

FLASH MEMORIES

edited by
Paulo Cappelletti
Carla Golla
Piero Olivo
Enrico Zanoni



KLUWER ACADEMIC PUBLISHERS

Micron Ex. 1008, p. 1

Micron v. YMTc

IPR2024-00795

FLASH MEMORIES

FLASH MEMORIES

By

Paulo Cappelletti

Carla Golla

Piero Olivo

Enrico Zanoni

KLUWER ACADEMIC PUBLISHERS
Boston/Dordrecht/London

Distributors for North, Central and South America:

Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 871-6528
E-Mail <kluwer@wkap.com>

Distributors for all other countries:

Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail <orderdept@wkap.nl>



Electronic Services <<http://www.wkap.nl>>

Library of Congress Cataloging-in-Publication Data

Flash memories / by Paulo Cappelletti ... (et al.).

p. cm

Includes bibliographical references.

ISBN 0-7923-8487-3

1. Flash memories (computers) I. Cappelletti, Paulo

TK7895.M4F58 1999

004.5--dc21

99-25278

CIP

Copyright © 1999 by Kluwer Academic Publishers

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061

Printed on acid-free paper.

Printed in the United States of America

Contents

1		
FLASH MEMORIES: AN OVERVIEW		1
<i>P. Olivo, E. Zanoni</i>		
1.1	Role of Non Volatile Memories in Microelectronic Systems and in Semiconductor Market	1
1.2	Evolution of Non-volatile Memories	3
1.3	The Floating Gate Device	4
1.4	Charge Injection Mechanisms	7
1.5	Erasable Programmable Read Only Memories	7
1.5.1	The Floating gate Avalanche-injection MOS transistor (FAMOS) Cell	7
1.5.2	The basic Erasable Programmable Read Only Memory (EPROM)	8
1.6	Electrically Erasable Programmable Read Only Memories	9
1.6.1	The FLOating gate Thin Oxide (FLOTOX) Memory Cell	9
1.6.2	Textured Polysilicon Cells	10
1.6.3	The EEPROM Architecture	12
1.6.4	Ferroelectric Memories	13
✓1.7	Flash Memories: The Basic ETOX Cell. Programming and Erasing Mechanisms	15
✓1.8	Memory NOR Architecture and Related Issues	16
✓1.9	The NAND Flash Mass Storage Concept	24
1.10	Embedded Flash Memories	26
1.11	The Future of Flash Memories	27
1.11.1	Evolution of Flash Memory Technology	27
1.11.2	Non Volatile Memories Market Development	29
References		33
2		
THE INDUSTRY STANDARD FLASH MEMORY CELL		37
<i>P. Pavan, R. Bez</i>		

vi FLASH MEMORIES

2.1	Introduction	38
2.2	Basic Structure	42
2.3	Operating Conditions	46
2.3.1	Read	46
2.3.2	Program	47
2.3.3	Erase	53
2.4	Technology and Process	58
2.4.1	Isolation	60
2.4.2	Well and Channel Doping	61
2.4.3	Cell Structure Definition	64
2.4.4	Interlevel Dielectrics	68
2.4.5	Interconnections	70
2.4.6	Final Passivation	74
2.5	Yield and Reliability	74
2.5.1	Retention	75
2.5.2	Endurance	75
2.5.3	Reading Disturbs	76
2.5.4	Programming Disturbs	77
2.5.5	Erasing Disturbs	79
2.6	Scaling Issues	81
	References	83

3
BINARY AND MULTILEVEL FLASH CELLS 91

B. Eitan, A. Roy

3.1	Introduction to Flash Cell Design	91
3.2	Binary Flash Cells	93
3.2.1	Figures of Merit	93
3.2.2	Cell Design Complication Hierarchy from ROM to Flash	94
3.2.3	Basis for Flash Cells/Array Classification	96
3.2.4	Detailed Description of Flash Cells and Architectures	98
3.2.5	Scaling and Conclusions	131
3.3	Multilevel Flash Cells	137
3.3.1	Introduction to the Concept of Multilevel Flash	137
3.3.2	Multilevel Programming Mechanisms	138
3.3.3	Architectures for Multilevel Flash Memories	142
3.3.4	Scaling and Trade-Offs for Multilevel	146
	References	147

4
PHYSICAL ASPECTS OF CELL OPERATION AND RELIABILITY 153

L. Selmi, C. Fiegna

4.1	Introduction	153
4.2	Electronic Properties of Carriers and MOS Structures	155
4.2.1	Electrons in Crystals	155
4.2.2	Electrons as Classical Particles	156

4.2.3	Silicon	157
4.2.4	Silicon Dioxide	157
4.2.5	Silicon - Silicon Dioxide Interface	159
4.2.6	Oxide and Interface Traps	159
4.3	Fundamentals of Tunneling Phenomena	161
4.3.1	Basic Concepts and the WKB Approximation	161
4.3.2	Transmission Coefficient	162
4.3.3	Tunneling Current	165
4.4	Tunneling Phenomena in MOSFETs	165
4.4.1	Fowler-Nordheim and Direct Tunneling Through Gate Oxides	166
4.4.2	Modeling the Tunnel Current of MOS Structures	167
4.4.3	Band-to-band and Trap-to-band Tunneling	170
4.4.4	Modeling the Band-to-band and Trap-to-band Tunneling Current	174
4.5	Fundamentals of Carrier Transport	176
4.5.1	The Distribution Function	176
4.5.2	The Boltzmann Transport Equation	178
4.5.3	Scattering	180
4.5.4	The Carrier Distribution in Thermal Equilibrium	181
4.5.5	Carrier Distributions in Homogeneous Electric Fields	182
4.5.6	The Effective Temperature Model	184
4.6	Hot Carrier Effects in MOSFETs	185
4.6.1	Carrier Heating in MOSFETs and Flash Cells	186
4.6.2	MOSFET Design and Carrier Heating	189
4.6.3	Simplified Models of Carrier Heating	190
4.6.3.1	Average Energy	190
4.6.3.2	Carrier Distribution	191
4.6.4	Impact Ionization	192
4.6.5	Substrate Current	194
4.6.6	Hot Carrier Injection into SiO ₂	197
4.6.6.1	Distribution Function	198
4.6.6.2	Injection Probability	198
4.6.7	Gate Current	199
4.6.7.1	Channel Hot Electron Injection	199
4.6.7.2	Drain Avalanche Hot Carrier Injection	201
4.6.7.3	Secondary Generated Hot Electron Injection	202
4.6.7.4	Substrate Hot Electron Injection	203
4.6.7.5	Implications for Device Operation	205
4.6.8	Hot Carrier Effects at Low Voltages	206
4.7	Oxide Degradation due to High Field Stress	207
4.7.1	Oxide Wear-out and SILC	207
4.7.1.1	Stress Induced Leakage Currents (SILC)	210
4.7.2	Oxide Breakdown	211
4.7.3	Lifetime Evaluation Models	213
4.7.3.1	SILC Lifetime Evaluation Model	214
4.7.3.2	Breakdown Lifetime Evaluation Model	215
4.8	Oxide and Interface Degradation due to Hot Carrier Injection	217
4.8.1	Homogeneous Hot Carrier Degradation	217
4.8.1.1	n-channel Devices	217

4.8.1.2	p-channel Devices	218
4.8.2	Non-homogeneous Hot Carrier Degradation	218
4.8.3	Lifetime Evaluation Models	221
References		223
5		
MEMORY ARCHITECTURE AND RELATED ISSUES		241
<i>M. Branchetti, G. Campardo, S. Commodaro, S. Ghezzi, A. Ghilardelli, C. Golla, M. Maccarrone, I. Martinez, R. Micheloni, J. Mulatti, M. Zammattio, S. Zanardi</i>		
5.1	Flash Architecture: General Overview	241
5.1.1	Flash Architecture Scenario	242
5.1.2	NOR Cell Operation and Array Organization	243
5.1.3	Flash Memory User Interface	250
5.1.4	Flash Memory Operations: Overview	253
5.1.4.1	Read Path Building Blocks Description	253
5.1.4.2	Program Path Building Blocks Description	256
5.1.4.3	Erase Path Building Blocks Description	256
5.2	Read Path: Decoding	257
5.2.1	Predecoding	259
5.2.2	Row Decoder	259
5.2.3	Column Decoder	263
5.2.4	Hierarchical Decoder	264
5.2.5	Low V_{CC} Problems	266
5.2.6	Boost Concept: Continuous Boost and "One-shot" Boost	267
5.2.7	A New Boost Approach: Miniboost	268
5.3	Read Path: Input and Output Buffers	270
5.3.1	Input Buffer	270
5.3.2	Output Buffer	272
5.3.3	Noise Issues	275
5.3.4	High Voltage Tolerance	277
5.4	Read Path: Sensing Techniques	280
5.4.1	Sensing Techniques: An Overview	281
5.4.2	Differential Sensing Technique	285
5.4.3	Differential Sensing Technique with Offset Current	289
5.4.4	Differential Semi-Parallel Sensing Technique	293
5.4.5	Reading Speed-up Techniques	295
5.4.6	From EPROM to Flash	304
5.4.7	Reading Flash Memories with Depleted Bits	305
5.4.8	Low Voltage Flash Read	308
5.4.9	Reference Problems	312
5.5	Program Operation Circuitry	314
5.5.1	Cell Programming Voltages: Optimum Choice	314
5.5.2	Typical Program Path	315
5.5.3	Drain Voltage Regulation: Principles and Basic Circuits	317
5.5.4	Gate Voltage Regulation Fundamentals	320
5.6	Erase Operation Circuitry	327
5.6.1	Double Supply Voltage Approach	329

