

**Nonvolatile
Semiconductor
Memory Technology**

IEEE Press
445 Hoes Lane, P.O. Box 1331
Piscataway, NJ 08855-1331

IEEE Press Editorial Board
Roger F. Hoyt, *Editor in Chief*

J. B. Anderson	S. Furui	S. Kartalopoulos
P. M. Anderson	A. H. Haddad	P. Laplante
M. Eden	R. Herrick	W. D. Reeve
M. E. El-Hawary	G. F. Hoffnagle	D. J. Wells

Kenneth Moore, *Director of IEEE Press*
John Griffin, *Senior Acquisition Editor*
Linda Matarazzo, *Assistant Editor*

Denise Phillip, *Associate Production Editor*

IEEE Solid-State Circuits Society, *Sponsor*
SSC Liaison to IEEE Press, Stu Tewksbury
IEEE Components, Packaging, and Manufacturing Technology Society, *Sponsor*
CPMT Liaison to IEEE Press, Joe Brewer

Technical Reviewers

Dennis Adams, *Northrop Grumman Corporation*
James D. Kinnison, *Johns Hopkins University*
Greg Koker, *Analog Devices*

Also of Interest from IEEE Press . . .

Nonvolatile Semiconductor Memories: Technologies, Design, and Applications

Chenming Hu

1991 Hardcover 464 pp IEEE Order No. PC2634 ISBN 0-87942-269-6

Semiconductor Memories: Technology, Testing, and Reliability

Ashok K. Sharma

1977 Hardcover 480 pp IEEE Order No. PC3491 ISBN 0-7803-1000-4

Compound Semiconductor Transistors: Physics and Technology

Sandip Tiwari

1993 Hardcover 336 pp IEEE Order No. PC3137 ISBN 0-7803-0417-9

Nonvolatile Semiconductor Memory Technology

*A Comprehensive Guide
to Understanding and Using
NVSM Devices*

Edited by

William D. Brown
University of Arkansas

Joe E. Brewer
Northrop Grumman Corporation



IEEE Press Series on Microelectronic Systems
Stu Tewksbury, *Series Editor*

IEEE Solid-State Circuits Council, *Sponsor*

IEEE Components, Packaging, and Manufacturing
Technology Society, *Sponsor*

The Institute of Electrical and Electronics Engineers, Inc., New York

This book and other books may be purchased at a discount from the publisher when ordered in bulk quantities. Contact:

IEEE Press Marketing
Attn: Special Sales
445 Hoes Lane, P. O. Box 1331
Piscataway, NJ 08855-1331
Fax: (732) 981-9334

For more information on the IEEE Press,
visit the IEEE home page: <http://www.ieee.org/>

© 1998 by the Institute of Electrical and Electronics Engineers, Inc.,
3 Park Avenue, New York, NY 10016-5997

*All rights reserved. No part of this book may be reproduced in any form,
nor may it be stored in a retrieval system or transmitted in any form,
without written permission from the publisher.*

Printed in the United States of America

10 9 8 7 6 5 4 3

ISBN 0-7803-1173-6

IEEE Order Number: PC5644

Library of Congress Cataloging-in-Publication Data

Nonvolatile semiconductor memory technology : a comprehensive guide to
understanding and using NVSM devices / edited by W. D. Brown, Joe
E. Brewer.

p. cm. — (IEEE Press series on microelectronic systems)

Includes bibliographical references and index.

ISBN 13: 978-0-7803-1173-2

ISBN 0-7803-1173-6 (cloth)

I. Semiconductor storage devices. I. Brown, W. D. (William D.).
(date) . II. Brewer, Joe (Joe E.) III. Series.

TK7895.M4N634 1997

621.39'732-dc21

97-19691

CIP

Floating Gate Flash Memories

4.0. INTRODUCTION

The semiconductor revolution started with the invention of the transistor in 1947. Bipolar transistor technology replaced the vacuum tube which, in turn, was replaced by MOSFET technology in the 1970s, starting the age of LSI/VLSI/ULSI (Fig. 4.1 and Table 4.1). The MOS/VLSI revolution was ushered in by the invention of DRAM, which replaced magnetic memory. In addition, several other semiconductor memories have been in use (Fig. 4.2), the most recent one being Flash memory, which appeared in 1985.

Many predict that Flash memory will be the fuel for the next stage of this revolution [4.1]. This belief is based on the observation that Flash memory has many attributes of an ideal memory (Table 4.2) needed in a computer system (Fig. 4.3). Flash memory's combination of density, rewritability, and nonvolatility (to name a few) makes it unique among semiconductor memories (Fig. 4.2), a superset of these key features.

Among the currently available memory technologies (floppy disk drive, hard disk drive, ROM, EPROM, OTP, EEPROM, DRAM, SRAM, shadow RAM, and Flash), Flash memory satisfies many of the characteristics of an ideal memory (Table 4.3).

Because of the combination of nonvolatility, in-system rewritability, and high density, Flash memory has already found many applications that can broadly be classified [4.2] as:

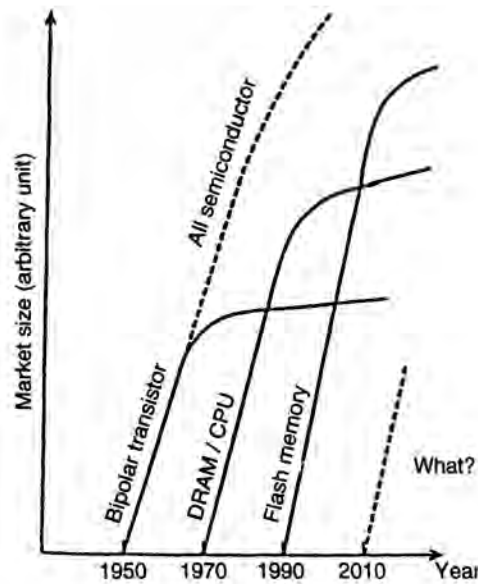


Figure 4.1 History of semiconductor business [4.1].

TABLE 4.1 HISTORY OF SEMICONDUCTOR TECHNOLOGY DRIVERS

Decade	New Technology	Replaced Technology	Market
1950s	Bipolar	Vacuum tubes	Radio/TV
1970s	MOS (DRAM)	Core memory	Computer
1990s	MOS (Flash)	Hard disk/floppy disk	Computer, personal systems

1. *Data Accumulation*—Medical instrumentation, flight recorders, point of sales terminals, handheld instruments, and remote sensing. The benefits of Flash memory in data accumulation applications include density, ruggedness, reliability, and nonvolatility.
2. *Data/Lookup Table Storage*—In this type of application, Flash memory devices store large amounts of infrequently updated system data or look-up tables, for example, PBX switcher and laser printers. Here, the advantages of Flash memory include in-system upgradability, nonvolatility, and density.
3. *Embedded Code Storage*—PC BIOS, digital cellular phones, control software in laser printers, and telecommunication bridges/routers.

Other areas in which Flash memory is finding application are high-end file storage, memory cards, smart cards, PLDs, digital audio machines, satellite systems, and neural nets.

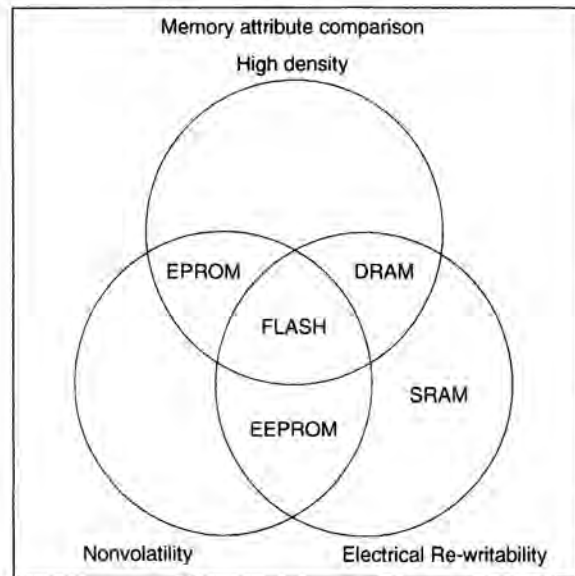


Figure 4.2 Memory attribute comparison. Flash memory is a functional superset [4.2].

TABLE 4.2 THE CHARACTERISTICS OF AN IDEAL MEMORY

1. Nonvolatility
2. Dense [consumes small space/bit]
3. Fast read/write/erase
4. In-system rewritability
5. Bit alterability
6. Endurance [highwrite/erase cycles]
7. Low power consumption
8. High bit count
9. Low cost
10. Highly scalable
11. Single-power supply
12. Ruggedness
13. Portability
14. Small form factor
15. Highly integrable with other system technologies

4.0.1 Why Flash?

Flash memories are designed to be the solution to the scaling problem of EEPROMs by trading off memory cell size with functionality. Standard

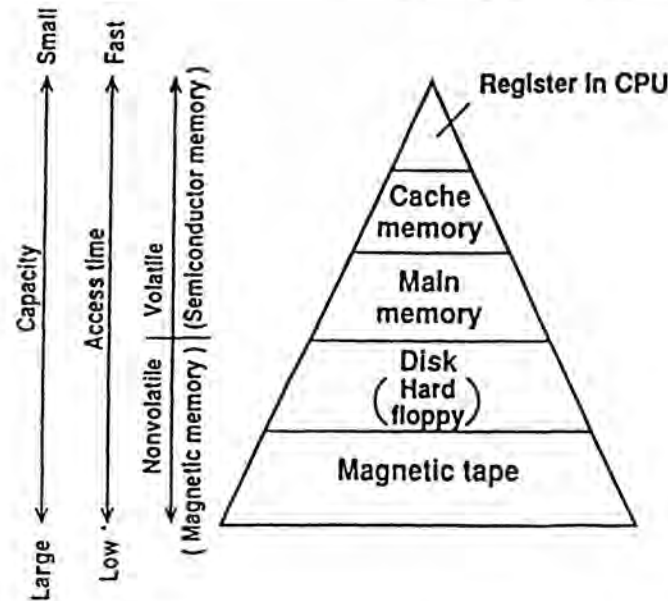


Figure 4.3 Hierarchy of memories in a computer system [4.1].

TABLE 4.3 MEMORY ATTRIBUTES

	Flash	DRAM	SRAM	EEPROM	UV-EPROM	Hard Disk	Floppy Disk
Nonvolatility	O	X	X	O	O	O	O
High density	O	O	X	X	O	O	X
I-T/cell	O	X	X	X	O	X	X
Low power	O	X	X	O	O	X	X
In-system rewritability	O	O	O	O	X	O	O
Bit alterability	X	O	O	O	X	O	O
Fast read	O	O	O	O	O	X	X
Fast write	O	O	O	O	O	O	O
High endurance	O	O	O	O	X	O	O
Low cost	O	O	X	X	O	O	O
Single-power supply	O	O	O	O	X	O	O

EEPROM memory cells have a select transistor that is either separated from or merged with the floating gate memory transistor. The select transistor is required for a number of reasons. The reason why it is needed for FLOTOX EEPROM is described in the following paragraphs. Other EEPROMs have similar limitations.

In a FLOTOX EEPROM, the select transistor minimizes program/erase disturb in an array. In Fig. 4.4, a standard cross-point array without a select transistor is

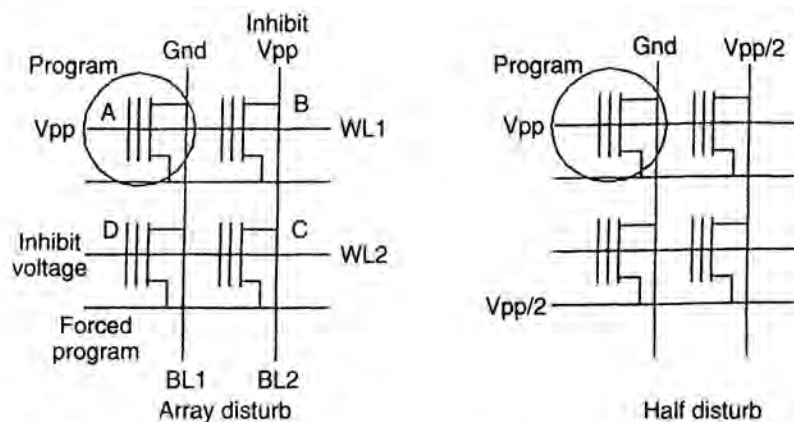


Figure 4.4 Disturb conditions in a typical cross-point single-transistor floating gate array.

shown. If transistor A is to be programmed (defined as electrons into the floating gate), high voltage is applied to wordline 1 and bitline 1 is grounded. If transistor B is not to be programmed, an inhibit voltage of V_{pp} has to be applied to bitline 2 so as not to create a net voltage across the tunnel oxide in transistor B. For the same reason, if transistor C is not to be disturbed, V_{pp} must be applied to the wordline. Given this bias condition, transistor D will be programmed whether or not it is desired to do so. In some designs, $V_{pp}/2$ is applied to the bitline of transistor B and the wordline of transistor C, giving a half-disturb condition where only half of the V_{pp} is applied across the cell. However, because of the logarithmic dependence of tunneling current on voltage, there will still be a small disturb for the half-inhibit condition when exercised over many cycles. Historically, a nondisturb condition has been realized by isolating the floating gate memory transistor from the array using a select transistor.

Another important consideration is that an erased (electrons removed from the floating gate) memory cell can have a net positive charge in the floating gate, giving a depletion state, which, in effect, shorts out the column in the array shown (Fig. 4.4). A select transistor serves to isolate the depletion transistor so as not to give a read error. The drawback of the extra select transistor is that it limits scaling.

4.0.2 A Brief History of Flash EEPROM

Table 4.4 shows the history of Flash EEPROM development. The first modern Flash EEPROM was proposed at the 1984 International Electron Devices Meeting by Masuoka et al. [4.3–4.5]. Figure 4.5 shows the cell structure of this first Flash EEPROM. This EEPROM structure was called a Flash EEPROM because the complete memory array is erased very quickly. At present, the name Flash

EEPROM is used for all EEPROM in which all or a large number of cells, called a block or a page, are erased at the same time.

The Flash EEPROM array reported is based on an EPROM array [4.6]. It has similar high-density, low-cost, and high-reliability advantages of EPROMs. The challenge in Flash memory is to learn how to remove electrons from the floating gate by an electrical process instead of the UV light illumination used for EPROMs. The advantage of Flash EEPROM is that the erase time is less than 1 second, while the erase time for an UV-EPROM is about 10 minutes. Moreover, the UV-EPROM requires an expensive package because of the UV transparent quartz window, and it must be taken out of the system for erasure. Flash EEPROM, however, can be packaged in small, inexpensive plastic packages and reprogrammed in-system (i.e., without removal of the package). The limitation of Flash EEPROM compared to traditional EEPROM is that many bytes are erased simultaneously, instead of a single byte at a time.

In 1985, Mukherjee et al. proposed a source-erase type of Flash memory cell [4.7] called the stacked gate cell (Fig. 4.6). The structure of this cell is the same as that of the stacked gate UV-EPROM, with two modifications: (1) the source junction is graded to support the high voltage during erase, and (2) the gate oxide is considerably thinner than UV-EPROMs to allow F-N (Fowler-Nordheim) carrier tunneling during erase. Since 1985, the stacked gate cell has become the volume shipment leader of Flash EEPROM and, at present, 16 MB Flash memories are in production [4.8]. Several variations of the stacked gate cell have been reported [for example, 4.9–4.12] and will be discussed in detail later in this chapter.

In 1987, Masuoka et al. [4.13] proposed a NAND structured cell. This structure reduces the cell size without scaling of the device dimensions. The NAND structure cell arranges a number of bits in series, as shown in Fig. 4.7. The current EPROM cell has one-half contact per bit. However, for a NAND structure cell, only one contact hole is required per two NAND structures. As a result, the NAND cell can realize a smaller cell area per bit than the current EPROM. At present, 16 MB NAND devices are in production, and 32 MB NAND devices have been discussed [4.15, 4.16].

Finally, in the last few years, several other Flash cell structures and array architectures have been disclosed and are in various stages of development. For example, AND [4.57], DINOR [4.58], and HICR [4.68] employ F-N tunneling for programming in order to achieve low power programming. These cell structures and others are discussed in Section 4.4 of this chapter.

4.1. BASICS OF PROGRAM AND ERASE OPERATIONS

4.1.1 Channel Hot-Electron (CHE) Programming

In this section, we briefly describe the program/erase operations of the majority of Flash memories described in the literature.

TABLE 4.4 HISTORY OF FLASH MEMORY

1978	EAROM	Guterman et al. [4.37]	TI
1984	Flash memory	Masuoka et al. [4.3]	Toshiba
1985	Flash memory (256 KB)	Masuoka et al. [4.4], [4.5]	Toshiba
1985	Source-erase type Flash	Mukherjee et al. [4.7]	Exel
1987	Drain-erase type Flash (128 KB)	Samachisa et al. [4.10], [4.11]	Seeq, UCB
1987	NAND structure EEPROM	Masuoka et al. [4.57]	Toshiba
1987	Source-erase type Flash	Kume et al. [4.57]	Hitachi
1988	ETOX-type Flash (256 KB)	Kynett et al. [4.9], [4.18]	Intel
1988	NAND EEPROM	Shirota et al. [4.19]	Toshiba
1988	ETOX-type Flash	Tam et al. [4.20]	Intel
1988	ETOX-type Flash, reliability	Verma et al. [4.21]	Intel
1988	NAND EEPROM	Momodomi et al. [4.22]	Toshiba
1988	Poly-poly erase Flash	Kazerounian et al. [4.23]	WSI
1988	Contactless Flash	Gill et al. [4.24]	TI
1989	Contactless Flash (256 KB)	D'Arrigo et al. [4.25]	TI
1989	Gate-negative erase	Haddad et al. [4.12]	AMD
1989	NAND EEPROM (4 MB)	Momodomi et al. [4.26]	Toshiba
1989	ETOX-type Flash (1 MB)	Kynett et al. [4.27]	Intel
1989	Side-wall Flash	Naruke et al. [4.28]	Toshiba
1989	Contactless Flash	Gill et al. [4.29]	TI
1989	Punch-through erase	Endoh et al. [4.30]	Toshiba
1990	Well-erase	Aritome et al. [4.31]	Toshiba
1990	NAND EEPROM	Iwata et al. [4.32]	Toshiba
1990	Contactless 1-T Cell Flash, ACEE	Riemenschneider et al. [4.33]	TI
1990	NAND EEPROM, well erase	Kirisawa et al. [4.34]	Toshiba
1990	FACE cell	Woo et al. [4.35]	Intel
1990	Gate-negative erase	Ajika et al. [4.36]	Mitsubishi
1990	Contactless Flash	Gill et al. [4.37]	TI
1990	Bipolarity write/erase	Aritome et al. [4.38]	Toshiba
1990	Negative gate, positive source erase	Miyawaki et al. [4.39]	Mitsubishi
1990	5 V-only NAND (4 Mb)	Tanaka et al. [4.40]	Toshiba
1990	1 Mb ETOX-type Flash	Seki et al. [4.86]	Hitachi
1990	Negative gate, positive source Flash	S. Haddad et al. [4.12]	AMD
1991	Virtual ground array with auxiliary gate	Yamauchi et al. [4.41]	Sharp
1991	5 V-only NAND (4 Mb)	Momodomi et al. [4.42]	Toshiba
1991	1-T Contactless 5 V-only Flash (4M)	McConnell et al. [4.43]	TI
1991	Stack Gate, high source impurity concentration cell (16 M)	Nakayama et al. [4.44]	Mitsubishi
1991	SCSG cell	Kuo et al. [4.45]	Motorola
1991	PB-FACE cell	Woo et al. [4.46]	Intel
1991	Burst-pulse erase	Kodama et al. [4.47]	NEC
1991	Sector-erase	Kume et al. [4.48]	Hitachi
1991	Flash cell, scaling	Yoshikawa et al. [4.49]	Toshiba
1991	Self-convergence erase	Yamada et al. [4.50]	Toshiba
1991	SSW-DSA 64 M Flash	Kodama et al. [4.119]	NEC

TABLE 4.4 HISTORY OF FLASH MEMORY (*Cont'd*)

1992	MEIT	Hori et al. [4.51]	Matsushita
1992	Polysilicon TFT	Koyama [4.52]	NEC
1992	2-Step erase	Oyama et al. [4.53]	NEC
1992	Bipolarity erase	Endoh et al. [4.54]	Toshiba
1992	RT ONO Flash	Fukada et al. [4.55], [4.56]	Oki
1992	FNT prgm/erase Flash	Kume et al. [4.57]	Hitachi
1992	DINOR	Onoda et al. [4.58]	Mitsubishi
1992	HIMOS	Keeney et al. [4.59]	NMRC/ IMEC
1992	SCSG cell	Kuo et al. [4.60]	Motorola
1992	Negative gate, positive source cell with row decoder in triple well (16 M)	Umezawa et al. [4.61]	Toshiba
1992	5 V-only Flash with DSA drain and channel erase	Jinbo et al. [4.62]	NEC
1992	5 V-only sector erase 16 M Flash	T. Jinbo et al. [4.120]	NEC
1992	Serial Flash (9 M)	Mehrotra et al. [4.63]	Sandisk
1993	Virtual ground array DINOR	Ohi et al. [4.64]	Mitsubishi
1993	NOR virtual ground Flash cell	Bergemont et al. [4.65], [4.66]	National
1993	3-D sidewall Flash cell	Pein et al. [4.67]	Stanford/ Phillips
1993	HICR cell	Hisamune et al. [4.68]	NEC
1994	3.3 V ETOX cell (16 M)	Baker et al. [4.8]	Intel
1994	3 V-only DINOR Flash (4 M)	Kobayashi et al. [4.69]	Mitsubishi
1994	Gate negative, source positive, erase Flash (16 M)	Atsumi et al. [4.70]	Toshiba
1994	3.3 V single-power supply HiCR Flash (64 M)	Takeshima et al. [4.71]	NEC
1994	Row redundancy (16 M)	Mihara et al. [4.72]	Mitsubishi
1994	High-speed programming (64 M)	Tanaka et al. [4.73]	Hitachi
1994	Serial Flash 18 MB	Lee et al. [4.74]	Sandisk
1994	Over-erase Detection	Miyawaki et al. [4.75]	Mitsubishi
1994	Quick Boosting Pump	Tanzawa et al. [4.76]	Toshiba
1994	Scaling tunnel oxide	Watanabe et al. [4.77]	Toshiba
1994	Source-side injection 5 V-only	Ma et al. [4.78]	Bright/ Sharp
1994	Subhalf micron NOR Flash cell	Mori et al. [4.79]	Toshiba
1994	Fully depleted floating gate	Sato et al. [4.80]	Sharp
1994	Buried channel Flash memory	Oda et al. [4.81]	Mitsubishi
1994	3 V-only S-injection	Kianian et al. [4.82]	SST
1994	TEFET-Trench Flash EEPROM	Kuo et al. [4.83]	Philips

Channel hot-electron (CHE) programming is achieved by applying high voltage to both drain and gate simultaneously [4.6]. With the control gate high, the high voltage across the drain to source gives a high channel current and channel field which generates hot electrons. The high voltage on the control gate couples a voltage

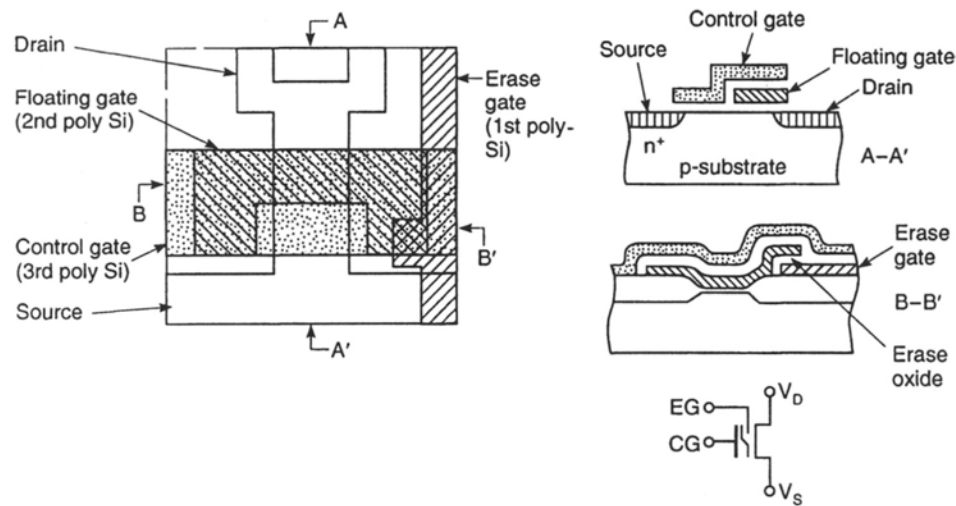


Figure 4.5 Top and cross-sectional view of the Flash EEPROM cell [4.3].

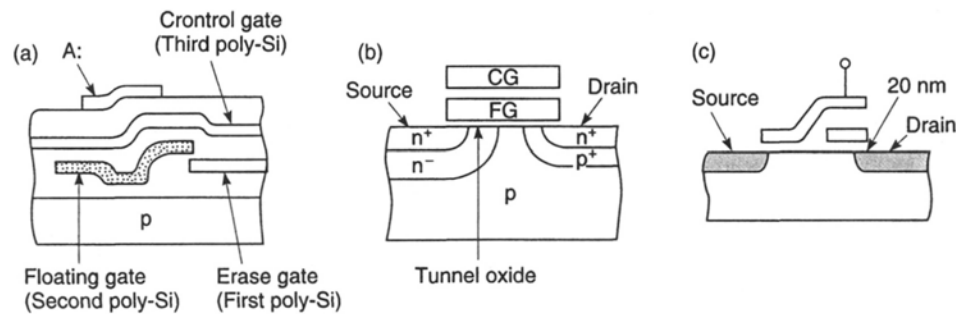


Figure 4.6 Stacked gate Flash EEPROM cell variations: (a) [4.3], (b) [4.7], and (c) [4.10].

to the floating gate and attracts hot electrons to the floating gate. Other cells along the same wordline and bitline are not programmed because no hot electrons are available with only the gate or the drain high. This programming mechanism is well proven, and EPROMs are known to be reliable and manufacturable. Program-disturb conditions in EPROMs are well understood and, in a well-designed EPROM, are insignificant. It is because of these advantages that early Flash memories used an EPROM-like structure [4.3, 4.9].

4.1.2 Source-Side Hot-Electron Programming

Instead of programming on the drain side, it is possible to design a cell that makes use of a narrow gap produced by a stepped gate or poly spacer on the source side to obtain a local high field for programming. For programming, both the floating gate and the stepped gate or poly spacer are biased with positive voltages

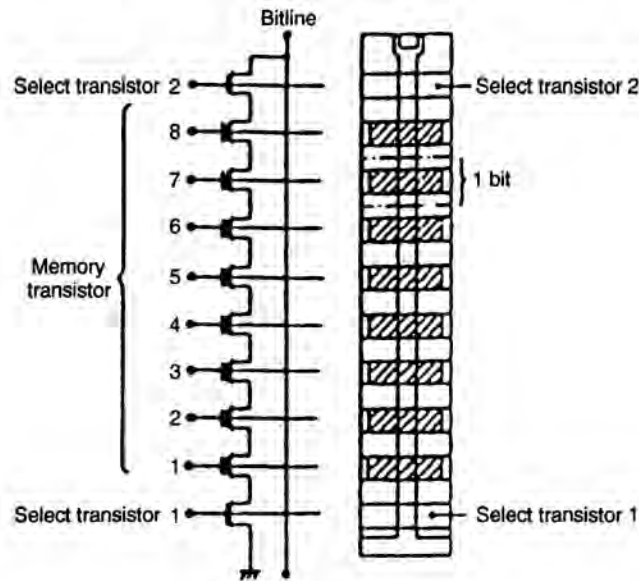


Figure 4.7 Top view and equivalent circuit of the NAND EEPROM [4.13].

to turn on the channel underneath. As a result, most of the source to drain voltage is dropped across the narrow gap. It is important to have the floating gate coupled to a higher positive potential compared to the stepped gate or poly spacer. The hot electrons generated are transported to the floating gate. This hot-electron process is more efficient, compared to channel hot electrons, requiring programming currents an order of magnitude lower. Variations of this technique have been used for a number of different cell structures [4.28, 4.41, 4.78, 4.82, 4.88]; for an example, see Fig. 4.8.

4.1.3 Fowler–Nordheim (F–N) Tunneling

Fowler–Nordheim (F–N) tunneling has been discussed in Chapter 2 with regard to the FLOTOX EEPROM. We will give a brief description here. Tunneling is a process whereby electrons are transported through a barrier. For silicon dioxide, with a 3.2 eV barrier, significant tunnel current can be observed when the oxide thickness is reduced to less than 4 nm. For thicker oxide, it is still possible to have a significant tunneling current when a high field of > 10 MV/cm is applied across the oxide, turning the barrier from a trapezoidal barrier into a triangular barrier of small tunnel distance. The tunneling process in oxide was first reported by Fowler and Nordheim and has the following functional form [4.160]:

$$I = A E^2 \exp(-B/E) \quad (4.1)$$

Fowler–Nordheim tunneling is the key transport mechanism for the FLOTOX EEPROM.

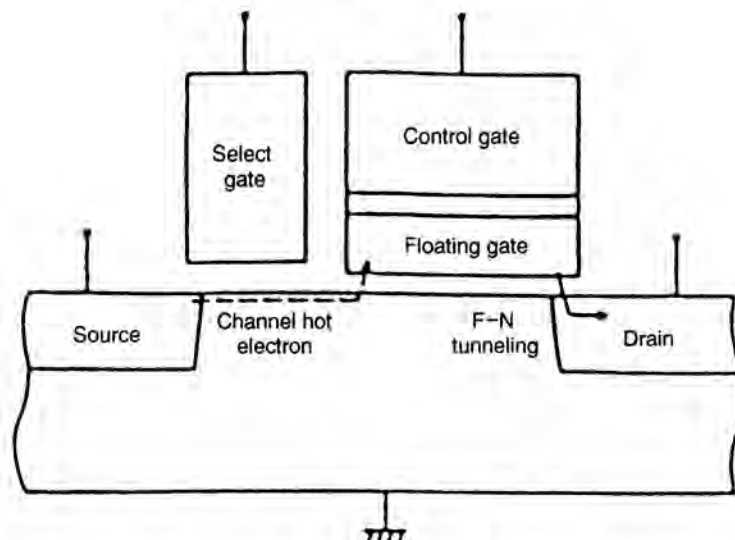


Figure 4.8 Schematic cross-sectional view of SISOS cell [4.28].

4.1.4 Tunnel Programming

Instead of channel hot-electron programming, we can program (electrons into a floating gate) a floating gate cell by Fowler–Nordheim tunneling by applying a high voltage to the top control gate, which couples a high voltage to the floating gate. For a tunnel oxide thickness of about 10 nm, the control gate voltage is 16 to 20 V. The program mechanism is the same as for the FLOTOX EEPROM. Examples of Flash memories based on F–N tunnel programming are: NAND [4.14], DINOR [4.58], AND [4.57], and ACEE [4.25].

As discussed previously, tunnel programming is a half-select process, and historically, a select transistor in series with the floating gate has been used to avoid disturb. More recently, the need for a select transistor has been obviated in tunnel programmable Flash memories by on-chip V_{pp} trimming, array sectoring techniques, and erase control techniques [4.14, 4.57, 4.58, 4.68].

4.1.5 Tunnel Erase Through Thin Oxide

When the first gate oxide in the memory cell is a thin tunnel oxide, it is possible to erase (electrons out of the floating gate) the memory cell by F–N tunneling through the tunnel oxide. The erase process can be accomplished in different ways. With the control gate grounded, erase can be accomplished by applying a high voltage to the source [4.7], to the drain [4.10], or to the substrate in a well [4.14]. It can also be accomplished by applying a negative voltage to the control gate with the other terminals connected to different positive voltages [4.12, 4.65, 4.78].

4.1.6 Tunnel Erase Through Poly-to-Poly Oxide

Instead of erasing to the silicon through the thin tunnel oxide, it is also possible to erase to the poly through the poly-to-poly oxide. Because of texturing in poly-oxide, it is possible to obtain significant tunnel current even with a thicker oxide [4.3, 4.63]. Although the poly-to-poly erase provides flexibility in the memory cell design, it requires a considerably higher voltage compared to the erase operation through thin oxide.

4.2. FLASH MEMORIES WITH CHANNEL HOT-ELECTRON (CHE) PROGRAM AND TUNNEL OXIDE ERASE

A schematic description of Flash cell programming by CHE is shown in Fig. 4.9. Some of the array architecture and erase conditions of channel hot-electron programmable Flash memories described in the literature will be discussed here. Programming is from the drain in a manner identical to that for an EPROM, and the cell is programmed to a high threshold state as shown in Fig 4.10.

4.2.1 Positive Source Erase

Tunnel erase is achieved by making use of the source to the floating gate overlap area as shown in Fig. 4.11. When the top gate is grounded, a high voltage applied to the source region (with the other junction floating) will pull electrons out of the floating gate by tunneling. The cell is erased to a low-threshold state. A graded source junction allows the application of a high voltage [4.17].

Under a positive source erase condition, the source junction generates significant substrate current. With a sufficiently high voltage, the breakdown process draws significant current and acts as a voltage clamp, limiting any further increase of the junction voltage. When the substrate current is sufficiently high, hot holes generated by the breakdown may start to erase the memory cell. Hot-hole erase is a process that is difficult to control and is avoided in a well-designed memory cell.

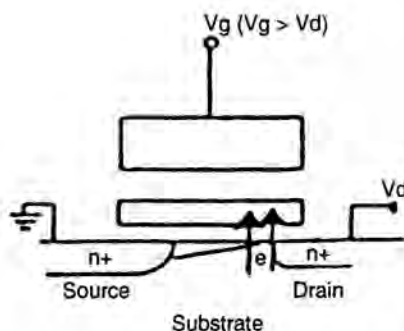


Figure 4.9 Schematic description of Flash EEPROM programming by CHE injection [4.14].

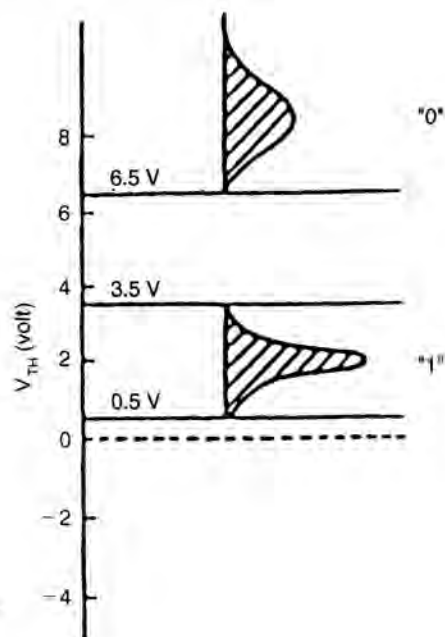


Figure 4.10 V_{th} distribution in programmed and erased state [4.14].

The source junction is graded by the addition of a phosphorous implant [4.7, 4.9, 4.17, 4.18, 4.20, 4.21, 4.27]. Also, because erase and program occur over different areas of the memory cell, the stresses of program and erase are spread over different regions, resulting in improved reliability. Since the tunnel area is the small source to floating gate overlap determined by lateral dopant diffusion, the coupling ratio for electrical erase is low and there is less need for a high coupling area over the field oxide, favoring a small and compact cell design. Finally, since the source is common to all the memory cells in the memory array, all the cells are erased simultaneously, providing block erase.

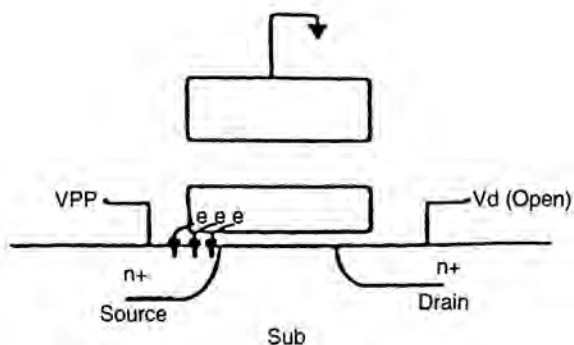


Figure 4.11 Schematic description of Flash EEPROM electrical erase by Fowler-Nordheim tunneling of electrons from the floating gate to the source [4.7, 4.9, 4.14, 4.17, 4.18, 4.20, 4.21, 4.27].

4.2.2 Positive Drain Erase

A split gate, CHE program, drain erase cell is shown in Fig. 4.12 [4.10, 4.11, 4.87].

If it is necessary to erase a smaller number of memory cells using source erase, the sources of the memory cells need to be separated. It is relatively simple to do a small number of sub-blocks by dividing up the array, but it is very area-intensive to do a large number of small blocks. Drains of memory cells are common in a given column, and erasing from the drain (Fig. 4.12) will give natural column blocking. However, since the junction is now used for both program and erase, it can only be optimized for one, and typically, it has to be optimized for erase because of the higher voltage requirement. As a result, programming performance is sacrificed. Second, now the junction is subjected to both program and erase stress so there is also more damage with cycling. Consequently, the cycling performance is limited compared to the case of separate program and erase junctions.

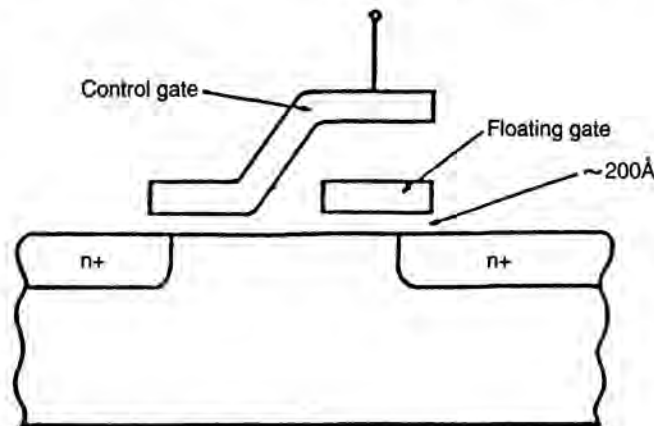


Figure 4.12 Drain erase Flash memory cell [4.10, 4.11, 4.87].

4.2.3 Negative Gate Erase

4.2.3.1 Negative (Control) Gate, Floating Gate to Source Erase. As mentioned previously, both source erase and drain erase suffer from a large substrate current that is not efficient for low power operation. However, the same erase field can be developed across the tunnel oxide by applying a negative voltage to the control gate. The source and drain can either be grounded or kept at a lower voltage, for example, V_{cc} . However, generation and switching of negative voltages increase circuit complexity.

The negative control gate, floating gate to source erase approach with the gate at a high negative voltage and the source typically at V_{cc} is shown in Fig. 4.13 [4.12, 4.48, 4.101, 4.119]. In this approach, the F-N erase current remains constant, and the adverse effects from holes generated by band-to-band tunneling are reduced.

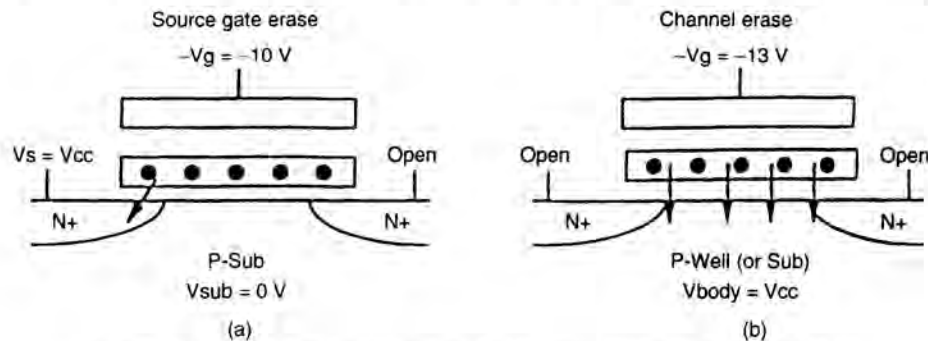


Figure 4.13 Cross-sectional drawing of (a) negative gate, floating gate to source erase, (b) negative gate, floating gate to channel erase [4.104].

4.2.3.2 Negative (Control) Gate, Floating Gate to Drain Erase. A triple-poly, split-gate Flash memory cell based on source-side injection for programming and negative control gate for erase (electrons tunnel from floating gate through the oxide to the drain) is shown in Fig. 4.14 [4.78]. The cell consists of a pair of stacked gates and a polycide blanket split gate. The select gate controls the weakly-on region during programming and prevents unselected cells from conduction in the program and read modes. The select gate runs along the channel direction. This eliminates the need for below-feature-size poly, reduces select gate delay, and improves the speed over select gate cells of the side-wall type [4.28, 4.88]. The use of a select gate allows the cell to operate in the depletion mode. To achieve a high read current in the erased state, the cell is put into the depletion mode.

The source-side reverse read is used to eliminate read-disturb induced by hot electrons. Figure 4.15 shows the array schematic and its operating conditions. To avoid high-voltage stress in the tunnel oxide during write, the high-voltage control gate (V_{pg}) and bitline (V_{cc}) are arranged to run parallel so that, in the program mode, both drain voltage and control gate voltage are present in the selected column. The self-imposed counterbiasing effect reduces the write-disturb.

The bitlines consist of source-lines and drain-lines arranged in alternating columns. An erase sector is naturally formed along each drain-line without any layout overhead. Each sector consists of two columns of opposing cells that share a common control gate.

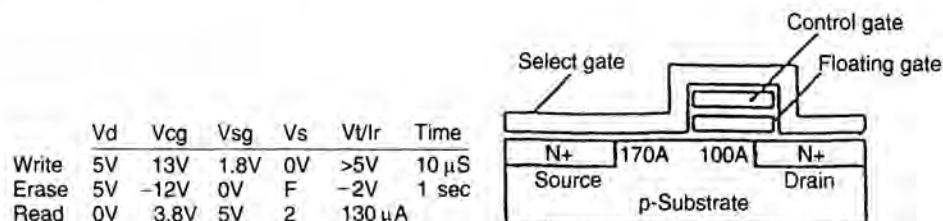


Figure 4.14 Cell cross-sectional view and operating conditions [4.78].

