the EPROM cell of the previous example, assuming the potential on Gate 1 is -5 V with 0 V applied to all other electrodes.

8.8 For a floating-gate EPROM, the lower gate oxide has a dielectric constant = 4, and a thickness of 100 nm. 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 $J = \sigma E$, where $\sigma = 10^{-7}\ (\Omega\text{-cm})^{-1}$, 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 (Gate 2) for (a) 0.25 μsec, and (b) a sufficiently long time that J in the lower insulator becomes negligibly small.