5.6.1.1	Source Erase Circuitry	329
5.6.1.2	Slow Discharge of Critical Nodes	330
5.6.2	Single Supply Voltage Approach	331
5.6.2.1	Charge Pumping	332
5.6.2.2	Voltage Regulators	335
5.6.2.3	Source Switch	336
5.7	Control Logic and Embedded Algorithms	337
5.7.1	Logic Architecture	339
5.7.2	Embedded Algorithms	343
5.7.2.1	Sequencer (Pseudo-Microcontroller)	344
5.7.2.2	Finite State Machine	344
5.7.3	Program Flow	344
5.7.4	Erase Flow	346
5.7.5	Erase Suspend - Erase Resume	348
5.7.6	Testability Issues	348
5.8	Redundancy and Error Correction Codes	350
5.8.1	The Yield	350
5.8.2	Static Redundancy	352
5.8.3	Wafer Yield	353
5.8.4	A Real Case	354
5.8.5	Error Correction Codes	356
6	MULTILEVEL FLASH MEMORIES	361
	<i>G. Torelli, M. Lanzoni, A. Manstretta, B. Riccò</i>	
6.1	Introduction	362
6.1.1	The Multilevel Approach	362
6.1.2	Basic Issues for ML Storage	364
6.2	Array Architectures for Multilevel Flash Memories	368
6.2.1	NOR Architecture with CHE Programming	369
6.2.2	NOR Architecture with FN Programming	371
6.2.3	NAND Architecture	371
6.3	Multilevel Sensing	373
6.3.1	Signal Production and Recognition	374
6.3.2	Sensing Schemes	376
6.4	Multilevel Programming	382
6.4.1	Program-and-Verify Approaches	384
6.4.2	Self-Controlled Approaches	387
6.5	Conclusions	389
	References	391
7	FLASH MEMORY RELIABILITY	399
	<i>P. Cappelletti, A. Modelli</i>	
7.1	Introduction	399
7.2	Memory Array V_t Distributions and Tunnel Oxide "Defects"	401
7.3	Main Yield and Reliability Issues	409

x FLASH MEMORIES

7.3.1	Over-Erasing	409
7.3.2	Program Disturbs	411
7.3.3	Read Disturb	414
7.3.4	Program/Erase Endurance	415
7.3.5	Data Retention	416
7.4	Testing for Reliability	417
7.5	Failure Modes Induced by Program/Erase Cycling	418
7.5.1	Memory Cell Intrinsic Endurance	418
7.5.2	The Behavior of Tail Bits	422
7.5.3	Single Bit Failure Mechanisms	422
7.5.3.1	The Erratic Erase Phenomenon	423
7.5.3.2	Single Bit Data Loss after Program/Erase Cycling	426
7.5.3.3	Gain Degradation	430
7.6	Multilevel Storage Reliability	436
7.7	Conclusion	439
References		439
8		
FLASH MEMORY TESTING		443
<i>G. Casagrande</i>		
8.1	Introduction	443
8.1.1	Impact of Testing on Product Cost	443
8.1.2	Impact on Product Life Cycle	444
8.1.3	Objectives of Production Testing	445
8.1.4	Testing Versus Quality and Reliability	445
8.2	Flash Testing Aspects	446
8.2.1	Flash Functional Model	446
8.2.2	Oxide Stress in a Flash	446
8.2.3	Flash Testing Aspects	447
8.2.4	Conceptual Test Flow	448
8.3	Flash Testability Tools	450
8.3.1	Focus on Cell and Technology	450
8.3.1.1	Direct Memory Access	451
8.3.1.2	Vt Measurement	452
8.3.1.3	Stress Modes	454
8.3.1.4	Depletion/Low-Vt Test	455
8.3.2	Focus on Test Productivity	457
8.3.3	Focus on Design	457
8.3.4	Flash Design Testability: an Example	458
8.4	Fault Repairing	460
8.4.1	Error Correction	462
8.4.2	Redundancy	462
8.4.2.1	Diagnosis and Repairing	464
8.4.2.2	Testability Tools for Redundancy	465
8.5	Production Testing	466
8.5.1	DC Tests	467

8.5.2	Functional Testing	468
8.5.3	AC Read/Command Interface	469
8.5.4	Erase/Program Performance; Endurance	470
8.5.5	Reliability	470
8.6	Test Productivity	470
8.6.1	Impact on Tester Structure	471
8.6.2	Parallel Testing Final Test	473
8.6.3	Parallel Testing at EWS	473
8.7	Product Characterization	474
8.8	Conclusions	478
References		479
9		
FLASH MEMORIES: MARKET, MARKETING AND ECONOMIC CHALLENGES		481
<i>B. Beverina, P. Bergé, with contributions by C. Kunkel, G. Moy, A. Damiano, R. Ferrara, A. Re</i>		
9.1	Introduction	482
9.2	Market Segmentations	483
9.2.1	Application Segments and Subsegments	485
9.2.2	Technology, Performances and Applications	488
9.2.3	Segment Dynamics	492
9.2.4	Commodity or Non-Commodity?	493
9.3	Customer/Supplier Relationship	495
9.4	The Development of the Flash Market	496
9.5	Flash Memory and the "Economy"	499
9.6	Applications More in Detail	500
9.6.1	Survey	500
9.6.2	Flash in Mobile Phones and Terminals	503
9.6.3	Flash in the BIOS	508
9.6.4	Flash in Automotive	515
9.7	Conclusions	525
References		526

1 FLASH MEMORIES: AN OVERVIEW

Piero Olivo¹, Enrico Zanoni²

¹ Dipartimento di Ingegneria, Università di Ferrara
Via Saragat 1, 44100 Ferrara, Italy
olivo@ing.unife.it

² Dipartimento di Elettronica e Informatica, Università di Padova
Via Gradenigo 6/A, 35131 Padova, Italy
zanoni@dei.unipd.it

1.1 ROLE OF NON VOLATILE MEMORIES IN MICROELECTRONIC SYSTEMS AND IN SEMICONDUCTOR MARKET

Solid-state memory devices which retain information once the power supply is switched off are called “nonvolatile” memories. For instance, using standard digital technology, a nonvolatile memory can be implemented by writing permanently the data in the memory array during manufacturing (mask-programmed Read Only Memories, ROM). As an alternative, the user can program the information by blowing fusible links or antifuses, thus changing permanently the cell content (i.e. obtaining a Programmable ROM or PROM). In both cases, the memory array can not be erased, thus making these solutions viable only for a limited number of applications.

In the course of the years, several technological solutions have been developed, which have led to the availability of non-volatile memories which can be electrically written and erased. Erasable Programmable Read Only Memories (EPROM) can be electrically programmed, but have to be removed and exposed to ultra-violet (UV) radiation for about 20 minutes in order to be

erased. Electrically Erasable Programmable ROMs (EEPROM) are electrically erasable and programmable in-system, byte by byte, but use larger areas than EPROMs, and have therefore higher costs and lower densities.

System designers have long dreamt of a non volatile memory which could be electrically erased and programmed in-system, offering at the same time very high-density and low cost-per-bit, random access, bit alterability, short read/write times and cycle times, excellent reliability. In most of the current system applications, these features should be also combined with low power consumption and single, low-voltage, power supply operation. If available, a solid-state memory technology having these characteristics would not only become dominant in the nonvolatile memory market, but could also make possible an unprecedented design flexibility, replacing all other kinds of memory in many applications. At the moment, this "ideal" memory chip has still to be invented.

The Flash memory technology has many of the characteristics of the "ideal" memory concept and is consequently considered as a driver for the semiconductor industry in the next decade. In 1996 it was forecasted that nonvolatile memories are going to be 12% of the worldwide memory market by the year 2000; Flash memories will occupy 50% of this nonvolatile memory market. Currently (1998), the Flash memory market is approximately \$2.5B [1].

Flash memories are non volatile memories in which a single cell can be electrically programmed and a large number of cells – called a block, sector or page – are electrically erasable at the same time. The word "flash" itself is related to the fact that since the whole memory can be erased at once, erase times can be very fast. Flash technology combines the high density of the UV EPROM (it has basically a single transistor cell like EPROMs) with electrical in-system erasability of EEPROMs.

There are two major applications for Flash memories that should be pointed out. One is the possibility of nonvolatile memory integration in logic systems – mainly, but not only, microprocessors – to allow software updates, store identification codes, reconfigure the system on the field, or simply have smart cards. The other application is to create storing elements, like memory boards or solid-state hard disks, made by Flash memory arrays which are configured to create large-size memories to compete with miniaturized hard disks. Flash solid-state disks are very useful for portable applications, since they have small dimensions, low power consumption, and no mobile parts, therefore being more robust.

Flash memories also combine the capability of nonvolatile storage with an access time comparable to Dynamic Random Access Memories (DRAM), which allows direct execution of microcodes. Many programs can be stored in Flash chips, without being continuously loaded and unloaded from the hard disk, and directly executed. Moreover, the realization of new generations of Flash memo-

ries that can be erased by blocks of different sizes, emulating EEPROMs in some applications, and with a single power supply, widens the field of applicability for Flash memories and encourages new uses. Besides voice recorders, answering machines and portable audio guides, Flash memories find wide applications in personal computers and peripherals, automotive engine control units, digital cordless telephones, and in emerging applications, such as personal digital assistants (PDAs), digital set-top boxes, digital still cameras, portable medical diagnostic systems and many others, also taking advantage of the recent possibility of storing more bits on a single cell (in "multilevel" Flash memories).

In this introduction the reader will find a description of the basic concepts and characteristics which have led to the development of Flash memories. The basic cell structures, device physics, memory chip architecture, reliability and testing issues of Flash memories will be analyzed in details in the following Chapters.

1.2 EVOLUTION OF NON-VOLATILE MEMORIES

Two are the parameters describing how "good" and reliable a nonvolatile memory cell is: *endurance* (capability of maintaining the stored information after erase/program/read cycling) and *retention* (capability of keeping the stored information in time). The need of information modification, however, always contrasts with that of a good data retention; cells with different characteristics have different applications according to the relevance of some device functional parameters (absorbed power, programming/erasing speed and selectivity, capacity...).

To have a memory cell which can commute from one state to the other, and which can store the information independently of external conditions, the storing element needs to be a device whose conductivity can be changed in a non-destructive way.

One solution is to have a transistor with the threshold voltage which can change repetitively from a high to a low state, corresponding to the two states of the memory cell. Following the P1005 IEEE Draft Standard for Definitions, Symbols and Characterization of Floating Gate Memory Arrays [2], the low- and the high-threshold states in Flash memories are generally called as "erased" and "programmed", respectively. It must be noticed, however, that this standard definition is not followed for all kinds of cells; for some implementations, it is common practice to distinguish the "program" and "erase" operations on the basis of the memory array organization; as a consequence, "programmed" does not always correspond to "high threshold". Deviations from the standard definitions will be pointed out in Chapter 3, which describes advanced Flash memory cells.

The threshold voltage V_T of a MOS transistor can be written as:

$$V_T = K - Q/C_{ox} \quad (1.1)$$

where K is a constant that depends on the gate and substrate material, on the channel doping, and gate oxide thickness, Q is the charge weighted with respect to its position in the gate oxide, and C_{ox} is the gate oxide capacitance. As can be seen, the threshold voltage of a MOS memory cell can be altered by changing the amount of charge present between the gate and the channel, i.e. by changing Q/C_{ox} . Two are the most common solutions used to store charge:

1. in traps which are present in the insulator or at the interface between two dielectric materials. The most commonly used interface is the silicon oxide/nitride interface. Devices obtained in this way are called MNOS (Metal-Nitride-Oxide-Silicon) cells;
2. in a conductive material layer between the gate and the channel and completely surrounded by insulator; this is the "floating gate" (FG) device.