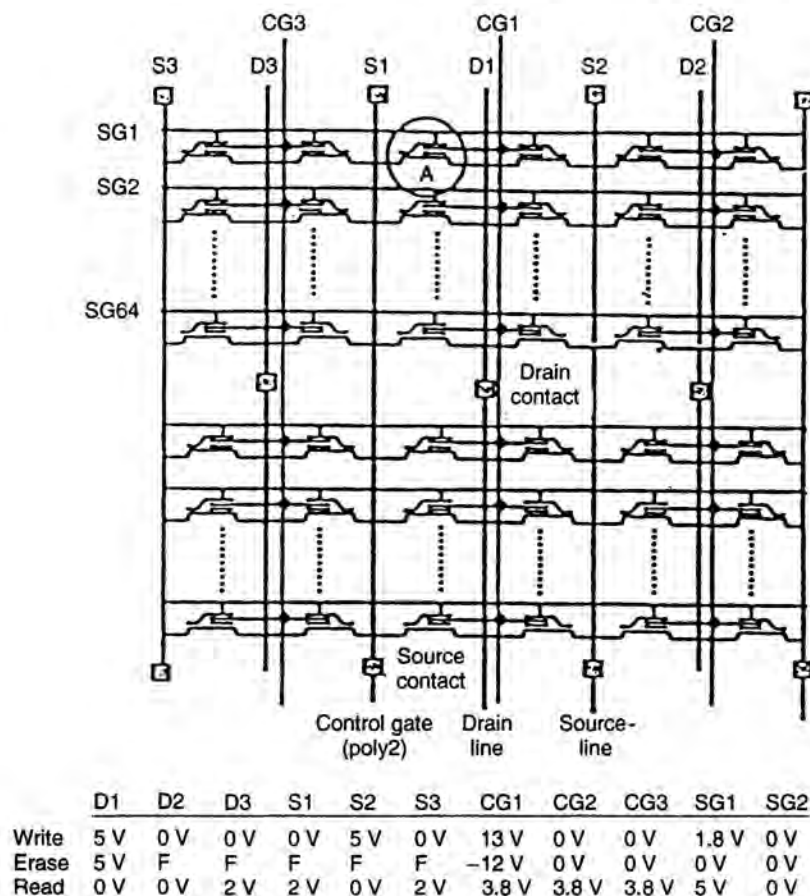


Figure 4.15 Array schematic and selection condition for cell "A" [4.78].

4.2.3.3 Negative (Control) Gate, Floating Gate to Channel Erase. The Nor Virtual Ground (NVG) array concept [4.65, 4.66] combines the Alternate Metal Gate (AMG) programming [4.157] with floating gate to channel erase [4.47]. The floating gate to channel erase avoids the problems associated with source erase discussed previously.

The erase concept operation is shown schematically in Fig. 4.16. The array schematic is shown in Fig. 4.17, and the array operation in Fig. 4.18.

The NVG array concept minimizes the drain turn-on during programming by the presence of a pass gate between the source and the bitline in a block. A column is segmented into blocks by the pass gate.

To achieve erase inhibit, the array well is biased to V_{cc} , necessitating a well-within-well structure. The floating gate to channel erase (F-N tunneling) requires a high electric field between the floating gate and channel. This is provided by a combination of a negative voltage on the control gate and a positive voltage on the array substrate. This necessitates an array well isolated from the substrate.

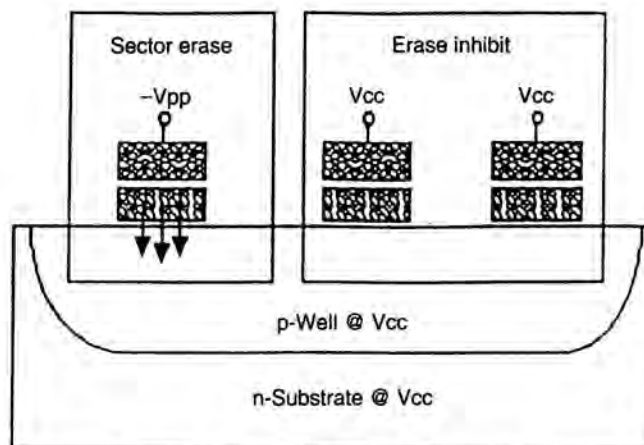


Figure 4.16 Sector erasing scheme [4.65].

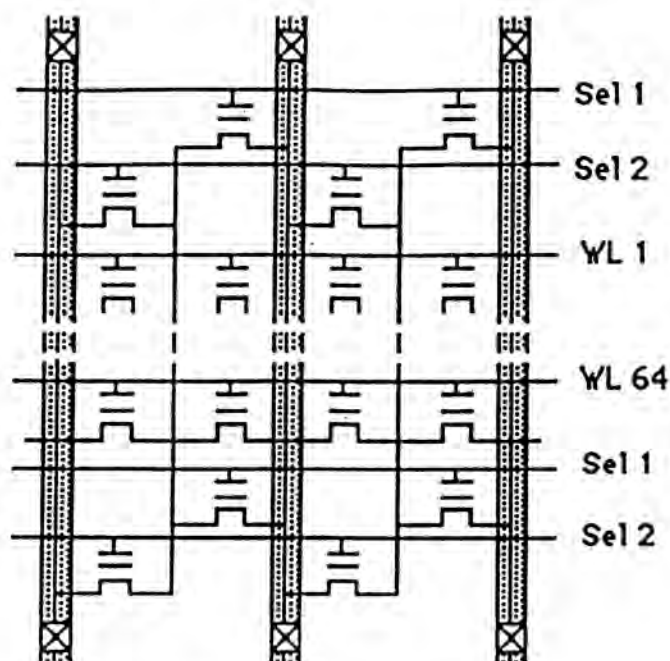


Figure 4.17 NVG Flash array schematic [4.65].

4.2.4 Erase Threshold Control

Electrical erasure presents new challenges compared to UV erase. First, instead of a natural self-limiting UV process, the electrical erase can continue beyond the neutral point to give a net positive charge on the floating gate and a net negative

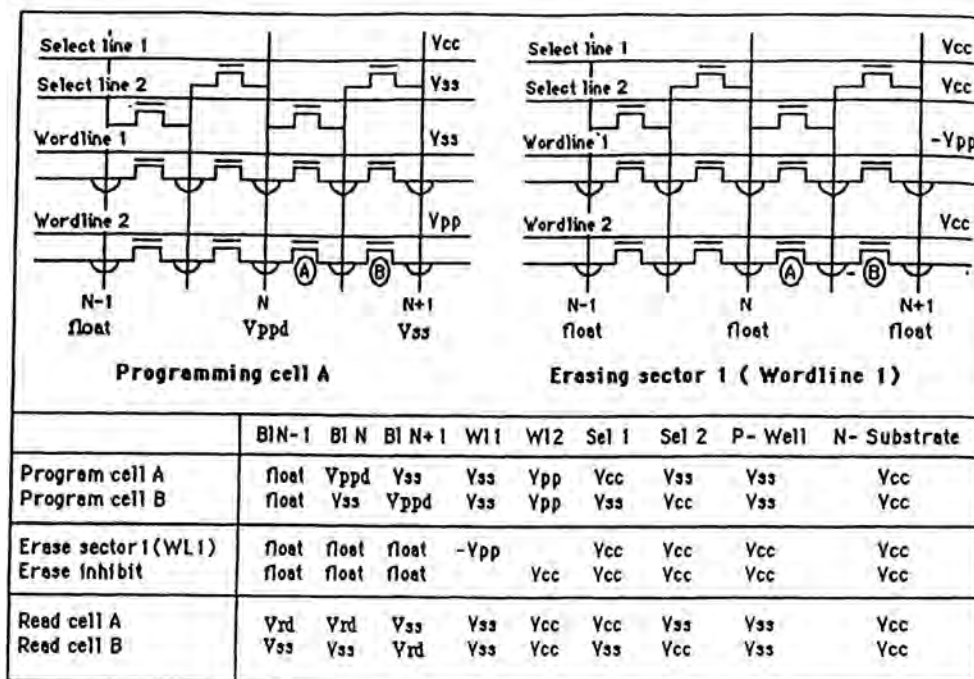


Figure 4.18 NVG Flash operations table [4.65].

threshold voltage. In the simple EPROM-like array structure, a negative threshold can short out columns, giving false ones for any column read. In the worst case, the conducting cell can draw so much current that it is not possible to program other cells in the same column. Second, any cell-to-cell variation, like coupling ratio and tunneling probability, will change the final threshold of the cell. This potentially results in a much wider threshold distribution after electrical erasure unless the process is well controlled. Several approaches have been proposed to solve this over-erase problem [4.9, 4.27, 4.28].

The verified-erase method [4.27] is the key method for controlling placement of erase cells (Fig. 4.19). An erase step is carried out by applying a suitable combination of voltages to source junctions and control gates. Subsequently, a read operation is performed with a read voltage of, for example, 3.2 V, applied to the control gate. This 3.2 V is to be the upper limit for the threshold voltage of a cell in the erased state. It is derived from a worst case read voltage of $V_{cc} = 4.5$ V and a gate drive of 1.2 V for sufficient read current. If some bits require more time to reach the erased state, erasing is performed again. The erase verify sequence is repeated until all cells in the array have a threshold voltage less than or equal to the 3.2 V maximum. This verified erase method has been shown [4.27] to be very effective in accurately placing (locating) the threshold of erased cells.

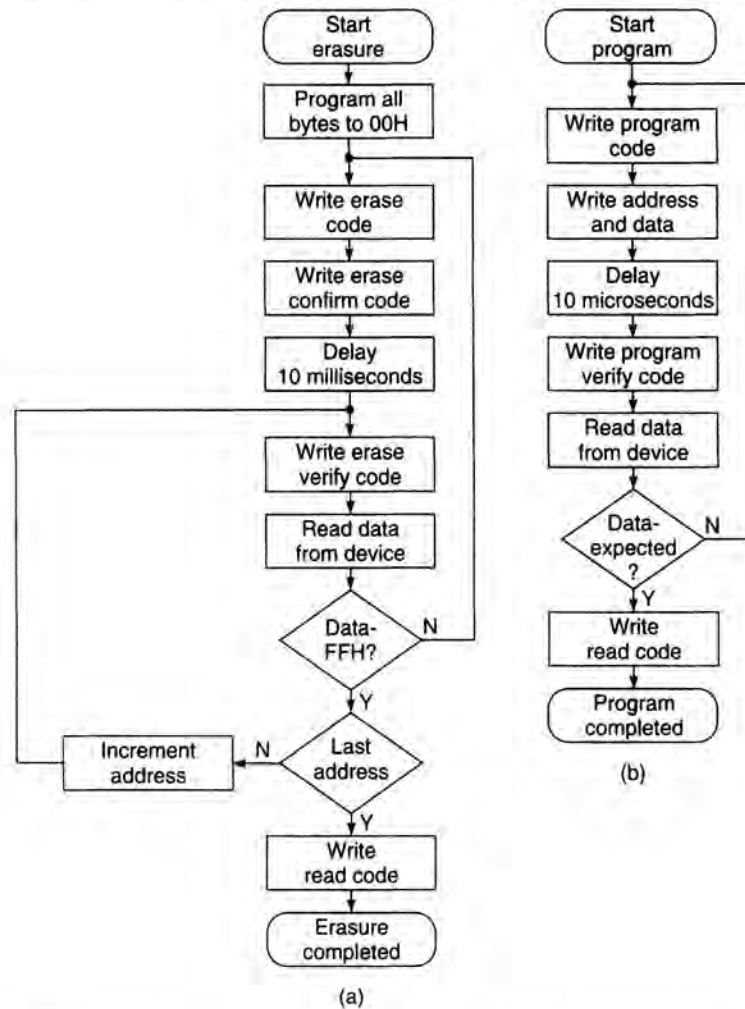


Figure 4.19 (a) Erase algorithm for source erase V_t control [4.27];
(b) program algorithm [4.27].

On the low-threshold end, the V_t limit is determined by the amount of cell leakage current under worst case conditions so that there won't be a false read on other cells. The low-end threshold is not verified by the algorithm, but is guaranteed by process control and product testing. Any arrays with leaky cells are either corrected by redundancy or discarded as bad. The current process can give erase times as short as 100 msec to as long as 30 secs, with the tunnel oxide thickness being the dominant factor in determining the erase time (Fig. 4.20). A typical erase time is 1 sec.

In the erase algorithm, if there are memory cells that are never programmed, the cells will be erased repeatedly over many cycles. Given enough cycles, these cells will eventually become leaky cells. In order to prevent this problem, all cells need to be pre-programmed before erase. The pre-program resets all cells to a programmed

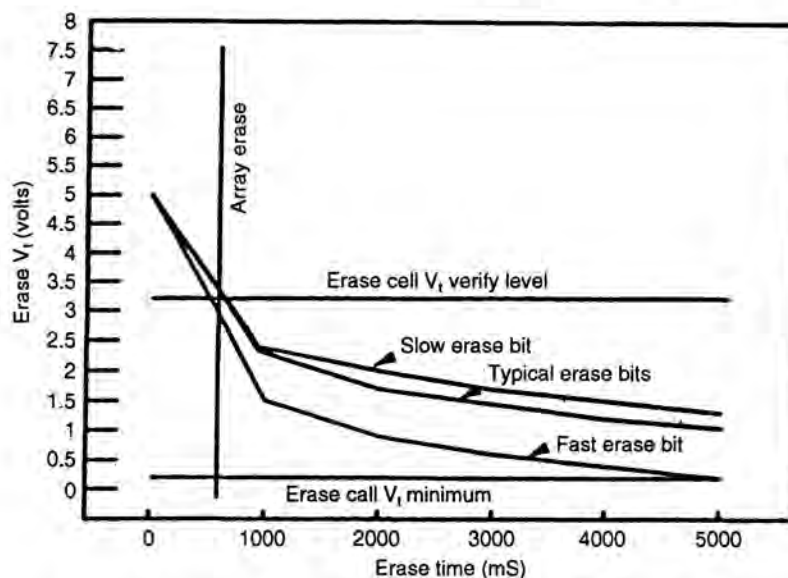


Figure 4.20 Array erase V_t profile as function of erase time [4.27].

charge state and assures good threshold control over a large number of write/erase cycles.

As noted previously, a series enhancement transistor, a merged passed gate—for example, Fig. 4.43 [4.24], Fig. 4.12 [4.10], Fig. 4.40 [4.82], or a side-wall select gate Fig. 4.8 [4.28]—can be used to prevent the leakage current. However, any of these approaches increases the cell size, as well as the process complexity.

4.2.5 Erased V_t Spread Reduction

The erase algorithm described previously controls the placement of the erased cell V_t s. It does not change the spread of the V_t s. The erased V_t spread can be reduced from both process optimization and cell-operating conditions.

4.2.5.1 Post-gate Oxide Poly Process. The fluence, or Q_{bd} , for a tunnel oxide versus annealing time for various annealing temperatures is shown in Fig. 4.21 [4.128]. More recently, careful material analysis studies [4.129, 4.130] have revealed that local oxide thinning (Fig. 4.22) and oxide surface roughness at the poly-oxide interface (Fig. 4.23) are greatly influenced by post gate growth temperature treatment. These are the direct consequence of the grain growth of the polysilicon gate and viscous flow of the oxide, which are enhanced with increasing annealing temperature and time. In addition, poly gate dopant level and species can play a major role in the local variation in tunneling current (Fig. 4.24) [4.130]. Phosphorous atoms segregated at the grain boundaries of the poly gate were found to diffuse to the poly-oxide interface and react with the oxide to form a phosphorous-rich SiO_2 , or so-

called oxide ridges, as shown in Fig. 4.23. The phosphorous-doped SiO_2 , due to the change in stoichiometry, exhibits a barrier height lowering effect for tunneling (Figs. 4.25 and 4.26), thereby causing the F-N current to vary, depending on the amount of dopant incorporation. It was found that, in the sample implanted with phosphorus ($1 \times 10^{16} \text{ cm}^{-2}$) and annealed at 950°C , oxide ridges were not observed with a scanning electron microscope (SEM). Thus, post gate oxide process optimization can play a major role in the control of the erased V_t distribution.

4.2.5.2 Two-Step Erase [4.53]. Since the erase current has more variation when electrons are emitted from the poly- SiO_2 interface, a two-step erasing scheme employing wordline post-stress has been proposed [4.53]. In the first step, which is equivalent to the conventional channel erase, a negative high voltage is applied to the control gate of the cell (first F-N erasing step). In the second step, a positive high voltage is applied to program the cell back (the second F-N programming-back step) to decrease the erased V_t distribution. The scheme is based on the observation that the distribution of F-N current emitted from the substrate, or the channel, is narrower than that emitted from polysilicon (Fig. 4.27) for the reasons discussed previously. Regardless of the wide erased V_t distribution after the conventional erase (the first step erase), the F-N tunneling from the channel during the programming-back step (the second step erase) tends to tighten the V_t distribution (Fig. 4.28). It was demonstrated [4.53] that, by applying this two-step erasing scheme in a 16 KB array, the erased V_t distribution could be tightened from 2.0 V to 0.9 V (Fig. 4.29).

4.2.5.3 Self-convergence Erasing Scheme [4.50]. Another technique that employs drain or source post-stress to tighten the erased V_t distribution is called the self-convergence erasing scheme [4.50, 4.134]. It uses drain or source bias with the

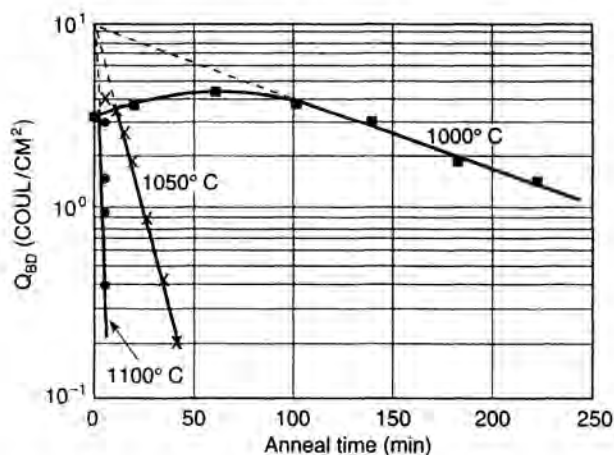


Figure 4.21 Q_{bd} versus annealing time for various annealing temperatures. Q_{bd} was measured using a current density of 50 mA/cm^2 (gate injection). $T_{ox} = 80 \text{ Å}$ [4.128].

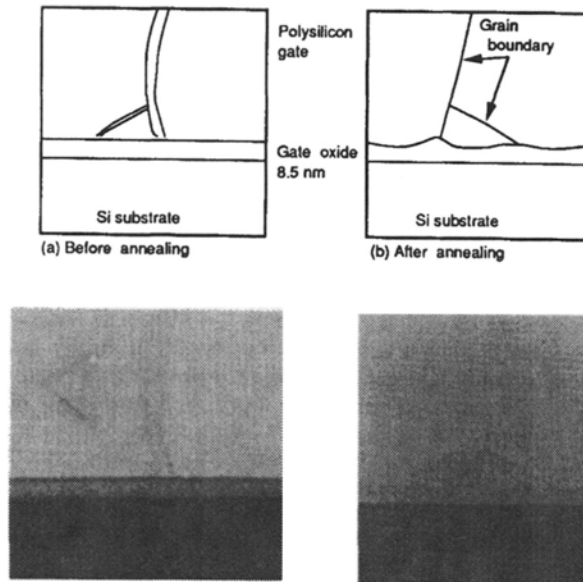


Figure 4.22 Cross-sectional TEM photographs of the polysilicon gate electrode/ultra thin tunneling oxide interface (a) before and (b) after the post-annealing [4.129].

other terminal grounded to produce channel hot-carrier injection to tighten the V_{ts} after the conventional erase. Figure 4.30a presents typical characteristics of the channel hot-carrier-induced gate current versus V_{fg} of a typical cell in which V_g^* corresponds to the saddle point where channel hot-hole injection is in balance with channel hot-electron injection. Using this device phenomenon, cells starting with

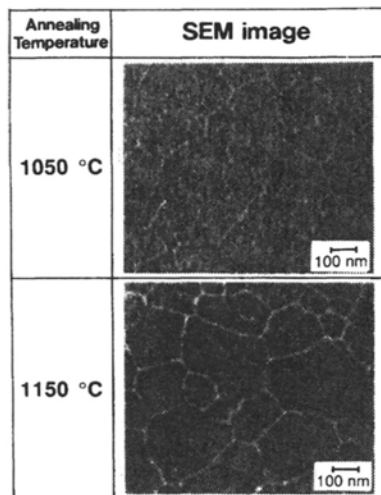


Figure 4.23 Oxide ridge structure dependence on annealing temperature (phosphorus dose: $1E16 \text{ cm}^{-2}$) [4.130].

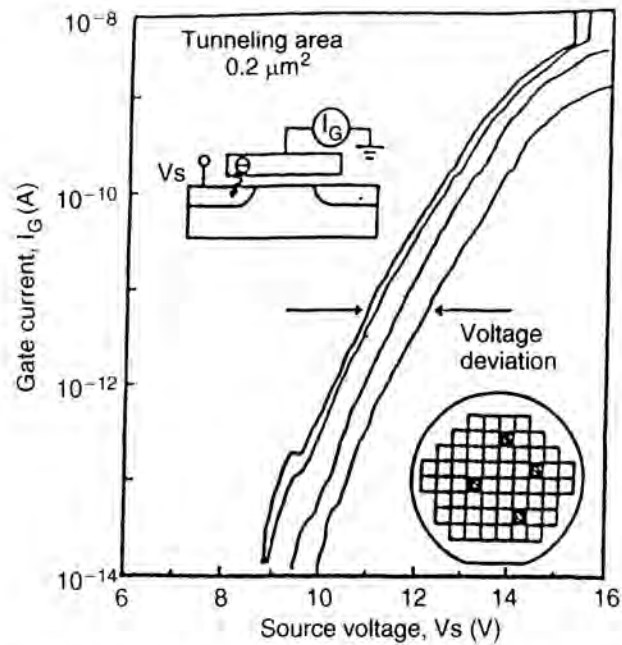


Figure 4.24 Relationship of gate current to source voltage (annealing temperature: 1100°C , phosphorus dose: $1\text{E}16 \text{ cm}^{-2}$) [4.130].

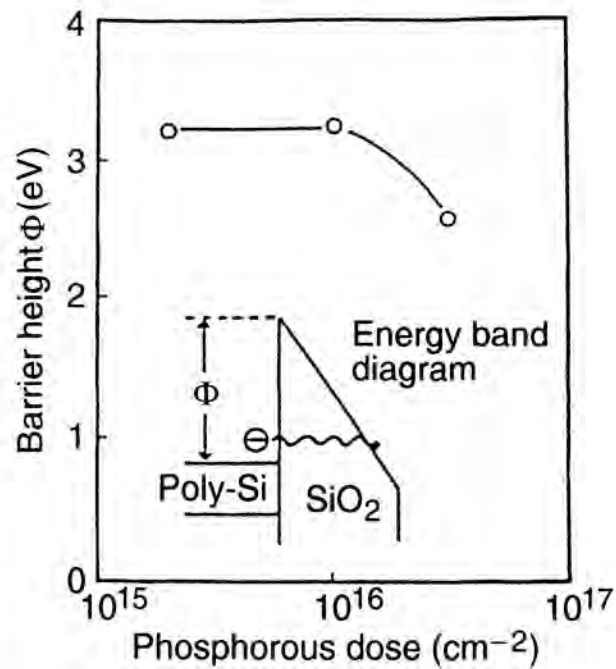


Figure 4.25 Barrier height dependence on phosphorous dose (annealing temperature = 1050°C) [4.130].

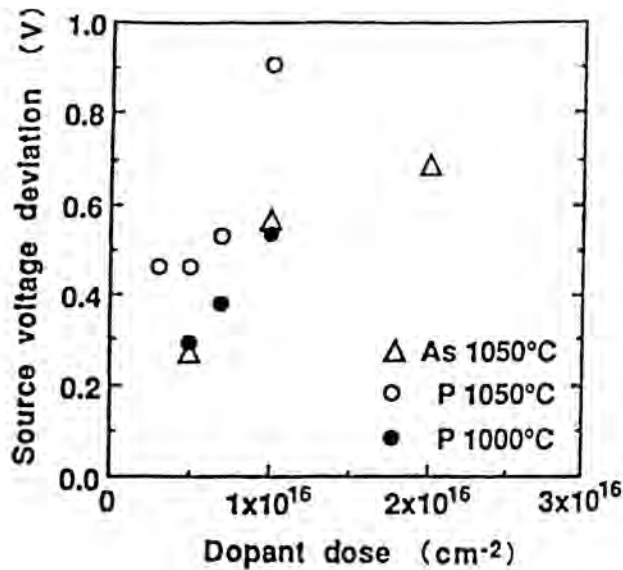


Figure 4.26 Voltage deviation dependence on annealing temperature, dopant dose, and dopant species [4.130].

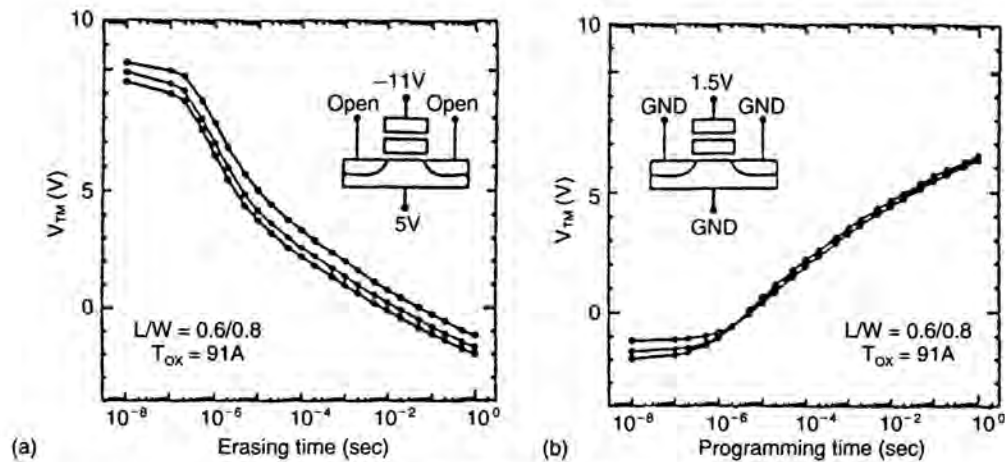


Figure 4.27 Three typical curves of F-N erasing characteristics (a) electrons emitted from the poly-SiO₂ interface, (b) electrons emitted from the Si/SiO₂ interface [4.53].

different erased V_t s will converge to a common threshold voltage, V_{th}^* , with an appropriate drain bias (Fig. 4.30b) [4.50]. The self-convergent V_{th}^* is a function of channel doping, and the total voltage is capacitively coupled to the floating gate. Figure 4.31 demonstrates the tightening of erased V_t s before and after the source voltage is used for convergence [4.50].

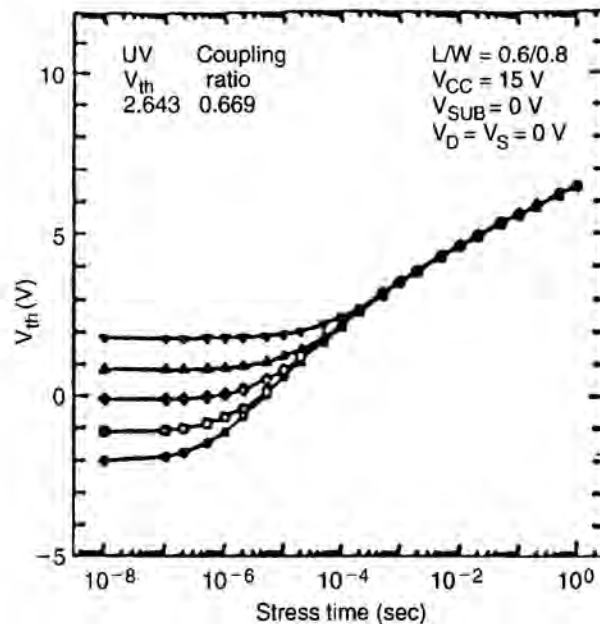


Figure 4.28 F-N programming (electrons emitted from the Si/SiO₂ interface) characteristics of a cell. F-N programming characteristics have good convergence of the programmed V_{th} regardless of the initial V_{th} [4.53].

4.2.6 Program-Disturb Mechanisms

The two principal memory cell-disturb mechanisms that can occur during programming of an array are called gate-disturb (DC program) and drain-disturb (program disturb) [4.21]. These mechanisms can occur in memory cells sharing a common wordline (WL) or a common bitline (BL) while one of the cells is being programmed. The effect of these disturbs on the different cells of the memory array is shown in Fig. 4.32.

Gate-disturb occurs in unprogrammed or erased cells that are connected to the same wordline as the cell being programmed. These cells have a low cell threshold voltage. During the programming operation, the common wordline is connected to a high voltage. The electric field across the tunnel oxide becomes high and may cause tunneling of electrons to the floating gate from the substrate. The threshold voltage of the cell will increase and reduce the sense margin. In severe cases, the cell is programmed unintentionally.

Drain-disturb occurs in programmed cells that are on the same bitline as the cell being programmed. These cells will experience a high electric field between the floating gate and the drain. This may cause electrons to tunnel from the floating gate to the drain and lead to a reduced cell threshold voltage.

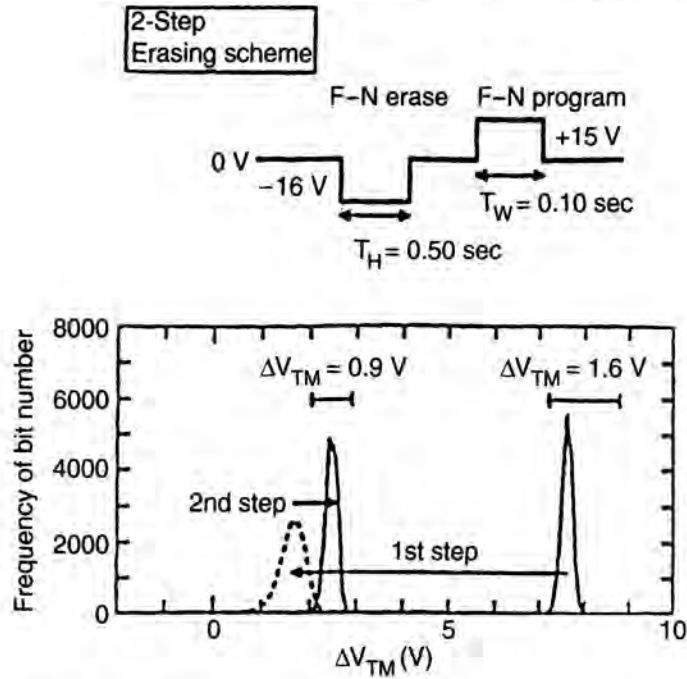


Figure 4.29 The erased- V_{tm} distribution of 16 Kbit cell array using two-step erasing. The erased V_{tm} distribution after two-step erasing operation is suppressed from $\Delta V_{tm} = 2.0 \text{ V}$ to 0.9 V in width [4.53].

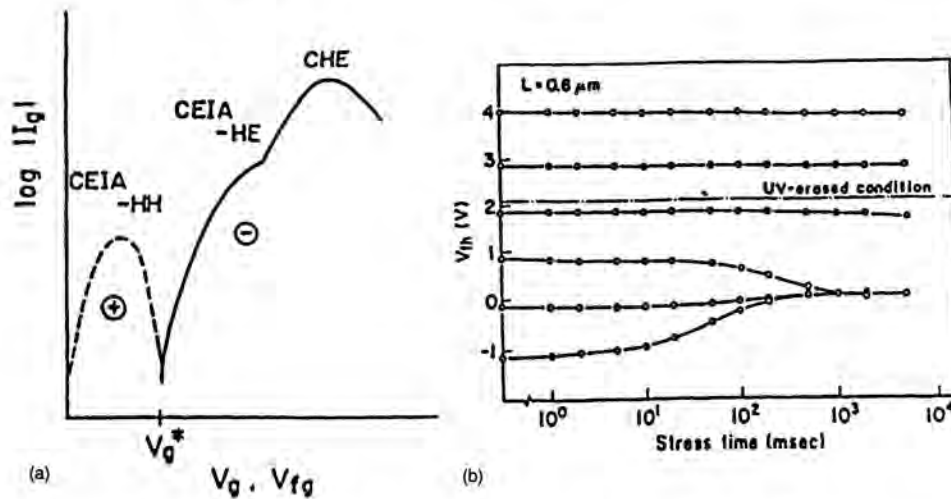


Figure 4.30 (a) Characteristics of gate current, I_g , versus gate voltage, V_g , for a NMOSFET, (b) threshold voltage, V_{th} , versus drain-stress time with different starting threshold voltages as parameters [4.50].

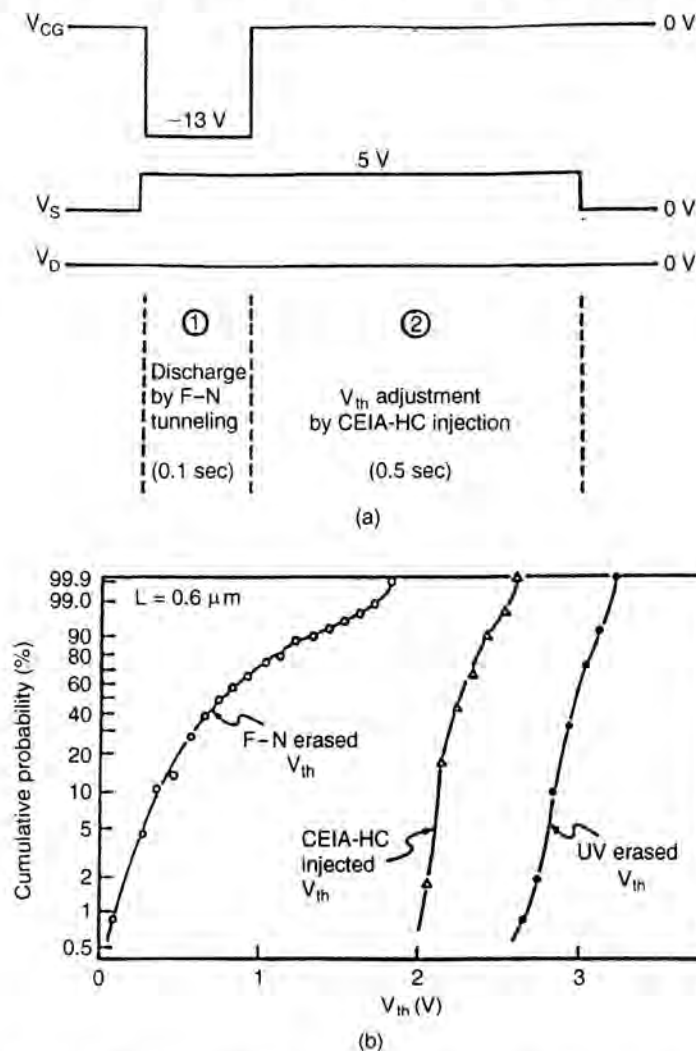


Figure 4.31 (a) Timing diagram used in experiments on new erase sequence, (b) threshold voltage distribution before and after adjustment by CEIA-HC injection [4.50].

An important design consideration is the proper selection of the programming voltages to minimize these disturbs. Write/erase cycling of the cell also affects these disturbs. Verma and Mielke [4.21] showed that the drain-disturb characteristics of Flash memory devices are excellent with no measurable change until several thousand cycles. However, write/erase cycling has an influence on the gate-disturb behavior. Figure 4.33 shows the gate-disturb time as a function of cycling. Before cycling, the disturb time is about 100 seconds. After 100 cycles, this margin is decreased about two to three orders of magnitude. This degradation in gate-disturb sensitivity is caused by hole trapping during erase [4.21].

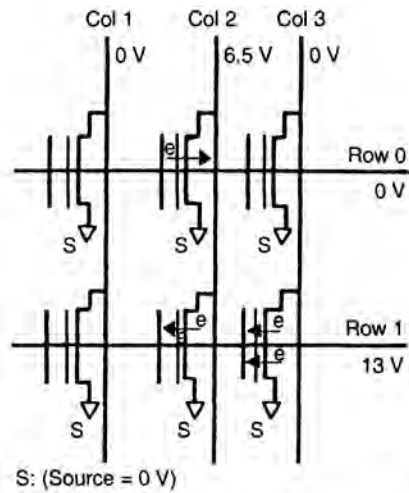


Figure 4.32 Schematic description of disturb during programming [4.21].

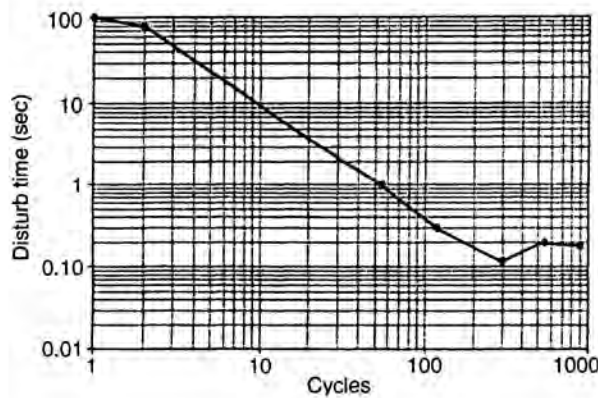


Figure 4.33 Gate-disturb time as a function of cycling [4.21].

4.3. FLASH MEMORIES WITH CHANNEL HOT-ELECTRON PROGRAM AND POLY-TO-POLY ERASE

Even though erasing through the source or drain does not require an extra transistor terminal, thereby giving the most compact cell, the use of junctions does create the problem of breakdown current. Also, a thin oxide is required for the cell gate oxide. Instead of erasing through the silicon, it is also possible to have an extra polysilicon

electrode that connects through an oxide to the floating gate. This extra poly electrode results in a slightly larger memory cell or a more complicated process.

The poly-to-poly erase (floating gate to control gate in a double-poly structure, or floating gate to an erase gate in a triple-poly structure) avoids stressing of the gate oxide during erase. The first modern Flash EEPROM [4.3–4.5] from Toshiba used poly-to-poly erase and will be described first, followed by a contactless, virtual ground array version from Sandisk. In both cases, erase occurs from the floating gate to an erase gate, although the locations of the erase gate in the cell are different.

4.3.1 Triple-Poly NOR

Triple polysilicon Flash memory cells [4.3–4.5] are shown in Figs. 4.34, 4.35, and 4.36. The select gate is integrated with the floating gate (merged pass gate). The first polysilicon layer is used as an erase gate and is located between the field oxide and the floating gate. The second polysilicon layer is used as a floating gate in a manner similar to an UV EPROM. The third polysilicon layer is used as a control gate and is, in fact, used as a bit select-line for programming and reading.

The cell is programmed by a channel hot-carrier injection mechanism similar to EPROM. The erasure is accomplished using the first level of polysilicon which serves as an erase electrode causing field emission of electrons from the bottom of the floating gate. In the triple-poly NOR cell, as shown in Fig. 4.34, one part of the channel is controlled by the control gate. The third polysilicon layer is used both as a gate of the selection transistor and a control gate of the triple-poly NOR cell. This selection transistor enables the triple-poly cell to remain in the enhancement mode even if the floating gate transistor becomes a depletion mode device after erasure. The erase gate of the memory cell is supplied with the boosted voltage (V_{EG}), which enables field emission from the floating gate. This boosted voltage (V_{EG}) is produced from the program voltage (V_{PP}). All of the memory cells are erased simultaneously by the erase gate because they are commonly connected to each other.

The read data depend on the stored electric charges on the floating gate similar to the case for EPROM. The erased memory cell (the floating gate not charged with electrons) conducts by applying voltage as shown in Fig. 4.36. The virgin memory cell, prior to being programmed, has the same threshold voltage as the erased memory cell. However, their g_m values are different. The threshold voltage of the F-E2PROM cell is controlled by the selection transistor. However, the g_m depends on the value of the stored electric charges in the floating gate, even if the memory cell is erased.

4.3.2 Triple-Poly, Virtual Ground Contactless [4.63]

The Flash memory cell uses a triple-polysilicon, single-metal, split-channel structure with buried n^+ diffused source-drain, integrated into a contactless, virtual ground array architecture. The buried diffusion bitlines are contacted periodically

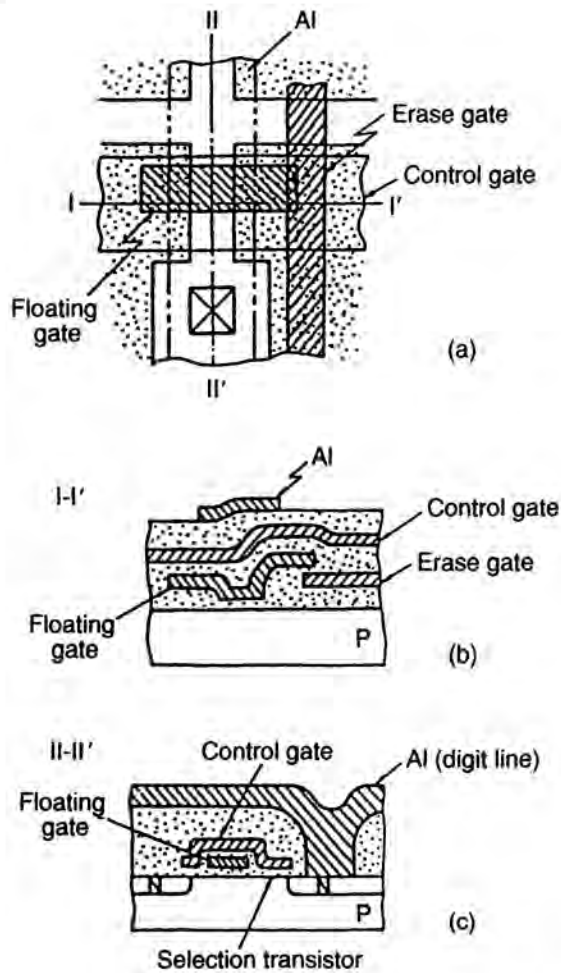


Figure 4.34 (a) Top view of F-E2PROM cell, (b) cross-sectional view along I-I' line in (a), (c) cross-sectional view along II-II' line in (a) [4.3–4.5].

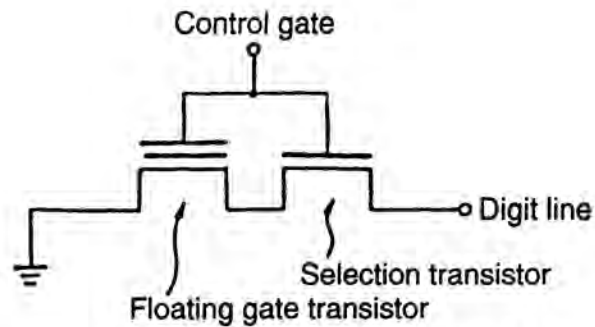


Figure 4.35 Equivalent circuit of Fig. 4.34 [4.3–4.5].