Because of their lower endurance and retention, MNOS devices are used only in specific applications (such as in military, thanks to their radiation hardness). Their modern counterpart, the SONOS (Silicon - Oxide - Nitride - Oxide - Silicon) nonvolatile memory technology, is still based on electron trapping in the nitride layer, but exploits the achievement of a better control of the processing of the ONO (Oxide-Nitride-Oxide) layer. By using a relatively thin (5-10 nm) dielectric layer low programming voltages (5 to 10 V) can be achieved. Despite these improvements, however, nonvolatile memories based on charge trapping are still a very low fraction of the total nonvolatile memory production.

Floating gate devices, on the contrary, are at the basis of every modern nonvolatile memory, and are used in particular for Flash applications.

1.3 THE FLOATING GATE DEVICE

The schematic cross section of a generic floating gate device is shown in Fig. 1.1a: the upper gate is the control gate, while the lower one, completely surrounded by dielectric, is the floating gate. The basic concepts and the functionality of a FG device can be easily understood by determining the relationship between the FG potential, that physically controls the channel conductivity, and the control gate potential, controlled by external circuitry. This can be done by using the simple electrical model of Fig. 1.1b, where C_C , C_S , C_D and C_B are the capacitance between FG and control gate, source, drain and bulk regions, respectively. The FG potential (V_F) is:

$$V_F = \frac{C_C}{C_T} V_C + \frac{C_S}{C_T} V_S + \frac{C_D}{C_T} V_D + \frac{C_B}{C_T} V_B + \frac{Q}{C_T} \quad (1.2)$$

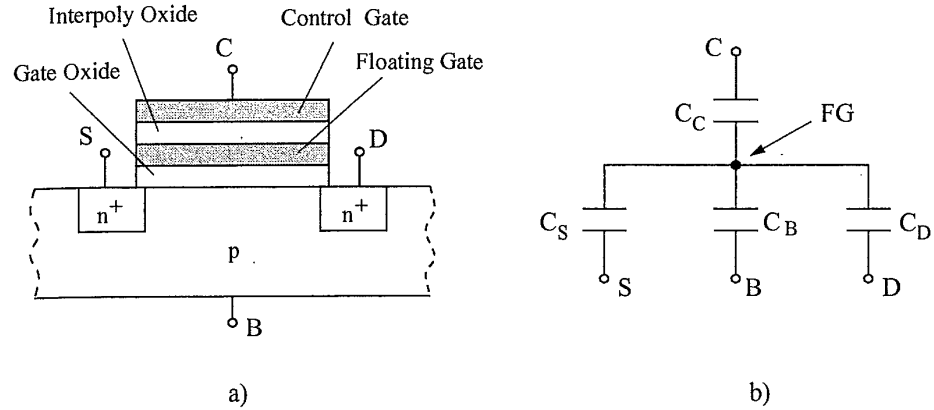


Figure 1.1 a) Schematic cross section of a generic floating gate device; b) electrical model of a floating gate device (junction capacitances are neglected).

where V_C , V_S , V_D , and V_B are the control gate, drain, source and bulk potentials, respectively; Q is the charge within the FG, while $C_T = C_C + C_S + C_D + C_B$ is the total capacitance.

Eq. (1.2) shows that the FG potential does not depend on the control gate voltage only, but also on source, drain and bulk potentials. If the source and bulk are both grounded and all potentials are referred to the source, (1.2) can be rearranged as

$$V_{FS} = \frac{C_C}{C_T} V_{CS} + \frac{C_D}{C_T} V_{DS} + \frac{Q}{C_T} . \quad (1.3)$$

By defining $\alpha_C = C_C/C_T$ as the "coupling factor" and $f = C_D/C_C$, (1.3) can be written as

$$V_{FS} = \alpha_C \left(V_{CS} + f V_{DS} + \frac{Q}{C_C} \right) . \quad (1.4)$$

The characteristics of a FG device depend on the threshold voltage, that is the potential ($V_{T_{FS}}$) that must be applied to the FG (with $V_{DS} = 0$) to reach the inversion of the surface population. Since the floating gate cannot be accessed, $V_{T_{FS}}$ is applied to the floating gate when a suitable voltage ($V_{T_{CS}}$), to be derived from (1.4), is applied to the control gate:

$$V_{T_{CS}} = \frac{1}{\alpha_C} V_{T_{FS}} - \frac{Q}{C_C} . \quad (1.5)$$

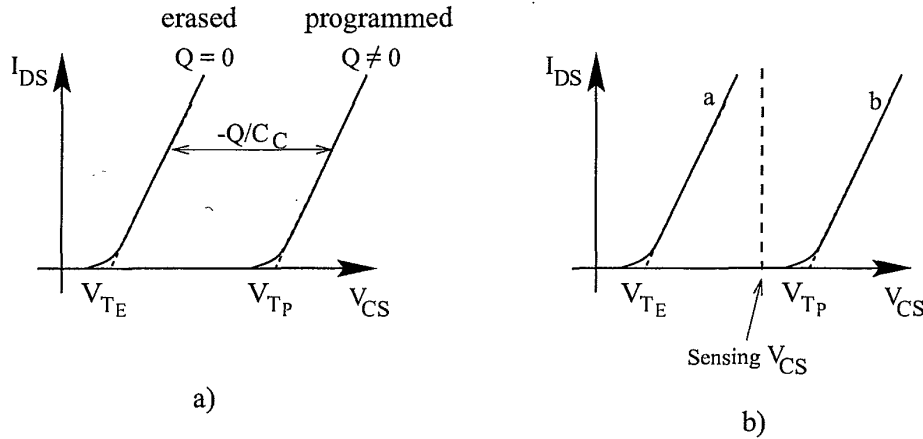


Figure 1.2 a) I-V trans-characteristics of a FG device for two different values of charge stored within the FG ($Q = 0$, and $Q < 0$), denoting two different states, respectively: erased and programmed; b) reading operation of a FG device: a suitable control gate voltage ($V_{TE} < V_{CS} < V_{TP}$) is applied to the device to determine whether it is conductive or not.

While $V_{T_{FS}}$ depends only on the device technology (and on the possible charge trapped within the gate oxide), $V_{T_{CS}}$ varies with the charge within the FG and this is the key result explaining the success of the FG device as the basic cell for nonvolatile memories applications. Fig. 1.2a shows two different I-V trans-characteristics obtained by modifying the FG charge. In particular, by choosing a suitable “threshold shift” ($|Q/C_C|$), it is possible to define two different (and separate) device states: *erased* for $Q = 0$, and *programmed* for $Q \ll 0$. The corresponding threshold voltages applied to the control gate are

$$V_{T_{CS}} = \frac{1}{\alpha_C} V_{T_{FS}} = V_{TE}, \quad (1.6)$$

$$V_{T_{CS}} = \frac{1}{\alpha_C} V_{T_{FS}} - \frac{Q}{C_C} = V_{TP}, \quad (1.7)$$

and they are denoted as “erased threshold” and “programmed threshold”, respectively.

The device state can be read by applying an appropriate “sensing” voltage to the control gate, as shown in Fig. 1.2b. When the FG device I-V curve corresponds to curve *a* ($Q = 0$), then $V_{CS} > V_{TE}$ and the device is ON; when the device has been previously programmed (curve *b*), $V_{CS} < V_{TP}$ and the device is OFF.

1.4 CHARGE INJECTION MECHANISMS

There are many solutions used to transfer electric charge from and into the floating gate. For both erase and program, the problem is making the charge pass through a layer of insulating material. The different physical phenomena which contribute to determine the behavior of a nonvolatile memory cell are analyzed in depth in Chapter 4.

The hot-electron injection and the Fowler-Nordheim tunneling mechanisms are generally used to write Flash memories. In the former, a lateral electric field (between source and drain) "heats" the electrons, and a transversal electric field (between channel and control gate) promotes the injection of the carriers through the oxide. The latter starts when there is a high electric field through a thin oxide. In these conditions, the energy band diagram of the oxide region is very steep; therefore, there is a high probability of electrons passing through the energy barrier itself.

Hot electrons and tunneling effects have been extensively studied since they can induce reliability problems in scaled MOS transistors. In nonvolatile memory cells, the very same mechanisms are controlled and exploited to become efficient program/erase mechanisms.

1.5 ERASABLE PROGRAMMABLE READ ONLY MEMORIES

1.5.1 *The Floating gate Avalanche-injection MOS transistor (FAMOS) Cell*

In 1967 D. Khang and S. M. Sze at Bell Laboratories [3] proposed a MOS-based nonvolatile memory cell based on a floating gate in a metal-insulator-metal-insulator-semiconductor structure. The lower insulator had to be thin enough (< 5 nm) to allow quantum-mechanical tunneling of electrons from the substrate to the floating gate and viceversa. At that time, however, it was almost impossible to deposit such a thin oxide layer without introducing fatal defects.

As a consequence, the tunneling mechanism was initially abandoned, and the first operating floating gate device, which adopted a fairly thick oxide layer, was developed at Intel in 1971 by Frohman-Bentchowsky [4]. This cell had no control gate, and was programmed by applying a highly-negative voltage at the drain, thus avalanching the drain/substrate junction, and creating a plasma of highly energetic electrons underneath the gate. The electrons were injected into the oxide and reached the floating gate, thus programming the cell.

Due to the absence of a control gate, however, the operation was extremely inefficient, and enormous voltages were needed. In order to inject electrons in the floating gate, p-channel devices had to be used. Erasure was obtained by providing externally the energy required by electrons to be re-emitted from the

floating gate. This was accomplished by exposing the cell to ultra-violet (UV) radiation.

1.5.2 The basic Erasable Programmable Read Only Memory (EPROM)

The FAMOS concept eventually evolved into a double polysilicon stacked gate n-channel cell, as schematically shown in Fig. 1.1, which constitutes the basic cell of an EPROM.