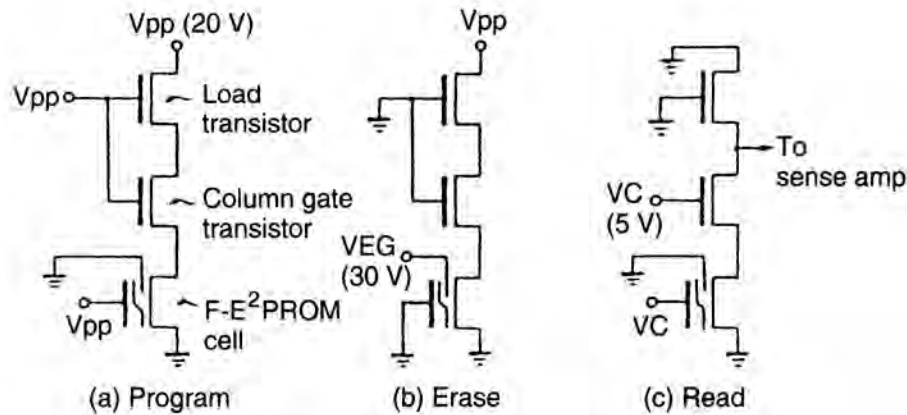


Figure 4.36 Operating conditions (a) program, (b) erase, (c) read [4.3–4.5].

with metal to reduce series resistance. The cell is programmed using channel hot-electron injection and is erased using inter-polysilicon dielectric tunneling. The split-channel memory transistor structure provides an integrated fixed enhancement threshold select transistor in series with the floating gate channel, allowing the floating gate to be erased to negative threshold conditions without introducing channel leakage current when unselected. This eliminates the program-before-erase requirement of 1-T cell Flash memories, greatly simplifies the erase operation, and relieves the very stringent process controls associated with maintaining a tight erase distribution on a large population of bits.

4.3.2.1 Cell/array Architecture. The structure of the cell, represented in Fig. 4.37, consists of three levels of polysilicon that form the floating gate, control gate, and erase gate, and uses conventional isolation. No oxide in the process is thinner than 25 nm. Employing a contactless, virtual ground array and using only a single contact for every 32 cells improves the array density. This reduces the cell size relative to nonvirtual ground cell/arrays, which use one contact for every two cells.

The memory array architecture is shown schematically in Fig. 4.38. In the virtual ground array, the source (S) diffusion of one column of cells is also the drain (D) diffusion of the adjacent column of cells. The memory array is organized into sectors of 576 bytes each—512 user data bytes (compatible with most mass storage operating systems) and 64 bytes containing sector overhead and ECC information. The erase gate (EG) is used for tunnel erase between it and the floating gates of all cells within a sector. The top view and a cross-sectional view of the cell-array are shown in Fig. 4.39.

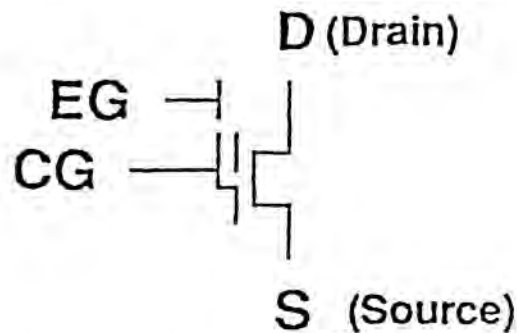
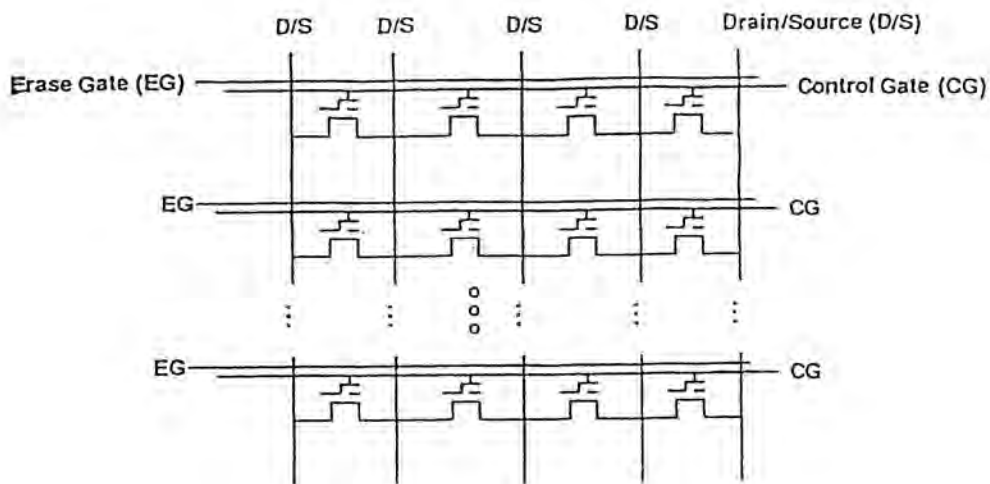


Figure 4.37 Flash memory cell representation [4.63, 4.158].

4.3.2.2 Programming. The conditions for programming are given in Fig. 4.38. To program, the control gate (CG) of a selected row (the wordline) is raised to 12 V, and the drain bitlines of the selected cells within this row are raised to 7 V (data conditional), while their source bitlines are held at 0 V. For cells whose data require them to be left in the erased state, the associated drain bitlines are held at 0 V, thereby preventing programming. The Flash cell is designed to program at relatively



	CG	EG	D	S
Program	12	0V	7V	0V
Erase	0V	12V to 22V	0V	0V
Read	5V	0V	1.5V	0V

Figure 4.38 Flash memory array configuration [4.63, 4.158].

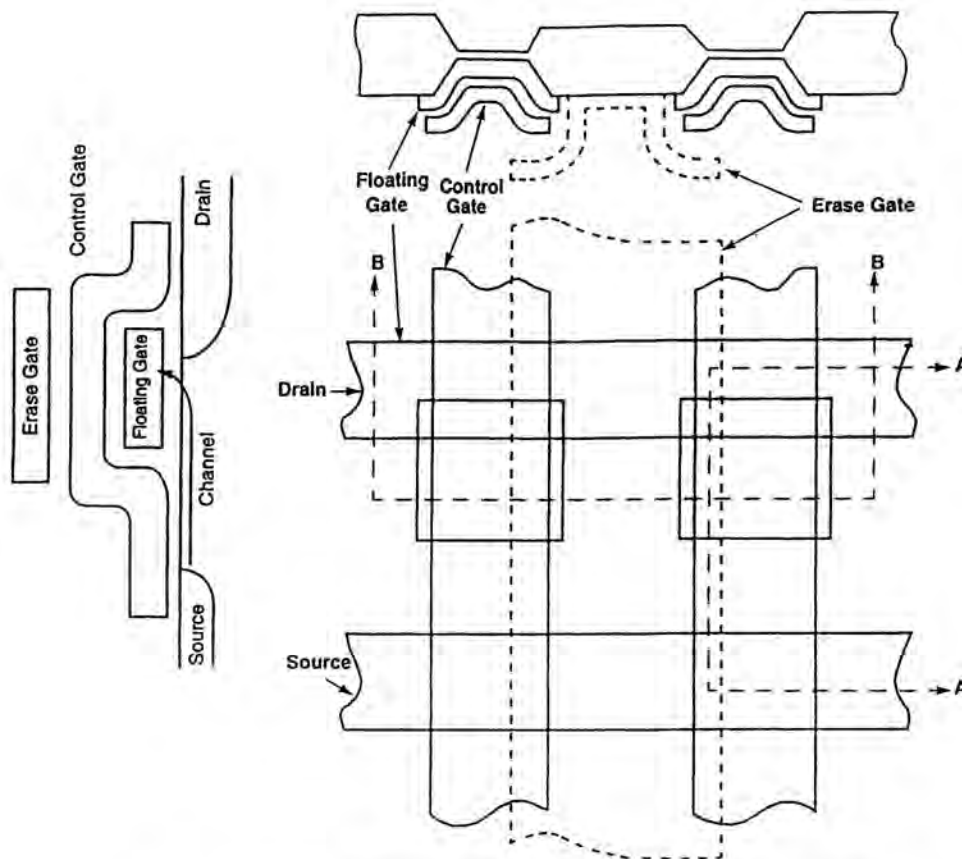


Figure 4.39 Memory array top view and cross-sectional views [4.63, 4.158].

low channel currents, thus allowing for programming 64 bits (termed a *chunk*) in parallel. A typical cell programming time is $5\ \mu\text{s}$, giving an effective programming time of $0.6\ \mu\text{sec/byte}$. An on-chip verify operation follows each programming operation to ensure full programming. While the majority of cells program in a single pulse, additional program/verify pulses are applied to each chunk, as needed, to accommodate the full programming of slower bits.

4.3.2.3 Erase. Cell erase uses polysilicon (floating gate side) to polysilicon (erase gate) tunneling through a relatively thick inter-polysilicon oxide to remove electrons from the floating gate. The smallest unit of erase is a sector. Typical sector erase time is 1 msec, giving an effective erase time of $2\ \mu\text{sec per byte}$. Because of the split-channel feature, there is no cell program-before-erase requirement and, therefore, no associated erase speed loss. The erase voltage, V_E , applied to the erase lines,

EG, ranges from 12 V to 22 V, depending on the sector requirements. The VE is automatically increased to compensate for charge trapped in the erase tunnel oxide during extended cycling. As with programming, a verify operation follows each erase operation.

To further reduce the effective erase time, simultaneous, random, multiple-sector erase capability is provided by a sector "tagging" feature that sets erase latches of selected sectors. During multiple sector, parallel erase, sectors verified to be properly erased are "untagged," thereby preventing further erasure, while other sectors in the group, needing additional erasure, continue the erase/verify sequence. In this way, erase oxide stress is minimized, achieving a sector endurance exceeding 100,000 program/erase cycles.

This cell/array approach offers improved erase performance and reliability by confining the relatively high erase voltages within the memory array to the erase polysilicon layer, rather than to diffusion regions within the substrate. This eliminates the relatively high erase currents and device degradation associated with high junction fields and leakages.

4.3.2.4 Read. Cell reading is accomplished by raising the selected wordline to 5 V and biasing the drain of the selected cell to 1.5 V via a current sense amplifier, with the source driven to 0 V. During read, a programmed cell will not conduct, resulting in a "1" state, while an erased cell will sink current, resulting in a "0" state. The 64 cells of the chunk are read simultaneously. The memory incorporates a 20 MHz, double input/double output path, serial interface with a read protocol designed to output data from the device at high rate under continuous read, with no intersector delay. While one chunk's data are being shifted out, the next chunk is addressed and sensed, absorbing all the associated memory access time, to give an effective read time of 200 nsec/byte.

4.3.2.5 Write Inhibit Conditions. Erasing of unselected sectors is inhibited by maintaining their erase-line voltage at 0 V. Programming of unselected sectors/rows is inhibited by maintaining their wordline voltage at 0 V. Programming of unselected cells in selected rows is inhibited by biasing approaches that limit the drain to source voltage differential, thereby eliminating channel hot-electron generation. A further advantage inherent in this split-channel cell over other Flash cells, which have only the floating gate channel, is the attribute of no cell programming when the bias conditions of its drain and source are interchanged. This facilitates integrating the cell into the virtual ground array, without incurring any disturb conditions under reverse drain-source bias conditions.

In all Flash memory arrays, unselected cells sharing wordlines or bitlines of cells being programmed are subjected to program-disturb conditions as their corresponding wordlines or bitlines are biased at programming voltage levels. In this Flash technology, no oxides are thinner than 25 nm, and the oxide fields associated with

the disturb conditions are lower than those which may exist in thin tunnel oxide technologies.

4.3.3 The Field-Enhancing Tunneling Injector EEPROM Cell

The field-enhancing tunneling injector EEPROM cell [4.82] is a double-poly, split gate memory cell using poly-to-poly (floating gate to control gate) F-N tunneling for erasing and source-side hot-electron injection for programming. Poly-to-poly tunneling is from a field-enhancing tunneling injector formed on the floating gate. Source-side, hot-electron injection is very efficient, thus allowing the use of small on-chip charge pumps from a single low-voltage power supply, for example, 3.0 or 2.5 V. Cells are normally erased prior to programming. The split gate memory cell size is comparable to traditional stacked gate memory cells because (1) the tunneling injector cell does not need the extra spacing to isolate the higher voltages required for programming the stacked gate array and (2) floating gate extensions are not needed to achieve the required stacked gate coupling ratios.

4.3.3.1 Cell Structure. A top view and cross-sectional view along the wordline and bitline directions are presented in Fig. 4.40. A common source is used for each page; that is, each pair of bits sharing a common source along a row pair (even plus odd row). A single wordline is referred to as a row. The combination of the even and odd rows form a page, which is erased as an entity. Programming may be either byte by byte individually or for all bytes within the same page simultaneously.

The drain region consists of an n^+ S/D diffusion, which is aligned with the edge of the second poly control gate. The source region consists of an n^+ S/D diffusion, which overlaps the floating poly. A cell implant beneath the floating gate is used to control the intrinsic cell threshold (V_t) and the punch-through voltage.

The select gate is separated from the channel by a 40 nm oxide. The floating gate is separated from the channel and source diffusion by a thermally grown 15 nm gate oxide. The floating gate is separated from the control gate by a 40 nm oxide on the side-wall and a 200 nm oxide vertically between the gates. The tunneling injector on the floating gate is formed by oxidation of the polysilicon, similar to the formation of the field oxide "bird's beak" on single-crystal silicon. A silicide or polycide can be formed on the control gate to reduce the poly wordline resistance. The cell array schematic is presented in Fig. 4.41. An equivalent circuit representation used to illustrate capacitive-coupling is presented in Fig. 4.42.

Table 4.5 gives the conditions for the memory cell terminals during erase, program, and read operations. V_{cc} is the power supply, for example, a nominal 3 or 2.5 V, V_{ss} is ground, and V_t is the nominal cell threshold. V_{REF} is the reference voltage used to access the memory cell during the read cycle. The high voltages on the wordline during erase and the source-line during programming are generated by on-chip charge pumps.

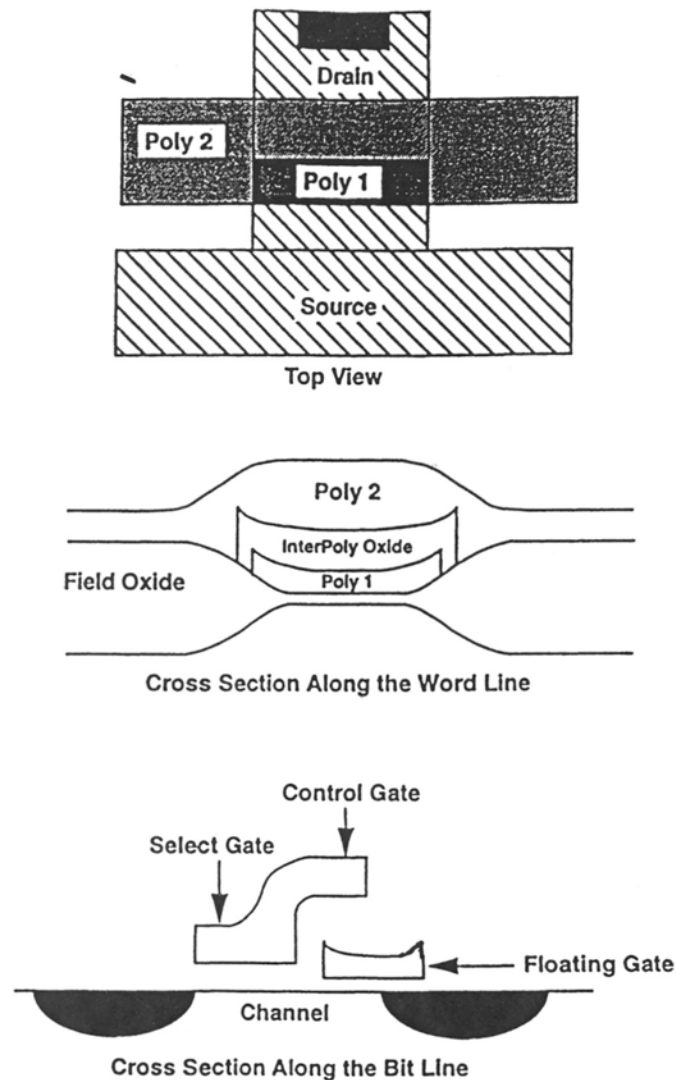


Figure 4.40 Top view, cross section along the wordline, and cross section along the bitline [4.82, 4.158].

During erase, the channel is inverted due to the wordline voltage. This increases the value of C_{lc} . During programming, the channel is in depletion and the value of C_{lc} is negligible. Therefore, the coupling ratios are different during erase and programming. During programming, the capacitive coupling ratio between the source and the floating gate is approximately 80%. This means that approximately 80% of the voltage at the source will be coupled to the floating gate.

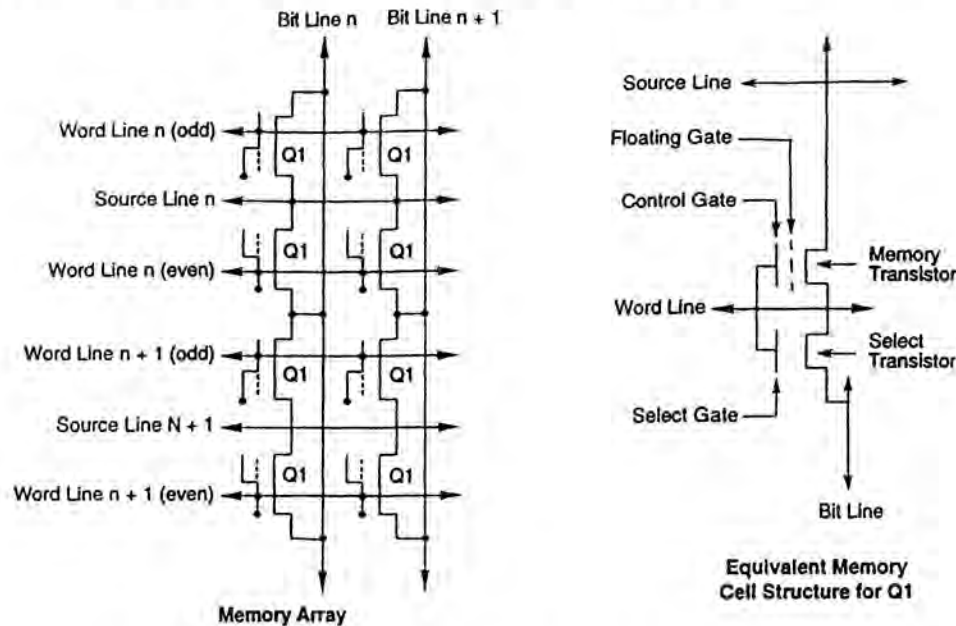


Figure 4.41 Representation of a section of a typical cross-point memory array arranged as eight memory cells in two columns (bitlines), two source-lines, and four wordlines. Note that the wordline is split into an even and odd row, which isolates the source-line from all other source-lines. On the right is an equivalent memory cell showing how the split gate cell provides the logical equivalent of a select transistor and a memory transistor [4.82, 4.158].

Figure 4.42 Equivalent capacitive-coupling circuit [4.158].

$$C10 = C1c + C1s; C12 = C12s + C12i + C12t$$

Coupling Ratios (CR) are defined as:

1. $CR10 = CR(\text{poly1 to substrate}) = \frac{(C1c + C1s)}{(C1c + C1s + C12)}$
2. $CR12 = CR(\text{poly1 to poly2}) = \frac{C12}{(C12 + C10)}$
3. $CR10 + CR12 = 1$

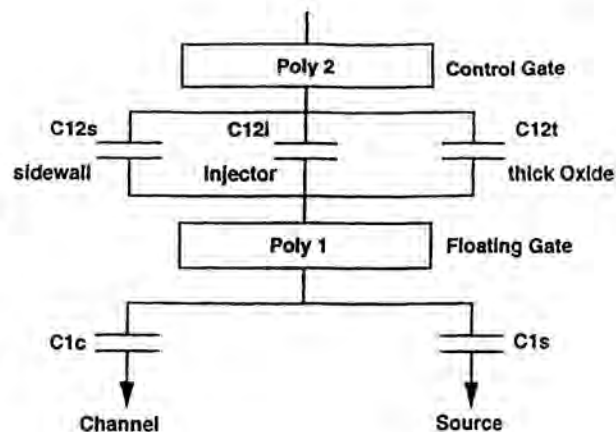


TABLE 4.5 MEMORY CELL OPERATION CONDITIONS

Operation	Wordline	Source-Line	Bitline
Erase	14 V	0 V(V_{ss})	0 V(V_{ss})
Program	V_t	11 V	0 V data "0" V_{cc} data "1"
Read	V_{cc}	0 V(V_{ss})	2 V

4.3.3.2 Erasing. The cell is erased by Fowler–Nordheim tunneling of electrons from floating gate to control gate through interpoly oxide. During erase, the source and drain are grounded and the wordline is raised to 14 V. The conditions for erase are given in Table 4.5. The low coupling ratio between the control gate and the floating gate provides a significant voltage drop across the interpoly oxide, which is the same everywhere between poly1 and poly2. A high field is generated primarily in the area of the tunneling injector. Charge transfer is very rapid and is eventually limited by the accumulation of positive charge on the floating gate. This positive charge raises the floating gate voltage such that there is insufficient voltage drop across the poly-to-poly dielectric to sustain Fowler–Nordheim tunneling.

The removal of charge can leave a net positive charge on the floating gate. The positive charge on the floating gate reduces the memory cell's threshold voltage to about the select gate V_t . The applied sense voltage is sufficient to turn on both the select transistor and the memory transistor in the addressed memory cell.

Erase can either be by fixed program pulses generated by an internal timer or algorithmically generated by an external controller to optimize erase conditions. Internal verify circuits assure an adequate erase margin.

4.3.3.3 Erase Disturb. Enhanced-field tunneling injector devices are internally organized by pairs (pages) of even and odd rows. Each row pair (page) shares a common source-line and has the wordline at the same voltage potential during erase. Thus, all bytes along the common wordlines are erased simultaneously. All other wordlines (pages) do not receive the erasing high voltage. Therefore, erase-disturb is not possible. The column leakage phenomenon caused by "overerase" in 1-T cells is avoided because the split gate provides an integral select gate to isolate each memory cell from the bitline.

4.3.3.4 Programming. The cell is programmed using high-efficiency source-side channel hot-electron injection. The conditions for programming are given in Table 4.5. During programming, a voltage of a cell threshold of approximately V_T volts is placed on the control gate via the wordline. This is sufficient to turn on the channel under the select portion of the control gate. The drain is at

approximately V_{ss} if the cell is to be programmed. If the drain is at V_{cc} , programming is inhibited. The drain voltage is transferred across the select channel because of the voltage on the control gate. The source is at approximately 12 V. The source to drain voltage differential (i.e., 11 V—approximately V_{ss}) generates channel hot electrons. The source voltage is capacitively coupled to the floating gate. The electric field between the floating gate and the channel sweeps the channel hot electrons that cross the Si–SiO₂ barrier height of approximately 3.2 eV to the floating gate very efficiently.

The programming effect is eventually self-limiting as negative charge accumulates on the floating gate. The programming source-drain current is low. Thus, the source voltage can be generated by an on-chip charge pump. The program time is fast because of the relatively high efficiency of source-side injection. Programming can be by fixed program pulses generated either by an internal timer or by an external controller to optimize program conditions.

4.3.3.5 Program-Disturb. There are two possible types of program-disturb with the field-enhanced tunneling injection cell: reverse-tunnel disturb and punch-through disturb.

Reverse-tunnel disturb can occur for unselected erased cells sharing a common source-line, but on the other row of the selected page to be programmed. Thus, the wordline is grounded. The source voltage is capacitively coupled to the floating gate of the unselected erased cell. If there is a defect in the oxide between the control gate and the floating gate, Fowler–Nordheim tunneling may occur. This could program the unselected erased cell.

Punch-through disturb can occur for selected erased cells, that is, those sharing a common source-line and wordline in an adjacent inhibited bitline. An inhibited bitline is taken high to prevent normal channel hot-electron injection. If there is a defect that reduces the bitline voltage or creates punch-through along the select gate channel, hot electrons could be available to program the inhibited erased cell.

Proper design and processing can prevent both mechanisms. Devices with this memory architecture do not have program-disturb caused by accumulated erase/programming cycles because each page is individually isolated. Each cell is only exposed to high voltage within the selected page along the row- or source-line. There is no high voltage on the bitline.

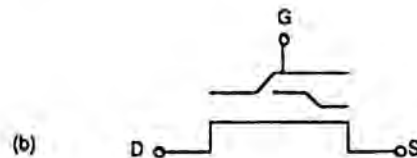
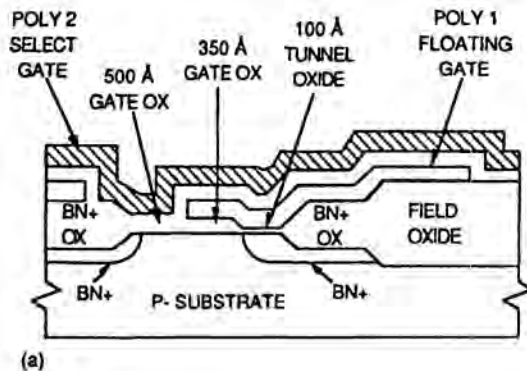
4.4. FLASH MEMORIES WITH FOWLER–NORDHEIM TUNNEL PROGRAM AND ERASE

Although EPROM-like cell structures offer density and historical learning advantages for Flash memory implementation, cell structures based on Fowler–Nordheim tunneling offer low power and single, low-voltage supply advantages. Examples are DINOR [4.58], AND [4.57], NAND [4.14], and HiCR [4.68]. Based on current understanding, the single-voltage supply Flash memories, below 3 V V_{cc} , will require

Fowler–Nordheim tunneling for programming. The half-select problems associated with F–N tunnel programming are minimized by small blocks, on-chip V_{PP} trimming, and process control.

4.4.1 Array Contactless EEPROM (ACEE)

In Array Contactless EEPROMs (ACEEs), the program and erase operations occur at the same node. The basic structure, schematic, and operation of the ACEE cell [4.24, 4.25, 4.29, 4.37, 4.159] are shown in Fig. 4.43. The floating gate is formed in first-level polysilicon, and the merged control and select gates are formed in second-level polysilicon. The ACEE cell is a Self-Aligned Thick Oxide (SATO) device, with the source and drain buried under a thick oxide. The oxide between the floating gate and the substrate near the source is 10 nm thick to allow for programming and erasing the cell by Fowler–Nordheim tunneling. A 35 nm gate oxide is employed under the floating gate structure to increase the poly2/poly1 capacitive-coupling ratio as a whole and to increase the gated diode breakdown voltage near the source. The pass gate, with 50 nm gate oxide, prevents the cell from drawing current when the floating gate structure is over-erased into depletion. Thus, the cell current is dominated by the pass gate. The buried source-drain of the



	D	G	S
ERASE	FLOAT	VEE	5V
PROGRAM	FLOAT	VPP	0
(c) READ	1.5 V	3 V	0

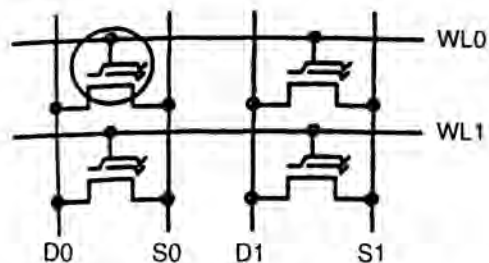
Figure 4.43 (a) Cross-sectional view through the ACEE cell, (b) schematic drawing of the cell, and (c) cell operating conditions [4.159].

cell constitutes a continuous buried bitline. Since negligible supply current is required during program or erase, the voltage drop along the buried bitline in a VLSI memory is small.

4.4.1.1 Charge Transfer. Charge transfer between the floating gate and the source region occurs in a manner similar to the FLOTOX memory cell. A high voltage, on the order of 20 V, is applied to the control gate, the source is held at ground, and the drain is left floating. Because of capacitive-coupling, a voltage of greater than 10 V is coupled to the floating gate. This establishes an electric field greater than 10 MV/cm across the tunnel oxide from the floating gate to the source, and electrons are transferred to the floating gate. Charge transfer to the floating gate raises the threshold voltage of the cell above that of the sense voltage, and the cell stops conducting current. To bring the cell to a conducting state, the charge stored on the floating gate must be removed; that is, a sufficiently high electric field must be applied across the tunnel oxide in the reverse direction.

4.4.1.2 Inhibit Condition. Programming of deselected cells is prevented by increasing the source-line voltage for cells sharing the same wordline and by applying a small positive voltage to unselected wordlines as is illustrated in Fig. 4.44.

4.4.1.3 Read. Reading of the ACEE cell is accomplished by raising the drain (the diffusion near the select gate with 50 nm gate oxide) to about 1.5 V, holding the source at ground potential, and raising the selected wordline to about 3 V. The read operation from the select transistor side provides a margin against read-disturb. The floating gate structure can be erased to a negative threshold voltage since the series select gate is an enhancement device. The threshold voltage of the select gate is lower than the sense voltage applied to the top gate. Thus, during the erased state, the cell current is determined predominately by the select gate and, during the programmed state, by the floating gate structure.



Cell (0,0) Operating conditions

	WL0	WL1	S0	D0	S1	D1
Erase	-11 V	-11 V	5 V	Float	5 V	Float
Program	17 V	7 V	0 V	Float	7 V	Float
Read	3.3 V	0 V	0 V	1.5 V	0 V	1.5 V

Figure 4.44 Schematic drawing and operating condition table showing the inhibit conditions for the ACEE cell array [4.37].

4.4.2 The AND Cell

The AND cell [4.57] program and erase operation is performed by F-N tunneling, as shown in Fig. 4.45 and Table 4.6. The memory cell is programmed through the drain by tunnel injection of electrons and is erased by tunnel injection from the whole channel region. The schematic diagram of this contactless memory array and the program/erase scheme are shown in Fig. 4.46 and Table 4.6, respectively. The memory cells are arranged in parallel between each pair of local data and source-lines, and ST1 and ST2 act as switches for connecting and disconnecting the target block to a global data-line and a common source-line. Both the local data-line and source-line are n^+ diffusions. Internal operating voltages applied to the selected wordline, -9 V for programming and $+13\text{ V}$ for erasing, are generated from the single 3 V power supply voltage by on-chip voltage converters, due to the low-current dissipation of tunneling. The rewrite size is as small as 512 bytes (which corresponds to memory cells connected to each wordline), suitable for various silicon file

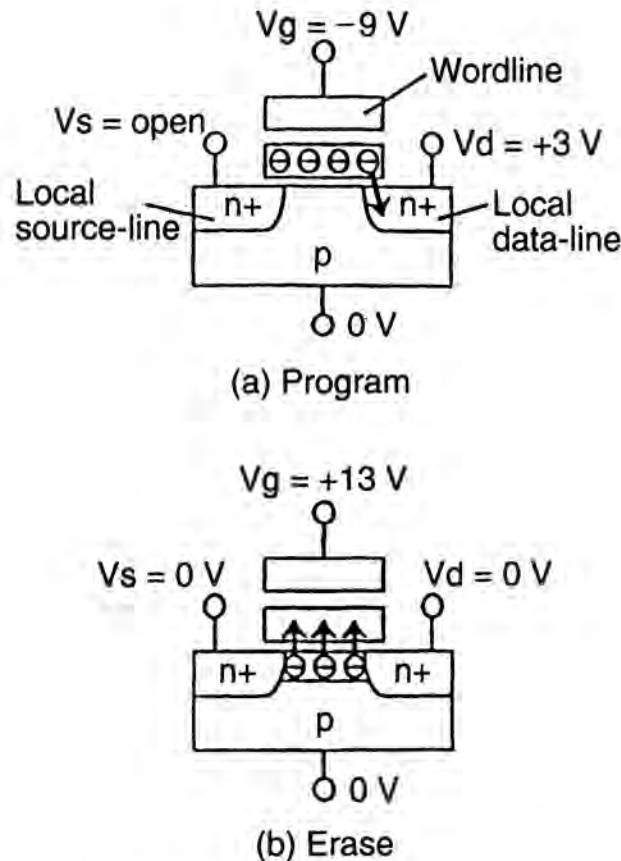


Figure 4.45 Program/erase diagrams and their bias conditions for the AND cell [4.57].

TABLE 4.6 OPERATING CONDITIONS FOR THE AND MEMORY ARRAY

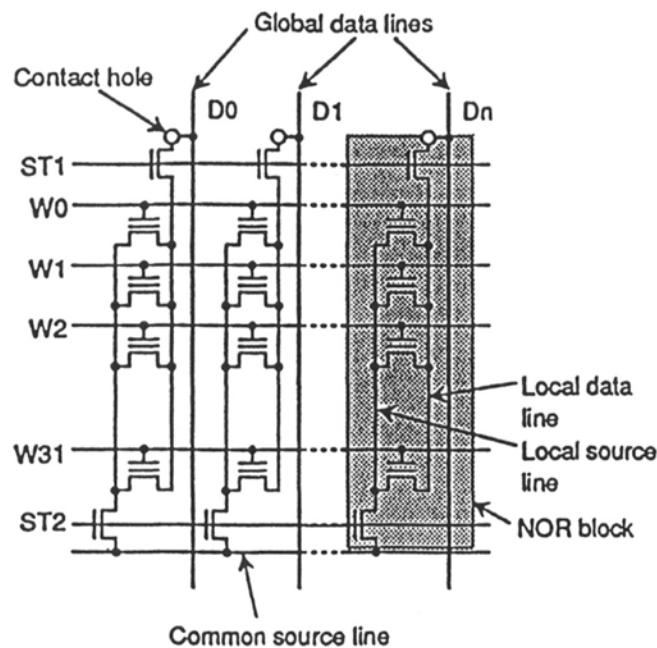
		Erase	Program	Read
Word	Selected	13 V	−9 V	3 V
	Deselected	0 V	3 V	0 V
Local data-line		0 V	3 V (program) 0 V (inhibit)	1 V
Local source-line		0 V	Floating	0 V

Erase: F–N tunneling

Program: F–N tunneling

applications. Each local source-line is disconnected from the common source-line when programming to allow program inhibit for the cells connected to the deselected wordlines. Adjacent bitlines are separated by field oxide isolation.

Scatter in the “low-level” threshold voltage of programmed memory cells can be suppressed below 0.5 V, as shown in Fig. 4.47 by using program and program-verify sequences. In page-mode programming for the selected wordline, the program is controlled bit by bit by setting the potential in each data-line at 3 V for successive program and at 0 V for program termination. Obtaining a tight distribution in the “low-level” threshold voltage would lead to a successful read operation at a 3 V

**Figure 4.46** Schematic diagram of the AND cell memory array [4.57].

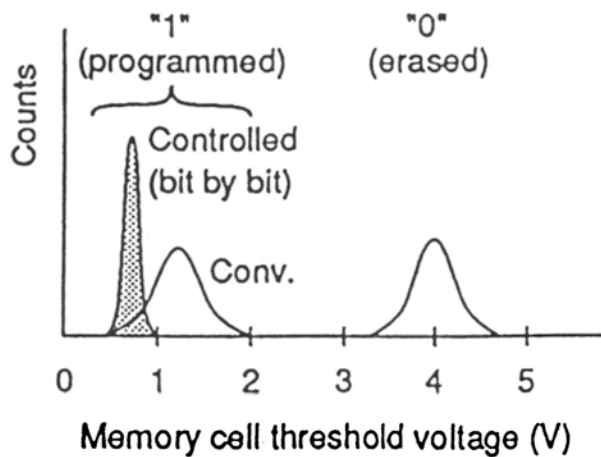


Figure 4.47 Schematic of the low-threshold voltage control for the programmed memory array [4.57].

power supply voltage without introducing a complicated wordline boost scheme [4.39].

Schematic cross sections of the memory array parallel to a wordline and a data-line are shown in Fig. 4.48. A floating gate consists of two layers of poly-Si films electrically combined to each other. The first polysilicon defines a channel length, and the second polysilicon realizes a large capacitive-coupling between the control gate and the floating gate.

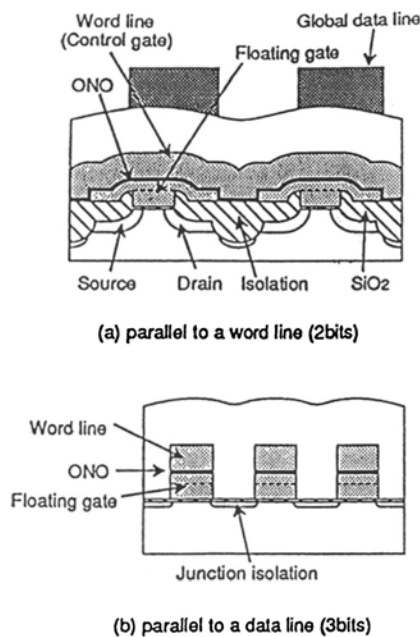
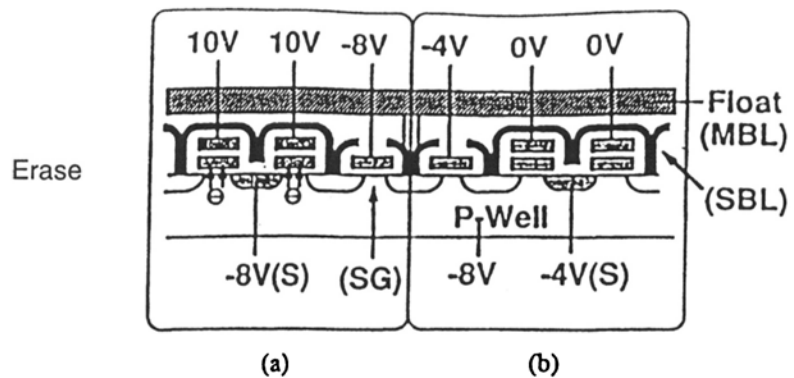


Figure 4.48 Schematic cross sections [4.57].

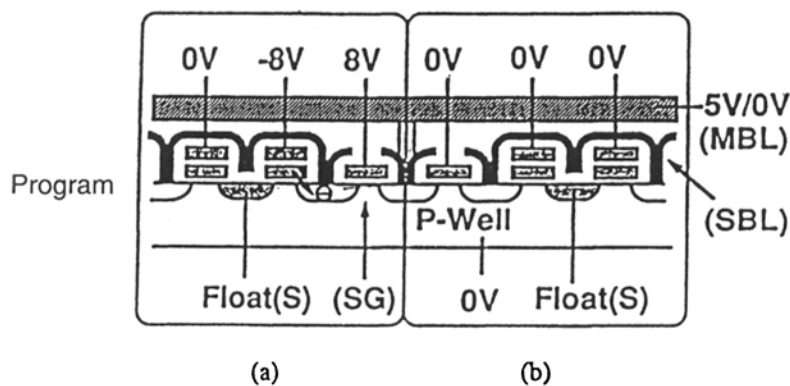
4.4.3 The DINOR Cell

The **D**ivided bitline **NOR** (DINOR) cell operation is schematically shown in Fig. 4.49. In the case of the DINOR [4.58, 4.69] cell, the “program” operation means to set the V_{th} of the selected cell to the low state, and the “erase” operation means to set the V_{th} of the cells of the selected sector to the high state. That is, the direction of the operation is opposite to the conventional NOR.

The “program” condition uses F–N electron ejection from the floating gate to the drain through the gate-drain overlapped area, while “erase” causes F–N electron injection from the channel area to the floating gate. By using the low V_{th} as the



Voltage condition of DINOR cell in the erase operation. (a) Selected sector. (b) Unselected sector.



Voltage condition of DINOR cell in the program operation. (a) Selected sector. (b) Unselected sector.

Figure 4.49 Program/erase diagrams and the bias conditions for DINOR cell [4.58].

programmed state, the DINOR cell is able to use the bit-by-bit verify programming sequence. This makes the V_{th} distribution tight and realizes high over-erase tolerance.

4.4.3.1 Program/Erase Operation. The DINOR cell utilizes the gate-biased F–N write/erase operation, which realizes a maximum internal voltage as low as 10 V. The conditions for various operations are summarized in Table 4.7, and a circuit diagram of the memory cell is shown in Fig. 4.50. Because of a low-current requirement, all the high voltages shown in Table 4.7 can be generated internally from a single external 3 V supply.

The DINOR cell uses a triple-well, triple-level polysilicon and a double layer of metal. The triple well is utilized to apply a negative bias voltage to the substrate in the erase operation. A symmetrical source-drain structure without a p^+ pocket is used for the memory cell.

TABLE 4.7 OPERATIONAL CONDITIONS FOR THE DINOR CELL

	Bitline	Control Gate	Source-Line	Substrate
Write	5 V	–8 V	Floating	0 V
Erase	Floating	10 V	–8 V	–8 V
Read	1 V	3 V	0 V	0 V

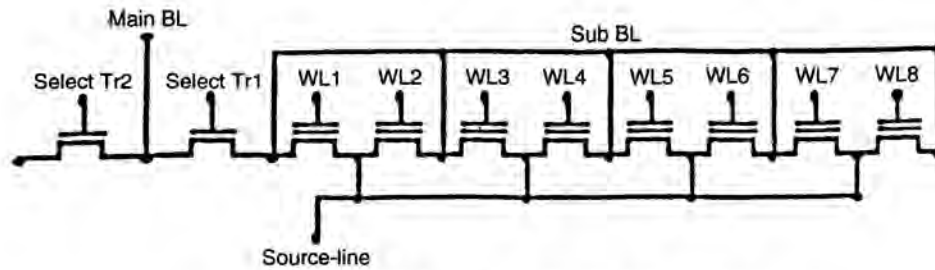


Figure 4.50 Circuit diagram of the DINOR cell [4.58].

4.4.3.2 Cell Structure. The DINOR cell is a unit consisting of one select transistor, eight stacked gate cells, and one sub-bitline connected through the select transistor to a main bitline. Figure 4.51 shows a schematic drawing of the DINOR cell.

The third polysilicon layer, which is used as the sub-bitline, is added to the conventional stacked gate cell structure formed by the first and second polysilicon layers. The low-current requirement of F–N operation makes it possible to use a polysilicon layer as the sub-bitline instead of a metal layer, and this is advantageous

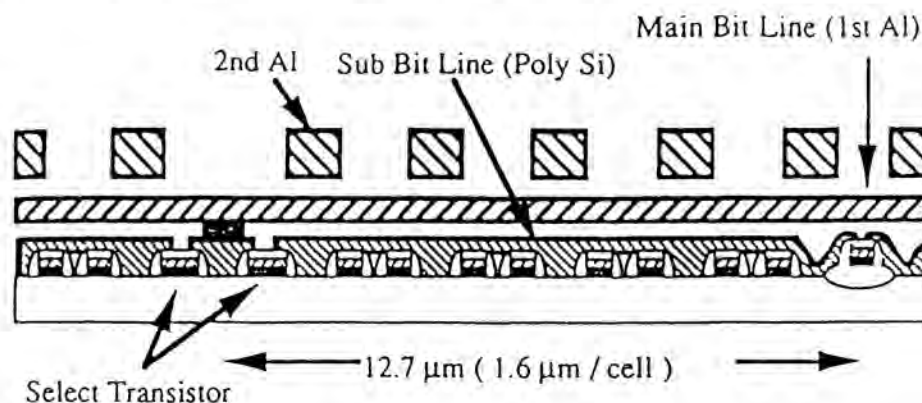


Figure 4.51 Schematic drawing of the DINOR cell [4.58].

in terms of wafer processing simplicity. The first aluminum layer forms the main bitline. The second aluminum layer in the cell array is used to stitch the gates of the select transistors and the main wordlines.

The sector size of the DINOR is 1 KB in which the cells are arrayed in 16 bits ($= 2$ basic units described above) $\times$ 512 bits (connected to the same wordline).

4.4.3.3 Disturb and Endurance Characteristics. The DINOR cell solves disturb problems by utilizing the divided bitline and parallel programming. By setting all select transistors off except for the one that is in the programming mode, the maximum drain-disturb is only seven write cycles from the cells connected to the same sub-bitline. The maximum gate-disturb is only one write cycle due to the use of a latch circuit. Characteristics of drain-disturb and gate-disturb, with sufficient margin for device operation, have been demonstrated [4.58].

In the case of the DINOR cell, one more new disturb mode, substrate disturb, exists. Substrate disturb is the undesired erasure caused by V_{sub} (substrate voltage) stress during the erasure of other sectors. The longest time over which substrate-disturb occurs is of the order of one year for actual 16 Mbit devices. By applying an erase inhibit voltage ($1/2$ of V_{sub}) to the source-lines of the unselected sectors, the effective V_{sub} can be reduced by one-half, making the substrate-disturb immunity longer than $1\text{E}10$ seconds.

4.4.3.4 Virtual Ground DINOR. A virtual ground DINOR array, utilizing F–N tunneling (for program and erase) has been described by using the asymmetrical offset source-drain structure [4.64] shown in Figs. 4.52 and 4.53. One side of the drain region has an offset against the floating gate, and the other side is overlapped with it. During programming, F–N tunneling current only flows to the overlapped drain region that is selected. Thus, only the selected cell is programmed. The operating conditions for the array are shown in Table 4.8.

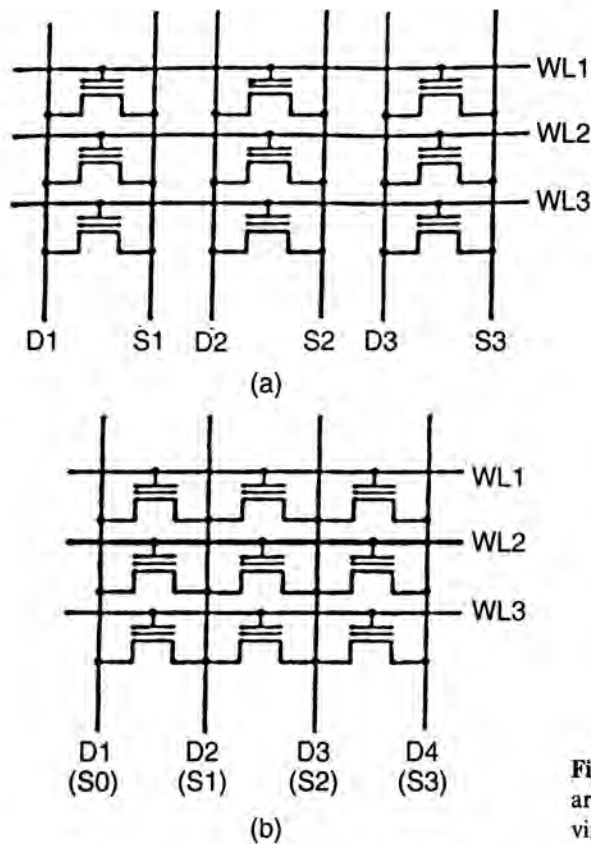


Figure 4.52 (a) Conventional contactless array with F-N operation, (b) simple virtual ground array [4.64].

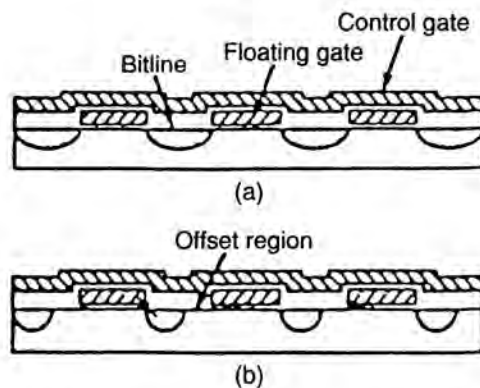


Figure 4.53 (a) Conventional virtual ground array that cannot use F-N mechanism, (b) virtual ground array using F-N mechanism with offset source/drain structure [4.64].

TABLE 4.8 OPERATING CONDITIONS FOR THE DINOR VIRTUAL GROUND ARRAY

	Bitline	Control Gate	Source-Line	Substrate
Program	5 V	-9 V	Floating	0 V
Erase	Floating	10 V	-9 V	-9 V
Read	1.5 V	3 V	0 V	0 V

4.4.4 HiCR Flash Memory Cell

The High Capacitive-coupling Ratio (HiCR) cell is a floating gate transistor, self-aligned with a buried n^+ source and drain, in a contactless array configuration [4.68]. The cell is programmed by F–N tunneling on the drain side and erased by F–N tunneling from the source-drain junction to the floating gate. The high capacitive-coupling ratio is realized by (1) a small tunneling region between the floating gate and the source-drain junction regions and (2) a 20 nm channel oxide and a 7.5 nm tunneling oxynitride. The floating gate consists of a lower polysilicon layer above the channel and an overlapping upper polysilicon layer above the tunneling oxide. The upper polysilicon layer is connected to the lower poly (Fig. 4.54).

The cell operation is shown in Figs. 4.55 and 4.56. Each pair of bitlines has a common source-line (1 1/2 rail array). Each pair of bitlines is isolated from the adjacent pairs of bitlines by field oxide.

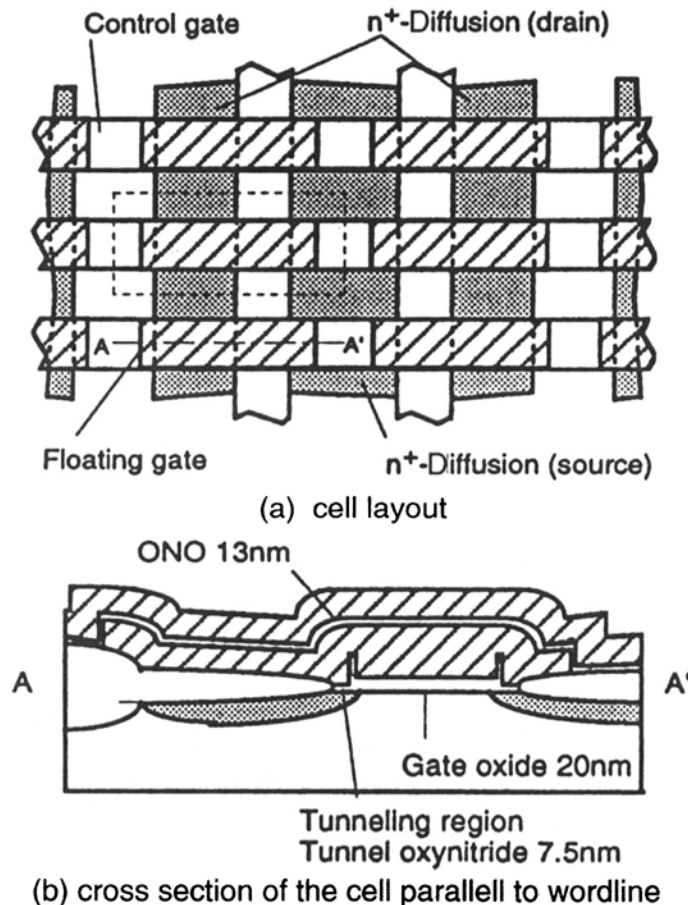


Figure 4.54 HiCR cell structure: (a) cell layout, and (b) cross section of the cell parallel to wordline [4.68].

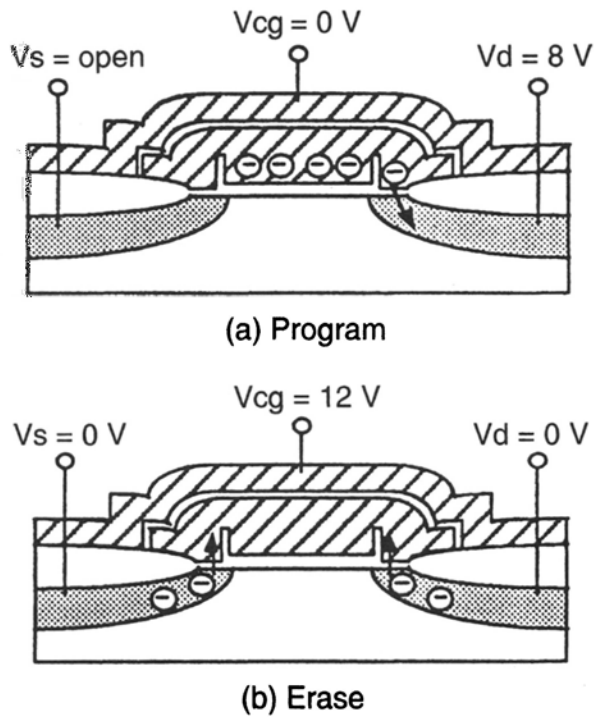
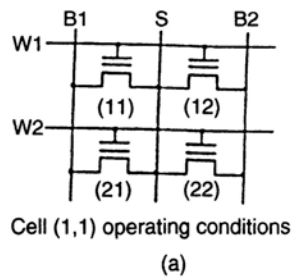


Figure 4.55 Schematic of HiCR cell operation: (a) program and (b) erase [4.68].



	W1	W2	B1	B2	S
Program	0V	5V	8V	open	open
Erase	12V	0V	0V	0V	0V
Read	3V	0V	1V	0V	0V
Program Verify	2.5V	0V	1V	open	0V
Erase Verify	6V	0V	1V	1V	0V

(b)

Figure 4.56 (a) Array schematic, (b) array operating condition [4.68].

During programming, the unselected bitline is kept open, and the unselected wordline is at 5 V to minimize program-disturb. During erase, 12 V is applied to the wordlines to be erased, and bitlines and source-lines are grounded. Because of the F–N tunnel program/erase operation, high voltage can be generated on-chip from the supply voltage. The memory cell array (Fig. 4.57) has the NOR structure. The select transistors act as switches for connecting and disconnecting the target block to a main data-line and a main source-line.

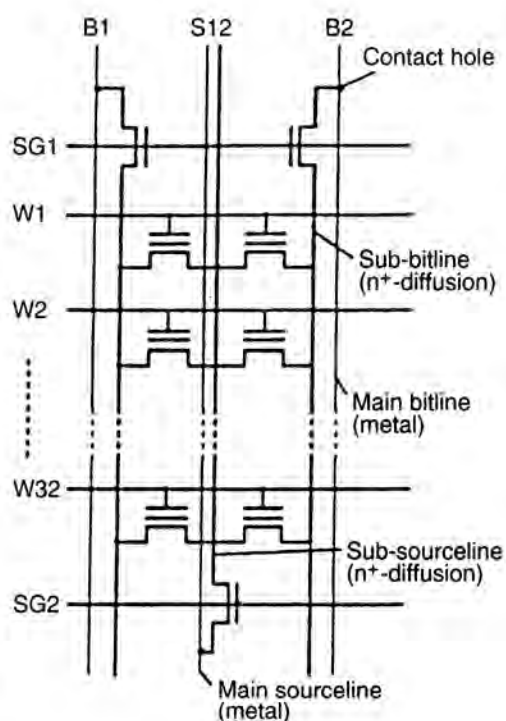


Figure 4.57 Schematic of HiCR memory array [4.68].

4.4.4.1 Program Disturb. Two principal disturbances can occur during programming of an array (Figs. 4.58 and 4.59). The first (I) is unexpected programming of the cell (12) connected to the same source-line (S) and to the selected wordline (W1). It is caused by the source voltage, $V_{CG} - V_{TM}$, of the cell, where V_{CG} is the program inhibit voltage of the unselected wordline (W2) and V_{TM} is the threshold voltage of the erased cells (cell 21 works as a pass transistor when it is erased). The second disturbance (II) is the unexpected programming of cells (21, 22, ...) connected to the selected bitline (B1).

These problems can be solved by utilizing parallel programming and sub-bitlines. Disturbance I, which has a maximum duration of only one write cycle (1 ms), can be solved by the use of a latch circuit [4.71]. The solution to disturbance II,

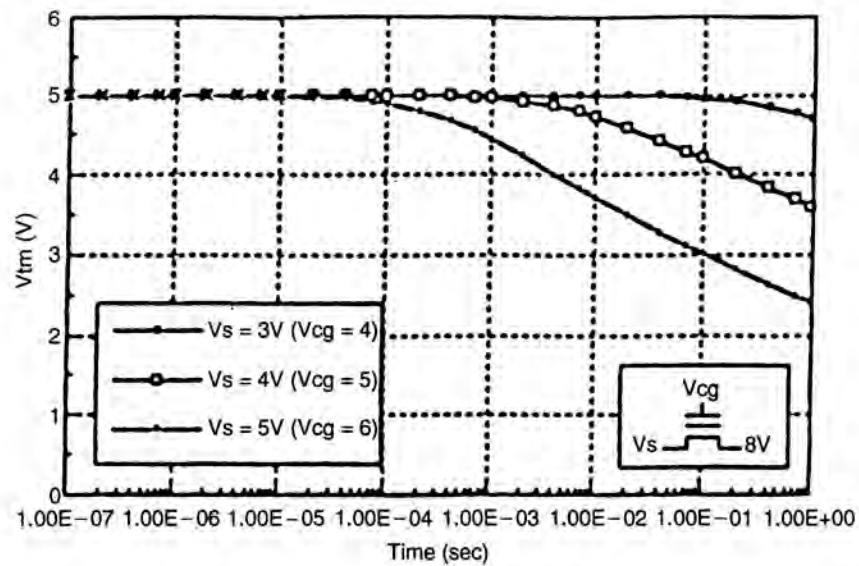


Figure 4.58 Threshold-voltage shifts during disturbance: I. The cell was programmed to a V_t of 5 V before disturbance measurement. Maximum disturbance I is 1 write cycle (1 ms) [4.68].

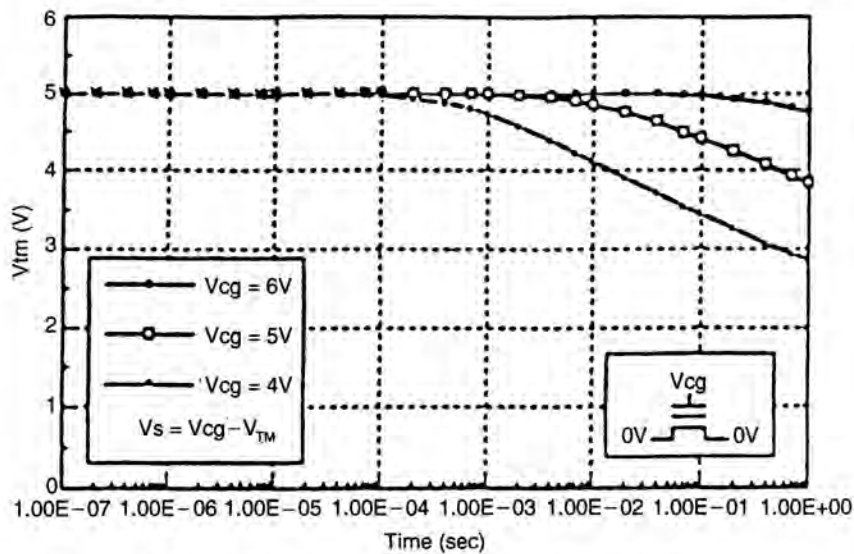


Figure 4.59 Threshold-voltage shifts during disturbance: II. Maximum disturbance II is 31 write cycles (31 ms) [4.68].

which has a maximum duration of 31 write cycles (31 ms), can be realized by setting the select transistors off except in the programming mode. Characteristics of disturbances I and II, which have sufficient margin for device operation, are shown in Figs. 4.58 and 4.59, respectively.

4.4.5 NAND

In the drive to reduce cell size, eliminating the select transistor in a full-feature EEPROM results in a bulk/block erasable Flash EEPROM. The Flash EEPROMs discussed so far are the NOR structure, in which memory cells are connected to a bitline in a parallel manner, and the NAND structure, which reduces the cell size by connecting the cells in series between a bitline and a source-line, thus eliminating the contact hole [4.13, 4.26, 4.32]. The resulting cell structure occupies 85% of the NOR cell area of a stacked gate array.

4.4.5.1 NAND Structure. Figure 4.60 shows the layout and the equivalent circuit of the NAND-structured cell. As shown in the figure, the NAND-structured cell arranges eight or sixteen memory transistors in series, sandwiched between two select gates, select gate 1 (SG1) and select gate 2 (SG2). The first gate (SG1) ensures selectivity, and the second (SG2) prevents the cell current from passing during a programming operation. The floating gates are made of first-level polysilicon. The control gates, which are wordlines in an array, are made of second-level polysilicon. The dielectric between the floating gate and the control gate is an ONO stack.

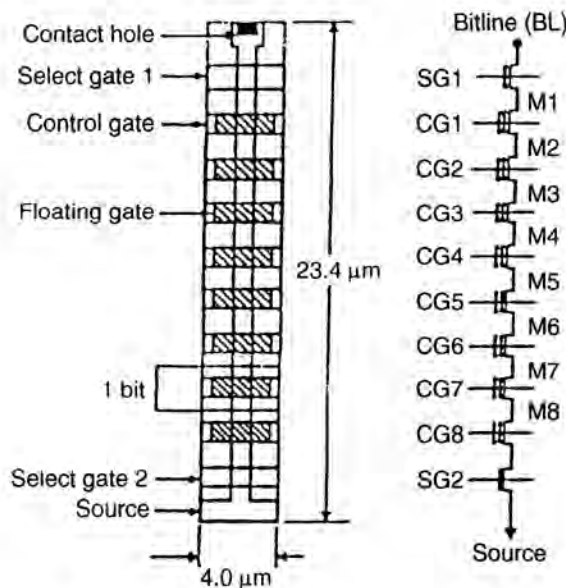


Figure 4.60 Top view and equivalent circuit of NAND-structural cell [4.13].

A NAND-structured cell array is positioned in a p-well region of the surface portion of the n-substrate as shown in Fig. 4.61. The peripheral CMOS circuitry and NAND cell array are located in different p-wells [p-well(1) and p-well(2)], which are electrically isolated from each other. The peripheral circuitry is fabricated in p-well(1). The NAND cell array is located in p-well(2). In an erase operation, p-well(2) is raised to 20 V and p-well(1) is always grounded. Thus, p-well(2) can be biased to V_{pp} during erase, while keeping p-well(1) and the n-well at V_{ss} and V_{cc} , respectively.

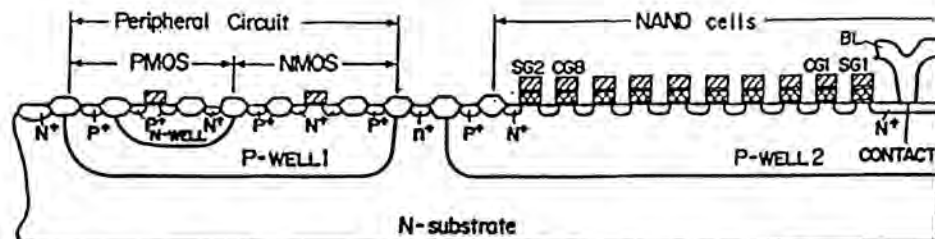


Figure 4.61 Cross-sectional view of NAND EEPROM [4.26].

4.4.5.2 NAND Operation. The operation mechanism of a single-memory transistor in a NAND structure is comparable to that of the conventional EEPROM. However, the programming and reading of the memory cells are more complex. Therefore, additional peripheral circuitry is required, and the reading speed is lower than that of the conventional type because a number of memory cells are connected in series.

The reading method is shown in Fig. 4.62 and is essentially the same as that of the NAND-type MASK ROM. Erased cells ("0") have a negative threshold voltage, and programmed cells ("1") have a positive threshold. Zero volts is applied to the gate of the selected memory cell, while 5 V is applied to the gate of the other cells. Therefore, all of the other memory transistors, except for the selected transistor, serve as transfer gates. As a result, in the case where a "0" is being written, the

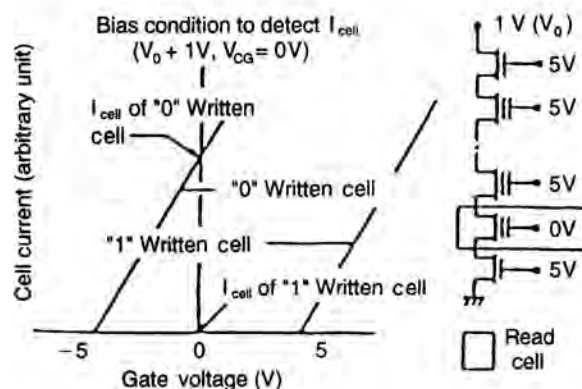


Figure 4.62 Reading of the NAND EEPROM [4.26].

memory transistor is in the depletion mode, and current flows. On the other hand, current does not flow in the case where a “1” is being written because the memory transistor is in the enhancement mode. The state of the cell is detected by a sense amplifier that is connected to the bitline. The difference between a “0” and a “1” depends on whether or not negative charge is stored in the floating gate. If negative charge is stored in the floating gate, the threshold voltage becomes higher and the memory transistor is in the enhancement mode. This is analogous to the conventional EEPROM.

Figure 4.63 shows the equivalent circuit and operating conditions during program/erase and read. During erase, the electrons tunnel from p-well(2) to the floating gate. During program, the electrons tunnel from the floating gate to p-well(2).

In the erase operation, all control gates are grounded, 20 V are applied to the n-substrate and p-well(2), and the source- and bitlines are floating. Erasing can be performed on the whole chip or selected blocks in the Flash manner. Electrons are emitted from the floating gate to p-well(2) by F–N tunneling, and the V_t of the erased cells becomes negative. During the erase operation, there is no voltage difference between the n^+ -drain and p-well(2). Therefore, gated-diode breakdown does not occur. The minimum erase block size is equal to the length of a wordline in bytes $\times$ number of series of bits between a drain and a source. For a 16 MB NAND, the minimum erase block is 4 KB. The typical erase time is 10 ms (2.5 μ s/byte). No preprogramming before erase is required. Hence, the NAND erase time is faster than many other Flash memories.

In the program operation, p-well(2) is grounded, 20 V is applied to the selected control gate, and 10 V is applied to the unselected control gate. The bitline of the

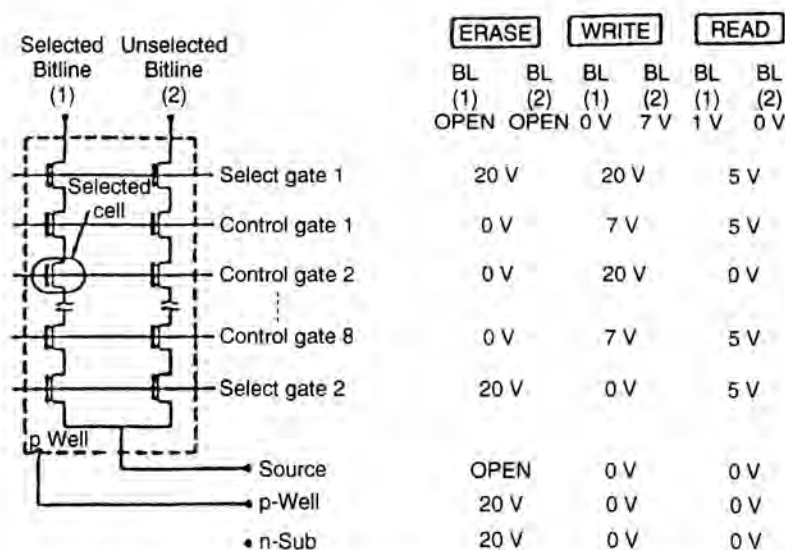


Figure 4.63 Equivalent circuit and operation voltage of NAND EEPROM [4.34].

ZERO data programming cell is grounded. Electrons are injected from p-well(2) to the floating gate by F-N tunneling. The V_t of the selected cell becomes positive. Ten volts is applied to the bitline of the ONE data programming cell. The V_t of the unselected cell remains negative because the voltage across the tunnel region is inadequate to start the tunneling current. In this case, while 10 V is applied between the drain and p-well(2), gated-diode breakdown does not occur because the control gate is biased at 10 V. In both erase and program operation, the gated-diode breakdown is suppressed entirely. As a result, all high voltages are internally generated easily from a 5 V power supply through charge-pump circuits because there is no breakdown leakage current. Furthermore, oxide degradation [4.9] from hole trapping, due to gated-diode breakdown, is avoided. This results in improved write/erase endurance.

4.4.5.3 Disturb Mechanisms. The operating conditions, including the inhibit conditions, are shown in Fig. 4.63. Erasure of deselected cells in the deselected NAND block is prevented by applying a high voltage of the order of 20 V to the wordlines. Thus, the tunnel dielectric field is zero in the deselected NAND block, and no erase-disturb condition exists.

In order to prevent programming of deselected cells sharing the same bitline in the same NAND block, the control gate voltage is raised to V_m (a medium voltage) to reduce the tunnel dielectric field between the floating gate and the channel. Programming of deselected cells sharing the same control gate is inhibited by applying a medium voltage (V_m) of around 7 V to the deselected bitline to reduce the electric field between the floating gate and the channel.

Programming of cells in each NAND block is performed in a serial order from the source-line side to the bitline side to avoid unintentional charging of deselected cells sharing the same control gate. This ordering prevents the existence of programmed cells on the drain side and avoids the channel potential decrease below 8 V of the deselected bitline voltage to minimize the disturb for deselected cells. Disturbs for deselected erased cells sharing the bitline exist, although the programming (disturb) time is fast enough to prevent the charge gain.

4.4.5.4 Special Features

Advantages

1. Since the number of bitline contacts in the NAND array is reduced by one-eighth to one-sixteenth compared to that of the standard T-cell Flash array, the unit cell size is smaller.
2. A NAND cell with n^+ source and drain junctions is more scalable than the E-Tox cell since the E-Tox cell has a graded-source junction.
3. Programming and erasing are achieved by F-N tunneling and need less power (i.e., low currents), thus allowing a single power supply operation by utilizing internal charge-pump circuits.

4. Uniform F-N tunneling used in the NAND cell does not generate band-to-band tunneling related hot holes, leading to high reliability. Typical endurance is 1 million write/erase cycles.

Disadvantages

1. Since the negative voltage erase is not used in erasing or programming, the required voltage for high-voltage transistors is of the order of 20 V, similar to that of FLOTOX cells. This is higher than that of E-Tox and E-Tox-like cells.
2. The random access time is slow. NAND applications are currently limited to hard disk replacement and to systems that do not require fast random access.

4.5. SPECIAL AND ADVANCED CELL STRUCTURES

This section covers Flash cell structures for embedded applications not covered earlier (SCSG and HIMOS) and advanced cell concepts (three-dimensional structures and multistate cells).

4.5.1 Source-Coupled Split-Gate (SCSG) Flash EEPROM Cell

The SCSG memory cell has a split gate configuration that is formed by merging an enhancement transistor (a split gate transistor) to the source side of a floating gate transistor. The floating gate is formed in the first-level polysilicon, and the merged control and split gates are formed in the second-level polysilicon. This serial configuration allows the memory cell threshold voltage to be determined by the floating gate transistor when the memory cell is programmed (floating gate charged with electrons) and by the split gate transistor, which typically has a threshold voltage of 1.5 V, when the memory cell is erased (floating gate charged with holes). A piece of the floating gate is extended into the source region to allow the erasure of the memory cell to take place at the source. In so doing, both the drain and the source junctions can be separately profiled for optimal performance.

4.5.1.1 Charge Transfer. The basic structure and operation of the SCSG memory cell are shown in Fig. 4.64. For programming, both a high-drain voltage, typically 7 V, and a high control gate voltage, typically 9 V, are applied with the source held at ground. Because of the gate coupling, the actual floating gate potential is less than the drain potential, which is in favor of the generation of hot electrons near the drain junction. A small fraction, typically 1 ppm, of electrons with the proper energy and momentum eventually enter the floating gate. Electrons transferred to the floating gate raise the memory cell threshold voltage, thereby reducing the cell current. For erasure, a high voltage, typically 12 V, is applied to the source, with the control gate grounded and the drain floating. A sufficient electric field is established across the tunnel oxide between the floating gate and the source, and

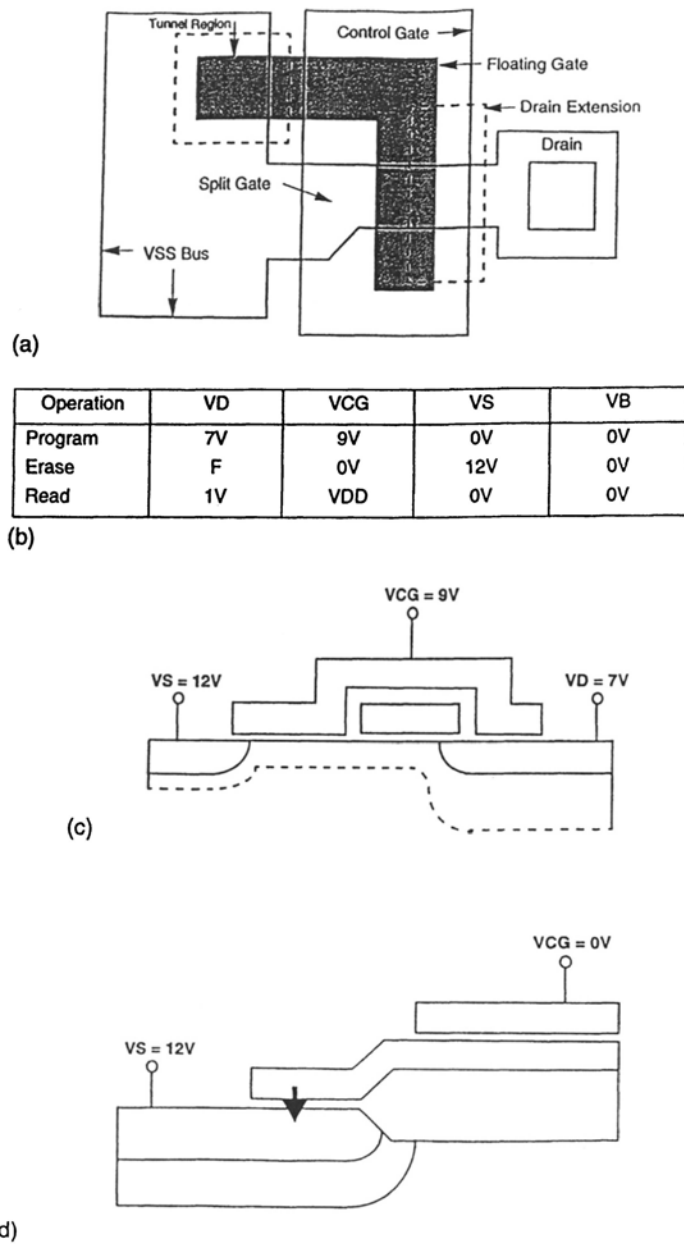


Figure 4.64 Structure and operation of SCSG cell [4.60, 4.158].

electrons are extracted from the floating gate. The extraction of electrons lowers the memory cell threshold voltage, thereby increasing the cell current. The incorporation of the split gate allows the floating gate transistor to be over-erased to depletion without shorting out the column. The region where the floating gate overlaps the source is doped n-type to eliminate the leakage current through the substrate. With

the drastically reduced current requirement during erasure, an optional on-chip charge pump can be built to enhance the erase speed to less than 1 msec.

4.5.1.2 Inhibit Conditions. A small section of the memory array during programming is depicted in Fig. 4.65. The selected bit to be programmed is encircled. The bits that share the same wordline will see the wordline voltage repeatedly during programming. The worst case bit will be subjected to this wordline voltage for a duration that equals the total time to program the other bits on the same wordline. The resultant threshold voltage gain is called the gate-disturb. The bits that share the same bitline will see the bitline voltage repeatedly during programming. The worst case bit will be subjected to this bitline voltage for a duration that equals the total time to program the other bits on the same bitline. The resultant threshold voltage loss is called the drain-disturb. As in most Flash EEPROM cells, the tunnel dielectric thickness in the floating gate channel region is around 100 Å. At this thickness and without a select transistor between the drain and the floating gate transistor, suppression of program-disturbs is achieved through the proper balance of the programming voltages and the programming speed. Typically, the duration of a particular half-selected memory bit, exposed to the programming bias conditions, is kept below 1 second.

4.5.1.3 Read. Reading of the SCSG cell is accomplished by raising the drain to about 1 V, grounding the source, and raising the control gate to the supply voltage. The charge state is determined by the relative magnitude of the cell current to a reference current in the sense amplifier. The split gate configuration allows the

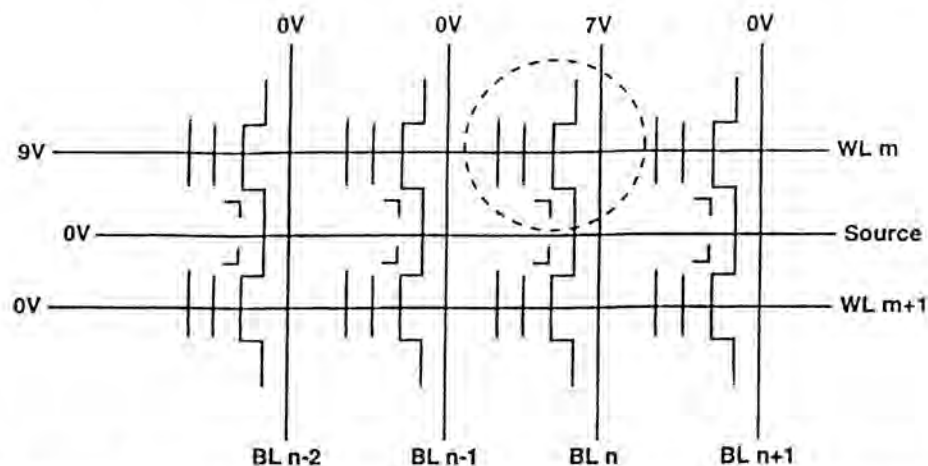


Figure 4.65 Program inhibit conditions [4.99]. Circled cell indicates the selected cell (row#m, column#n). The half-selected bits along the same wordline (WL m) will see the gate disturb, while the half-selected bits along the bitline (BL n) will see the drain disturb. Other bits are unaffected [4.158].

erased threshold voltage distribution of the entire memory array to be targeted and tightly controlled at around 1.5 V, provided that all the floating gate transistors are over-erased to depletion, which is easily achieved with an on-chip charge pump. This, in turn, allows the array to be read at low-supply voltages due to the enhanced gate drive.

4.5.2 High-Injection MOS (HIMOS)

High Injection MOS (HIMOS) [4.99] is a two-poly split gate structure. The presence of a coupling capacitor enhances the injection efficiency through source-side injection, suitable for 5 V-only applications. The erase occurs by electrons tunneling from the floating gate to the drain junction. The programming mechanism is by source-side injection.

Two perpendicular cross sections of the HIMOS device are shown in Fig. 4.66. The device consists of a control gate (CG) channel and a floating gate (FG) channel in series. An additional program gate (PG) is located on the field region in order to couple a voltage onto the FG during programming. During the write operation, the CG voltage is typically 1 to 1.5 V (just above the threshold voltage of the CG channel), and the PG voltage is typically 10 to 12 V, which can be generated on-chip by charge-pumping techniques. The programming efficiency is determined mainly by the interpoly dielectric thickness [4.100]. The interpoly dielectric layer is a thermally grown polyoxide with a thickness of 30 nm.

Moreover, because of this feature, drain engineering is strongly simplified. And 5 V-only (or 3.3 V-only) Flash erase and wordline-oriented sector erase are achieved by Fowler–Nordheim tunneling of electrons from the FG toward the drain junction

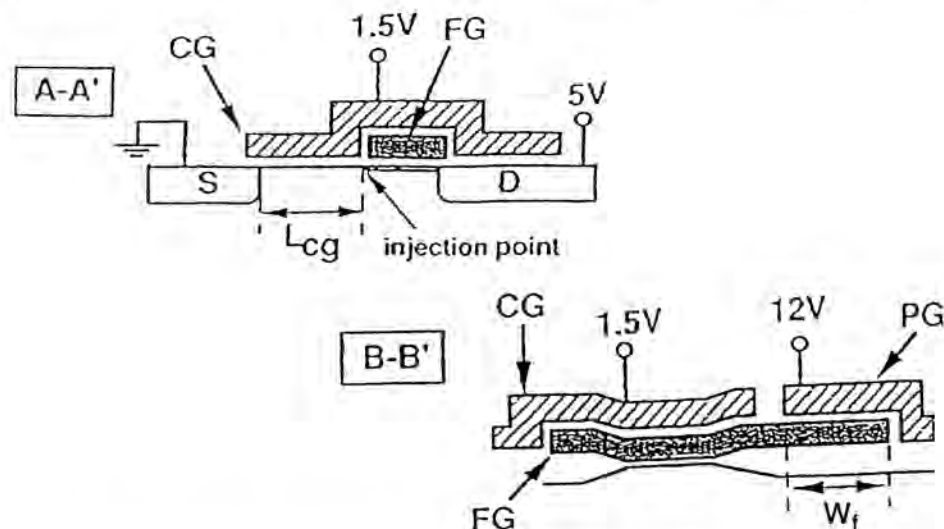


Figure 4.66 The HIMOS cross section with typical programming voltages [4.99].

by applying 5 V (or 3.3 V) to the drain junction and a negative voltage (typically -10 to -12 V) to the PG.

To minimize the additional area that is necessary for the PG, a virtual ground array (VGA) configuration has been chosen. The VGA layout principle and the circuit equivalent are shown in Figs. 4.67 and 4.68, respectively. Because of the strongly asymmetrical programming characteristics of the HIMOS device, the bitlines can be shared between adjacent bits in a row. The additional area for the PG can be reduced by sharing the Program Line between adjacent bytes. This is possible because of the triple gate structure of the device; the CG can be used as the wordline to access every byte separately during the write and read-out operations. The operating conditions of the HIMOS are shown in Table 4.9.

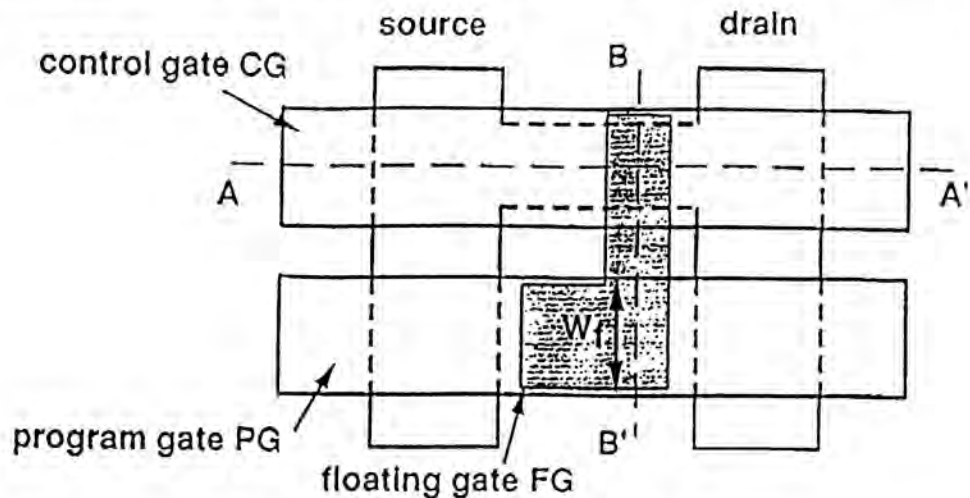


Figure 4.67 Virtual ground array principle [4.99].

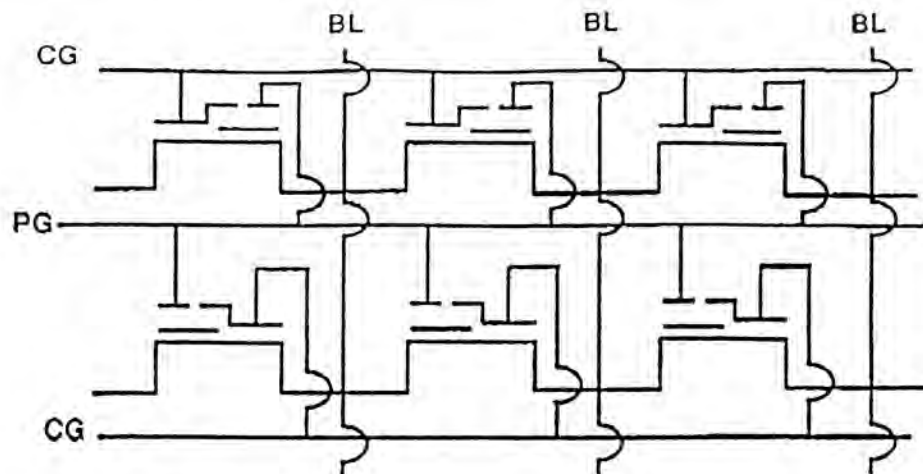


Figure 4.68 Virtual ground array equivalent circuit [4.99].

TABLE 4.9 TYPICAL OPERATING VOLTAGES FOR THE HIMOS DEVICE

	Source	Drain	CG	PG
Write	0	3.3/5	1/1.5	12
Erase	3.3/5	3.3/5	0	-12
Read-out	0	2	2	0

Figure 4.69 shows the programming characteristics of the cell for 5 V and 3.3 V applications. During programming, the CG voltage is very close to the threshold voltage of the built-in select device, and the optimal V_{cg} value is equal to 1 V for a 3.3 V drain voltage and equal to 1.5 V for a 5 V drain voltage. Therefore, the drain current is a constant during the entire programming cycle, which ensures a constant supply of electrons for the hot-carrier generation process. This drain current is also much smaller than in conventional devices where the optimal condition for hot-electron injection occurs at the onset of saturation (V_{fg} @ V_d). The latter feature reduces bitline voltage drops and strongly decreases the necessary programming power, which offers the possibility of page-mode programming and additional charge-pumping techniques for boosting the bitline voltage on-chip, in the case of 3.3 V-only operation.