This cell is programmed by injection of channel hot-electrons into the floating gate and is erased using UV radiation. The programming consists in raising both the control gate (wordline) and the drain (bitline) to high voltages, typically 12 V. There are several relevant features:

1. hot electron programming is a very inefficient process, which requires both high voltage and high current. The stacked gate EPROM can not work with a single, low voltage supply;
2. only the cell which has both the control gate and the drain at high voltage is programmed: the operation is bit-selective. The same applies to the reading operation;
3. both the bit-selective hot-electron programming mechanism and the UV erasure process, which is obviously carried out on the whole array, are self-limiting. In particular, by UV erasure one can not indefinitely remove electrons from the floating gate, thus obtaining a cell with a too low threshold, i.e. an overerased cell. An over-erased cell is a cell with excessive source-drain leakage current when unselected, due to the threshold of the cell itself being lower than the applied control gate voltage.

Since the programming and reading operations are automatically bit-selective, while erasure is carried out on the whole chip, the EPROM does not require a select transistor or a split-gate structure to carry out bit selection, and can be implemented as a one-transistor memory cell. Its T-shaped cell is therefore extremely compact, leading to densities of 64 Mbit and more. Low-density memories, often adopting special processes can reach access times as fast as 15 ns, and a single supply voltage of +5 V is required for all operations except programming, which requires 12 V.

If reprogramming is needed, the EPROM must be removed from the circuit board, UV erased and then reprogrammed. The UV erasure requires the adoption of expensive packages having a transparent quartz window over the chip. When reprogramming is not required, a cheaper, plastic-packaged version of the same EPROM chip is often available, called One Time Programmable (OTP) memory. Since UV erasure is impossible, the memory can be written

only once, and represents a more dense, reliable, and cheaper alternative to PROMs.

1.6 ELECTRICALLY ERASABLE PROGRAMMABLE READ ONLY MEMORIES

Since the very beginning of nonvolatile memories development, various methods to achieve in-system electrical erasure, thus obtaining an Electrically Erasable Programmable Read Only Memory (EEPROM), were developed.

In 1967 Wegener *et al.* [5] introduced the already mentioned MNOS cell. The MNOS cell resembles a standard MOS transistor in which the oxide has been replaced by a nitride-oxide stacked layer. Electrons and holes can be trapped in the nitride, which then behaves as a charge storage element. Programming is achieved by applying a high, positive bias V_G to the gate, thus inducing quantum-mechanical tunneling of electrons from the channel region into the nitride traps. Erasure is obtained by tunneling of holes from the semiconductor to the nitride traps when V_G is negative and sufficiently high.

In order to improve the charge retention of MNOS memories, new structures have been developed. The SNOS (Silicon Nitride Oxide Semiconductor) employs a nitride layer deposited by Low Pressure Chemical Vapor Deposition (LPCVD) and a hydrogen anneal which improves the quality of the interfaces. The retention of the SNOS improves as the thickness of the nitride is reduced; unfortunately this leads to enhanced hole injection from the gate. In order to eliminate this problem, a top oxide layer is used between the gate and the nitride layer, thus obtaining the SONOS (Silicon Oxide Nitride Oxide Semiconductor) structure. SONOS EEPROM have been reported to withstand erase/write cycling up to 10M cycles, with $1.0 \mu\text{m}^2$ cells suitable for 256MB memory arrays and PCMCIA cards [6].

1.6.1 The Floating gate Thin Oxide (FLOTOX) Memory Cell

In order to obtain an electrically erasable and programmable non volatile memory, one can adopt Fowler-Nordheim for both programming and erasing, as proposed for the first time by Harari *et al.* [7]. Figure 1.3 shows the schematic cross-section of the FLOTOX cell, including a "selection" transistor which is required due to the non-selectivity of the tunneling process, as explained below. This combination represents the basic cell of a byte-addressable EEPROM memory [8]. Programming is obtained by applying a high voltage to the control gate, with the drain at low bias. By capacitive coupling, the voltage on the floating gate is also increased, and tunneling of electrons from the drain to the floating gate is initiated through the thin (8-10 nm) oxide grown on top of the

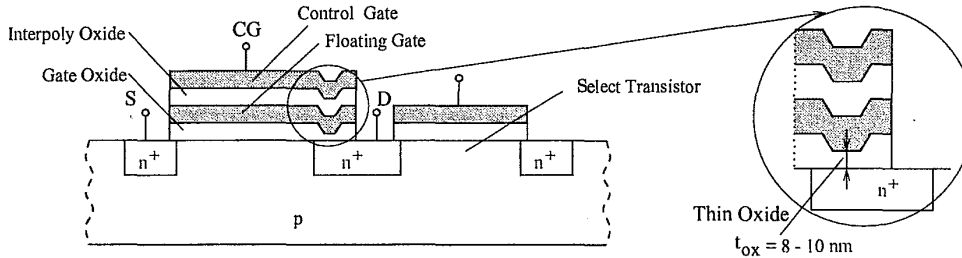


Figure 1.3 Schematic section of a FLOTOX cell including the select transistor.

drain, see the inset in Fig. 1.3. Erasing occurs when the drain is raised to a high voltage, and the control gate is grounded; the floating gate is capacitively-coupled to a low voltage, and electrons tunnel from the floating gate into the drain. The drain bias is controlled by the select transistor.

The process variations behind the implementation of a FLOTOX EEPROM cell starting from an EPROM process are relatively simple, so this kind of nonvolatile memory is widely diffused both as a stand alone memory product and within ASICs and logic. EEPROMs are erased on a bit or byte level, and can be reprogrammed in-system, with endurance up to 10^6 cycles. Since Fowler-Nordheim is a low current programming/erasing mechanism, the high voltage required can be generated within the chip, by specific charge-pumping circuits which multiply the supply voltage. The main limit of EEPROM is the cell size (3-4 times larger than that of a single transistor ROM cell) and the low density, due to the need of a select transistor (i.e. a two-transistors cell); typical densities of 4Mbit can be now reached.

Writing the EEPROM is also a rather slow process which has to take place in two steps. Referring to Fig. 1.3, the first step consists in programming all the byte cells by injecting electrons in the floating gate; this is carried out by raising the control gate and the gate of the select transistor, grounding the drain of the select transistor itself (the n^+ region at the extreme right in Fig. 1.3). Subsequently, selective erasing of the single bits is carried out. Electrons are removed only in selected cells (bits) by grounding the control gate, raising the select gate and applying a high voltage to the drain of the select transistor which transfers it to the floating gate transistor.

1.6.2 Textured Polysilicon Cells

Instead of exploiting tunneling through the thin oxide layer between the floating gate and the silicon drain area, textured polysilicon cells adopt tunneling through oxides thermally grown on polysilicon, with charge exchange between

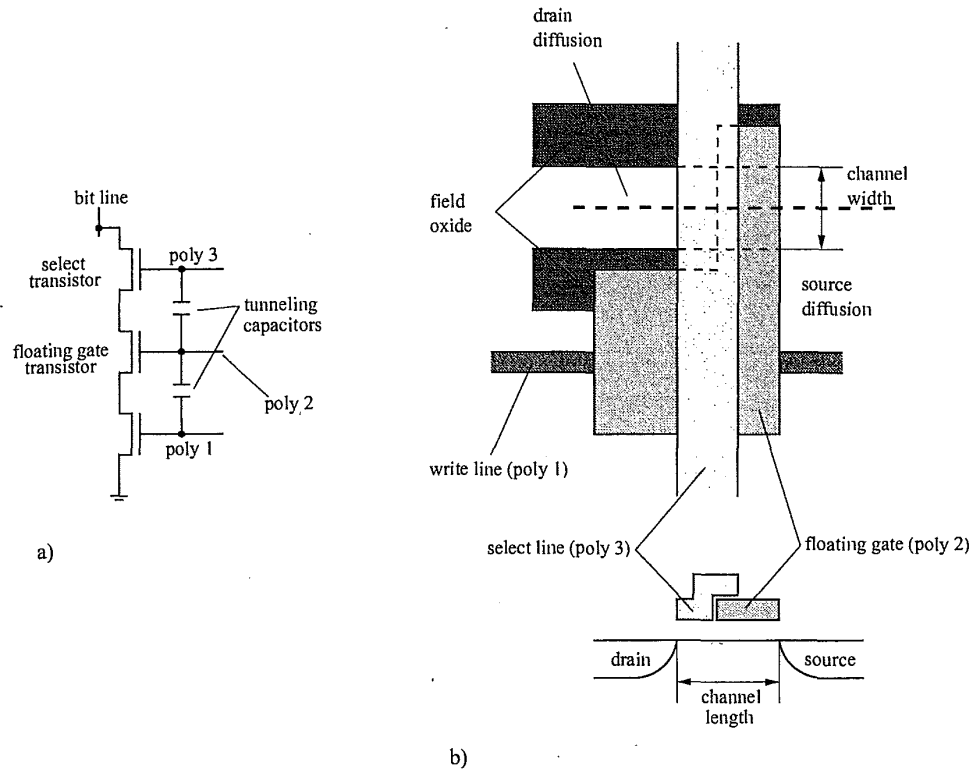


Figure 1.4 Equivalent circuit (a), layout and schematic cross-section (b) of a textured poly EEPROM cell

different polysilicon gates. The advantage is that, since the surface of polysilicon is rough (i.e. "textured"), electric field enhancement takes place, and tunneling is possible, for the same applied voltage, also using thick oxide layers. The cell consists in three transistors in series (see Fig. 1.4a), having their polysilicon gates partially overlapped, see Fig. 1.4b; the alignment of these layers may result critical for the adopted lithography. The floating gate is in the middle (poly2); electrons are injected from poly1 to the floating gate (programming) and from the floating gate to poly3 (erasing). The voltage on poly3 is always high, so erasing or programming is selected according to the voltage applied to the drain. The overlapping of the gates results in a compact vertical structure, which occupies an area lower than that of the FLOTOX EEPROM cell, resulting in an advantage for higher density EEPROMs. Moreover, this cell does not require programming before selective erasure, as in the case of the FLOTOX EEPROMs. The quality and reliability of this technology strongly

depends on the feature of the polysilicon and of the oxide thermally grown on it. Trapping in the poly oxide may result in device wearout.

1.6.3 The EEPROM Architecture

To understand why an EEPROM cell includes the actual storage element together with a series transistor, it is necessary to move our analysis to the architectural level. In a NOR architecture (see Fig 1.5 for the simplest 2×2 matrix without select gates), the same word line drives several control gates, while each bit line is shorted to the drain of many cells.