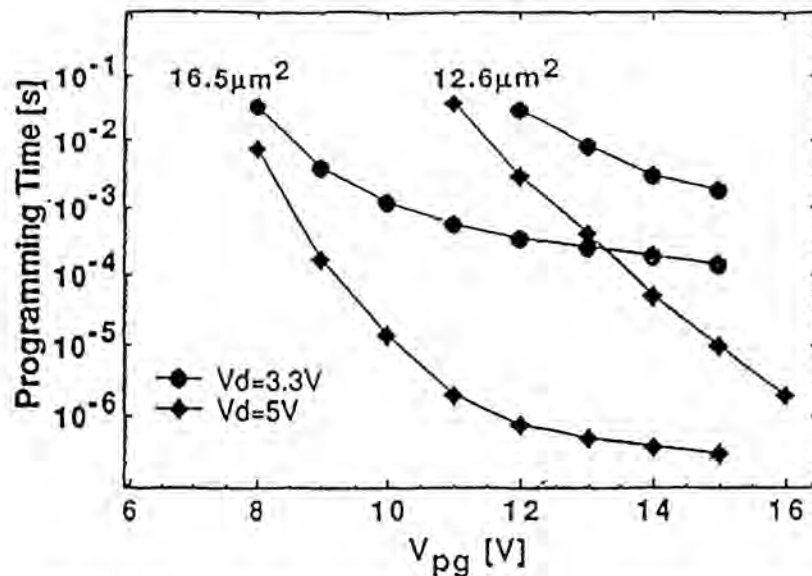


Figure 4.69 The programming time versus V_{pg} for a 12.6 and 16.5 μm^2 cell and for 3.3 V/5 V drain voltages [4.99].

4.5.3 3-D Cell Structures

Recently, three-dimensional (3-D) structures have been reported to realize small cell structures [4.67, 4.83]. Cell size reduction is realized by forming the channel in a vertical direction in contrast to conventional cells where the channel length is lateral, along the surface of the silicon. The major challenge in such structures is the uniformity of the tunnel oxide. Three-dimensional structures offer both promise and challenges.

4.5.3.1 Trench-Embedded Field-Enhanced Tunneling (TEFET). The TEFET cell is a three-dimensional, single-transistor, floating gate memory cell. As in a conventional planar Flash EEPROM cell [4.83], the TEFET cell is programmed by channel hot-electron injection and erased by F-N tunneling. The small cell size is achieved by exploiting trench technology to integrate the NMOS transistor vertically along the side-walls of a trench, as shown in Fig. 4.70. The etched trench extends from the planar surface through the n^+ drain, p-body, and a graded n-type region into the underlying heavily doped n^+ layer that serves as the source. By patterning this underlying n^+ layer, a memory array can be divided into erase sectors. Both the control gate and the floating gate are embedded within the interior of the trench. This vertical structure provides inherently smaller cell size than a conventional planar cell. For example, with $0.7\text{ }\mu\text{m}$ lithography, the cell size can be as small as $2.7\text{ }\mu\text{m}^2$, which is a factor of two smaller than a conventional planar cell. In addition, the trench EEPROM cell's vertical channel allows the cell pitch to be scaled inde-

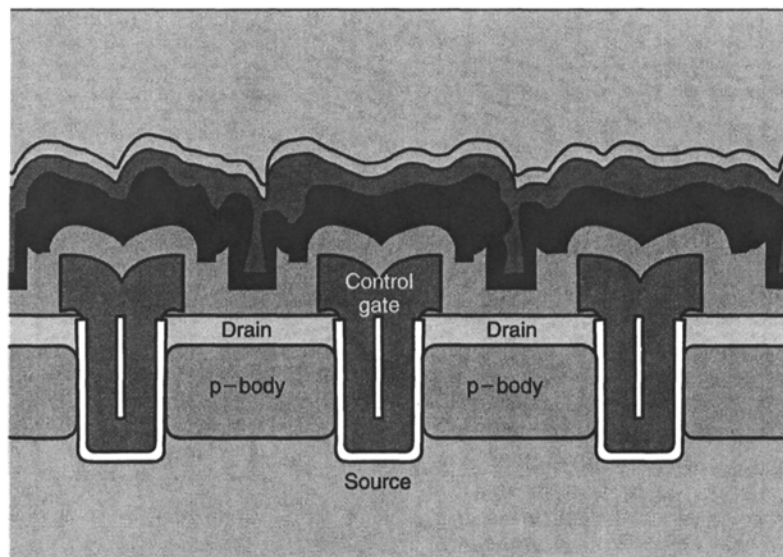


Figure 4.70 SEM cross section of a TEFET cell [4.83].

pendent of the channel length. Thus, the inherent hot-electron effects and punch-through limitation associated with the scaling of planar cells are eliminated, and the freedom of tailoring the channel doping profile for optimal cell performance need not be traded off for scaling. Consequently, the trench cell is suitable for scaling down to the very deep sub-micron regime.

Moreover, the trench cell has a substantially larger channel width than a planar cell of comparable lithography because the trench cell's channel width is equal to the circumference of the trench, whereas the planar cell's channel width is only comparable to one edge of the trench opening. The larger channel width leads to a higher read current and, in turn, a shorter access time. Furthermore, the enhanced electric field at the bottom of the trench allows a low erase voltage. The field enhancement also permits a Fowler–Nordheim (F–N) tunneling erase without inducing band-to-band tunneling and the associated hot-hole injection, which is the major limiting factor for a planar cell's cycling endurance [4.12]. By eliminating hot-hole injection, the trench cell's endurance can be greatly improved.

4.5.3.2 3-D Side-Wall Flash EEPROM Cell. The side-wall Flash EPROM cell is a 3-D, single-transistor, floating gate memory cell that is formed on the side-walls of a narrow silicon pillar (Fig. 4.71) [4.67]. The pillars are approximately $1\text{ }\mu\text{m}$ high, and each pillar forms a separate cell. Current flows vertically along the side-walls of the pillar from the top of the pillar (drain) to the base of the pillar (source) (Fig. 4.72). The floating gate completely encircles each pillar, and similarly, the wordline encircles the floating gate. The cell is programmed by hot-electron injection at the drain end (top) of the silicon pillar, while erase is performed by Fowler–Nordheim tunneling of electrons from the floating gate to the source. The key idea in implementing the cell within an array is to have a smaller space between pillars in the same wordline row than between pillars in adjacent wordline rows. When the wordline polysilicon is deposited, it completely fills the gaps between pillars in the wordline direction but does not fill the gaps between pillars in adjacent

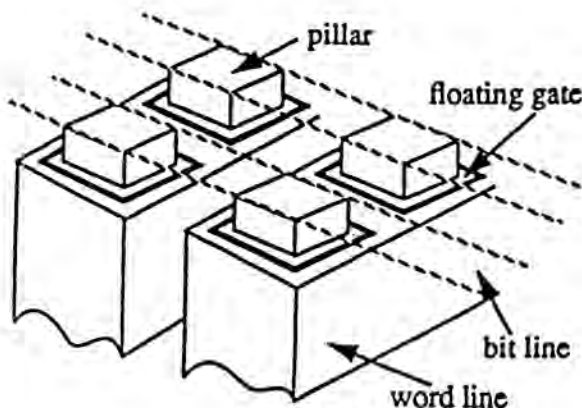


Figure 4.71 Schematic of the side-wall Flash EPROM array [4.67].

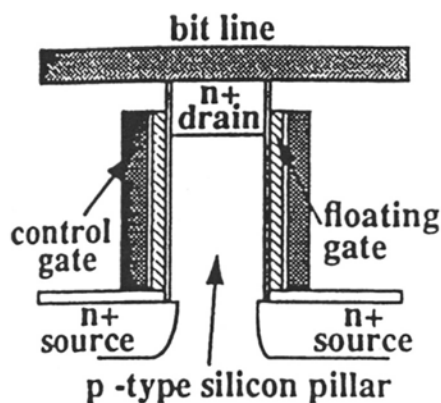


Figure 4.72 Cross-sectional diagram of 3-D cell [4.67].

wordline rows. Therefore, after blanket etching of the wordline polysilicon, the wordlines remain continuous in the wordline direction but are isolated from adjacent wordlines. Bitlines run perpendicular to the wordlines making contact to the top of each pillar. The wordline and floating gate edges are 300 nm below the top of the silicon pillar, which provides the margin that allows the bitline contact to overlap the edges of the pillar without shorting to the gates. Thus, the bitline does not have to have a separate contact opening to the top of every pillar, but can be formed by etching a continuous trench in the planarizing dielectric to expose the top of the pillars along a single bitline. These trenches are then filled with the bitline metal to form a continuous bitline. With this approach, no registration tolerance is required for the bitline contact, and thus, the bitline pitch can be equal to the minimum lithographic pitch.

The key advantages of the cell and array are (1) a cell size within 10% of the square of the minimum pitch with > 50% reduction when compared to the conventional NOR-type planar cell [4.101]; (2) a channel length independent of the lateral dimensions, thereby allowing aggressive lateral scaling without concern for short-channel effects; (3) inherent complete isolation between cells; (4) read current that is effectively three times that of an equivalent generation planar cell; (5) true cross-point array, thereby simplifying circuitry; (6) metal or silicide low-resistance bitlines; (7) simple optimization of the cell channel doping profile, and (8) low bitline diffusion capacitance.

4.5.4 Multilevel Cell

The primary technique for increasing the memory density is to reduce the size of the memory cell. Another approach to improving memory density is to increase the number of possible states in a cell. The multilevel cell concept has been discussed for volatile and nonvolatile memories [4.92–4.96], but has been difficult to implement on a commercial level. Recently, it has been reported that Intel, Sandisk, AMD, and Toshiba have been working on Flash memories based on multilevel cell technologies [4.97, 4.118, 4.125].

Multi-bit storage takes advantage of the analog nature of the Flash storage element. The conventional 1 bit/cell approach would place the cell in one of two states, a "0" or "1," using a program or erase operation. Erase might be denoted as the absence of charge and program as the presence of charge on the floating gate. Thus, the cell is placed in one of two discrete charge bands. If programming can be done accurately enough, the cell can be placed in one of four discrete charge bands achieving 2 bits/cell storage (Fig. 4.73). An example of state assignment is shown in Table 4.10.

The number of states is determined by

1. Total available charge range (window)
2. Ability to accurately program a state
3. Ability to accurately read a state
4. Disturbance of a state over time

Because of the high accuracy required for programming and sensing, multilevel storage is expected to be somewhat slower than conventional 1 bit/cell Flash for these operations. Multilevel storage also has a lower tolerance for disturbance of the

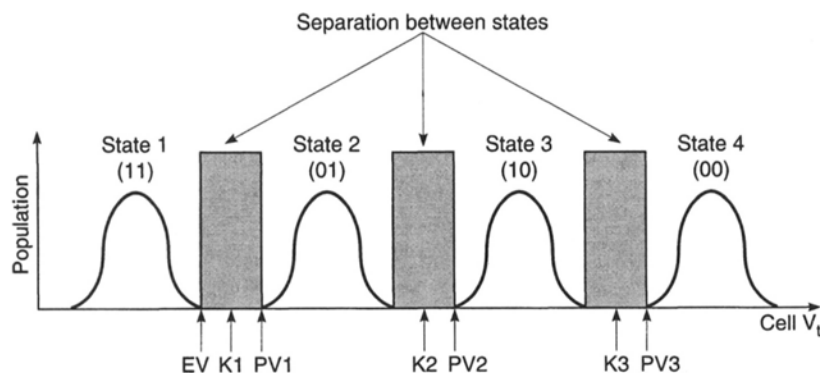


Figure 4.73 Multilevel-cell threshold voltage distribution [4.118].

TABLE 4.10 AN EXAMPLE OF STATE ASSIGNMENT FOR A 2 BIT/CELL

State	V_t Range
0	$< 3.0 \text{ V}$
1	$3.25 \text{ V} - 3.75 \text{ V}$
2	$4.25 \text{ V} - 4.75 \text{ V}$
3	$> 5.0 \text{ V}$

stored data (e.g., data retention) than conventional Flash, and thus will most likely require some form of error correction in the system or will be used in loss tolerant applications for the higher bit/cell densities. At what bit/cell density error correction is required is not yet clear, but it is expected to be a strong function of the process technology capability and the system requirements.

4.6. FLASH RELIABILITY ISSUES

As with other semiconductor-based memory devices, there are reliability issues with the programming, erasing, and read operation of Flash memory devices. For example, device degradation due to programming with channel hot-electrons and erasing with hot-holes lead to serious changes in the performance of Flash devices. This and other aspects of Flash reliability, such as dielectric breakdown, program/erase endurance, data retention, etc., are addressed in the following sections.

4.6.1 Channel Hot-Electron Programming

The channel hot electron (CHE) has been widely used for programming EPROM's and several Flash memories. This section describes the reliability issues of channel hot electron programming and how to minimize degradation effects associated with CHE programming.

4.6.1.1 Electron Trapping (N_{ox}) in $V_{fg} > V_d$ regime. Under typical programming conditions of $V_{cg} = 12V$, $V_d = 5.5V$, and $V_s = V_{sub} = 0V$, the EPROM cell is biased in the so-called $V_{fg} > V_d$ regime so that the direction of the oxide field near the drain favors electron injection [4.104, 4.121]. In MOS transistor degradation studies under similar stress conditions, Doyle et al. [4.106] found that the V_t shift was caused predominantly by electron charge trapping (N_{ox}) in the nearby oxide. Their experimental results also suggested that the trapped electrons could be subsequently detrapped by channel hot-hole injection when stressed at $V_g = 1/2 V_d$. In a floating gate structure, the presence of trapped electrons in the oxide near the drain edge can repel the channel hot electrons and reduce electron injection. This effect can be fairly significant in Flash memory cells where the programmability degrades substantially after repeated program/erase (P/E) cycles as seen in Fig. 4.74. That is, the cell program V_t decreases progressively versus P/E cycling when programming time duration is fixed, and the V_t versus cycling window closes (Fig. 4.74).

As a result, electron trapping in the gate oxide, possibly accompanied by electron trap generation, under $V_{fg} > V_d$ bias conditions is of real concern for cell reliability. Liang, et al. [4.107] observed that preexisting electron traps in the oxide would trap-up electrons to a saturation level independent of the oxide field when it is below 5 MV/cm (Fig. 4.75). At oxide fields larger than 5 MV/cm, electron trap generation and the amount of electron trapping can increase dramatically. This suggests that the voltage difference ($V_{fg}-V_d$) in a memory cell during programming needs to be kept reasonably low in order to avoid excessive electron trapping at the

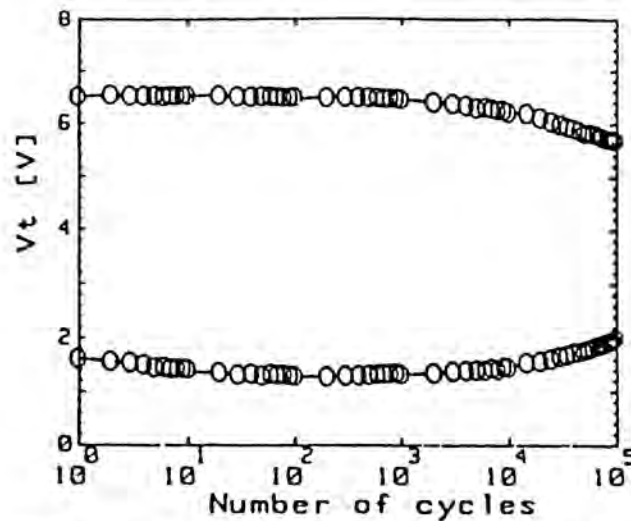


Figure 4.74 Program/erase endurance curve of a typical Flash memory cell [4.104].

drain edge. Nishida and Thompson [4.108] further observed that, for sub-10 nm oxides, there is a reduced trap generation and a lower saturation oxide charge density compared to thicker oxides as shown in Fig. 4.76. For the 10 nm gate oxides being used in Flash cells, electron trapping is reduced over its thick oxide EPROM counterpart.

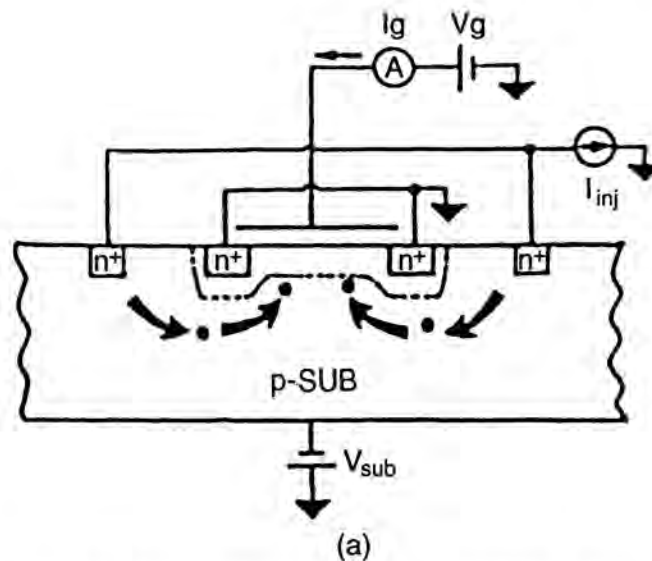
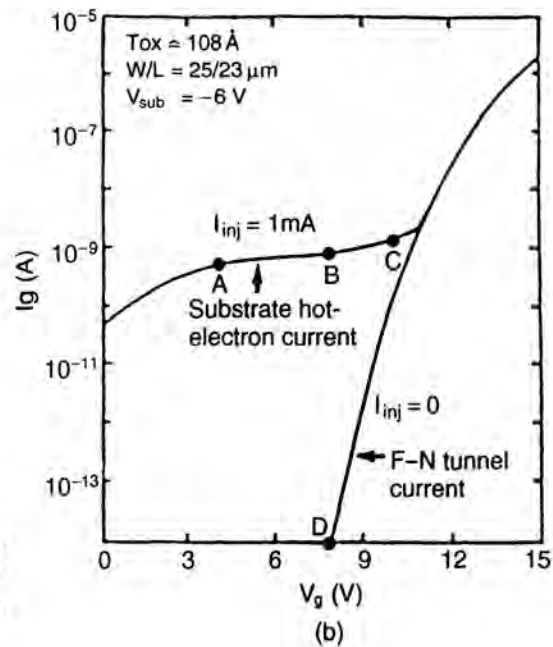
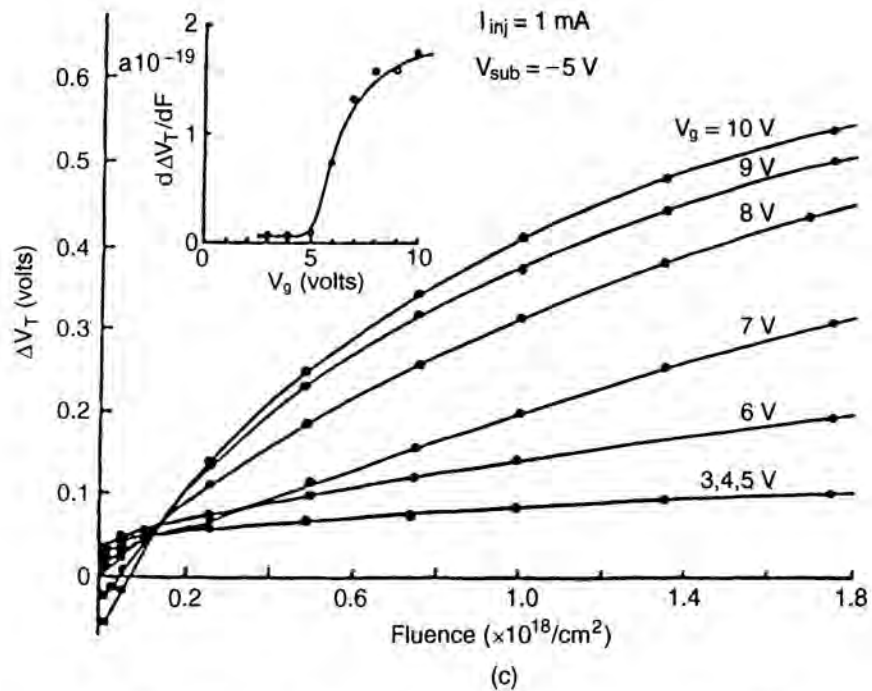


Figure 4.75 (a) Schematic diagram of substrate hot-carrier injection [4.107].



(b) I_g versus V_g for F-N tunneling (with $V_{sub} = 0$, $I_{inj} = 0$) and substrate hot-electron injection (with $V_{sub} = -5$ V, $I_{inj} = 1$ mA) [4.107].



(c) ΔV_T versus electron fluence for different gate biases. The insert shows the slope of the curves at high fluence versus V_g [4.107].

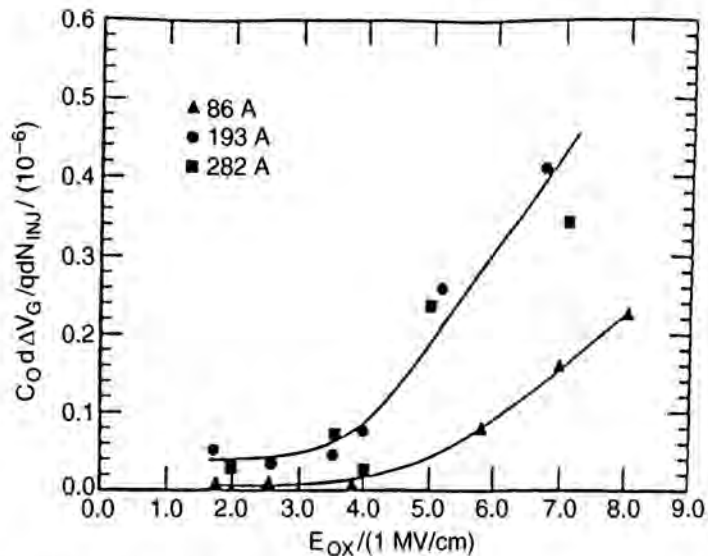


Figure 4.76 Oxide field dependence of slope of gate voltage shift versus electron fluence calculated at constant electron fluence of $6 \times 10^{17} \text{ cm}^{-2}$ for 900°C , 0% TCA dry oxide with thicknesses of 86 Å and 282 Å [4.108].

4.6.1.2 Surface State Generation (N_{ss}) in $V_{fg} < V_d$ Regime. As the cell is being programmed, the floating gate potential rapidly collapses. The lateral channel electric field, as well as the average electron energy, rises (Fig. 4.77) [4.104]. Moreover, near the drain end, the oxide field reverses direction and can attract hot holes generated in the channel. Upon co-injection of hot electrons and hot holes, it is well known that interface state (N_{ss}) generation can become important [4.106]. This can lead to device transconductance (G_m) degradation as a result of mobility reduction. Therefore, the ending condition of cell programming operation also needs to be carefully considered [4.104].

4.6.1.3 Trap Reduction Through Nitridation. Besides optimizing programming and erasing conditions, the composition of a gate oxide can also play a part in reducing the trap density in the tunnel oxide. Several studies have shown that nitridation of tunnel oxide makes the tunnel oxide more resistant to electron trap generation [4.131–4.133]. Fukuda et al. [4.56] showed that both electron and hole trap densities are greatly reduced in heavily nitrided tunnel oxide films (RTONO). Figure 4.78 shows trap density generation versus P/E cycles, and Fig. 4.79 presents the improved endurance characteristics of a heavily nitrided, tunnel oxide-based Flash memory cell. Figure 4.80 shows that $1 \times 10^{20} \text{ N-atoms/cm}^3$ are incorporated into the oxide and that the number of H-atoms, which are the origin of electron traps [4.132], decreases in RTONO films.

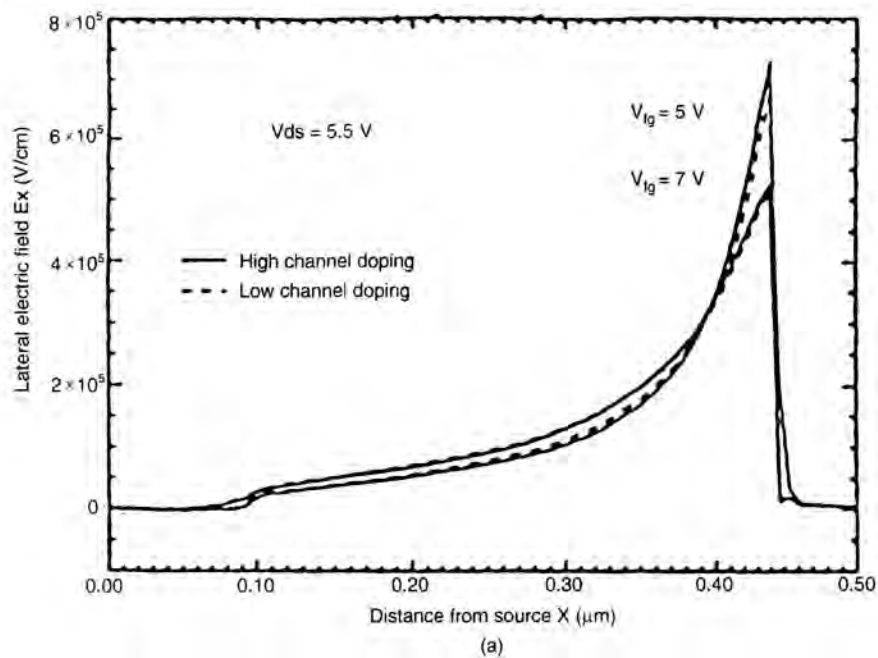
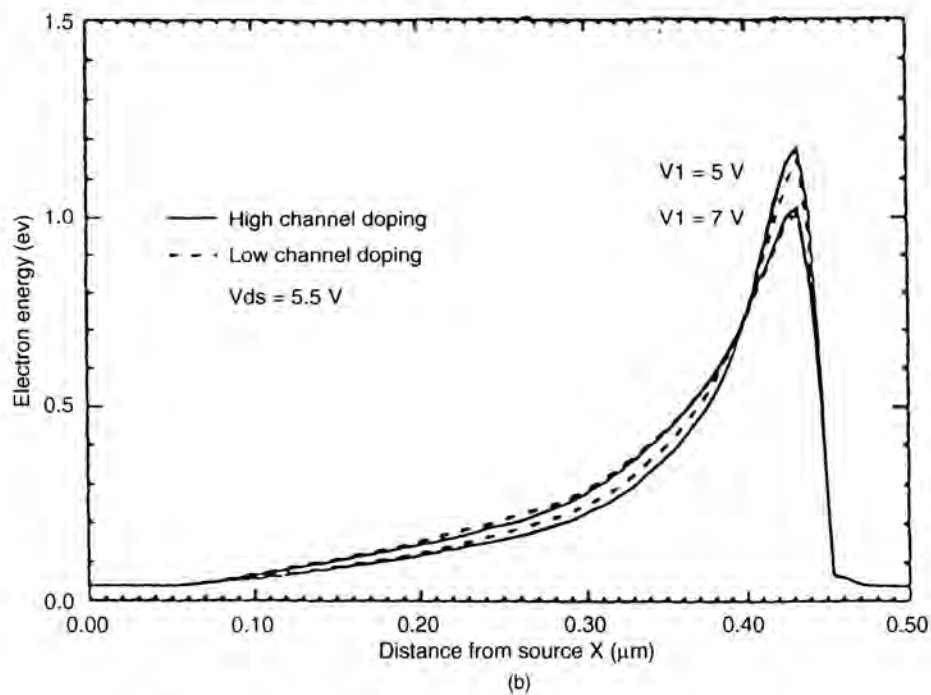


Figure 4.77 (a) Lateral electric field across the channel for $V_{fg} = 5$ V and $V_{fg} = 7$ V for two-channel doping calculated by 2-D device simulator UMDFET2 [4.104].



(b) Corresponding average electron energy of Fig. 4.77(a) [4.104].

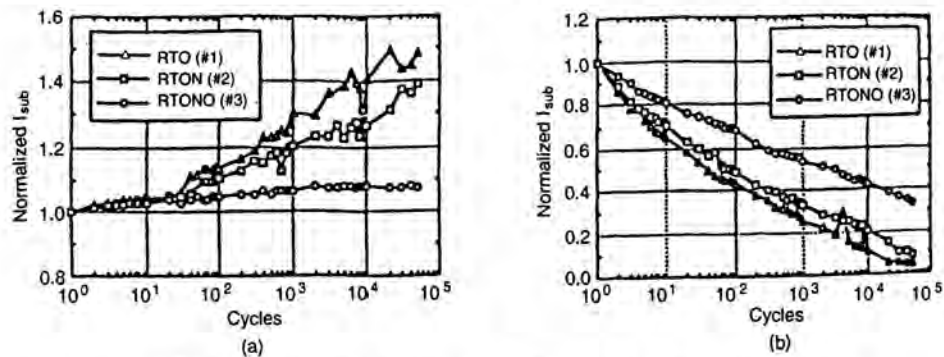


Figure 4.78 I_{sub} shift as a function of the P/E cycles for the (a) forward and (b) reverse biases [4.56].

Preparation sequences employed

	RTO	RTN (RTON)	RTON
RTO (#1)	$O_2, 1100^\circ C, 50\text{ s}$		
RTON (#2)	$O_2, 1100^\circ C, 40\text{ s} \rightarrow N_2 O, 1100^\circ C, 20\text{ s}$		
RTONO (#3)	$O_2, 1100^\circ C, 37\text{ s} \rightarrow NH_3, 900^\circ C, 60\text{ s} \rightarrow N_2 O, 1100^\circ C, 30\text{ s}$		

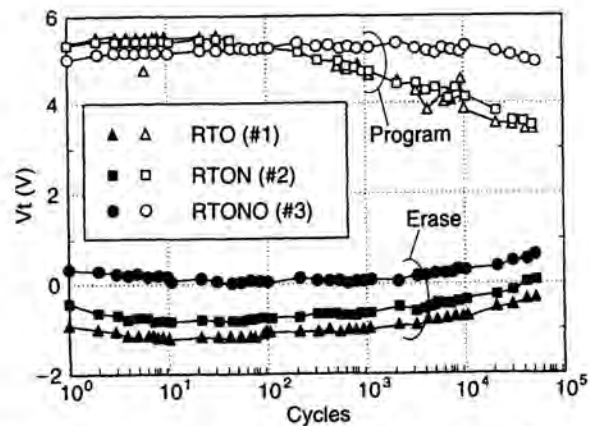


Figure 4.79 Endurance characteristics of the flash cells #1-#3 [4.56].

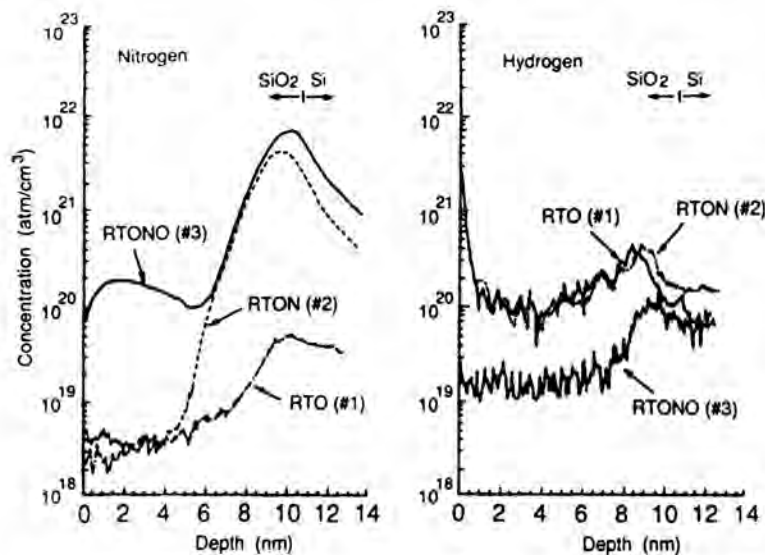


Figure 4.80 SIMS depth profiles of N and H atoms in the oxide films RTO (#1), RTON (#2), and RTONO (#3) [4.56].

4.6.2 Grounded-Gate Source-Erase-Induced Cell Degradation

Unlike the conventional EEPROM cell, the Flash EPROM cell almost always has to contend with the existence of band-to-band tunneling (BBT) current, as shown in Fig. 4.81 [4.113]. The BBT current is due to electron hole pairs generated in the gated-diode source junction during erase (Fig. 4.82). Electron hole pair generation occurs as a result of the presence of a high surface field in the n^+ junction depletion region [4.113–4.115, 4.121–4.123].

When high voltage is applied to the source junction with the control gate grounded, a deep depletion region is formed underneath the gate-to-source overlap region. Electron hole pairs are generated by the tunneling of valence band electrons into the conduction band. The electrons are collected by the source junction, and the holes are collected by the substrate. Since the minority carriers generated thermally or by band-to-band tunneling in the source region flow to the substrate due to the lateral field near the Si–SiO₂ interface, the deep depletion region remains present and the band-to-band tunneling process can continue without creating an inversion layer. The generated holes gain energy because of the electric field in the depletion region. While the majority of these generated holes flow into the substrate, some of them gain sufficient energy to surmount the Si–SiO₂ barrier and are trapped in the oxide [4.113–4.115].

4.6.2.1 Band-to-Band Tunneling Generated Hot-Holes. Band-to-band tunneling generated hot holes have been reported to cause device degradation

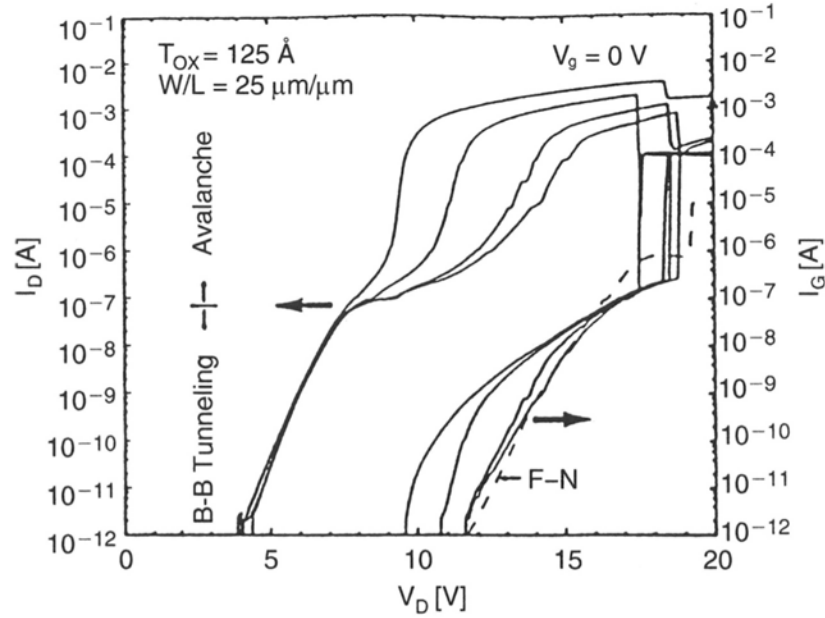


Figure 4.81 Measured drain and gate currents for the 125 Å oxide devices with channel surface concentrations of 10, 5, 2, and 1 E16 cm⁻³ (from left to right). The dashed curve represents the normalized F-N gate current [4.113].

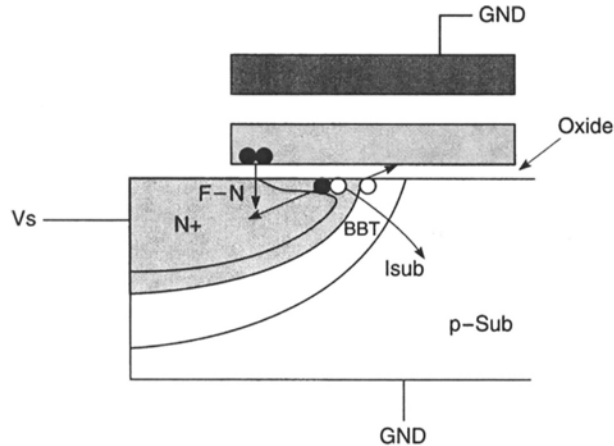


Figure 4.82 Cross-sectional drawing of the cell source diffusion with band-to-band (BBT) current and Fowler-Nordheim (F-N) current [4.104].

[4.113–4.117, 4.123], and hot-hole injection during erase has been reported to cause erased V_t nonuniformity (Fig. 4.83) in the memory array [4.17] and to speed up erase time (Fig. 4.84) versus cycling [4.104]. Trapped holes in the oxide have also been shown to degrade the charge retention (Fig. 4.85) of the memory cell [4.12, 4.21] and to speed up gate-disturb (Fig. 4.86) during the program mode [4.12, 4.21].

Band-to-band tunneling leakage current has also been reported in trench-capacitor DRAM cells [4.124]. This may pose a serious problem in scaling down the trench-capacitor cell since the leakage current from band-to-band tunneling increases drastically with decreasing capacitor oxide thickness [4.84].

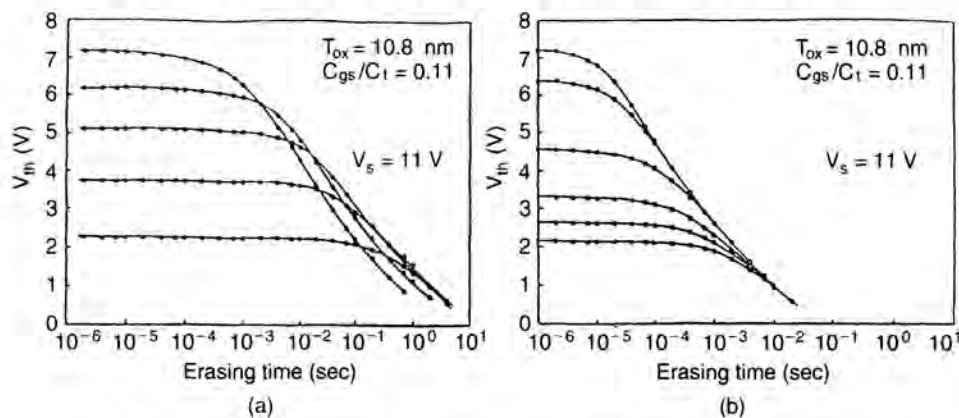


Figure 4.83 Erased V_{th} versus erasing time with programmed V_{th} as a parameter (type-a: n^+ doping $< 2E15 \text{ cm}^{-2}$, type-b: n^+ doping $> 5E15 \text{ cm}^{-2}$) [4.17].

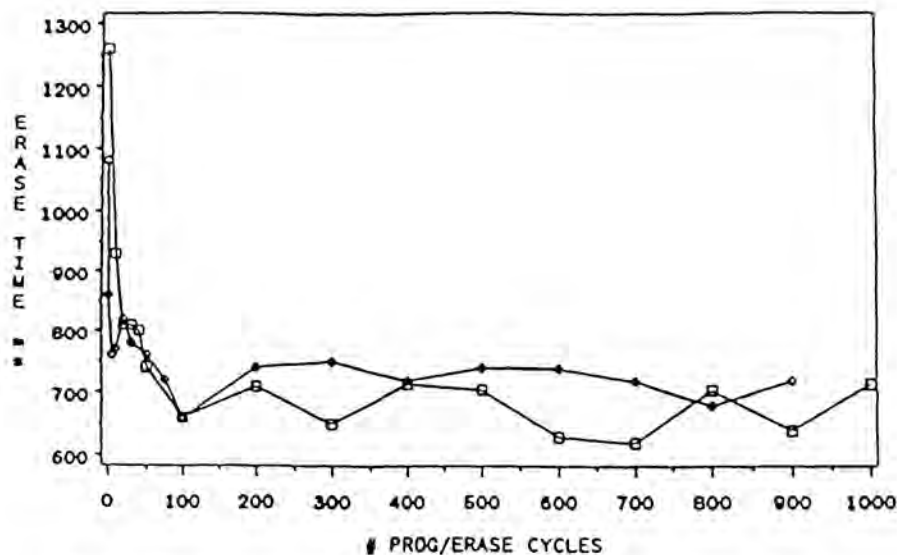


Figure 4.84 Erase time versus P/E cycles. Erase time speedup is due to trapped holes [4.104].

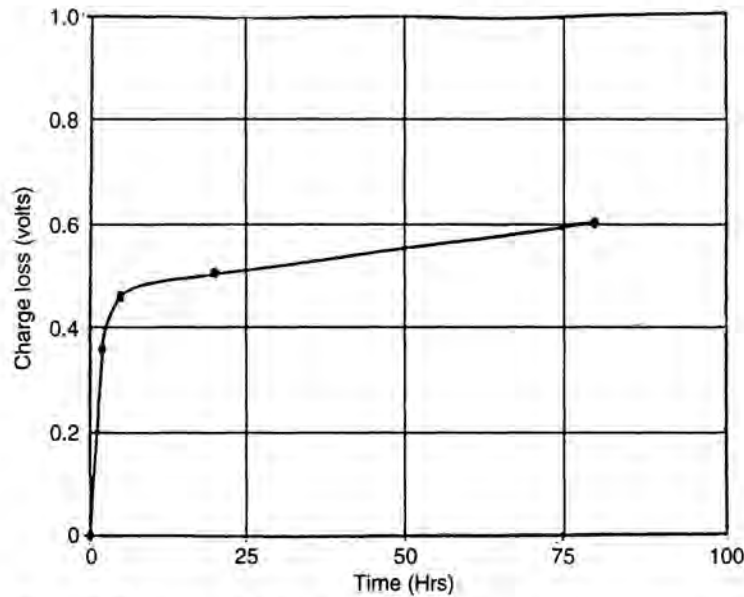


Figure 4.85 Temporal behavior of charge loss seen for system arrays at 250°C. Each point represents the mean worst bit margin shift in the arrays [4.21].

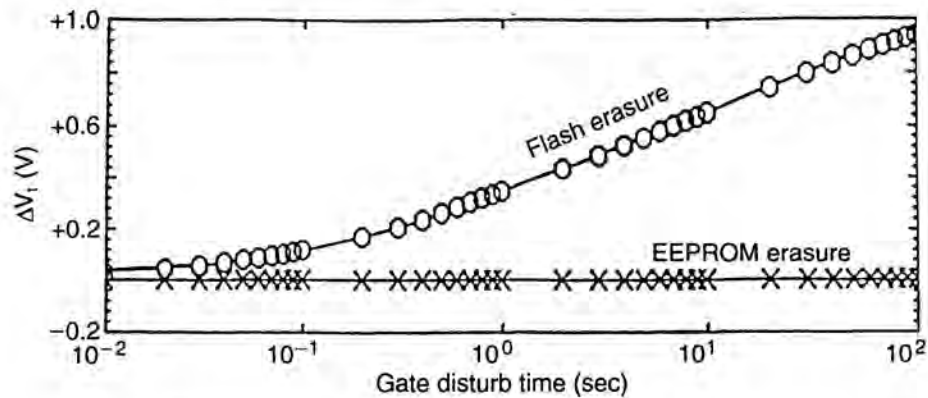


Figure 4.86 Threshold voltage shifts during gate-disturb after the cells having been cycles 20 times, using Flash (O) and EEPROM (X) erasure. The cells were erased to a V_t of 1 V before gate-disturb measurement [4.12].

4.6.2.2 Avalanche Breakdown-Induced Cell Wearout. Aside from BBT, avalanche breakdown can occur during erase if the source junction of the cell is not made robust relative to the erase voltage. In an arsenic n^+ to p-substrate abrupt junction, this event can be particularly detrimental to the cell. The avalanche-

generated hot holes can gain sufficient energy to cause severe interface damage (Fig. 4.87) [4.116, 4.117], which can lead to serious degradation of transconductance (G_m) (Fig. 4.88) [4.117].

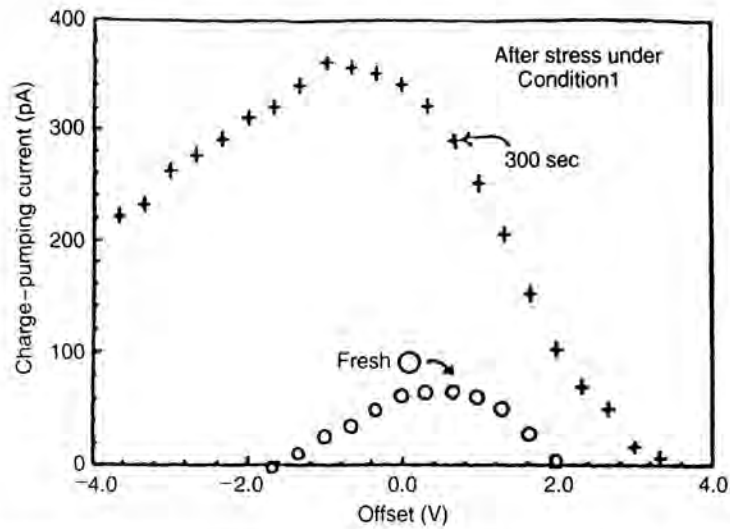


Figure 4.87 Interface states increase measured by charge-pumping current after 300 sec of stress at $V_g = 2.5$ V and $V_d = 9.6$ V with $V_{sub} = 0$ and V_s floating [4.117].

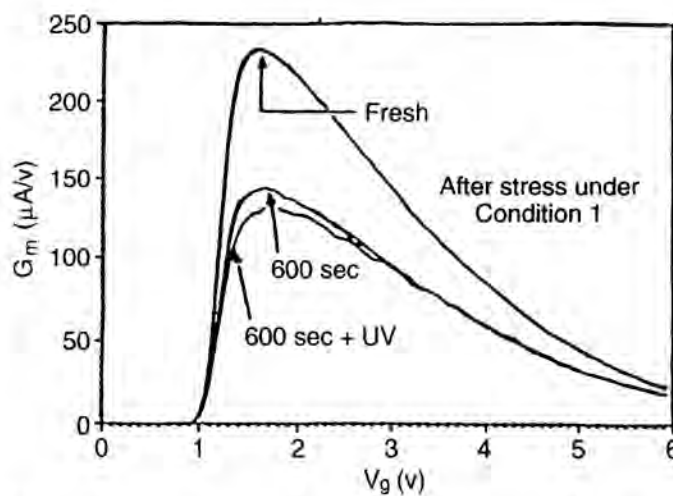


Figure 4.88 G_m reduction due to stress condition in Fig. 4.87 [4.117].

4.6.2.3 Reduction of Hot-Hole Injection. Generally speaking, junction-generated hot holes can be reduced through changes in the cell structure or changes in the erase operation.

4.6.2.3.1 DOUBLE-DIFFUSED JUNCTION. Since the heating of BBT-generated holes is greatly affected by the lateral electric field in the source junction, reducing the electric field by a graded source junction would reduce heating of holes. By introducing an additional phosphorous diffusion, not only is the BBT current shifted to a higher n^+ voltage, but the avalanche breakdown is also softened up (Fig. 4.89) [4.89]. This provides some margin against hot-hole injection during a memory cell erase operation. The optimization of the n-region is a tradeoff between the source junction breakdown voltage and the capacitive-coupling ratio between the source junction and the floating gate [4.17].

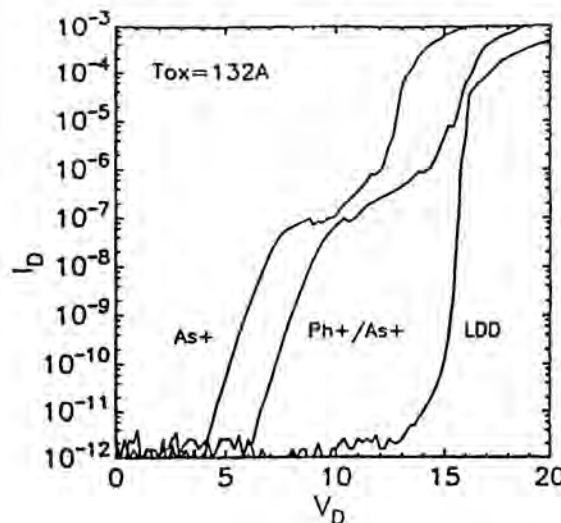


Figure 4.89 Comparison of BBT leakage currents in devices made with As^+ , P^+/As^+ , and LDD junction [4.89].

4.6.2.3.2 REDUCTION OF APPLIED SOURCE VOLTAGE. In a grounded-gated, source-erase configuration, one method of reducing hot-hole injection from band-to-band tunneling is to reduce the electric field by reducing the applied voltage to the source.

4.6.2.3.3 NEGATIVE GATE, FLOATING GATE TO SOURCE ERASE. Erase can be accomplished with the gate at a high negative voltage ($-10V$) and the source at V_{cc} ($+5V$) as shown in Fig. 4.13a [4.12, 4.48, 4.101, 4.119]. In this approach, the vertical field is kept constant, and the F-N erase current is unchanged. However, the lateral field is greatly reduced (only $5V$ at the source junction); in that way, the BBT-generated holes are not heated as much. In addition, the “apparent” avalanche breakdown voltage has been shifted up (Fig. 4.90) [4.12] so that the breakdown-related damage to the device is no longer an issue [4.12, 4.101].

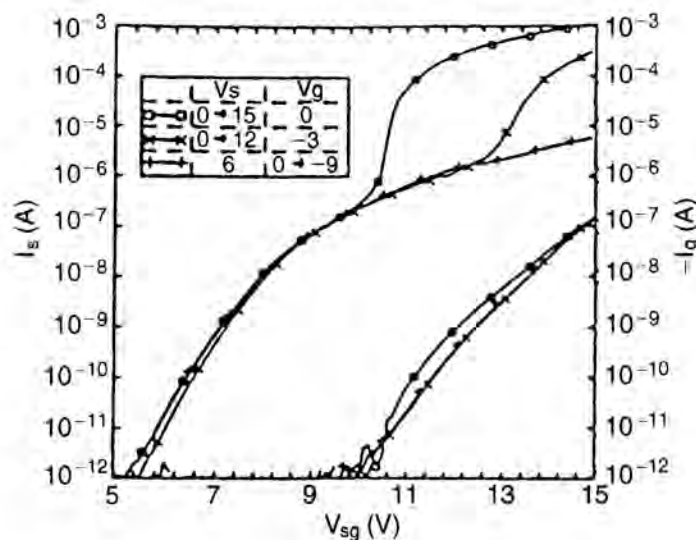


Figure 4.90 Source current, I_s , and gate current, I_g , versus source-to-gate voltage, V_{sg} , for three different bias conditions as defined in the inset. $V_{sg} = V_s - V_g$ defines the oxide field for tunneling [4.12].

4.6.2.3.4 NEGATIVE GATE, FLOATING GATE TO CHANNEL ERASE. This method, also called channel erase (Fig. 4.13b [4.119, 4.120]), is typically accomplished with the gate at a high negative voltage (-13 V) and the channel in the well at V_{cc} ($+5$ V). Since there is no reverse-biased junction in this case, BBT hot holes and junction breakdown are no longer a concern, and the cell is erased with uniform F-N tunneling across the gate oxide.

4.6.2.4 Drain-Leakage-Induced Over-Erase. If bits, which are outside the main V_t distribution, are over-erased and become sufficiently depleted, they can cause program verify failures in the following program cycle. For example, it has been reported that fast bit erase can occur if a low level of bitline leakage is present during high-voltage source erase (Fig. 4.91) [4.126]. This DC leakage path will cause the channel current to flow as the memory cell is sufficiently erased. A large amount of channel hot holes are easily generated under such weak channel turn-on and large V_{ds} conditions. The injection of channel hot holes will accelerate the cell erase and over-erase the fast bit. Such bits, which are outside the main distribution, can be screened out during wafer sort if they are well-behaved or stable.

4.6.2.5 Erratic Bits. It has also been reported [4.126, 4.127] that, in a large memory array, a small number of bits can behave erratically during cycling in a somewhat random fashion, as shown in Fig. 4.92. This phenomenon has been

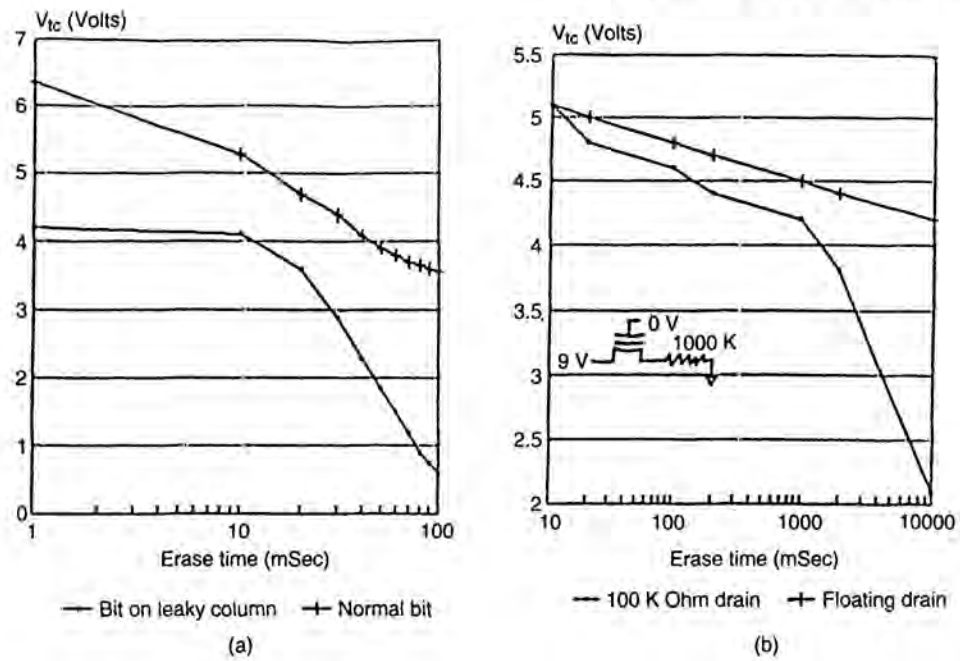


Figure 4.91 (a) Single fast erase bit on column with bitline leakage. (b) Single-cell erase with and without drain leakage [4.126].

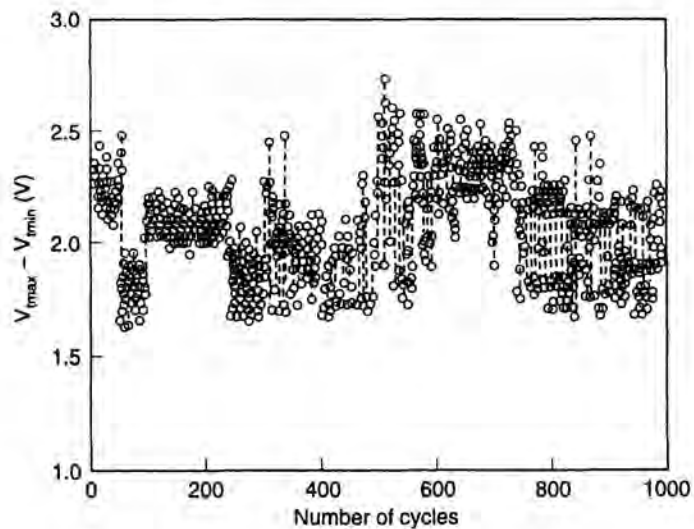


Figure 4.92 Variation in erase distribution (V_t spread) as a function of number of cycles [4.126].

attributed to dynamic hole trapping and detrapping/neutralization in the oxide, which leads to F-N tunneling current fluctuation. Under a worst case scenario, a bit can become so depleted during erase as to cause a soft or temporary failure in cycling. After reconditioning, the problem bit, by some external stimulus such as gate stress or temperature bake, can become a normal bit again [4.127]. This random bit instability effect, if not controlled, can actually lead to early endurance failure.

4.6.3 Source-Side Injection

While source-side injection Flash memory cells have a relatively higher injection efficiency than cells utilizing channel hot-electron injection, the injection is concentrated near the narrow gap. Because of this split gate gap structure, a small amount of electron trap-up in the spacer-like oxide between the floating gate edge and the control gate can change the local electric field and limit the endurance of these structures.

4.6.4 Stress-Induced Oxide Leakage

The reliability of the tunnel oxide is one of the fundamental reliability parameters for Flash memories. These devices are either erased (e.g., stacked gate) or both programmed and erased by high-field (Fowler–Nordheim) injection of electrons into a very thin dielectric film to charge and discharge the floating gate. Unfortunately, this current through the oxide degrades the quality of the oxide and eventually leads to breakdown. The wearout of tunnel oxide films during high-field stress has been correlated with the buildup of both positive or negative trapped charges.

Recently, it has been shown that high-field stress also induces a low-field leakage current in thin oxides, a thickness less than 10 nm [4.136, 4.137, 4.140]. The mechanism of this leakage current has been attributed to the generation of localized defects or weak spots [4.137] and trap states near the injecting interface [4.135, 4.140, 4.141]. These low-field leakage currents degrade the data retention of the EEPROM cell [4.140, 4.1138, 4.38]. Unfortunately, these oxide leakage currents increase with decreasing oxide thickness [4.137, 4.138] and make it difficult to scale down the oxide thickness of the memory cell. The read-disturb observed in Flash memories [4.135], as discussed in Section 4.6.9, is the direct result of high-field stress during P/E cycling.

Aritome et al. [4.38] studied the influence of the waveform of the stress voltage on the degradation of the tunnel oxide. Three types of high-field dynamic stress were used to study the thin oxide leakage currents. Figure 4.93 shows the applied dynamic stress waveforms. In the case of bipolarity stress, positive high voltages were applied to the gate or to the substrate and source-drain (S/D) regions. Fowler–Nordheim tunneling occurs alternately from the substrate to the gate and from the gate to the

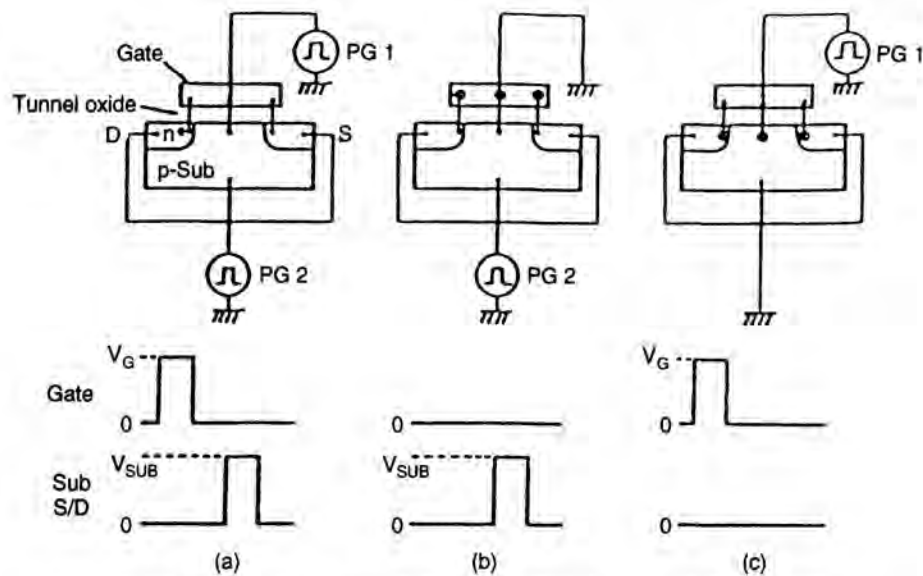


Figure 4.93 Setup and stressing waveform for (a) bipolarity stress, (b) electron-emitted stress, and (c) electron-injected stress [4.38].

substrate. Electron-emitted stress (emission of electrons from the gate) and electron-injected stress (emission of electrons from the substrate) are performed for comparison. The applied gate voltage is comparable to the floating gate voltage of a memory cell during the write operation. The substrate voltage (V_{sub}) is chosen in such a way that the tunnel current during emitted and injected stress is the same.

The $I_g - V_g$ characteristics before and after dynamic stress for a 5.6 nm oxide thickness are shown in Fig. 4.94. The thin oxide leakage currents, as a function of three types of high-field dynamic stressing, are compared. It is observed that the thin oxide leakage current induced by bipolarity stress is about one order of magnitude smaller than that induced by both the electron-emitted and electron-injected stresses. This result shows that the origin of the thin oxide leakage current can be suppressed by a reverse high-field stress.

4.6.4.1 Time-Dependent Dielectric Breakdown (TDDB) under High-Frequency Stress. It has been shown that TDDB characteristics (Fig. 4.95) are dependent on the stress waveform and its frequency [4.142, 4.143]. The stress waveforms are the electron-emitted stress, the electron-injected stress, and the bipolarity stress. The time-to-breakdown during bipolarity stress is longer than that for both electron-emitted stress and electron-injected stress. Moreover, it is shown that the TDDB characteristics depend on the frequency of the stress wave. The breakdown time under unipolarity stress does not depend on frequency. However, the breakdown time to bipolarity stress depends strongly on the frequency. The bipolarity

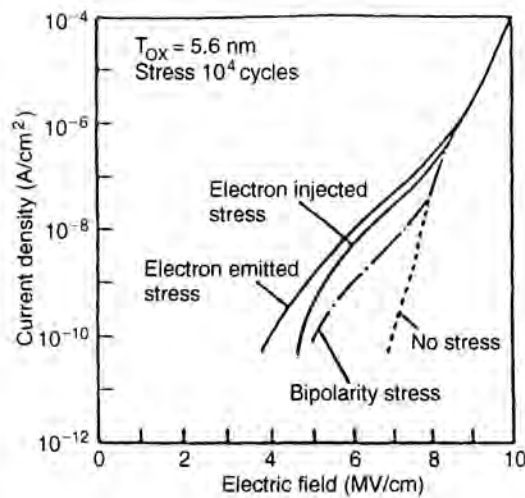


Figure 4.94 Leakage current of the thin oxide at low voltages for 5.6 nm oxide after bipolarity stress, electron-emitted stress, and electron-injected stress [4.38].

stress lifetime in the megahertz range is about 40 times higher than it is at 10 Hz for the same applied voltage. It is suggested that the improved TDDF obtained with bipolarity stress is due to the field-assisted detrapping of holes at the Si-SiO₂ interface [4.143].

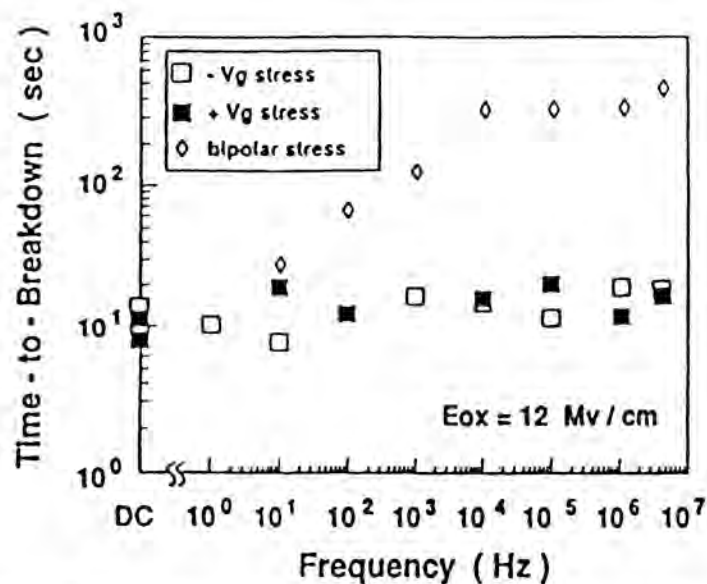


Figure 4.95 Dependence of time-to-breakdown on stress frequency [4.142, 4.143].

4.6.5 Poly-to-Poly Erase

While thicker oxide for poly-to-poly erase does not suffer from stress-induced oxide leakage, it does suffer from increased electron trapping compared to planar tunnel oxide. There are two reasons for this. First, the oxide grown on polysilicon is of lower quality than oxide grown on silicon. Typically, some phosphorus is embedded in the oxide. Second, with enhanced tunneling, the tunnel current flows primarily from the enhanced tunneling points. Electron trapping with the tunnel current tends to reduce the local field. The tunnel current reduction is larger as compared to planar tunneling. Typically, the window closing is about 10 times faster compared to FLOTOX EEPROM.

4.6.6 Reliability of Interpoly Dielectric

ONO (oxide–nitride–oxide) interpoly dielectrics are currently used for non-volatile memories. In this case, the important issues are not only low-defect density and long mean time to failure, but also charge retention capability. Mori et al. investigated the ONO interpoly dielectric thickness scaling effect on charge retention characteristics in nonvolatile memories [4.144].

For the top oxide layer, a certain thickness (3 nm), which can block hole injection, is required. Si_3N_4 thickness reduction leads to a reduction in initial rapid charge loss. However, too much reduction sometimes leads to enhanced charge loss in a long bake test. Bottom polyoxide thinning, down to 10 nm, does not result in degradation. The NO (nitride–oxide) double-layer structure shows an initially large rapid charge loss, which may be prevented by applying the ONO structure. However, if the bottom oxide layer is thin and its quality is poor, a portion of the bits may suffer from rapid initial charge loss. When scaling down the ONO thickness, one must consider these key factors concerning charge retention, as well as dielectric reliability.

4.6.7 Stacked Gate Write/Erase Endurance

As discussed previously, both channel hot-electron program and tunnel erase are degraded by repeated cycling. The typical result is a decrease in V_t window size when a fixed program and erase voltage and time are used. Figures 4.74 and 4.96 show V_t distribution versus write/erase cycles [4.27 and 4.104].

4.6.7.1 NAND Write/Erase Endurance. Write/erase endurance characteristics are shown in Fig. 4.97. The uniform write and uniform erase technology (charge transfer between floating gate and substrate) guarantees a wide cell threshold window (as large as 4 V), even after 1 million write/erase cycles (Fig. 4.97a). However, the threshold window obtained by the uniform erase and nonuniform write technology begins to decrease rapidly at approximately $1\text{E}2$ write/erase cycles, and fails at $1\text{E}5$ write/erase cycles (Fig. 4.97b) because the Fowler–Nordheim tunneling current during nonuniform writing is confined to a small region at the drain.

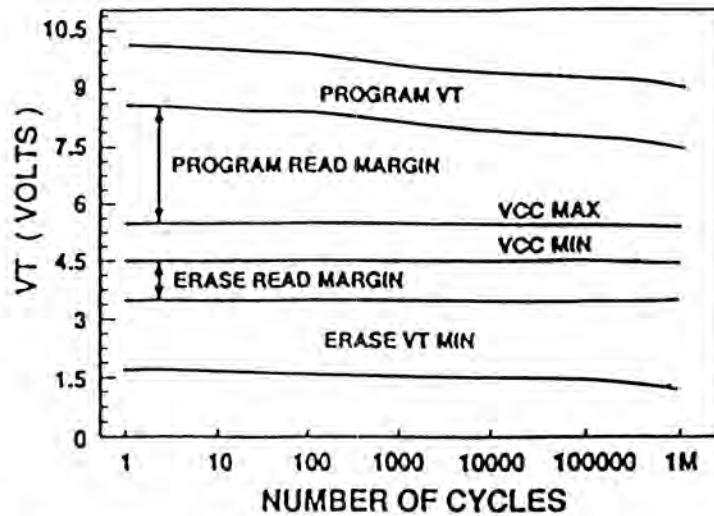


Figure 4.96 Stacked gate FLASH endurance characteristics [4.27].

Therefore, electron traps in the tunnel oxide are generated at a high rate near the drain area, and these electron traps impede electron injection and emission between the floating gate and substrate. This effect may be aggravated by holes that are generated by band-to-band tunneling.

In uniform write and uniform erase technology, the threshold voltage of the erased cell is dependent on the number of write/erase cycles. However, the threshold voltage of the written cell is not dependent on the number of write/erase cycles. This can be explained as follows. Oxide traps and interface traps are generated uniformly over the entire channel area because Fowler–Nordheim tunneling of electrons is

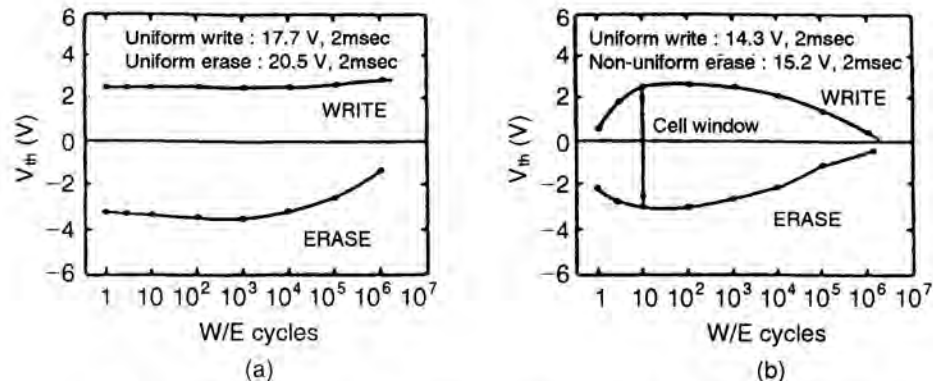


Figure 4.97 NAND endurance characteristics. (a) Uniform write and uniform erase technology; (b) uniform write and nonuniform erase technology [4.14].

performed uniformly during both the writing and erasing operations. The charges trapped uniformly over the channel area affect not only the electron-tunneling current through the oxide, but also the flatband voltage. The threshold voltage of the erased cell decreases slightly for up to 10^3 cycles due to hole trapping. Hole traps affect the increase of stored positive charge on the floating gate, as well as the decrease in flatband voltage. After 10^3 – 10^6 cycles, the threshold voltage of the erased cell increases due to electron detrapping. On the other hand, the threshold voltage of the written cell remains almost constant up to 10^6 cycles in spite of charge trapping in the oxide because of the influence of stored charge on the floating gate and trapped oxide charge on both the flatband voltage and injection field, canceling each other [4.103].

4.6.8 Data Retention

In floating gate memories, the stored charge can leak away from the floating gate through the gate oxide or through the interpoly dielectric. This leakage, caused by mobile ions, oxide defects, or other mechanisms, results in a shift of the threshold voltage of the memory cell. The threshold voltage shift can also be caused by detrapping of electrons or holes from oxide traps.

Different charge-loss mechanisms have been described [4.90, 4.91]. These mechanisms are briefly discussed here. The first loss mechanism is thermionic emission over the image force lowered potential barrier. However, this mechanism is not dominant since the barrier height for thermionic emission is considerably higher than the activation energies for the other charge-loss mechanisms. The second mechanism is called electron detrapping. The activation energy for this type of charge loss is about 1.4 eV. Because the intrinsic charge-loss rate decreases with time and stops after a threshold voltage drop of about 0.5 V [4.91], this type of charge loss is associated with the detrapping of electrons that are trapped in the oxide. The third mechanism is related to defects in the oxide and results in both charge loss and gain, depending on the biasing condition. The activation energy for this kind of charge loss is about 0.6 eV [4.91]. The fourth mechanism is related to contamination. Positive ions entering the memory cell may compensate a part of the negative charge stored on the floating gate. The activation energy for charge loss by contamination is about 1.2 eV [4.91].

Kynett et al. [4.27] have shown the results of data retention tests for a 1 MB Flash memory. Accelerated retention bake experiments performed at 250°C for 168 hours indicate that, after 10 K write/erase cycles, Flash memory will exhibit a 0.7 V program V_{th} shift [4.27]. Furthermore, a retention bake of 52 hours produced less than a 0.5 volt V_{th} shift in programmed devices which had been through more than 1 million write/erase cycles.

Table 4.11 shows the data loss of a Flash memory cell after various cycling conditions. The average charge loss is small in the case of a low applied voltage to the source because the hot-hole injection due to band-to-band tunneling is suppressed. The low electric field across the tunnel oxide during erasing also results in excellent write/erase endurance characteristics.

TABLE 4.11 CHARGE LOSS OF ETOX FLASH EEPROM CELL AFTER 100 CYCLES

Cycling Condition	Average Charge Loss
EEPROM program ($V_g > 0$ V) and erase ($V_g < 0$ V)	0.1 V
EEPROM program/Flash erase at $V_{\text{source}} = 12.5$ V; $V_{\text{cg}} = 0$ V	0.51 V
EEPROM program/Flash erase at $V_{\text{source}} = 11.5$ V; $V_{\text{cg}} = 1.5$ V	0.33 V
EEPROM program/Flash erase at $V_{\text{source}} = 8.5$ V; $V_{\text{cg}} = -6$ V	0.24 V
Uncycled	0.12 V

The data retention characteristics of Flash memory cells programmed by two different write/erase (W/E) technologies are compared in [4.38]. Figure 4.98 illustrates these two W/E techniques. The first is the bipolarity W/E technology, which is a uniform write and erase technology. During the write operation, a high voltage (V_{cg}) is applied to the control gate, with the substrate and source-drain regions grounded. Electrons are injected from the substrate to the floating gate over the whole channel area of the memory cell. In the erase operation, a high voltage (V_{sub}) is applied to the substrate and source-drain regions and the control gate grounded. Electrons are then emitted from the floating gate to the substrate. In this write/erase method, the high-field stress of the thin oxide corresponds to bipolarity stress. The other write/erase method utilizes channel hot-electron (CHE) for the write and F-N tunneling for the erase. This technology is a nonuniform write

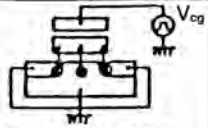
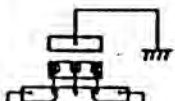
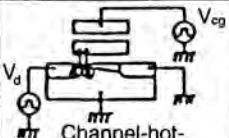
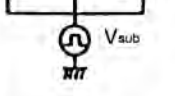
	Write (electron injection)	Erase (electron emission)	Oxide stress
Bipolarity FN-t Write/erase technology	 Fowler–Nordheim tunneling	 Fowler–Nordheim tunneling	Bipolarity stress
Channel-hot-electron Write and FN-t erase technology	 Channel-hot- electron injection	 Fowler–Nordheim tunneling	Electron emitted stress

Figure 4.98 Comparison between (a) bipolarity F–N tunneling write/erase technology and (b) channel hot-electron (CHE) write and F–N tunneling erase technology [4.38].

and uniform erase technology. The erase operation is the same as that in the bipolarity W/E technology. However, during the write operation, high voltages are applied to the control gate and the drain. Thus, channel hot electrons are generated by the lateral electric field, and electrons are injected from the substrate to the floating gate. In this case, the high-field stress of the thin oxide is electron-emitted stress. The data retention will be different for these two W/E technologies because the thin oxide leakage current is different for bipolarity and electron-emitted stresses of thin oxide. Moreover, in the conventional Flash memory erasing method [4.19, 4.135], a high voltage is applied to the source. However, in this experiment [4.38], a high voltage is applied to the substrate, as well as the source-drain regions, in order to prevent the degradation of the thin oxide due to hole injection caused by band-to-band tunneling [4.19, 4.135].

Figure 4.99 shows the write/erase endurance characteristics of both W/E technologies [4.38]. No closure of the cell threshold window occurs up to 10^5 write/erase cycles in either technology. Data retention characteristics are measured under various gate voltage conditions in order to accelerate the retention test. In the case of the CHE write and F-N tunneling erase technology, the stored positive charge rapidly decays as a function of time. As a result, the threshold window decreases (Fig. 4.100). However, in the case of the bipolarity F-N tunneling W/E technology, data loss of the stored positive charge is significantly reduced. This phenomenon can be explained by the fact that the thin oxide leakage current is reduced by the bipolarity F-N tunneling stress. Figure 4.101 shows the data retention time after write/erase cycling as a function of the tunnel oxide thickness. The data retention time is defined by the time it takes V_{th} to reach -1.0 V during the gate voltage stress. In devices with a 7.5 nm tunnel oxide, the data retention time obtained for the bipo-

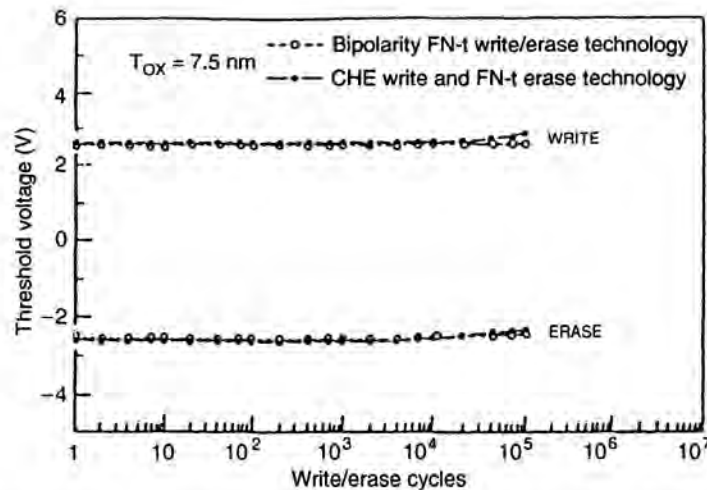


Figure 4.99 Write/erase endurance characteristics of Flash memory cell with 7.5 nm oxide [4.38].

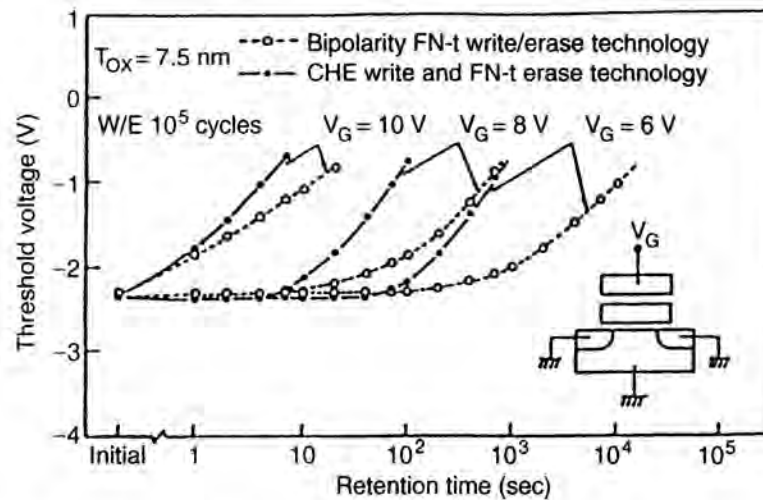


Figure 4.100 Data retention characteristics under gate voltage stress for bipolarity F-N tunneling write/erase technology and channel hot-electron (CHE) write and F-N tunneling erase technology [4.38].

larity F-N tunneling write/erase technology is 50 times longer than that of CHE write and F-N tunneling erase technology, after $1\text{E}5$ write/erase cycles. However, in devices with a 9 nm tunnel oxide thickness, the data retention time is almost the same for both technologies. For very thin ($<9\text{ nm}$) tunnel oxides, the bipolarity write/erase technology offers improved data retention times in comparison with the CHE write and F-N tunneling erase technology. Therefore, this technology may facilitate the downscaling of tunnel oxides. Reducing the tunnel oxide thickness results in lower programming voltages and in faster read operations because the read current is increased.

NAND data retention characteristics of the programmed memory cell, measured at 300°C after different numbers of write/erase cycles from 10 to 10^6 were carried out, are shown in Fig. 4.102 [4.14]. The stored positive charge effectively remains for up to 100 minutes baking time due to the detrapping of electrons from the gate oxide to the substrate during the retention bake (Fig. 4.102). This retention bake time becomes longer with an increasing number of cycles because the number of trapped negative charges in the thin oxide increases.

The effect of detrapping electrons is equivalent to the effect of trapping holes in the gate oxide. As a result, detrapping of electrons suppresses the data loss of the positively charged cell. Consequently, the stored positive charge, which effectively increases at the beginning of the bake, extends the data retention time of the memory cell programmed by uniform write and erase. Figure 4.103 [4.14] shows the data retention time after write and erase cycling. Thus, the data retention time can be extended by using uniform write and uniform erase technology, especially beyond 10^5 write/erase cycles.

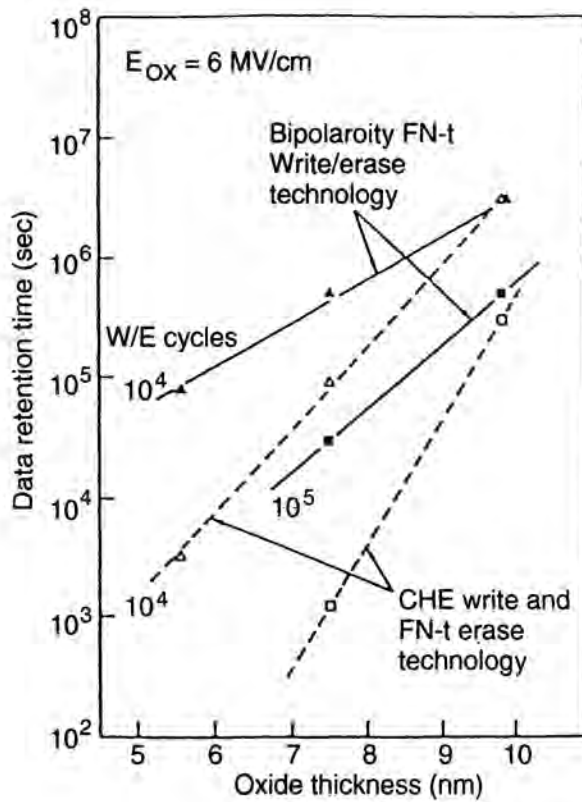
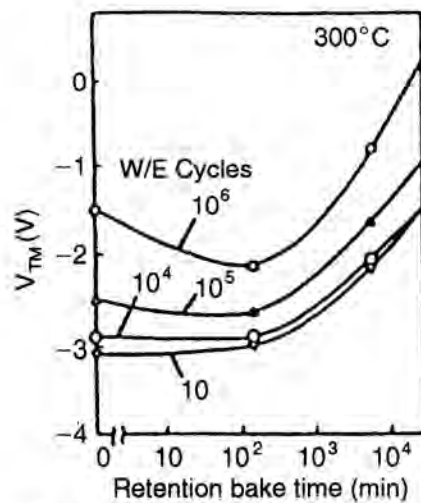


Figure 4.101 Data retention time of Flash memory cell after write and erase cycling as a function of tunnel oxide thickness. The data retention time is defined by the time at which V_{th} reaches -1.0 V during the applied gate voltage stress [4.38].



Uniform write and
Uniform erase technology

Figure 4.102 NAND data retention characteristics [4.14].

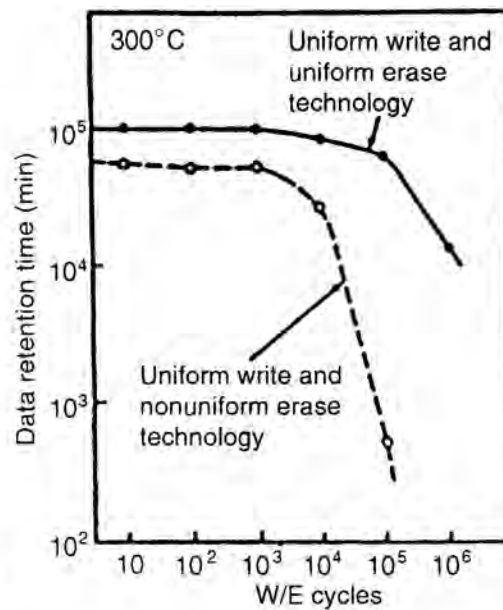


Figure 4.103 NAND data retention time [4.14].

4.6.9 Read-Disturb

The read-disturb stress condition occurs on erased bits that share a wordline with a bit that is being read. The common wordline places the control gate of the erased cells at 5.0 V (V_{cc}). The selected device's drain is driven to about 1.0 V. The unselected bits have their source, drain, and substrate at 0.0 V. In high-density Flash memories, many bits are put into a low-field stress condition when one bit is read. This condition is shown in Fig. 4.104 [4.135].

It has been reported that thin gate oxides exposed to high-field stress and high levels of charge injection can develop a pre-breakdown leakage condition [4.136, 4.137]. As program/erase cycling requirements increase, and the Flash tunnel oxide thickness is decreased, the EEPROM cell becomes more susceptible to tunnel oxide leakage [4.138]. During P/E cycling, some Flash bits can exhibit a leakage condition that permits charge gain on erased bits due to the low electric field present during a read operation. Under prolonged or DC read conditions, the defective cells can appear programmed.

It was shown that the charge gain takes place by electron tunneling through a corrupted oxide barrier [4.135]. The barrier reduction is caused by positive charge trapping at the tunnel oxide to source junction. The charge trapping is due to hole generation during F-N erase. Since the effective barrier takes on various levels, it was proposed that the configuration of trapped charge determines the extent of barrier lowering. The effective barrier of leaking cells was determined by tracking the cell threshold voltage during read-disturb stress conditions.

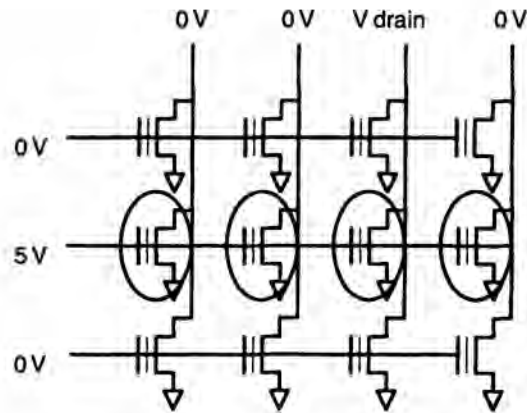


Figure 4.104 The read-disturb stress is applied to all cells sharing the gate select wordline with the cell being read. While the solidly circled cell is read, all of the dotted circled cells are stressed [4.135].

The distribution of calculated effective barrier heights is shown in Fig. 4.105. The distribution of barrier heights suggests that there is variation in the physical configuration of the trapped charge responsible for the read-disturb leakage current.