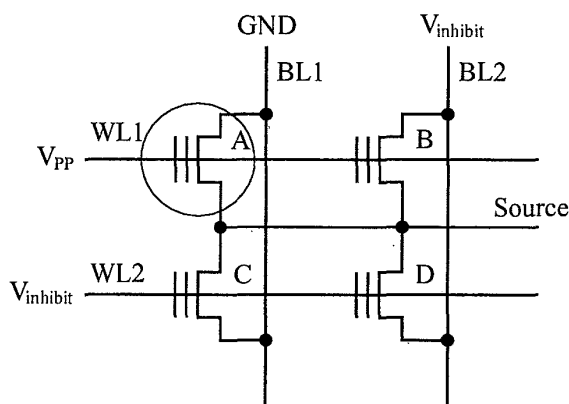


Figure 1.5 Simple 2×2 NOR-EEPROM architecture showing a program disturb on cell C when cell A is programmed.

If, as shown in Fig. 1.5, cell A is to be programmed, its control gate must be driven at the high programming voltage V_{PP} , while its drain must be kept at ground, so that a high oxide field allows for electron tunneling from the drain to the floating gate FG. Cell B, that shares the same word line WL1, may be also programmed, if its drain is also kept at a low voltage. Therefore, to inhibit an undesired programming of B, its drain, i.e. BL2, must be raised to a high voltage $V_{inhibit}$ (possibly equal to V_{PP} , to reduce the number of different biases) guaranteeing that no electrons may tunnel into its FG.

At this point, however, the problem of undesired writing is moved to cells C or D: if WL2 is kept at a low voltage, cell D may be erased (or even overerased if not already programmed, thus becoming depleted) since the high oxide field forces electron tunneling from the FG into the drain while, if WL2 is driven at a inhibiting voltage $V_{inhibit} = V_{PP}$, cell C is automatically programmed.

These write disturbs, as well as those occurring during erasing operations, may be reduced by using a lower inhibiting voltage, for instance $V_{inhibit} = V_{PP}/2$, with the need of an internal generation of such a voltage.

However, the general solution consists in the integration of a selecting transistor in series to any storage element, so that writing operations can affect only the selected cell. As discussed previously, such a solution implies a larger area occupied by the cell.

1.6.4 Ferroelectric Memories

Charge storage on a floating gate is not the only way to obtain a nonvolatile memory element. Other electrical properties of materials can be exploited in order to obtain an EEPROM-like memory cell; in particular, various approaches adopting ferroelectric materials have been demonstrated in recent years.

Ferroelectric materials are composed by small crystals characterized by a charge dipole which tends to align in parallel to an externally applied electric field. On increasing the electric field, the polarization level increases and saturates (see Fig. 1.6); the same applies when the electric field is reversed, so that there are two stable polarization states. These states do not require external electric field or current to be maintained and can be used for nonvolatile digital data storage.

The memory element is usually (but not always) represented by a capacitor, using a ferroelectric thin film as dielectric (see right-hand side of Fig. 1.7a). One of the most commonly adopted materials is lead zirconate titanate, $\text{Pb}(\text{Zr,Ti})\text{O}_3$ (PZT), which can be deposited by sol-gel or RF sputtering as an add-on of a standard CMOS process.

When the voltage applied to the capacitor exceeds a certain positive value V_{coerc} , the polarization becomes positive and increases up to a saturation value P_{sat} (see Fig. 1.6). The same applies for negative voltage lower than V_{ncoerc} , leading to a saturated polarization $-P_{sat}$. When the electric field is removed, the ferroelectric film maintains its state of polarization, but the value of polarization is somewhat reduced to a relaxation value P_{rel} (or $-P_{rel}$ if a negative voltage has been applied). Two logic states are therefore possible, corresponding to the P_{rel} and $-P_{rel}$ polarizations. When, during the read operation, a positive voltage is applied to the ferroelectric capacitor the polarization changes from P_{rel} to P_{sat} , thus requiring a low current (logic state 0) or from $-P_{rel}$ to P_{sat} , which corresponds to a high current, or to the logic state 1. The difference in current between two memory states is sensed to generate the output. After reading the 1 state, the $-P_{rel}$ negative polarization must be regenerated by applying a negative voltage to the capacitor.

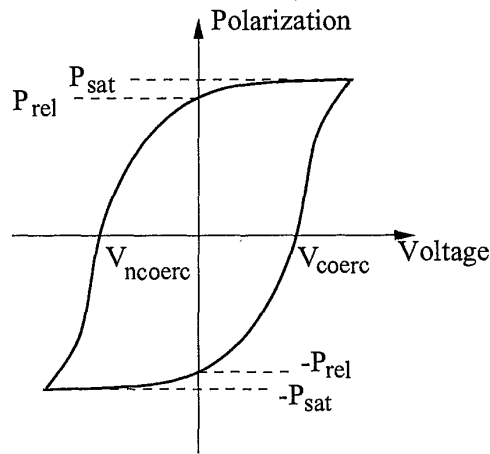


Figure 1.6 Typical hysteresis curve of a ferroelectric capacitor, identifying polarization states.

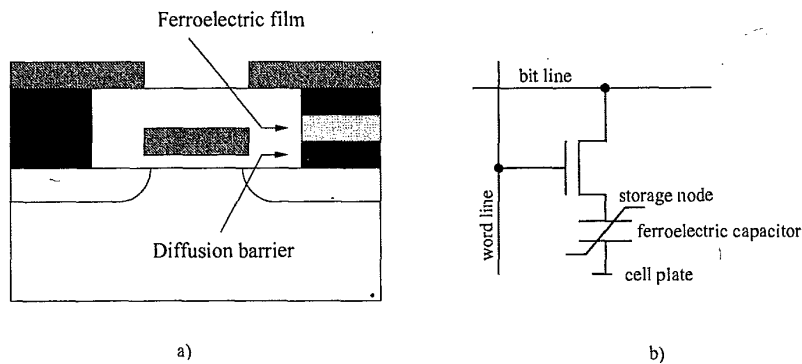


Figure 1.7 a) Schematic cross-section of a ferroelectric nonvolatile DRAM; b) equivalent circuit.

Potential advantages of ferroelectric memories can be summarized as follows: *i)* ferroelectric memories can use a single low voltage supply for all operations [9]; *ii)* fast read/write operations can be achieved, with access times of the order of 50 ns and cycle times lower than 100 ns; *iii)* ferroelectric memories are characterized by an excellent endurance, of over 10^{12} cycles, which would be impossible to achieve with Flash memories [10]; *iv)* ferroelectric films can maintain their characteristics in a very wide range of temperatures, up to 350°C; *v)* they can achieve excellent radiation tolerance characteristics, suitable for space and military applications.

The basic ferroelectric random access memory cell is composed by the series connection of one transistor and one ferroelectric capacitor, as shown in Fig. 1.7b, and its size is still much larger than that of DRAMs. Single-transistor cells adopting the ferroelectric film as dielectric in a Metal - Ferroelectric - Semiconductor Transistor (MFST) are affected by fabrication difficulties, poor data retention, high write voltage and read disturb [10]. Despite other solutions such as NAND organization (sacrificing random access) or vertical transistors (at the cost of very difficult fabrication) have been tried, the cell size limitations still has to be solved, though a 60 ns, 1 Mb nonvolatile ferroelectric memory has been presented in 1996 [11].

Ferroelectric Random Access Memories have a great potential as a future high density non volatile memories, because of the advantages previously listed; however, the ferroelectric films used as memory storage elements have some reliability concerns, such as aging effects after extensive cycling, thermal stability, degradation due to electric field, time-dependent breakdown phenomena. Extensive research activity concerning ferroelectric materials properties is being carried out; quaternary compounds such as $\text{SrBi}_2\text{Ta}_2\text{O}_9$ (SBT) currently provide best performances in terms of fatigue-free, low-voltage operation.

1.7 FLASH MEMORIES: THE BASIC ETOX CELL. PROGRAMMING AND ERASING MECHANISMS

Flash memories represent the synthesis of EPROM and EEPROM, since they are programmed and erased electrically but composed by single transistor cells. Programming is carried out selectively by means of the hot electron mechanism; erasing is based on tunneling, and is carried out in blocks of different sizes, from 512 bytes to full chip [12]. The first cell based on this concept was presented in 1979 [13]; the first commercial product, a 256-K memory chip, was presented by Toshiba in 1984 [14]. The market did not take off until this technology was proven to be reliable and manufacturable [15].

Figure 1.8 shows the cross-section of an industry-standard Flash cell. This cell structure was presented for the first time by INTEL in 1988 and named ETOXTM (EPROM Tunnel OXide; ETOX is a trademark of INTEL) [16]. Though it is derived from an EPROM cell, there are a few meaningful differences.

First, the oxide between the substrate and the floating gate is very thin (of the order of 10 nm). Therefore if a high voltage is applied at the source when the control gate is grounded, a high electric field exists in the oxide, enabling electron tunneling from the floating gate to the source. This bias condition is dangerously close to the breakdown of the source-substrate junction. Therefore, the source diffusion is realized differently from the drain diffusion, which

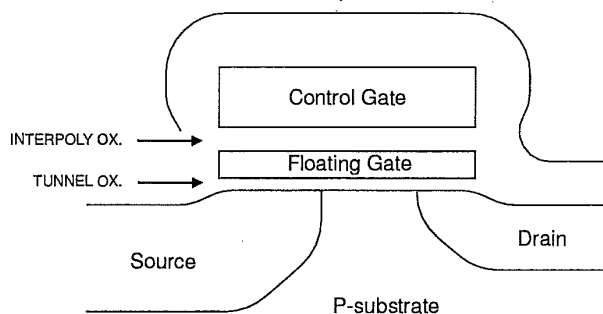


Figure 1.8 Schematic cross-section of an industrial Flash cell.

does not undergo such bias conditions. To do so, a new mask is added to the technological process to discriminate source and drain implants. The cell is not symmetrical, but this is the only difference with respect to the standard EPROM process. This is a great advantage, since all the accumulated experience in process development can be used to produce these devices.

The first Flash prototypes needed an external supply voltage for programming and external management of the erasing algorithm. They featured only a bulk-erase capability and their endurance was very poor (less than 10000 cycles). As an advantage versus EPROM's, they offered just an electrical-erase capability. Modern Flash memories have an embedded microcontroller to manage the erasing algorithms and offer sector erase capability and single power supply. In the following it will become clear that the correct operation of a Flash memory requires the design of a complex electronic circuit; due to the interaction between the various cells of the array, also the yield, quality, testing and reliability of the memory depend not only on the cell technology, but also, and in a more subtle way, on its architecture, which will be specifically discussed in Chapter 5. Flash reliability and testing will be addressed in Chapters 7 and 8, respectively.