The leakage current is due to tunneling, and the generation of leakage sites is related to program/erase cycling. It has previously been shown that Flash cycling causes hole trapping [4.20] in the vicinity of the source to tunnel oxide junction. The actual tunneling mechanism could be barrier lowering due to a positive charge cluster located close to the source-oxide junction [4.136] or trap-assisted tunneling [4.139], with traps distributed over the cross section of the oxide. If the mechanism is barrier lowering, the degree of effective barrier reduction depends on the number of trapped charges and their positions. Positive charges, clustered within 2–4 nm of the

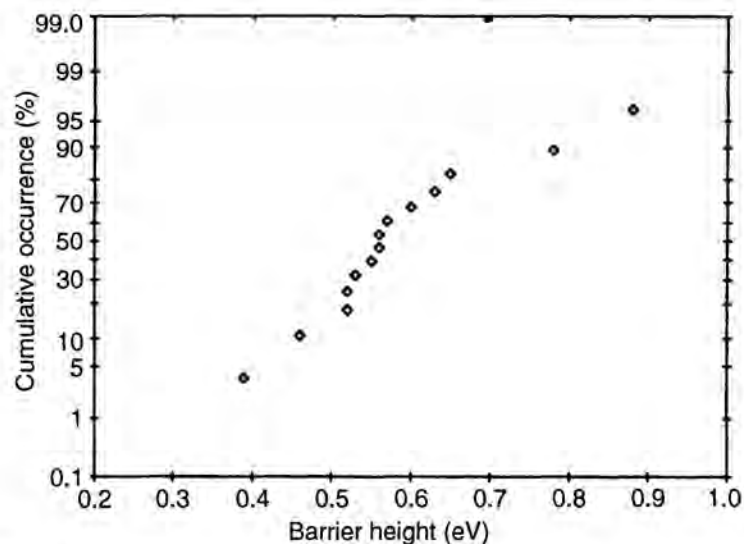


Figure 4.105 The distribution of effective barrier height has a median of 0.56 eV, although there are some higher and lower points [4.135].

source, will have the greatest effect on barrier reduction [4.135], as sketched in Fig. 4.106, by reducing the tunneling distance into the conduction band of the floating gate. It is speculated that some trap configurations are more likely to occur, therefore favoring the appearance of the corresponding barrier height. The effective barrier ranged from 0.39 to 0.88 eV [4.135]. Figure 4.106 shows the conduction band of source, tunnel oxide, and floating gate regions under a read-disturb stress.

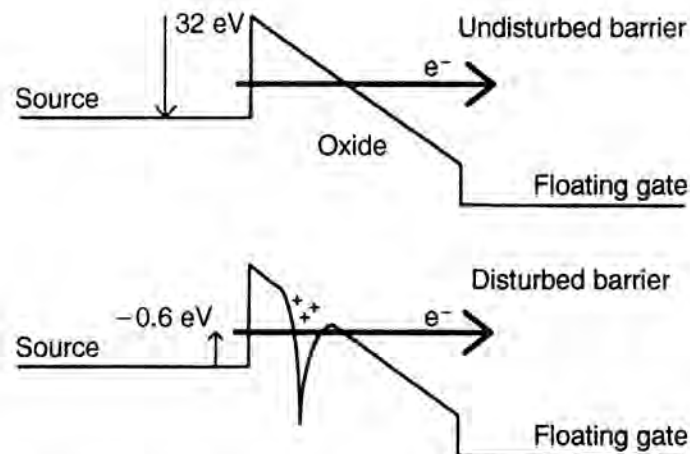


Figure 4.106 Sketch of the conduction band of the Flash source, tunnel oxide, and floating gate region under read-disturb stress. The disturbed barrier has a reduced tunneling distance due to trapped positive charges located near the source region [4.135].

4.7. PROCESS TECHNOLOGY

The process technology required for floating gate devices is derived from standard polysilicon gate technology that has been used so successfully in the manufacturing of SRAMs, DRAMs, and logic devices. The most significant addition is floating gate technology, which requires the highest quality insulating dielectric. Another important consideration is the high voltage that is required for EPROM programming and tunnel erase. It puts special requirements on isolation, junction breakdown, and transistor technology. A further reliability consideration is the quality of the dielectric used in the transistors and in the floating gates.

4.7.1 Floating Gate Technology

In a typical EPROM technology, poly1 is used exclusively for the floating gate. After poly1 is defined, a poly-to-poly dielectric is formed that surrounds the poly1

floating gate completely. In the case of poly-to-poly tunneling, the polyoxide is a specially processed oxide that gives poly-texturing to enhance tunneling probability. If the cell is not designed for poly-to-poly tunneling, then an oxide-nitride-oxide (ONO) composite dielectric is typically used. The composite layer gives a much lower defect density. Much thinner ONO layers can be formed, compared to oxides, with much lower leakage currents. This, in turn, gives a higher coupling-capacitance and smaller cell sizes.

4.7.2 High-Voltage Technology

The high voltage required to operate Flash memories ranges from a low of 12 V for stacked gate Flash to nearly 25 V for poly-to-poly tunnel erase, even though the voltage may be generated internally. Furthermore, the EPROM programming voltage is not expected to be scalable much below 10 V. This is because the gate program voltage is determined by the intrinsic silicon-silicon dioxide barrier (3.2 eV), which is not scalable, as well as the floating gate voltage change which may actually be growing with ever increasing density.

In order to handle these voltages, the field oxide isolation process has to give a sufficiently high-field turn-on voltage, as well as a sufficiently high junction breakdown voltage. The standard approach is to use a much thicker field oxide compared to logic technology of a comparable generation. However, there is a limit on how much a standard field oxide isolation can be scaled. It is believed that eventually more exotic isolation types, like trench isolation, will be used for the smallest geometry.

In addition to isolation technology, the transistor technology also has to handle the high voltage. This is contrary to transistor scaling for performance. Special high-voltage devices are required. In the case of CMOS, for example, it is difficult to have a full CMOS circuit operate above 20 V, and a depletion device is generally used as a load device for higher voltages. A 20 nm transistor can operate reliably at 12 V, but, with further scaling, a 15 nm transistor will not be reliable for 12 V operation. Thus, for scaled technologies or for EEPROM-based Flash technologies, dual gate oxide technologies are used—one to handle high voltages and the other, a high-performance, low-voltage transistor for high-speed logic.

4.8. MEMORY CIRCUITRY

The standard array architecture for Flash memory is just an EPROM array. In a conventional EPROM array, the cells are arranged in an X-Y cross-point array with an EPROM cell at each cross-point. In the horizontal X-direction, control gates of the EPROM cells are connected together by poly to form the wordline. In the vertical Y-direction, the drains of the EPROM cells are connected together by metal through a contact shared by two memory cells. The sources of the cells are connected together horizontally by diffusion. Spaced every 16 or 32 cells, the sources are connected together by a metal line, called a source strap, running in the Y-direction. This is a well-proven EPROM array architecture.

The wordlines are selected by X-decoders, which pass a high V_{pp} to selected wordlines during programming and V_{cc} during read. Nonselected wordlines are held at ground. The columns are selected by Y-select, which passes V_{dd} to a cell being programmed and V_{drain} for read. The array is organized into eight blocks for each bit in a byte. During program or read, one column in each block is activated. If the programmed bits are not separated, but are adjacent to each other, there will be an excess voltage drop when they are programmed at the same time. Finally, there is a sense amplifier for each bit for reading the state of the memory cell.

4.8.1 Row Decoder Circuits

The row decoder circuits are standard circuits derived from EPROM X-decoders. For Flash memories using negative gate voltage erase, the circuit requires modifications in the row decoders. During the erase operation, the negative voltage generated by a negative charge pump is applied to the wordline. The n^+/p^- substrate diode is forward-biased if a negative voltage is applied to the n^+ region with the p-substrate at ground level. P-MOS transfer gates have been used in the first generation of Flash memories using negative gate erase as one way to overcome this problem. In this scheme, the p-MOS transfer gate isolates the negative voltage from the wordline driver n-MOS transistors during the erase operation as shown in Fig. 4.107.

In this row decoder, p-MOS transfer gates are inserted between each wordline and an n-MOS driver, and a gate-drain connected p-MOS transistor is utilized as a diode between each wordline and the negative charge pump. A positive voltage is applied to the p-MOS gate during the erase operation to isolate the negative voltage. The p-MOS gate is negatively biased during other operations.

There are several problems with this scheme. First, a severe stress voltage is applied to the p-MOS gate. The p-MOS gate has to be deeply negatively biased in order to supply zero volt to a nonselected wordline during the read and program

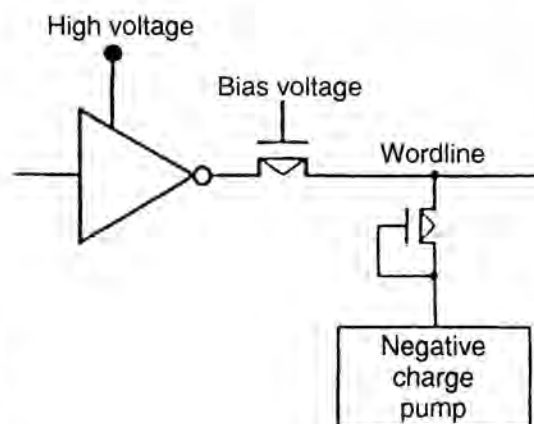


Figure 4.107 Conventional row decoder circuit for negative gate, positive source erase [4.61].

operations. Therefore, a large electric field is applied to the p-MOS transistor gate oxide in the selected wordline. The stress is about 15 V, which is higher by 50% than a normal programming voltage. The high voltage prevents gate oxide scaling. Second, the negative voltage required for the output of the negative charge pump is very large because the negative voltage applied to the wordlines is reduced by the diode-like p-MOS transistor. Finally, this conventional scheme requires a standby current because a negative voltage is needed during standby operation. This is not preferable for portable systems. At the same time, the resistance of the transfer gate increases the wordline delay time.

To overcome these problems, a row decoder utilizing a triple-well process technology was developed [4.61] and is shown in Fig. 4.108. In this circuit, n-MOS transistors used in the row decoder circuit are fabricated in the p-well region,

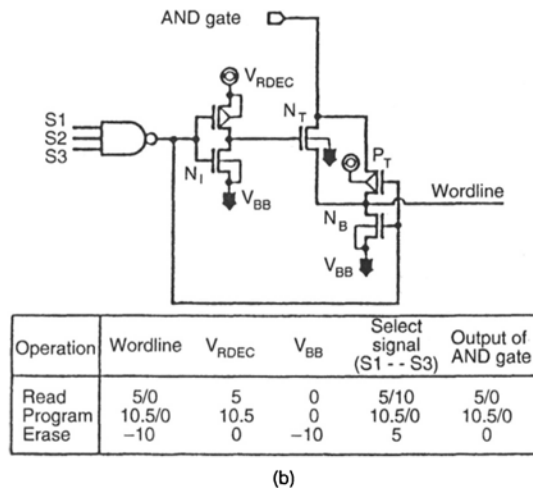
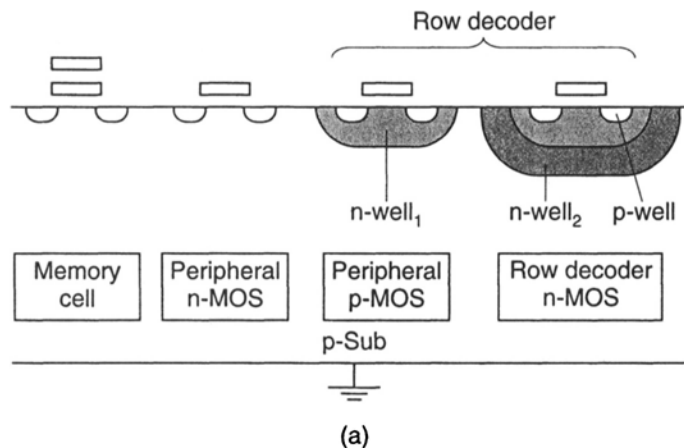


Figure 4.108 Row decoder: (a) cross-sectional view, (b) circuit and operation [4.61].

which is isolated from the p-substrate by a deep n-well region. During the erase operation, a negative voltage of -10 V , generated by the negative charge-pump circuit, is applied to the p-well, and all wordlines are set to -10 V by the n-MOS transistors in the p-well. During other operations, the p-well is set to ground level. This scheme prevents the forward-bias problem. As shown in Fig. 4.108, the selected wordline is set to $10.5\text{ V}/5\text{ V}$ for programming and reading operations, respectively, and the nonselected wordlines are set to ground level. During the erase operation, all the wordlines are set to the negative voltage. The row decoder circuit consists of a NAND gate, an inverter, and transfer gates. It is designed to minimize the electric field on the gate oxide. During the erase operation, VRDEC is set to ground level, while all the decoder select signals of S1–S3 are set to a high level and the output of an AND gate is set to ground level. Thus, the maximum stress voltage on the gate oxide is reduced to a value as small as 10 V , which is almost the same as the voltage stress during programming. Consequently, a thick gate oxide is no longer needed.

This row decoder is appropriate for a relatively large-size block erasure. A compact row decoder circuit [4.70], in which individual wordlines are selectively biased to a negative voltage during the erase operation, is shown in Fig. 4.109. This circuit consists of a decoder select NAND gate A and wordline select transmission gates Q_1 – Q_n . All the NMOS transistors in this row decoder are formed in the p-well region and biased to an internal voltage V_{BB} . During the erase operation, V_{BB} is

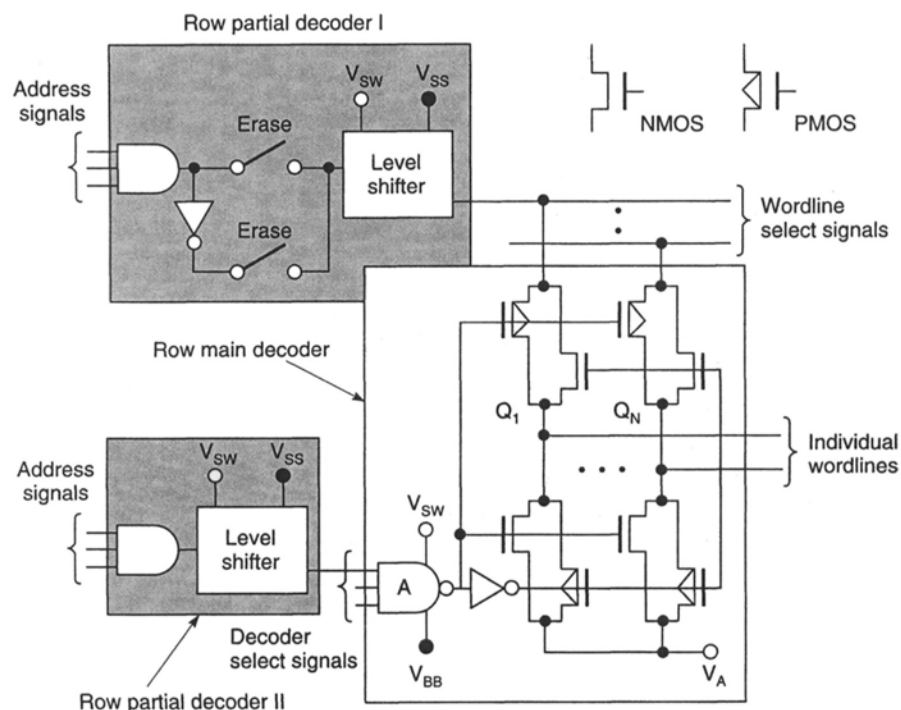


Figure 4.109 Row-decoder circuit in a triple-well technology [4.70].

set to a negative voltage, and during other operations V_{BB} is set to ground. In the selected row decoder, the individual wordlines are connected to the partial decoder's outputs. In the unselected row decoders, the individual wordlines are connected to the signals V_A . The V_A is set to a positive level, V_{SW} , during the erase operation to avoid erase-disturb problems [4.48] and is set to ground during other operations. A parallel-arranged NMOS and PMOS transmission gate scheme enables the application of any voltage, for both selected and unselected wordlines. Both the decoder select signals and wordline select signals are shifted to V_{SW}/V_{BB} in the partial decoders. During the erase operation, the wordline select signals are inverted in the partial decoder to apply a negative voltage to the selected wordline. A minimum sector erase operation, consisting of a single wordline, is possible in principle. A compact row decoder has been obtained by introducing the parallel arranged transmission gate scheme.

During the erase operation, the high-level V_{SW} is set to about 4 V and the low level V_{BB} is set to -7.5 V. The maximum voltage stress applied to the gate oxide during the erase operation is limited to less than that for the programming operation.

In order to accomplish a selective erase in a sector (one wordline) with negative-gate-bias channel erase, a row decoder [4.62], shown in Figs. 4.110 and 4.111, is used. Table 4.12 shows its operating voltages. This row decoder consists of two row-main

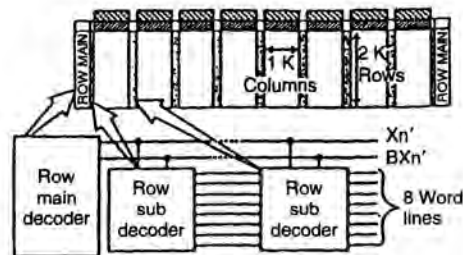


Figure 4.110 Block diagram of row decoder [4.62].

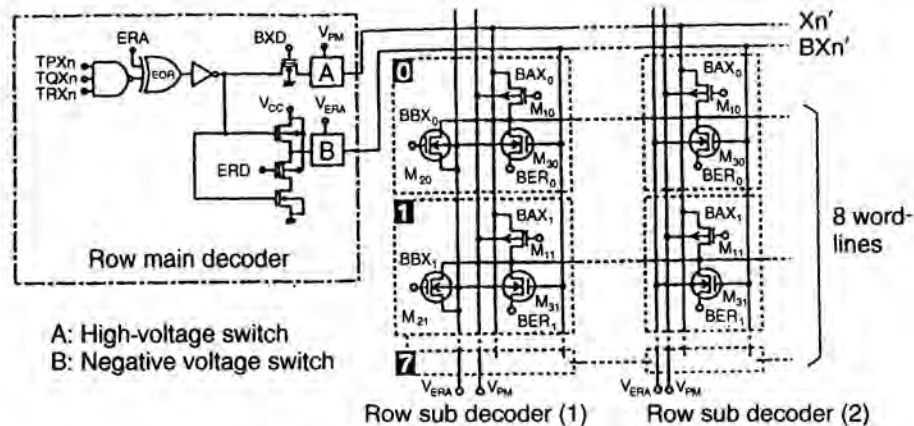


Figure 4.111 Circuit diagram of row decoder [4.62].

TABLE 4.12 ROW-DECODER OPERATING VOLTAGES

Signals	Read		Program		Erase	
	Select	Deselect	Select	Deselect	Select	Deselect
TPXn, TQXn, TRXn	V_{cc}	V_{ss}	V_{cc}	V_{ss}	V_{cc}	V_{ss}
BAXm	V_{ss}	V_{cc}	V_{ss}	12 V	V_{ss}	V_{ss}
BBXM	V_{ss}	V_{cc}	V_{ss}	V_{cc}	-13 V	-13 V
BERm	V_{ss}	V_{ss}	V_{ss}	V_{ss}	-13 V	V_{ss}
Xn ⁿ	V_{cc}	V_{ss}	12 V	V_{ss}	V_{ss}	V_{cc}
BXn ⁿ	V_{ss}	V_{cc}	V_{ss}	V_{cc}	V_{cc}	-13 V
Wordline	V_{cc}	V_{ss}	12 V	V_{ss}	-13 V	V_{cc} OR V_{ss}
ERA, ERD		V_{ss}		V_{ss}		V_{cc}
BXD		V_{cc}		V_{cc}		V_{cc}
V _{PM}		V_{cc}		12 V		V_{cc}
V _{ERA}		V_{ss}		V_{ss}		-13 V
p-Well for cell		V_{ss}		V_{ss}		V_{cc}

decoders and nine row-sub decoders. The row-main decoders are placed on both sides of the cell array and select eight of 2,048 wordlines. Its outputs, Xn and BXn, are wired to the row-sub decoders using second metal interconnection. The row-sub decoders are placed on each side of the cell array blocks and select a single wordline of the eight wordlines. This row decoder organization has the advantage of reducing wordline delay and its occupied area.

In a row-main decoder, an NMOS depletion transistor for high-voltage isolation and a high-voltage switch (block A) are the same as those used in conventional EPROM. They are active during the programming mode. The negative voltage switch (Fig. 4.112) in block B becomes effective during the erase mode. The row-sub decoder (1) consists of one PMOS transistor, M1m, and two negative voltage NMOS transistors, M2m and M3m, as shown in Fig. 4.113. In the row-sub decoder (2), the negative voltage NMOS transistor M2m, which is used to discharge the wordline, is omitted in order to reduce the area occupied by the row-sub decoders.

During the reading operation, the high-voltage switch and the negative voltage switch are unavailable. For the selected row-main decoder, the outputs Xn'' (active high) and Bxn'' (active low) become 5 and 0 V, respectively. At the same time, one of the PMOS transistors (M1m) in the row-sub decoder turns on by a row pre-decoder; thus, 5 V is applied to the target wordline.

During the programming operation, the gate signal of the NMOS depletion transistor (BXD) is set to 0 V and the high-voltage switch is active. The output Xn' is converted to 12 V by the high-voltage switch and is applied to the target wordline through one of the PMOS transistors (M1m) in the row-sub decoder.

During the erasing operation, one input signal of the exclusive OR (EOR) circuit (ERA) is set to 5 V, and the negative voltage switch becomes effective. For

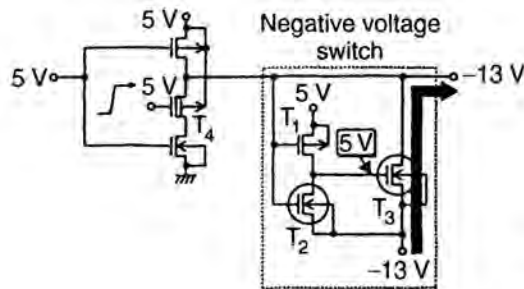


Figure 4.112 Circuit diagram of negative voltage switch [4.62].

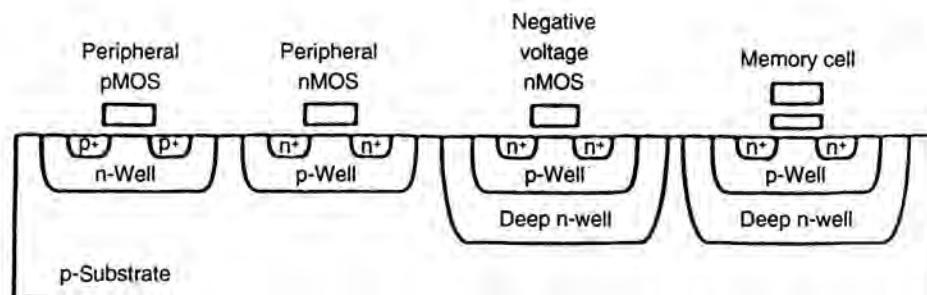


Figure 4.113 Schematic cross-sectional view of the triple-well CMOS structure.

the selected row-main decoder, the output of the EOR circuit becomes 5 V. Then, outputs X_n and BX_n become 0 and 5 V, respectively. All of the eight NMOS transistors, M30–M37, in the row-sub decoder turn on because BX_n is at 5 V. At the same time, the row pre-decoder supplies -13 V to one of the source signals (BER_m) of the NMOS transistors (M3m), and seven others are at 0 V. Therefore, only the target wordline is charged to -13 V, while seven others are at 0 V. On the other hand, for the deselected row-main decoders, the outputs X_n and BX_n become 5 V and $Z - 13$ V, respectively. Thus, the deselected wordlines wired to the deselected row-main decoders are 5 V through the PMOS transistors M10–M17, and any sector erase and block erase are possible by controlling the pre-decoding signals for the row decoder.

4.8.2 Erase Circuits

Electrical erase circuits are the new circuits required for Flash, compared to EPROM. As discussed previously, the exact circuitry depends on the cell type. For source erase, instead of connecting the source straps to ground, these are connected through a switching circuit that switches between ground for normal operation and V_{pp} for erase. For drain erase, the same switch circuits can be used to switch V_{pp} . Finally, for poly-to-poly erase, new circuits are incorporated into the X-decoders or Y-select.

4.8.3 Charge-Pump Circuits

Charge-pump circuits have been discussed in previous chapters. The concept of a quick boosting charge-pump circuit for high-density and low-voltage Flash memories [4.102] is shown in Fig. 4.114.

For the conventional charge-pump circuit, the number of stages connected in series between the power supply and the output of the charge-pump circuit is fixed. This fixed number of stages is designed so as to generate a certain high voltage required for programming. However, the charge-transfer efficiency, I_{PP}/I_{CC} , for the charge-pump circuit is given by $1/(n+1)$, where I_{CC} and I_{PP} are the mean input and output currents for the charge-pump circuit, respectively, and n is the number of stages connected in series between the power supply and the output of the charge-pump circuit. As a result, many stages connected in series are so redundant that the conventional charge-pump circuit not only consumes too much power, but also lowers the charge-transfer efficiency, while the boosted voltage is not much higher than the power supply voltage at the beginning of the operation. Thus, it takes a longer time for the conventional charge-pump circuit to generate a desired voltage for programming when a higher density EEPROM increases the load capacitance for the charge-pump circuit and the power supply voltage is lowered.

The charge-pump scheme conceptually presented in Fig. 4.114 can change the number of stages and the capacitance used for charge pumping. While the output voltage is not much higher than the power supply voltage in the beginning, the number of stages is controlled to be small and the capacitance for charge pumping is controlled to be large in order to increase the charge-transfer efficiency. Accordingly, as the output voltage is boosted, the number of stages is increased and the capacitance for charge pumping is decreased step by step. As a result, the charge-pump scheme shown in Fig. 4.114 offers a high charge-transfer efficiency and reduces the rising time of the output voltage.

The switches SW1 and SW2 are controlled by a set of complementary voltages of a high and low voltage converted by a level shifter circuit (LS).

4.9. FLASH APPLICATIONS

In this section, the current and potential applications of Flash memory are discussed and a brief discussion on DRAM versus Flash is presented.

4.9.1 EPROM Replacement

EPROM was important in the product development phase but was replaced by Mask ROM as the product went into volume production. On the other hand, with Flash, product designers and end-users can continue to update the product more easily and in situ. Another advantage of Flash over EPROM is a smaller footprint, thus saving board space. In embedded microcontroller applications, Flash is replacing EPROM, as well as EEPROM, Boot RAM, and battery RAM. Low-voltage

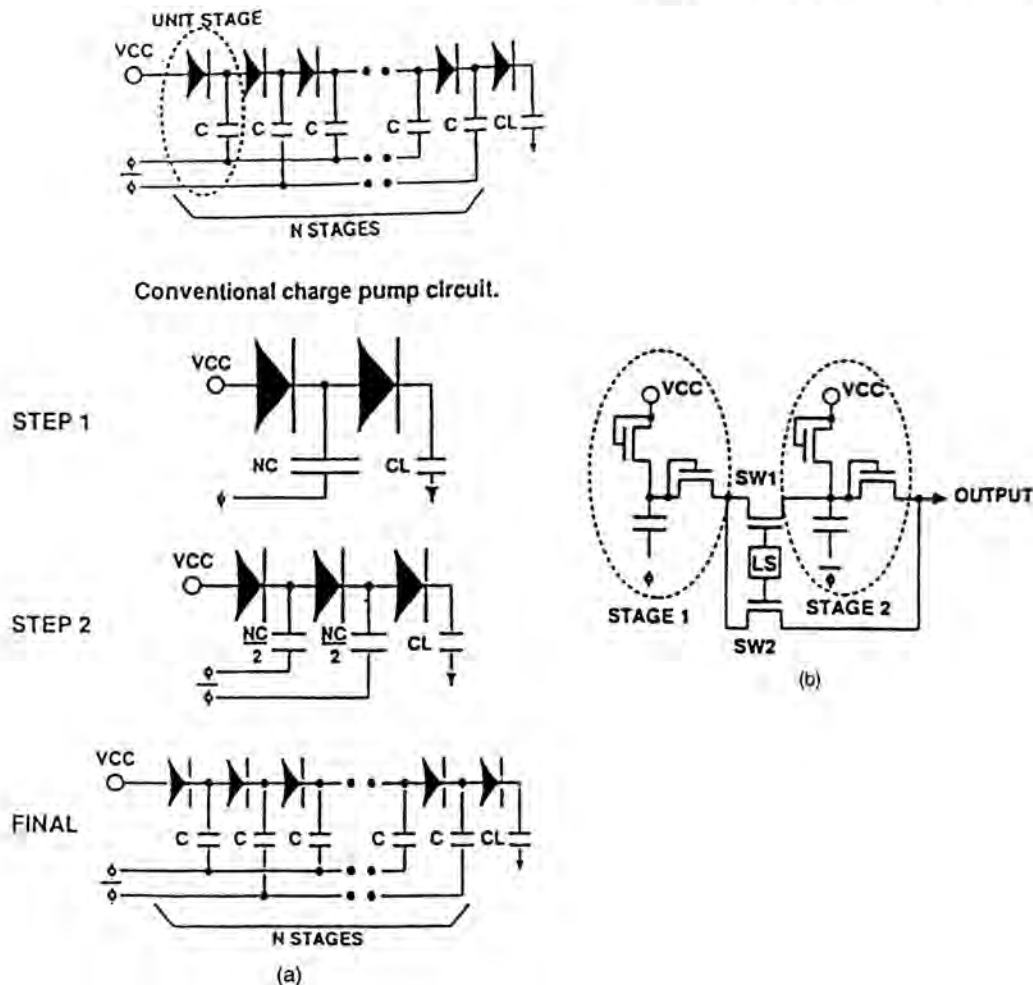


Figure 4.114 (a) Charge-pump scheme; (b) charge-pump circuit [4.102].

Flash is also expected to replace EPROMs in mobile computer microcontroller applications.

4.9.1.1 Code Storage. Except in cases where disk or tape memory is practical, most embedded computers store code in ROM. But as embedded systems have evolved from 8 bit single-chip controllers to 16 and 32 bit computers, the amount of code in a typical system has increased enormously. Along with the increase in code size, the need to provide code changes has also grown. Bigger programs contain more errors, and systems with more features tend to be upgraded more often.

When code is stored in Mask ROM or UV EPROM, each update can mean that vendors must visit the system, physically take it apart, remove ROMs, and replace them with new ones. Both the cost and the opportunity for damaging the system can be prohibitively large. Boards can be designed so that Flash EPROMs can be erased

and reprogrammed from an edge connector, reducing the update problem to opening the cabinet, connecting the programmer, and waiting a few minutes. No physical disassembly of the system is necessary.

Boards can be designed to let the embedded central processing unit (CPU) erase and write the Flash parts in-circuit, under the control of the embedded CPU. The system contacts a remote update source. Then, the local CPU selectively erases blocks of Flash EPROM, writes updated code into them, and continues operating as modified. So, Flash offers a cost advantage over EEPROMs and in-system rewritability over UV EPROMs.

4.9.2 Automotive Applications

In-system reprogrammability makes Flash well suited to many automotive applications.

1. In-system code change for development and optimization without hardware changes.
2. Just-in-time manufacturing flexibility to select appropriate codes and programs for particular modules for vehicles.
3. Inventory control—Minimization of part numbers.
4. In-vehicle selection of features—engine, transmission, ride options.
5. Testability using a test bus and system service diary chips to store service history and codes.
6. Continuous system optimization through the vehicle life cycle by real-time system programmability; in-field repair/alteration, maintainability, and recall avoidance by in-vehicle alteration at the dealer.

4.9.3 Joint Testability Action Group (JTAG)

Another class of system applications for reprogrammable memory elements is storage of system intelligence—typically in a few bytes of EEPROM, distributed in various parts of the system. Using such an approach, for example, we can place system security codes, fault information, system service diaries, manufacturer's system IDs, code IDs for overall system testability, and, so on, throughout the system. The Joint Testability Action Group (JTAG) has, for example, defined a system bus that can include serial Flash EEPROM memory to store information related to board service history. As the electronic content within a future system expands, such testability features will take on additional urgency during development, manufacturing, and maintenance (i.e., the field service environment). The EEPROM element can also be used in applications for system identification and vehicle location (as on toll roads, parking lots, garages, etc.), and eventually will be used as part of intelligent sensors for remote diagnosis and servicing.

4.9.4 Cards

Until recently, Flash memory cards were used primarily in specialized, less cost-sensitive applications in which their ruggedness, removability, small size, low power consumption, and high performance were valued. For example, factory-floor equipment, robots, and industrial control systems use Flash memory cards to increase the flexibility and functionality of precise electromechanical systems. But lower memory media prices and new packaging technology have reduced costs and enhanced the density and functionality of Flash memory cards, making them attractive in mainstream portable equipment and mobile computers. Flash memory cards are now feasible in data-file and application-program storage or transfer.

In mobile designs, the nonvolatility, updateability, and removability of the Flash memory card makes it suitable for file exchange between systems. Solid-state memory access speeds increase system performance over that of rotating disk technology. In addition, the low power consumption of Flash memory enables systems to be developed that can run on common batteries. It is feasible to increase the system battery life of handheld computers using Flash memory storage.

Personal notebook and sub-notebook computers, with standard disk interfaces, can also derive benefit from Flash memory in a hard disk drive configuration. A Flash drive is defined as a mass storage device based on Flash memory technology and integrated with such industry-standard disk interfaces as integrated drive electronics (IDE), small computer systems interface (SCSI), or PCMCIA/AT attachment. A Flash drive differs from a Flash memory card in its system software and hardware requirements, as well as in its relative cost.

4.9.4.1 Smart Cards. Smart cards consist of rewritable, nonvolatile memory and a microprocessor. The microprocessor makes the card "smart." The microprocessor controls the communication between the system and memory, allowing multiple applications, and provides security via interpretation of personal ID number [PIN] codes. Power is supplied to the card externally.

Smart cards are carried in purses or wallets and have many potential applications, such as financial transactions, pay phones, employee ID cards, authentication, and storage of personal information. The smart card concept can utilize high-density Flash in supermarket cards, current credit cards, driver's licenses, access control devices, and car maintenance history, to name a few.

4.9.5 Look-Up Tables/Data Acquisition

Flash can be used in many other applications, including field updateable look-up tables, capture and recording of important data, updateable firmware as BIOS, and the more sophisticated disk drive-like solid-state memory subsystems that have attained sizes up to 200 MB.

Another application, data acquisition, takes full advantage of Flash memory's byte-write capability and full nonvolatility. Any kind of data logger or sensor that depends on some form of battery-backed RAM can benefit from using Flash

memory. Flash memory offers cost savings, higher density, and critical data security by eliminating the battery for retaining memory contents.

4.9.6 Personal Systems

Notebooks, laptops, and other portable PCs, such as the luggable, which are light enough to be portable are called portable systems. On the other hand, a mobile computer is a portable computer that can be operated while in motion, for example, a handheld data acquisition computer, a palm-type scheduler, or palm-type and/or pen-based personal computing systems. Flash will be a preferred solid-state disk in mobile applications in the form of PCMCIA cards, for example.

4.9.7 Analog Applications

Flash memories, based on Fowler–Nordheim programming (as in EEPROMs) can be used in mixed-signal chips to store calibration data. Laser trimming for precision can be replaced by Flash trimming. The Flash cells would drive small, precise D/A converters that would control offset, filter parameters, and the like. One application could be in equalization filters for disk-read channels and communication circuits.

Analog memories are also possible candidates for Flash application. Analog memories provide direct audio record/playback without digital conversion. For example, ISD uses 2 bits/cell EEPROM and combines these with audio microphones, speakers, amplifiers, and other components [4.145, 4.161].

4.9.8 Logic

While the advantages of high density Flash in programmable logic are evolving, low to moderate density PLDs have been implemented in Flash. Even PCMCIA designs can almost be debugged in real time using Flash reprogrammability in combination with field programmable gate array (FPGA) technology. PCMCIA I/O cards often use a combination of chipsets, glue-logic, and memory to accomplish their function. In many cases, the glue-logic portion of the design is implemented using programmable logic, such as FPGAs. The FPGA configuration memory can be stored in Flash memory on the card and modified if needed [4.146].

4.9.9 New Architectures

Mobile PC and communications manufacturers are discovering new ways to exploit Flash memory's capabilities in their system designs. Placing applications on the motherboard in a resident Flash array configuration, directly connected to the processor, dramatically speeds up system performance. Frequently used software becomes immediately available, eliminating the lengthy download step required by RAM/disk-based system architectures. A resident Flash disk, acting as a nonvolatile RAM drive, emulates disk drive functionality with no spin-up time or seek/rotation

delay. Finally, in systems designed with advanced power management, the contents of system DRAM can be stored in resident Flash memory during the system sleep mode for longer battery life (DRAM refresh can be disabled) and fast system wake-up.

Using Flash memory, a new system architecture is possible (Fig. 4.115). Flash memory can replace system RAM in the RAM's traditional role of direct execute code storage, while some RAM is retained for data manipulation and storage. This allows a reduction in RAM content, with no impact on the total system's memory budget.

For small form-factor systems, converting the application portion of system RAM to nonvolatile memory reduces overall system cost and size, since a disk drive is no longer needed for nonvolatile application storage. The resident application can be upgraded by means of an interface port or Flash memory plug-in cards.

The recent introduction of low-voltage Flash memories with synchronous or DRAM-like host interfaces can replace code-RAMs and nonvolatile code storage in 16 or 32 bit microprocessors or microcontroller-based systems [4.149].

4.9.10 DRAM Versus Flash

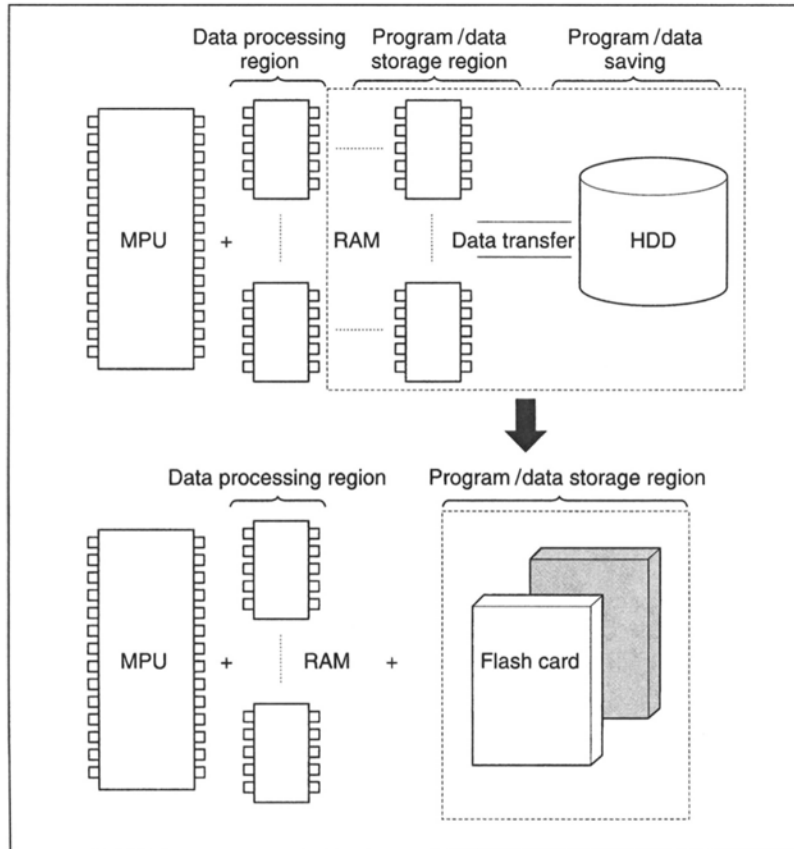
Modern DRAMs are based on a 1-transistor MOSFET charge storage cell. The cell consists of a transistor and a capacitor. The information is stored on the capacitor, in the form of charge, and must be refreshed after every read. On the other hand, a Flash memory cell is a single MOS transistor consisting of a double poly-stack structure, and the stored information is nonvolatile, in contrast to DRAM.

The DRAM cell is about one and a half times larger than the 1-T Flash memory cell. Thus, Flash chips are smaller than DRAM memory chips of comparable feature size and density. Furthermore, it is anticipated that, as the manufacturing volume of Flash increases, the cost of Flash memory will become lower than the cost of DRAM (Fig. 4.116).

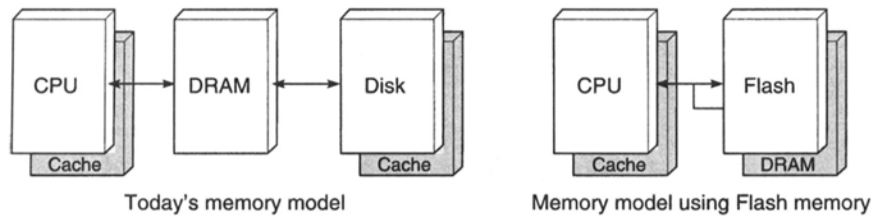
DRAM and Flash cells have different limitations in scaling (Table 4.13). For DRAM, because of soft errors, the capacitance must remain relatively constant for each successive generation. To maintain the constant capacitance for a small area, 3-D structures, such as trenches or stacked capacitors, are required, complicating the manufacturing process. For Flash, the scaling limitation is the high voltage (~ 10 V minimum) that is required to be applied to the gate for channel hot electron or > 10 V for the Fowler–Nordheim tunneling process. This means that the technology has to support a relatively high voltage.

The performance of first-generation 16 Mbit NOR Flash memories is in the same range as 16 Mbit DRAMs. Recent Flash memories, with synchronous or DRAM-like host interfaces, provide DRAM-like data accesses [4.149] and can replace shadow-RAM in 16 and 32 bit microprocessor or microcontroller-based systems. In general, the rate of acceptance of and demand for Flash memory for varied and sundry applications has been much greater than expected, thanks to the marketing efforts of major manufacturers.

At present, most of the major DRAM manufacturers and other silicon suppliers are pursuing the development of Flash memories. According to various projections,



(a)



CPU = central processing unit; Dram = dynamic RAM.

(b)

Figure 4.115 Flash memory's nonvolatility and updateability displace the combination of dynamic RAM and hard disk drive used for code storage, creating a new memory architecture (a) [4.148] and (b) [4.147].

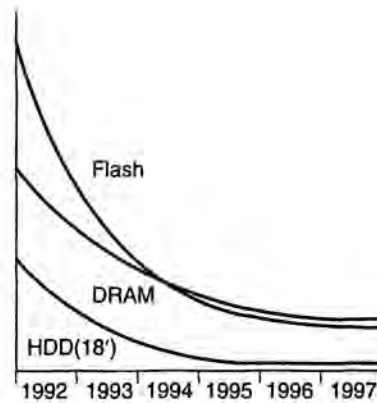


Figure 4.116 Projected prices of Flash, DRAM, and HDD [4.148].

TABLE 4.13 SCALING RULES FOR CONSTANT-FIELD SCALING

	Scaling Factor
Surface dimension	$1/K$
Vertical dimensions [T_{ox} , X_j]	$1/K$
Impurity concentration	K
Currents/voltages	$1/K$
Current density	K
Capacitance/area	K
Transconductance	1
Circuit delay time	$1/K$
Power dissipation	$1/K^2$
Power density	1
Power-delay product	$1/K^3$

Flash memory will become cheaper than DRAMs sometime between 1995 and 2000. This will further fuel the growth of Flash memory and help make it a commodity memory.

Flash memory cannot compete with DRAMs in applications where bits or bytes must be altered individually or frequently. Flash's slow write speed and limited number of rewrite cycles are serious disadvantages, compared to DRAMs, for many applications.

4.10. CONCLUSIONS AND A LOOK INTO THE FUTURE

Since the emergence of a 256 Kbit Flash chip in 1987 [4.3, 4.9], Flash memory has rapidly evolved into a high-density memory (Table 4.14). Today, several makers offer 16 Mbit Flash. As of 1995, 32 Mbit Flash products and cell technologies up to 256 Mbit had been presented at the IEDM-94 and ISSCC-95.

TABLE 4.14 FLASH MEMORY EVOLUTION AND PROJECTIONS

Year	1987	1991	1993	1994	1998 (Proj.)	2000 (Proj.)
Density (bits)	256 K	1 M	4 M	16 M	64 M	256 M
$V_{CC}(V)/V_{PP}(V)$	5/12	5/12	5/12	3/12	—	—
	—	—	5/5	5/5	5/5	—
	—	—	—	3/5	3/5	—
	—	—	—	—	3/3	3/3
	—	—	—	—	—	2/3
	—	—	—	—	—	2/2
Access time (ns)	150	150	120	60	50	50
Erase block size	Chip	Chip	16 KB	512 bytes	—	—
Erase time (ms)	10^3	10^3	—	10 ms/sector	—	—
Program time [μ s/byte]	—	25	—	2.5	2.5	0.25
Write/erase cycles	10^2	10^3 – 10^4	10^4	10^5	10^6	10^7
Major density breakthrough	—	—	—	—	—	Multistate cell memory
Major application	EROM replacement	EROM replacement	Memory card	PC	Handheld and mobil system	Consumer electronics

Flash memory development work by DRAM manufacturers in the United States, Japan, Korea, and Europe continues. Many alliances have been formed between memory makers, original equipment manufacturer (OEM), and system houses.

Sub-notebook computers, palmtop and pen computers, personal digital assistants (PDAs), cellular phones, and other mobile devices are gaining popularity. It is anticipated that the personal/mobile market will explode as computers become smaller, lighter, more rugged, and less expensive. Flash memory is expected to be the optimum choice for mass storage in personal mobile systems. In mobile and handheld computers, for example, Flash will replace DRAM in order to contain operating systems, BIOS, and applications software.

Because of Flash's attributes, this new market segment will prefer it, while hard disk and floppy disk will remain the medium of choice in most nonportable applications. Although the innovative designs continue to shrink the Flash cell size (e.g., $0.4 \mu\text{m}^2$ cell by Hitachi @ $0.25 \mu\text{m}$ rules [4.151] and $0.67 \mu\text{m}^2$ cell by Toshiba [4.152] for 256 Mbit Flash memory), the cost/bit of hard disk continues to improve unabated [4.153]. Therefore, it is becoming clear that Flash cost/bit may overtake hard disk for high-density systems in spite of the many projections over the last few years [4.13, 4.57, 4.58, 4.68, 4.154]. In addition to cost, Flash memory must match the hard disk in reliability (write/erase endurance and data retention) and programming speed to replace it.

At the present time, Flash has already displaced EPROM in many applications, and Flash memory will be the technology of choice in many portable applications in

the future. The Flash memory will replace hard disks in low-end applications (1 MB–100 MB), but will not be able to replace hard disks in most applications, primarily because of cost. Solid-state disks and hard disks will be complementary technologies.

DRAM, which has been the fuel for LSI/VLSI/ULSI, is intrinsically more expensive per bit because of its relatively larger cell size and more complex processing, compared to Flash. Flash memory speeds are in the same range as DRAM, and with a DRAM-like interface [4.149], Flash memory can perform many of DRAM's functions [4.145, 4.147, 4.153]. On the other hand, research efforts are underway to improve the cost and nonvolatility of DRAM, for example, NAND-structured DRAM cells to reduce the cell size [4.155] and a nonvolatile RAM (NVRAM) using ferroelectric materials [4.156].

So the semiconductor memory revolution continues unabated.

4.10.1 Flash Market Development Trends

- PC BIOS
- PC operating system and application storage
- Portable computer main memory and mass storage
- Cost/MB crossover with DRAM
- Dominance in mobile disk drive applications
- Faster read/write, lower power
- Application specific Flash
- Commodity memory

4.11. ACKNOWLEDGEMENTS

Flash memory technology and applications continue to explode at an ever increasing rate. Thanks are due to many colleagues in the industry who have contributed the material for this chapter. This list includes, but is not limited to, Dave Sweetman of SST, Kuniyoshi Yoshikawa of Toshiba, Greg Atwood of Intel, Chi Chang of AMD, and Rob Frizzell of National Semiconductor.

References

- [4.1] K. Sakui and F. Masouka, "Sub-half micron Flash memory technologies," *IEICE Trans. Electron.*, vol. E-77, no. 8, pp. 1251–1259, 1994.
- [4.2] B. Dipert and M. Levy, "Designing with Flash memory," Published by Annabooks, 1993.
- [4.3] F. Masuoka, M. Assano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new Flash EEPROM cell using triple polysilicon technology," *IEEE IEDM Tech. Dig.*, pp. 464–467, 1984.
- [4.4] F. Masuoka et al., "A 256K Flash EEPROM using triple polysilicon technology," *IEEE ISSCC Dig. Tech. Pap.*, pp. 168–169, 1985.

- [4.5] F. Masuoka et al., "A 256K Flash EEPROM using triple polysilicon technology," *IEEE J. Sol. St. Cir.*, vol. SC-22, pp. 548–552, 1987.
- [4.6] The following is an example from a large literature base on EPROMs:
 M. Van Buskirk et al., "EPROMs graduate to 256-k density with scaled n-channel process," *Electronics*, pp. 89–93, February 24, 1983.
 J. J. Barnes et al., "Operation and characterization of N-channel EPROM cells," *Solid State Electr.*, vol. 22, pp. 521–529, 1978.
 M. Wada et al., "Limiting factors for programming EPROM of reduced dimensions," *IEEE IEDM Tech. Dig.*, pp. 38–41, 1980.
 N. R. Mielke, "New EPROM data loss mechanisms," *Proc. IRPS*, pp. 106–113, 1983.
- [4.7] S. Mukherjee, T. Chang, R. Pan, M. Knecht, and D. Hu, "A single transistor EEPROM cell and its implementation in a 512K CMOS EEPROM," *IEEE IEDM Tech. Dig.*, pp. 616–619, 1985.
- [4.8] A. Baker et al., "A 3.3V 16Mb Flash memory with advanced write automation," *IEEE ISSCC Dig. Tech. Pap.*, pp. 146–147, 1994.
- [4.9] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, and S. Wells, "An in-system reprogrammable 256K CMOS Flash memory," *Proc. IEEE ISSCC Dig. Tech. Pap.*, pp. 132–133, 1988.
- [4.10] G. Samachisa, C. S. Su, Y. S. Kao, G. Smarandoiu, T. Wong, and C. Hu, "A 128K Flash EEPROM using double polysilicon technology," *IEEE ISSCC Dig. Tech. Pap.*, pp. 76–77, 1987.
- [4.11] G. Samachisa, C. S. Su, Y. S. Kao, G. Smarandoiu, C. Y. M. Wang, T. Wong, and C. Hu, "A 128K Flash EEPROM using double polysilicon technology," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 10, pp. 676–683, 1987.
- [4.12] S. Haddad, C. Chang, B. Swaminathan, and J. Lien, "Degradation due to hole trapping in Flash memory cells," *IEEE Elect. Dev. Lett.*, vol. EDL-10, no. 3, pp. 117–119, 1989.
 S. Haddad et al., "An investigation of erase-mode dependent hole trapping in Flash EEPROM memory cell," *IEEE Elect. Dev. Lett.*, pp. 514–516, 1990.
- [4.13] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirola, "New ultra high density EPROM and Flash EEPROM cell with NAND structure cell," *IEEE IEDM Tech. Dig.*, pp. 552–555, 1987.
- [4.14] S. Aritome et al., "Reliability issues of Flash memory cell," *Proc. IEEE*, vol. 81, no. 5, pp. 776–788, 1993.
- [4.15] Y. Iwata et al., "The internal voltage system of the 32 Mbit NAND Flash EEPROM for low-voltage power supply," *Proc. IEEE Nonvolatile Semiconductor Memory Workshop*, Paper #4-2, 1995.
- [4.16] K. D. Suh et al., "A 3.3V 32 Mb NAND Flash memory with incremental step pulse programming scheme," *IEEE ISSCC Dig. Tech. Pap.*, pp. 128–129, 1995.
- [4.17] H. Kume, H. Yamamoto, T. Adachi, T. Hagiwara, K. Komori, T. Nishimoto, A. Koike, S. Meguro, T. Hayashida, and T. Tsukada, "A Flash-erase EEPROM cell with an asymmetric source and drain structure," *IEEE IEDM Tech. Dig.*, pp. 560–563, 1987.

- [4.18] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, S. Wells, and M. Winston, "An in-system reprogrammable 32Kx8 CMOS Flash memory," *IEEE J. Solid-State Circuits*, vol. SC-23, no. 10, pp. 1157–1162, 1988.
- [4.19] R. Shirota, Y. Itoh, R. Nakayama, M. Momodomi, S. Inoue, R. Kirisawa, Y. Iwata, M. Chiba, and F. Masuoka, "A new NAND cell for ultra high density 5V-only EEPROMs," *Symp. VLSI Tech.*, pp. 33–34, 1988.
- [4.20] S. Tam, S. Sachdev, M. Chi, G. Verma, L. Ziller, G. Tsau, S. Lai, and V. Dham, "A high density CMOS 1-T electrically-erasable non-volatile (Flash) memory technology," *Symp. VLSI Tech.*, pp. 31–32, 1988.
- [4.21] G. Verma and N. Mielke, "Reliability performance of ETOX-based Flash memories," *Proc. IRPS*, pp. 158–166, 1988.
- [4.22] M. Momodomi, R. Kirisawa, R. Nakayama, S. Aritome, T. Endoh, Y. Itoh, Y. Iwata, H. Oodaira, T. Tanaka, M. Chiba, R. Shirota, and F. Masuoka, "New device technologies for 5V only 4 Mb EEPROM with NAND structure cell," *IEEE IEDM Tech. Dig.*, pp. 412–415, 1988.
- [4.23] R. Kazerounian, S. Ali, Y. Ma, and B. Eitan, "A 5 volt high density poly-poly erase Flash EPROM cell," *IEEE IEDM Tech. Dig.*, pp. 436–439, 1988.
- [4.24] M. Gill, R. Cleavelin, S. Lin, S. D'Arrigo, G. Santin, P. Shah, A. Nguyen, J. Esquivel, B. Riemenschneider, and J. Paterson, "A 5-volt contactless array 256K Flash EEPROM technology," *IEEE IEDM Tech. Dig.*, pp. 428–431, 1988.
- [4.25] S. D'Arrigo, G. Imondi, G. Santin, M. Gill, R. Cleavelin, S. Spagliccia, E. Tomasetti, S. Lin, A. Nguyen, P. Shah, G. Savarese, and D. McElroy, "A 5V-only 256K bit CMOS Flash EEPROM," *IEEE ISSCC Dig. Tech. Pap.*, pp. 132–133, 1989.
- [4.26] M. Momodomi, Y. Itoh, R. Shirota, Y. Iwata, R. Nakayama, R. Kirisawa, T. Tanaka, S. Aritome, T. Endoh, K. Ohuchi, and F. Masuoka, "An experimental 4-Mbit CMOS EEPROM with a NAND structure cell," *IEEE J. Solid-State Circuits*, vol. SC-24, no. 10, pp. 1238–1243, 1989.
- [4.27] V. N. Kynett, M. Fandrich, J. Anderson, P. Dix, O. Jungroth, J. Kreifels, R. A. Lodenquai, B. Vajdic, S. Wells, M. Winston, and L. Tang, "A 90-ns one-million erase/program cycle 1-Mbit Flash memory," *IEEE J. Solid-State Circuits*, vol. SC-24, no. 10, pp. 1259–1264, 1989.
V. N. Kynett et al., "A 90 ns 100K erase/program cycle mega bit Flash memory," *IEEE ISSCC Dig. Tech. Pap.*, pp. 140–141, 1989.
- [4.28] K. Naruke, S. Yamada, E. Obi, S. Taguchi, and M. Wada, "A new Flash erase EEPROM cell with sidewall select-gate on its source side," *IEEE IEDM Tech. Dig.*, pp. 603–606, 1989.
- [4.29] M. Gill, R. Cleavelin, S. Lin, S. D'Arrigo, G. Santin, P. Shah, A. Nguyen, R. Lahily, P. DeSimone, G. Piva, and J. Paterson, "A 5-volt only Flash EEPROM technology for high density memory and system IC applications," *IEEE CICC*, pp. 1841–1844, 1989.
- [4.30] T. Endoh, R. Shirota, Y. Tanaka, R. Nakayama, R. Kirisawa, S. Aritome, and F. Masuoka, "New design technology for EEPROM memory cells with

- 10 million write/erase cycling endurance," *IEEE IEDM Tech. Dig.*, pp. 599–602, 1989.
- [4.31] S. Aritome, R. Kirisawa, T. Endoh, N. Nakayama, R. Shirota, K. Sakui, K. Ohuchi, and F. Masuoka, "Extended data retention characteristics after more than 10^4 write and erase cycles in EEPROMs," *Proc. IRPS*, pp. 259–264, 1990.
- [4.32] Y. Iwata, M. Momodomi, T. Tanaka, H. Oodaira, Y. Itoh, R. Nakayialtla, R. Kirhiawa, S. Aritome, T. Endoh, R. Shirota, K. Ohuchi, and F. Masuoka, "A high-density NAND EEPROM with block-page programming for microcomputer applications," *IEEE J. Solid-State Circuits*, vol. SC-25, no. 4, pp. 417–424, 1990.
- [4.33] B. Riemenschneider, A. L. Esquivel, J. Paterson, M. Gill, S. Lin, D. McElroy, P. Truong, R. Bussey, B. Ashmore, M. McConnell, H. Stiegler, and P. Shah, "A process technology for a 5-volt only 4 Mb Flash EEPROM with an $8.6 \mu\text{m}^2$ cell," *IEEE Symp. VLSI Tech.*, pp. 125–126, 1990.
- [4.34] R. Kirisawa, S. Arilome, R. Nakayama, T. Endoh, R. Shirota, and F. Masuoka, "A NAND structured cell with a new programming technology for highly reliable 5V-only Flash EEPROM," *Symp. VLSI Tech.*, pp. 129–130, 1990.
- [4.35] B. J. Woo, T. C. Ong, A. Fazio, C. Park, G. Atwood, M. Holler, S. Tam, and S. Lai, "A novel memory cell using Flash array contactless EPROM (FACE) technology," *IEEE IEDM Tech. Dig.*, pp. 91–94, 1990.
- [4.36] N. Ajika, M. Ohi, H. Arima, T. Malsukawa, and N. Tsubouchi, "A 5 volt only 16 Mbit Flash EEPROM cell with a simple stacked gate structure," *IEEE IEDM Tech. Dig.*, pp. 115–118, 1990.
- [4.37] M. Gill, R. Cleavelin, S. Lin, M. Middendorf, A. Nguyen, L. Wong, B. Huber, S. D'Arrigo, P. Shah, E. Kougianos, P. Hefley, G. Satin, and G. Naso, "A novel sublithographic tunnel diode based 5-volt Fash memory," *IEEE IEDM Tech. Dig.*, pp. 119–122, 1990.
- [4.38] S. Aritome, S. Shirota, R. Kirisawa, T. Endoh, N. Nakayama, K. Sakui, and F. Masuoka, "A reliable bi-polarity write/erase technology in Flash EEPROMs," *IEEE IEDM Tech. Dig.*, pp. 111–114, 1990.
- [4.39] Y. Miyawaki et al., "A new erasing and row decoding scheme for low supply voltage operation 16 Mb/64 Mb Flash EEPROMS," *Symp. VLSI Tech.*, pp. 85–86, 1990.
- [4.40] T. Tanaka et al., "A 4-Mbit NAND-EEPROM with tight programmed V_t distribution," *Symp. VLSI Circuits*, pp. 105–106, 1990.
- [4.41] Y. Yamauchi et al., "A 5V-only virtual ground Flash cell with an auxiliary gate for high density and high speed applications," *IEEE IEDM Tech. Dig.*, pp. 319–322, 1991.
- [4.42] M. Momodomi et al., "A 4-Mb NAND EEPROM with tight programmed V_t distribution," *IEEE J. Solid-State Circuits*, vol. SC-26, no. 4, pp. 492–496, 1991.
- [4.43] M. McConnell et al., "An experimental 4-Mb Flash EEPROM with sector erase," *IEEE J. Solid-State Circuits*, vol. SC-26, no. 4, pp. 484–491, 1991.

- [4.44] T. Nakayama et al., "A 60ns 16Mb Flash EEPROM with program and erase sequence controller," *IEEE J. Solid-State Circuits*, vol. SC-26, no. 11, pp. 1600–1604, 1991.
- [4.45] C. Kuo et al., "A 512Kb Flash EEPROM for a 32 bit microcontroller," *Symp. VLSI Circuits*, pp. 87–88, 1991.
- [4.46] B. J. Woo, T. C. Ong, and S. Lai, "A poly-buffered FACE technology for high density Flash memories," *Symp. VLSI Tech.*, pp. 73–74, 1991.
- [4.47] N. Kodama, K. Saitoh, H. Shirai, T. Okazaki, and Y. Hokari, "A 5V only 16 Mbit Flash EEPROM cell using highly reliable write/erase technologies," *Symp. VLSI Tech.*, pp. 75–76, 1991.
- [4.48] H. Kume, T. Tanaka, T. Adachi, N. Miyamoto, S. Saeki, Y. Ohji, M. Ushiyama, T. Kobayashi, T. Nishida, Y. Kawamoto, and K. Seiki, "A $3.42\text{ }\mu\text{m}^2$ Flash memory cell technology comformable to sector erase," *Symp. VLSI Tech.*, pp. 77–78, 1991.
- [4.49] K. Yoshikawa, S. Mori, E. Sakagami, N. Arai, Y. Kaneko, and Y. Ohshima, "Flash EEPROM cell scaling based on tunnel oxide thinning limitations," *Symp. VLSI Tech.*, pp. 79–80, 1991.
- [4.50] S. Yamada, T. Suzuki, E. Obi, M. Oshikiri, K. Naruke, and M. Wada, "A self-convergence erasing scheme for a simple stacked gate Flash EEPROM," *IEEE IEDM Tech. Dig.*, pp. 307–310, 1991.
- [4.51] T. Hori et al., "A MOSFET with Si-implanted gate-SiO₂ insulator for non-volatile memory applications," *IEEE IEDM Tech. Dig.*, pp. 17.7.1–4, 1992.
- [4.52] S. Koyama, "A novel cell structure for giga-bit EPROMs and Flash memories using polysilicon thin film transistors," *Symp. VLSI Tech.*, pp. 44–45, 1992.
- [4.53] K. Oyama et al., "A novel erasing technology for 3.3V Flash memory with 64Mb capacity and beyond," *IEEE IEDM Tech. Dig.*, pp. 607–610, 1992.
- [4.54] T. Endoh et al., "New write/erase operation technology for Flash EEPROM cells to improve the read disturb characteristics," *IEEE IEDM Tech. Dig.*, pp. 603–606, 1992.
- [4.55] H. Fukuda et al., "Heavy oxynitridation technology for forming highly reliable Flash-type EEPROM tunnel oxide films," *Electr. Letts.*, vol. 28, no. 19, pp. 1781–1783, 1992.
- [4.56] H. Fukuda et al., "High performance scaled Flash-type EEPROMs with heavily oxynitrided tunnel oxide films," *IEEE IEDM Tech. Dig.*, pp. 465–468, 1992.
- [4.57] H. Kume et al., "A $1.28\text{ }\mu\text{m}^2$ contactless memory cell technology for a 3V only 64Mb EEPROM," *IEEE IEDM Tech. Dig.*, pp. 991–993, 1992.
- [4.58] H. Onoda et al., "A novel cell structure suitable for a 3V operation, sector erase Flash memory," *IEEE IEDM Tech. Dig.*, pp. 599–602, 1992.
- [4.59] S. Keeney et al., "Simulation of enhanced injection split gate Flash EEPROM device programming," *Microelectronic Engineering*, vol. 18, pp. 253–258, 1992.
- [4.60] C. Kuo et al., "A 512Kb Flash EEPROM embedded in a 32b microcontroller," *IEEE J. of Solid State Circuits*, vol. SC-27, pp. 574–582, 1992.

- [4.61] A. Umezawa et al., "A 5V only operation 0.6 μ m Flash EEPROM with row decoder scheme in triple-well structure," *IEEE J. Solid State Circuits*, vol. SC-27, no. 11, pp. 1540–1546, 1992.
- [4.62] T. Jinbo et al., "A 5V only 16Mb Flash memory with sector erase mode," *IEEE J. of Solid State Circuits*, vol. SC-27, no. 11, pp. 1547–1554, 1992.
- [4.63] Mehrotra et al., "Serial 9 Mb Flash EEPROM for solid state disk applications," *Symp. VLSI Cir.*, pp. 24–25, 1992.
- [4.64] M. Ohi et al., "An asymmetrical offset source/drain structure for virtual ground array Flash memory with DINOR operation," *VLSI Technology Symp.*, pp. 57–58, 1993.
- [4.65] A. Bergemont et al., "NOR virtual ground (NVG) — A new scaling concept for very high density Flash EEPROM and its implementation in a 0.5 μ m process," *IEEE IEDM Tech. Dig.*, pp. 15–18, 1993.
- [4.66] A. Bergemont, "Status, trends, comparison and evolution of FLASH technology for memory cards applications," *IC Card Conf. Proc.*, pp. 63–71, 1993.
- [4.67] H. B. Pein and J. D. Plummer, "Performance of the 3-D sidewall Flash EEPROM cell," *IEEE IEDM Tech. Dig.*, pp. 11–14, 1993.
- [4.68] Y. S. Hisamune et al., "A high capacitive coupling ratio (HiCR) cell for 3V only 64 Mbit and future Flash memories," *IEEE IEDM, Tech. Dig.*, pp. 19–22, 1993.
- [4.69] S. Kobayashi et al., "Memory array architecture and decoding scheme for 3V-only sector erasable DINOR Flash memory," *IEEE J. of Solid State Circuits*, vol. SC-29, no. 4, pp. 454–460, 1994.
- [4.70] S. Atsumi et al., "A 16-Mb Flash EEPROM with a new self data refresh scheme for a sector erase operation," *IEEE J. of Solid State Circuits*, vol. SC-29, no. 4, pp. 461–469, 1994.
- [4.71] T. Takeshima et al., "A 3.3V single-power-supply 64Mb Flash memory with dynamic bit-line latch (DBL) programming scheme," *IEEE ISSCC Dig. Tech. Pap.*, pp. 148–149, 1994.
- [4.72] M. Mihara et al., "Row-redundancy scheme for high-density Flash memory," *IEEE ISSCC Dig. Tech. Pap.*, pp. 150–151, 1994.
- [4.73] T. Tanaka et al., "High speed programming and program-verify methods suitable for low voltage Flash memories," *Symp. VLSI Cir.*, pp. 61–62, 1994.
- [4.74] D. J. Lee et al., "An 18 Mb serial Flash EEPROM for solid-state disk applications," *Symp. VLSI Cir.*, pp. 59–60, 1994.
- [4.75] Y. Miyawaki et al., "An over-erase detection technique for tightening V_{th} distribution for low voltage operation NOR type Flash memory," *Symp. VLSI Cir.*, pp. 63–64, 1994.
- [4.76] T. Tanzawa et al., "A quick boosting charge pump circuit for high density and low voltage Flash memories," *Symp. VLSI Cir.*, pp. 65–66, 1994.
- [4.77] H. Watanabe et al., "Scaling of tunnel oxide thickness for Flash EEPROMS realizing stress-induced leakage current reduction," *Symp. VLSI Tech.*, pp. 47–48, 1994.

- [4.78] Y. Ma et al., "A novel high density contactless Flash memory array using split-gate source-side-injection cell for 5V-only applications," *Symp. VLSI Tech.*, pp. 49–50, 1994.
- [4.79] S. Mori et al., "High speed sub-half micron Flash memory technology with simple stacked gate structure cell," *Symp. VLSI Tech.*, pp. 53–54, 1994.
- [4.80] S. Sato et al., "An ultra-thin fully depleted floating gate technology for 64 Mb Flash and beyond," *Symp. VLSI Tech.*, pp. 65–66, 1994.
- [4.81] H. Oda et al., "New buried channel Flash memory cell with symmetrical source/drain structure for 64-Mbit or beyond," *Symp. VLSI Tech.*, pp. 69–70, 1994.
- [4.82] S. Kianian et al., "A novel 3 volts-only, small sector erase, high density Flash E²PROM," *Symp. VLSI Tech.*, pp. 71–72, 1994.
- [4.83] D. Kuo et al., "TEFET—A high density, low erase voltage, trench Flash EEPROM," *Symp. VLSI Tech.*, pp. 51–52, 1994.
- [4.84] J. Chen, T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu, "Subbreakdown drain leakage current in MOSFET," *IEEE Elect. Dev. Lett.*, vol. EDL-8, p. 515, 1987.
- [4.85] T. Kamata, K. Tanabashi, and K. Kobayashi, "Substrate current due to impact ionization in MOSFET," *Jpn. J. Appl. Phys.*, vol. 15, p. 1127, 1976.
- [4.86] K. Seki, H. Kume, Y. Ohji, T. Tanaka, T. Adachi, M. Ushiyama, K. Shimohigashi, T. Wada, K. Komori, T. Nishimoto, K. Izawa, T. Hagiwara, Y. Kubota, K. Shoji, N. Miyamoto, S. Saeki, and N. Ogawa, "An 80 ns 1Mb Flash memory with on-chip erase/erase-verify controller," *ISSCC Dig. Tech. Pap.*, pp. 1147–1151, 1990.
- [4.87] R. A. Cernea, G. Samachisa, C. S. Su, H. F. Tsai, Y. S. Kao, C. Y. Wang, Y. S. Chen, A. Renninger, T. Wong, J. Brennan, and J. Haines, "A 1Mb flash EEPROM," *ISSCC Dig. Tech. Pap.*, p. 138, 1989.
- [4.88] A. T. Wu et al., "A novel high-speed, 5-V programming EPROM structure with source-side injection," *IEEE IEDM Tech. Dig.*, pp. 584–587, 1986.
- [4.89] C. Chang and J. Lien, "Corner-field induced drain leakage in thin oxide MOSFETs," *IEEE IEDM Tech. Dig.*, pp. 714–717, 1987.
- [4.90] R. E. Shiner et al., "Data retention in EPROMs," *Proc. IRPS*, pp. 238–243, 1980.
- [4.91] R. N. Mielke, "New EPROM data-loss mechanisms," *Proc. IRPS*, pp. 106–113, 1983.
- [4.92] K. C. Smith, "The prospects for multivalued logic: a technology and application view," *IEEE Trans. Computers*, vol. C-30, pp. 619–634, 1981.
- [4.93] S. J. Wei and H. C. Lin, "Multivalued SRAM cell using resonant tunnelling diodes," *IEEE J. Solid-State Circuits*, vol. SC-27, no. 2, pp. 212–216, 1992.
- [4.94] M. Horiguchi et al., "An experimental large-capacity semiconductor file memory using 16-level/cell storage," *IEEE J. Solid-State Circuits*, vol. SC-23, no. 1, pp. 27–33, 1988.
- [4.95] T. Furuyama et al., "An experimental 2-bit/cell storage DRAM for macro-cell or memory-on-logic applications," *IEEE J. Solid-State Circuits*, vol. SC-24, no. 2, pp. 388–393, 1989.

- [4.96] E. Harari, "Flash EEPROM memory system having multi-level storage cells," U.S. Patent #5,043,940, 1991.
- [4.97] "Intel working on multilevel Flash," *EE Times*, p. 1, August 1, 1994.
- [4.98] R. M. Anderson and D. K. Ker, "Evidence for surface asperity mechanism of conductivity in oxides grown on polycrystalline silicon," *J. Appl. Phys.*, vol. 48, no. 4, p. 483, 1977.
- [4.99] J. Van Houdt et al., "A 5V/3.3V-compatible Flash E²PROM cell with a 400 ns/70 μ s programming time for embedded memory applications," *Proc. 5th Biennial Nonvolatile Memory Technology Review*, Linthicum Heights, Md., pp. 54-57, 1993.
- [4.100] J. Van Houdt et al., "Analysis of the enhanced hot-electron injection in split-gate transistors useful for EEPROM applications," *IEEE Trans. Elect. Dev.*, vol. ED-39, p. 1150, 1992.
- [4.101] N. Ajika et al., "A 5V-only 16 Mbit Flash EEPROM cell with a simple stacked gate structure," *IEEE IEDM Tech. Dig.*, pp. 115-118, 1990.
- [4.102] T. Tanazawa et al., "A quick boosting charge pump circuit for high density and low voltage Flash memories," *Symp. VLSI Cir.*, pp. 65-66, 1994.
- [4.103] J. S. Witters et al., "Degradation of tunnel-oxide floating gate EEPROM device and the correlation with high field current-induced degradation of thin gate oxide," *IEEE Trans. Elect. Dev.*, vol. ED-36, no. 9, pp. 1663-1682, 1989.
- [4.104] Chi Chang, "Flash memory reliability," a tutorial at IRPS, 1993.
- [4.105] Stefan Lai, "Flash memory reliability," a tutorial at IRPS, 1993.
- [4.106] B. Doyle, M. Bourcier, J. C. Marchetaux, and A. Boudou, "Interface state creation and charge trapping in the medium-to-high gate voltage range ($V_d/2 < V_g < V_d$) during hot-carrier stressing of n-MOS transistors," *IEEE Trans. Elect. Dev.*, vol. ED-37, no. 3, p. 744, 1990.
- [4.107] M.-S. Liang, C. Chang, W. Yang, C. Hu, and R. W. Brodersen, "Hot carriers-induced degradation in thin gate oxide MOSFET's," *IEEE IEDM Tech. Dig.*, p. 186, 1983.
- [4.108] T. Nishida and S. E. Thompson, "Oxide field and temperature dependent gate oxide degradation by substrate hot electron injection," *Proc. IRPS*, p. 310, 1991.
- [4.109] C. S. Jeng, T. R. Ranganath, C. H. Huang, H. S. Jones, and T.T.L. Chang, "High-field generation of electron traps and charge trapping in ultra-thin SiO₂," *IEEE IEDM Tech. Dig.*, p. 388, 1981.
- [4.110] M.-S. Liang and C. Hu, "Electron trapping in very thin thermal silicon dioxides," *IEEE IEDM Tech. Dig.*, p. 396, 1981.
- [4.111] M.-S. Liang, C. Chang, Y. T. Yeow, C. Hu, and R. W. Brodersen, "MOSFET degradation due to stressing of thin oxide," *IEEE Trans. Elect. Dev.*, vol. ED-31, no. 9, p. 1238, 1984.
- [4.112] J. S. Witters, G. Groeseneken, and H. E. Maes, "Programming mode dependent degradation of tunnel oxide floating gate devices," *IEEE IEDM Tech. Dig.*, p. 544, 1987.

- [4.113] C. Chang, S. Haddad, B. Swaminathan, and J. Lien, "Drain-avalanche and hole trapping induced gate leakage in thin-oxide MOS devices," *IEEE Elect. Dev. Lett.*, vol. EDL-9, no. 11, p. 588, 1988.
- [4.114] Y. Igura, H. Matsuoka, and E. Takeda, "New device degradation due to 'cold' carriers created by band-to-band tunneling," *IEEE Electron Device Letters*, vol. 10, no. 5, p. 227, May 1989.
- [4.115] K. Yoshikawa, S. Mori, E. Sakagami, Y. Ohshima, Y. Kaneko, and N. Arai, "Lucky-hole injection induced by band-to-band tunneling leakage in stacked gate transistors," *IEEE IEDM Tech. Dig.*, p. 577, 1990.
- [4.116] R. Rakkhit, S. Haddad, C. Chang, and J. Yue, "Drain-avalanche induced hole injection and generation of interface traps in thin oxide MOS devices," *Proc. IRPS*, p. 150, 1990.
- [4.117] Y. Tang, C. Chang, S. Haddad, A. Wang, and J. Lien, "Differentiating impacts of hole trapping vs. interface states on TDDDB reduction in MOS transistors," *Proc. SRC Topical Res. Conf. on Floating Gate Non-Volatile Memory Research*, Berkeley, Calif., 1992.
- [4.118] M. Baur et al., "A multilevel cell 32 Mb Flash memory," *IEEE ISSCC Tech. Dig. Pap.*, pp. 132–133, 1995.
- [4.119] N. Kodama, K. Oyama, H. Shirai, K. Saitoh, T. Okazawa, and Y. Hokari, "A symmetrical side wall (SSW)-DSA cell for a 64Mbit Flash memory," *IEEE IEDM Tech. Dig.*, p. 303, 1991.
- [4.120] T. Jinbo, H. Nakata, K. Hashimoto, T. Watanabe, K. Ninomiya, T. Urai, M. Koike, T. Sato, N. Kodama, K. Oyama, and T. Okazawa, "A 5V-only 16 Mb Flash memory with sector-erase anode," *IEEE ISSCC. Dig. Tech. Pap.*, p. 154, 1992.
- [4.121] E. Takeda, N. Suzuki, and T. Hagiwara, "Device performance degradation due to hot-carrier injection at energies below the Si-SiO₂ energy barrier," *IEEE IEDM Tech. Dig.*, p. 396, 1983.
- [4.122] T. Endoh, R. Shiota, M. Momodomi, and F. Masuoka, "An accurate model of subbreakdown due to band-to-band tunnelling and some applications," *IEEE Trans. Elect. Dev.*, vol. ED-37, no. 1, pp. 290–296, 1990.
- [4.123] H. Matsuoka, Y. Igura, and E. Takeda, "Device degradation due to band-to-band tunnelling," *Intl. Conf. Sol. State Dev. and Mat.*, Tokyo, pp. 589–592, 1988.
- [4.124] T. Ozaki, A. Nitayama, T. Hamamoto, K. Sunouchi, and F. Horiguchi, "Analysis of band-to-band tunnelling leakage current in trench-capacitor DRAM cells," *IEEE Elect. Dev. Lett.*, vol. EDL-12, no. 3, pp. 95–97, 1991.
- [4.125] R. Shiota et al., "A new programming method and cell architecture for multilevel NAND Flash memories," *IEEE Nonvolatile Semiconductor Memory Workshop*, Paper #2.7, 1995.
- [4.126] S. K. Lai, "Oxide/silicon interface effects in E²PROMs and ETOX™ Flash," *Proc. SRC Topical Res. Conf. on Floating Gate Non-Volatile Memory Research*, Berkeley, Calif., 1992.
- [4.127] T. C. Ong, A. Fazio, N. Mielke, S. Pan, G. Atwood, and S. K. Lai, "Instability of erase threshold voltage in ETOX™ Flash memory array,"

- Proc. SRC Topical Res. Conf. on Floating Gate Non-Volatile Memory Research, Berkeley, Calif., 1992.
- [4.128] S. Lassig and M.-S. Liang, "Time-dependent degradation of thin gate oxide under post-oxidation high-temperature anneal," *IEEE Elect. Dev. Lett.*, vol. EDL-8, no. 4, p. 160, 1987.
- [4.129] K. Yoneda, Y. Fukuzaki, K. Satoh, Y. Hata, Y. Todokoro, and M. Inoue, "Reliability degradation mechanism of the ultra-thin tunneling oxide by post-annealing," *Dig. Symp. VLSI Tech.*, p. 121, 1990.
- [4.130] M. Ushiyama, Y. Ohji, T. Nishimoto, K. Komori, H. Murakoshi, H. Kume, and S. Tachi, "Two dimensionally inhomogeneous structure at gate electrode/gate insulator interface causing Fowler-Nordheim current deviation in nonvolatile memory," *Proc. IRPS*, p. 331, 1991.
- [4.131] H. Fukuda, M. Yasuda, T. Iwabuchi, and S. Ohno, "Novel N₂O-oxynitridation technology for forming highly reliable EEPROM tunnel oxide films," *IEEE Elect. Dev. Lett.*, vol. EDL-12, no. 11, pp. 587-589, 1991.
- [4.132] S. K. Lai, J. Lee, and V. K. Dham, "Electrical properties of nitride-oxide systems for use in gate dielectric and EEPROM," *IEEE IEDM Tech. Dig.*, pp. 190-193, 1983.
- [4.133] H. Fukuda, T. Arakawa, and S. Ohno, "Thin-gate SiO₂ films formed by in-situ multiple rapid thermal processing," *IEEE Trans. Elect. Dev.*, vol. ED-39, no. 1, pp. 127-133, 1992.
- [4.134] K. Yoshikawa, S. Yamada, J. Miyamoto, T. Suzuki, M. Oshikiri, E. Obi, Y. Hiura, K. Yamada, Y. Ohshima, and S. Atsumi, "Comparison of current Flash EEPROM erasing methods: stability and how to control," *IEEE IEDM Tech. Dig.*, p. 595, 1992.
- [4.135] A. Brand, K. Wu, S. Pan, and D. Chin, "Novel read disturb failure mechanism induced by Flash cycling," *Proc. IRPS*, pp. 127-132, 1993.
- [4.136] J. Maserjian and N. Zamani, "Behavior of the Si/SiO₂ interface observed by Fowler-Nordheim tunneling," *JAP*, pp. 559-567, January 1982.
- [4.137] P. Olivio, T. N. Nguyen, and B. Ricco, "High-field-induced degradation in ultra-thin SiO₂ films," *Trans. Elect. Dev.*, vol. 35, no. 12, pp. 2259-2267, December 1988.
- [4.138] K. Naruke, S. Taguchi, and M. Wada, "Stress-induced leakage current limiting to scale down EEPROM tunnel oxide thickness," *IEEE IEDM Tech. Dig.*, pp. 424-427, 1988.
- [4.139] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," *IEEE IEDM Tech. Dig.*, pp. 139-142, 1992.
- [4.140] D. A. Baglee and M. C. Smayling, "The effects of write/erase cycling on data loss in EEPROMs," *IEEE IEDM Tech. Dig.*, pp. 624-626, 1985.
- [4.141] R. Rofan and C. Hu, "Stress-induced oxide leakage," *IEEE Elect. Dev. Lett.*, vol. EDL-12, pp. 632-634, 1991.
- [4.142] E. Rosenbaum and C. Hu, "High-frequency time-dependent breakdown of SiO₂," *IEEE Elect. Dev. Lett.*, vol. EDL-12, pp. 267-269, 1991.
- [4.143] E. Rosenbaum, Z. Liu, and C. Hu, "The effect of oxide stress waveform on MOSFET performance," *IEEE IEDM Tech. Dig.*, pp. 719-722, 1991.

- [4.144] S. Mori, E. Sakagami, H. Araki, Y. Kaneko, K. Narita, Y. Ohshima, N. Arai, and K. Yoshikawa, "ONO inter-poly dielectric scaling for nonvolatile memory applications," *IEEE Trans. Elect. Dev.*, vol. ED-38, no. 2, pp. 386–391, 1991.
- [4.145] R. D. Hoffman, "Speciality memories gain currency," *EBN*, p. 25, April 18, 1994.
- [4.146] C. Westmont and B. Wall, "Reprogrammability eases PCMCIA designs," *Electronic Design*, pp. 110–113, April 18, 1994.
- [4.147] B. Bipert and L. Hebert, "Flash memory goes mainstream," *IEEE Spectrum*, pp. 48–52, October 1993.
- [4.148] Suguru Kuki, "Speed of Flash memories sparks maker interest," *JEE*, pp. 40–43, November 1992.
- [4.149] D. Bursky, "Flash memory interfaces simplify systems," *EDN*, pp. 168–171, November 7, 1994.
- [4.150] R. H. Dennard et al., "Design of Ion-Implanted MOSFET's With Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, vol. SC-9, no. 10, pp. 256–268, 1974.
- [4.151] M. Kato et al., "A 0.4- μm^2 self-aligned contactless memory cell technology suitable for 256-Mbit Flash memories," *IEEE IEDM Tech. Dig.*, pp. 921–923, 1994.
- [4.152] S. Aritome et al., "A 0.67 μm^2 self-aligned shallow trench isolation cell, [SA-STI cell] for 3V-only 256 Mbit NAND EPROMs," *IEEE IEDM Tech. Dig.*, pp. 61–64, 1994.
- [4.153] Data storage report, Jonas Press Publishing Company, 1994.
- [4.154] S. Aritome et al., "Advanced NAND-structure cell technology for reliable 3.3V 64 Mb electrically erasable and programmable read only memories (EEPROM)," *Jap. J. Appl. Phys.*, vol. 33, no. 1B, pp. 524–528, 1994.
- [4.155] M. Aoki et al., "Triple density DRAM cell with silicon selective growth channel and NAND structure," *IEEE IEDM Tech. Dig.*, pp. 631–634, 1994.
- [4.156] S. Onishi, "A half-micron ferroelectric memory cell technology with stacked capacitor structure," *IEEE IEDM Tech. Dig.*, pp. 843–846, 1994.
- [4.157] B. Eitan et al., "Alternate metal virtual ground (AMG)—a new scaling concept for very high density EPROMS," *IEEE Elect. Dev. Lett.*, vol. EDL-12, no. 8, pp. 450–452, 1991.
- [4.158] *IEEE Floating Gate Standards – to be published in 1997.*
- [4.159] *IEEE Floating Gate Standards, IEEE Std.-1005, IEEE Standard Definitions and Characterization of Floating Gate Semiconductor Arrays, 1991.*
- [4.160] M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling in thermally grown SiO_2 ," *JAP*, vol. 40, p. 278, 1969.
- [4.161] F. Goodenough, "IC holds 16 seconds of audio without power," *Electronic Design*, pp. 39–44, January 31, 1991.