1.8 MEMORY NOR ARCHITECTURE AND RELATED ISSUES

There are three basic operations in a Flash memory: read (a byte or a word), program (a byte or a word), erase (one or more sectors). Among these operations, the read one is the most frequent and it is also the simplest. To illustrate this basic operation, we consider that only one cell (rather than a byte or a word) is read at a time, as shown in Fig. 1.9, where a NOR organization has been considered. The extrapolation of the read procedure to a real case is quite simple since all the cells belonging to the same byte or word share the

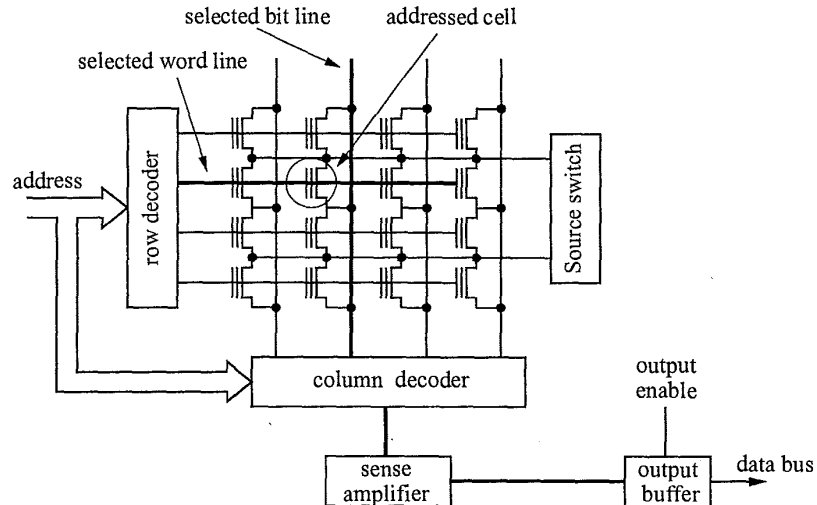


Figure 1.9 Schematic structure of the read path in a NOR organization. Only one bit at a time is here considered as addressable.

same word line, while 8 or 16 bit lines, as well as 8 or 16 sense amplifiers, are activated simultaneously. Once the cell address has been provided, the row decoder activates the selected word line by raising its voltage while keeping all the others at ground. The addressed bit line is connected to the sense amplifier. If the addressed cell is programmed (high threshold voltage, corresponding to the logic state "0"), no current flows through the cell and the bit line. On the contrary, if the cell is erased (low threshold voltage, corresponding to a logic "1"), the cell is ON and its current is detected by the sense amplifier.

The activation of the output enable signal transfers the read data to the data bus and the read operation ends. As it will be extensively illustrated in Chapter 5, the basic role is played by the sense amplifier, whose design must take into account several constraints.

The write operations (program and erase) are much more complex, as it can be understood by the following schematic description.

When programming, the word line is raised to activate the cell to be programmed. If the input data is 0, the bit line is driven at a high voltage, to allow for channel hot electrons to raise the cell threshold voltage. After the application of the programming pulse it is necessary to verify whether the cell has been correctly programmed, i.e. to verify whether the threshold voltage is larger than a minimum acceptable value V_{Tp_m} . This basic task is performed by reading the cell with a gate voltage higher than that usually applied during normal reading and by comparing the read data with that to be programmed,

that has been latched in a dedicated register. If the two data coincide, it means that the threshold voltage of the cell has risen from the erased value to the programmed one and that, since this read operation has provided a correct result with a higher reading voltage, the correct value is expected to be detected even in a conventional reading. If the verification fails, another programming pulse is applied to the cell (always by raising both gate and drain), until the cell is correctly programmed or a maximum number of pulses has been reached, so that a fail signal is produced.

For several reasons the erase procedure is even more complicated. First of all, it is performed on an entire sector, so that the verification process requires that all the cells of the sector are read in sequence. In addition, it is important to check whether the threshold of some cells become too low and, in case, to raise their threshold to a higher value. A schematic behavior of the thresholds' distribution for cells belonging to the same sector is shown in Fig. 1.10, starting from a typical situation occurring before erasing (Fig. 1.10a). Once the erase procedure has been activated, all the cells of the sector are programmed with a 0, so that their threshold is raised (Fig. 1.10b). This normalization task reduces the possibility of overerasing cells written with a 1 (that could become leaky when unaddressed), and it allows for a more uniform distribution of the erased thresholds, since all the initial thresholds belong to the same range.

To erase a single sector, a high electric field must be applied between the sources and the gates of the cells belonging to the sector, to allow for Fowler-Nordheim current to discharge the floating gate of the cells. This task is accomplished in two different ways: i) by applying a high voltage (in the range of 12 V) to the source of the cells to be erased while grounding their gates (*source erase*), or ii) by splitting the biasing voltage between source (at $5 \div 7$ V) and gate (at $-8 \div -10$ V) (*negative gate erase*). Both solutions present a major drawback: the highly negative bulk-source voltage drop can activate avalanche injection in the former case, while the generation of a negative voltage is required in the latter.

As illustrated in Chapter 4, tunneling current depends on many physical and technological parameters, so that even adjacent cells can discharge at different rates. After a single erase pulse has been applied, the threshold distribution may be similar to that depicted in Fig. 1.10c. In particular, many cells are not fully erased ($V_T > V_{TeM}$), while others may feature threshold voltages below the minimum allowed ($V_T < V_{Te_m}$).

As for the programming operation, it is then mandatory to check for the correctness of the erase procedure, by reading the entire sector with a gate voltage lower than that usually applied during normal reading. If the data read is "1", it means that the threshold voltage of the cell has been lowered from the programmed value to the erased one and that, since this read operation

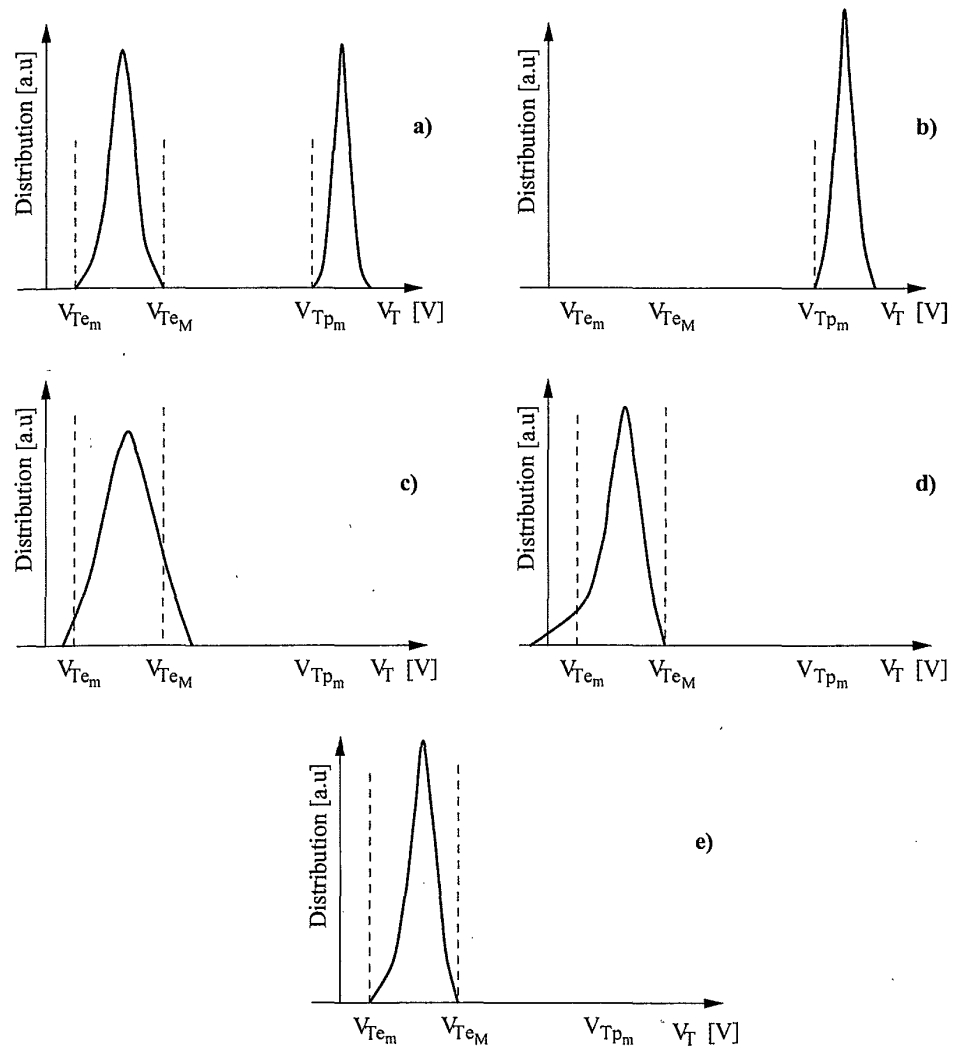


Figure 1.10 Schematic distribution of the threshold voltages during an erase operation: a) before erase; b) after a "program all-0" operation; c) after a single erase pulse; d) after the erase verify procedure has been successfully performed; e) after soft-programming.

has provided a correct result with a low reading voltage, the correct value is expected to be detected even in a conventional reading. If the verification fails for at least one cell, another erasing pulse is applied to the sector until all the cells are correctly erased or a maximum number of pulses has been reached, so that a fail signal is produced. After the erase verify procedure

(see Fig. 1.10d), it is important to check whether some cells are overerased ("depletion verify") and, in case, their thresholds must be driven to the correct range ("soft programming"). The former operation must check whether some cells feature low or even negative threshold (*depleted* cells), so that they would draw current even if not biased, thus preventing from a correct reading of cells belonging to the same bit line. With the latter operation, these cells are written with suitable gate and drain voltages that, lower than those used during the normal program procedure, allow for a slight increase of the threshold voltage. The final threshold distribution is then bounded between $V_{T_{em}}$ and $V_{T_{eM}}$ (see Fig. 1.10e).

From the schematic description of the three basic operations it is possible to stress some peculiar features concerning the architecture and the reliability of Flash memories:

Line biasing. Word lines (cell gates), bit lines (cell drains) and common source must be biased at different voltages, depending on the selected operation and even during the same procedure (for example, during erasing, the gate of a single cell must be driven at the programming voltage during "all-0 programming", at ground or at a negative voltage when an erase pulse is applied, at a low reading voltage during "erase verify", at suitable voltages during "depletion verify" and "soft programming").

As it will be extensively shown in Chapter 5, these requirements make the circuitry controlling these voltages quite complex, in particular in terms of decoders, switches and charge/discharge of highly-capacitive lines, noise disturbs due to capacitive coupling.

High voltage requirements. Since writing requires high fields applied to the cells, high voltages must be provided to cell terminals. These writing voltages are higher than those used during normal operations and to bias logic circuits. Therefore, two solutions are adopted: i) double voltage supply devices, in which the high voltages to be applied during programming are provided at an external pin; ii) single voltage supply devices, in which high voltages are generated internally, by means of charge pumps.

Charge pumps present several limitations making the overall performance of a Flash memory critically dependent on their design, in particular when low-voltage supplies are used. The main problems and solutions, detailed in Chapter 5, can be summarized as follow: relevant power consumption, large area and time requirements for high voltage generation, limited maximum output current.

In addition, when the negative gate erase scheme is chosen, both positive and negative charge pumps must be integrated.

Low voltage requirements. To reduce the total power consumption, that is proportional to the square of the power supply V_{cc} , there is a general trend towards low power supplies, with two basic implications:

1. the design of high-performance charge pumps is complicated by the fact that the final output voltage required for hot electron injection and Fowler-Nordheim tunneling cannot be scaled because of effectiveness and reliability problems (see Chapters 2, 4 and 7);
2. for the same reasons the range of the thresholds for erased and programmed cells ($[V_{Te_m}, V_{Te_M}]$, $[V_{Tp_m}, \infty[$, respectively) cannot be significantly modified. Consequently, while a reading gate voltage equal to $V_{cc} = 5V$ or to $V_{cc} = 3.3V$ allows for a direct reading of cells with $V_{Te_M} \leq 2.5V$ and $V_{Tp_m} \geq 5.5V$, the reading gate voltage must be boosted too when lower power supplies are used, thus increasing the complexity of the reading path.

Program disturbs. A Flash memory, as well as other non-volatile memories, is inherently prone to disturbs induced by the high voltages used for programming.

Program disturbs affect cells sharing either the word line or the bit line of a cell addressed for programming, as depicted in Fig. 1.11, where the marked cell is to be programmed.

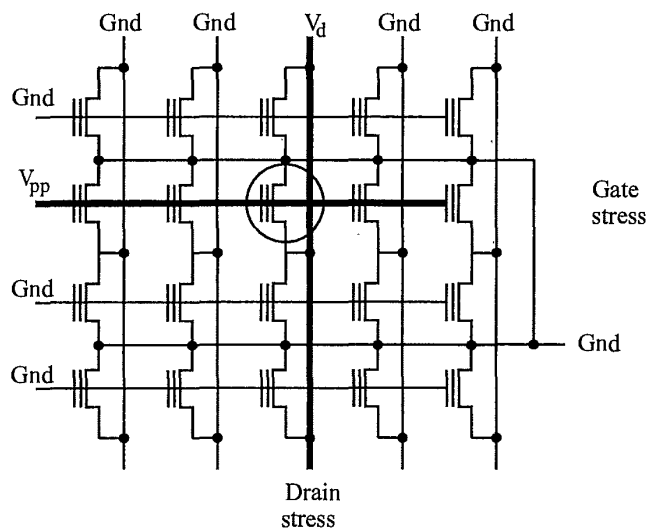


Figure 1.11 Program disturbs.

All cells sharing the word line at $V_{pp} \simeq 12V$ suffer for the so-called *gate disturb*, that can induce charge loss in programmed cells because of tunneling from the floating gate to the control one (DC-erasing), or charge gain in erased cells because of tunneling from the substrate to the floating gate (DC-programming). Similarly, cells sharing the bit line at $V_d \simeq 5 \div 7V$ can suffer for *drain disturbs*, caused by tunneling from the floating gate to the drain and by substrate hot holes injection.

These disturbs, affecting the overall reliability of the matrix, become more and more important with the number of programming cycles and must be minimized with a tailored choice of the programming voltages and an optimized circuit design.

Read disturbs. Read disturbs, on the contrary, only affect cells sharing the same bit line of a read cell (see Fig. 1.12).

The relatively low voltage applied to the word line ($\simeq 5V$) does not allow for electron tunneling from the substrate to the floating gate or from the floating gate to the control gate, while the high drain voltage may induce charge transfer to the floating gate, thus giving rise again to drain disturbs. Since the number of read operations is potentially infinite, the reliability of a Flash memory may be strongly affected by read disturbs, whose effects must be carefully limited. The best approach to tackle this problem is the reduction of the bit line potential:

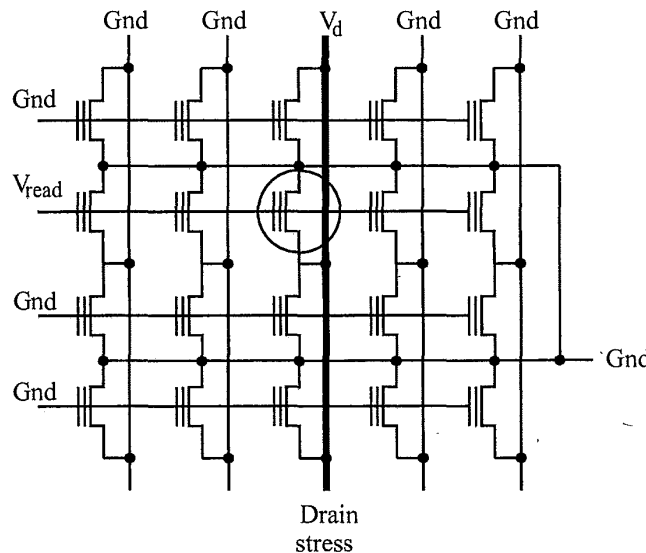


Figure 1.12 Read disturbs.

as it will be shown in Chapter 5, a maximum value of $\simeq 1V$ is considered, with a direct impact on the design of sense amplifiers.

Endurance and Data Retention. The capability of performing cycles of write operations without affecting the memory performance (*endurance*) and the possibility of maintaining unaltered the written data for long times with no power supply (*data retention*) represent two of the basic characteristics of a non-volatile memory. The main limitation to these parameters is due to the reliability of the thin gate oxide, which can break down when the number of cycles is strongly increased, or that can become leaky, even with low biases applied, as a consequence of the high fields applied during erasing. This last phenomenon, which does not allow for a further reduction of the oxide thickness (with the consequent benefit of lower writing voltages), affects both endurance and data retention. It must be noticed that Flash memories do not suffer for programming window closing due to charge trapping within the gate oxide: since the write operations always end with appropriate checks (program and erase verify), there is a guarantee that the threshold shift will remain unaltered, while the number of pulses required to perform the two operations will be adapted to deal with possible charge trapped within the oxide.

Embedded controller. The times required for read, program, and erase operation are quite different. For instance, while reading is very fast (in the order of $50 \div 100$ ns) and represents a major feature of a specific product, writing times are much longer, varying from tens of μs (typical programming time) up to few seconds (typical erasing time). An important consequence is that the board/system microcontroller cannot be fully dedicated to the Flash for several seconds just to accomplish one single operation, because it would introduce both an unacceptable slow-down of the overall performance and a considerable complication of the control software.

To solve this problem, all the logic circuitry necessary to handle slow operations is embedded entirely inside the Flash memory. Another advantage of this choice is the simplicity of the interface: all the timings required to perform the operation correctly (setting of different voltages, counting of the programming and erasing pulses, verify, soft programming,...) are transparent to the user. Therefore, to control the memory operation it is sufficient to provide the operation code, the data to be programmed (and the corresponding address) or the sector to be erased and check, when the microcontroller is not busy in other tasks, the current status of the memory.

The main consequence is that a Flash memory is not just a huge matrix of cells with some logics (decoders, output multiplexers,...), and some sense

amplifiers, but a complex system in which the impact of embedded logics is continuously increasing.

1.9 THE NAND FLASH MASS STORAGE CONCEPT

Besides the common NOR, parallel architecture, Flash memories can also be organized in NAND arrays, by connecting 16 cells in series between a bit line and the source line. The main advantage of this solution relies upon a reduced matrix area obtained thanks to a word line pitch scaling. This feature is possible because of: *i*) a decreased number of contacts, from one contact to the bit line for every cell to one contact for 16 cells, and *ii*) scaled source and drain junctions with respect to the standard ETOX cell, made possible by the physical mechanism used for cell writing.

The basic structure is depicted in Fig. 1.13, showing two different bit lines. In series with the 16 cells, 2 select transistors are present. Data are stored as charge within the floating gate: a positive threshold denotes a programmed cell, while a negative threshold indicates an erased cell. When reading a cell, its control gate is kept at 0 V, while all other cells in series are driven at a high voltage, thus acting as ON pass gates independently of their actual thresholds. The current, to be detected by a sense amplifier, flows through the series if and only if the selected transistor presents a negative threshold, thus behaving as a depleted transistor.

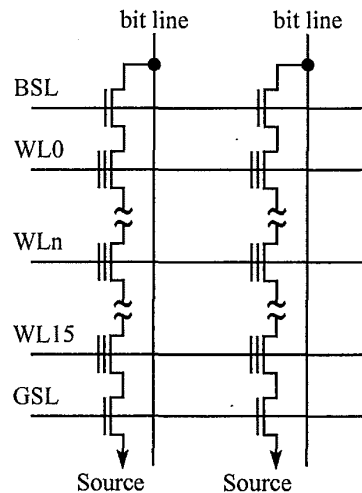


Figure 1.13 Basic structure of a NAND architecture.

To program a cell (see Fig. 1.14), its control gate is driven at a high voltage V_H (in the range of $15 \div 20$ V), while the corresponding bit line is biased at 0 V. Such a value is transferred to the channel through the cells in series, forced to act as pass gates, so that the high voltage drop between the floating gate and the channel allows for Fowler-Nordheim electron tunneling. The voltage of the unselected word lines ($V_m \approx 10$ V) derives from a trade-off between the need for good pass gates independently of their thresholds and that of preventing program disturbs, achieved by limiting the thin oxide field in unselected cells. The oxide field is also limited in unselected cells sharing the same word line of the selected cell to prevent undesired programming: this task is accomplished by raising the bit lines of these cells, and therefore their channels, to V_{CC} .

Micron Ex. 1008, p. 36

Channel Fowler-Nordheim tunneling, used both for programming and erasing, involves low coupling ratio and, therefore, very high voltages, thus complicating the design of charge pumps required to raise the single voltage supply value. Electron tunneling, however, does not require high currents, allowing for reduced power consumption with respect to cells programmed by hot electrons and for weakened requirements for charge pumps design.

1.10 EMBEDDED FLASH MEMORIES

The request for high-performance systems including both high-speed logics and large memories arrays has increased in last decade. Masked ROMs have been intensively embedded in digital systems to store data and control parameters to be used by the logic controller.

Following the evolution that brought to the development of memories that can be electrically programmed and erased, EEPROMs and, successively, Flash have replaced embedded ROMs, thanks to their great flexibility, allowing for "on-the-fly" data and parameters modification.

At the beginning, low-density EEPROMs were integrated in smart cards and chips for several applications (TV, video controllers, recorders, car stereo, hard disk drivers, cellular communications, automotive,...). Now, large Flash arrays are embedded for system personalization and reconfiguration.

Besides the flexibility introduced by reconfigurable memories and the obvious impact on board dimensions, there are also several advantages in terms of performance and reliability with respect to the standard connection of stand-alone chips, that push towards a strong development of systems integrating both logics and memories: *i*) faster access times because of a reduced capacitive connection between microprocessor and memories; *ii*) strongly reduced ground bouncing effects present in stand-alone systems and caused by parasitic inductance when outputs are switching; *iii*) increased number of memory outputs (since the reduction of ground bouncing effects allows for a large number of simultaneously switching outputs); *iv*) optimized bus, clock and control signals design; *v*) reduction of the power consumption, since output buffers driving interconnections can be removed; *vi*), reduction of ElectroMagnetic Interferences (EMI) at board level.

The integration of NV memories together with microprocessors, however, presents several difficulties in terms of technology and larger costs, making such a solution less appealing than theoretically expected.

In terms of technology, two different approaches can be followed, depending on the memory dimension. If the requested memory size is comparable to that of stand-alone chips, it is convenient to design the logic circuits on the basis of the memory technology: in such a way, the reproducibility and the reliability

of a standard process for Flash memories are guaranteed, with the drawback of a reduction of logic circuits performance that cannot reach those of *ad hoc* designs.

When high-end microprocessors are considered, where computation capability is the dominant figure of merit, and when the required memory size is not comparable with that of stand-alone devices, it may be convenient to design *ad hoc* cells and memory organization. In such a case, the performance in terms of speed, power consumption and area occupation of the logic blocks are not degraded, and the increment of memory occupation and reliability degradation with respect to a consolidated stand-alone memory design can be acceptable.

1.11 THE FUTURE OF FLASH MEMORIES

1.11.1 Evolution of Flash Memory Technology

There are several challenges which have to be confronted by Flash memories developers in order to fulfill future application requirements in terms of densities and performances; just to mention some of them, we can quote *i*) multilevel cell development, *ii*) cell scaling and scaling limitations, *iii*) low-voltage compatibility, *iv*) product diversification.

Increasing the number of possible states in a cell, thus obtaining a multilevel memory, is a viable strategy for increasing the density and reducing the cost per bit, due to the intrinsic analog nature of charge storage in the Flash memory. To store 2 bits per cell, four separated threshold voltages need to be correctly identified; tolerance for disturbs, data retention, charge sensing, accuracy of programming become more critical with the number of bits per cell. Multilevel cells are also more affected by manufacturing conditions, temperature, supply voltage, and aging effects. The issues related to multilevel Flash memories are discussed in Chapters 3 and 6. Depending on processing capability and on the specific application in system, multilevel Flash memories have been proposed as direct plug-in replacement of single level ones, without the need for error correction, or with some form of error correction at system/component level; it is envisaged that error correction will become mandatory for four bits per cell memories, which require 16 threshold levels.

Multilevel implementation is a critical task, but it is becoming attractive also due to the difficulties of the traditional way of increasing the memory density, i.e. technology scaling. Scaling a Flash cell is a completely different problem with respect to scaling a MOS transistor for logic applications. For instance, while CMOS technology scaling requires the reduction of operating voltages, the program/erase operations of the Flash are based on physical mechanisms whose major parameters do not scale (3.2 eV energy barrier for channel hot

electrons and 8-9 MV/cm oxide field for Fowler-Nordheim data alteration in 0.1-1s).

Other constraints come from manufacturability or reliability considerations. The threshold voltage distribution resulting from Fowler-Nordheim data alteration is very difficult to scale. Retention constraints limit the scalability of the tunnel and interpoly dielectrics. Due to the direct tunneling mechanism, the tunnel oxide can not be reduced below 6 nm to guarantee 10 years of charge retention; however, if one takes into account array disturb effects and trap-assisted electron tunneling caused by oxide ageing, or stress-induced leakage current [17], a more realistic minimum tunnel oxide thickness can be placed around 8 nm [18]. Current tunnel oxide thickness in production is larger than 9 nm, almost unchanged in the last generations of Flash processes [1]. The scalability limit of the interpoly dielectric (ONO) has been reported to be around 12 nm [18].

Cell punch-through and drain turn-on limit the scaling of the effective length of the cell, which is also affected by architectural considerations: a) channel hot-electron requires some minimum drain-gate overlap and abrupt junction to maximize the injection efficiency; b) Fowler-Nordheim tunneling carried out through the gate/diffusion region requires an overlap with high n^+ concentration below the gate; c) Fowler-Nordheim tunneling carried out through the channel region requires smaller gate/diffusion overlap. Finally, given the limit in channel length, the width scaling is limited by the minimum read current. A possible figure for the 0.1 μm lithography generation is a cell size of 0.1 μm^2 [1].

Low voltage is another critical issue. Low-voltage/low-power circuits currently operate at supply voltages of 3.3 V or less, comparable to the critical thresholds of injection over the Si/SiO₂ energy barrier. The internal voltages currently adopted by Flash memory cells are in the 10 V - 20 V range and, even including the use of both positive and negative voltages, there will be no dramatic scaling within the conventional program/erase scheme. For this reason, low-voltage operation has received increasing attention in the last few years [1, 19, 20, 21], and additional programming mechanisms have been proposed [22] and analyzed [23].

The versatility of the Flash technology is encouraging a differentiation of memory products: for EPROM-like storage of microcodes, which is still the widest application of Flash memories, high density and fast random byte programming and random reading are required; EEPROM emulation is important for specific embedded applications; small erase block size (512 byte sectors like in magnetic disks), relatively fast erase and high read/write throughput are essential for data storage memory cards in order to compete with magnetic media. Even if conventional stacked gate Flash is used in memory cards, optimization

of these performances requires dedicated solutions. Cost-per-bit becomes extremely important for all memory card applications, mostly oriented to the consumer market.

1.11.2 Non Volatile Memories Market Development

Flash memories represent an excellent trade-off between cost and functionality of EPROM and EEPROM. The capability of nonvolatile storage, coupled with an access time comparable to DRAMs has made Flash memories one of the fastest growing semiconductor product: Flash memories currently represent 9% of the total memory market, see Fig. 1.15, a share which is thought to increase up to 12% of the total memory market (50% of the nonvolatile memory market) by the year 2000. The Flash market size has been approximately three billion US\$ in 1997. Fig. 1.16 shows the worldwide memory market forecast for the 1993-2001 time period.

Several effects have boosted this performance: not only the flexibility of Flash memories, but also the diffusion of widely accepted industry specifications, and the availability of increased chip densities, suitable for a variety of applications. The list of electronic products which now include Flash memories is almost infinite, the three major markets being related to personal computers, telecom and wireless apparatus, automotive electronics. The various applications require different Flash solutions in terms of architecture (NOR vs. NAND or other implementations), multilevel implementation, density, access times. The needs of the various applications are schematically indicated in Fig. 1.17, which shows the projected Flash market for the year 2002 divided into the main application segments. Many applications use Flash memories for code storage. In the computer environment, the most established applications are devoted to the storage and update of the PC BIOS and of the HDD operating system, which usually require low densities (2 - 4 Mb).

The computer application segment of the Flash memory market will further expand with the diffusion of Personal Digital Assistants (PDAs). PDAs have no hard disk drive storage and store the whole of their operating system in a non-volatile memory, with an increased need for embedded Flash arrays and memory cards. Moreover, almost all peripherals, HDD, CDROMs, DVDs and most add-on boards like video and sound cards require 1-4 Mb of upgradeable nonvolatile memory.

On computer network equipments, the frequent software upgrades required for network equipment can be easily carried out remotely if the code is stored in nonvolatile Flash memories; this requires densities which increase from a few Mb for modems and interface cards up to 128 Mb and more, as in the case of network routers.

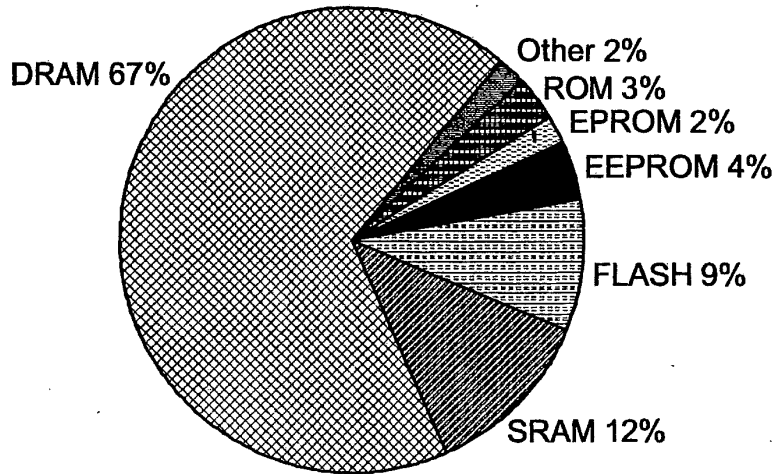


Figure 1.15 1997 worldwide MOS memory market. The total market is 33 B\$

	FLASH	EEPROM	EPROM	ROM	SRAM	DRAM	Other	All MOS Mem.
1993	0.725	0.492	1.417	1.731	3.485	14.411	0.266	22.528
1994	0.992	0.558	1.444	2.174	4.067	22.864	0.318	32.414
1995	1.846	0.680	1.337	2.135	6.132	41.755	0.534	54.418
1996	2.828	1.062	1.271	1.368	4.709	25.843	0.577	37.659
1997	2.994	1.364	0.713	1.085	3.988	22.060	0.642	32.836
1998	3.417	1.725	0.634	0.911	4.840	26.643	0.835	39.006
1999	4.163	2.147	0.642	0.990	6.046	40.088	0.956	55.033
2000	5.192	2.601	0.659	1.029	8.479	68.657	0.102	87.636
2001	6.280	3.150	0.704	1.025	11.309	62.591	0.107	86.129

Figure 1.16 1993-2001 memory market (figures in B\$).

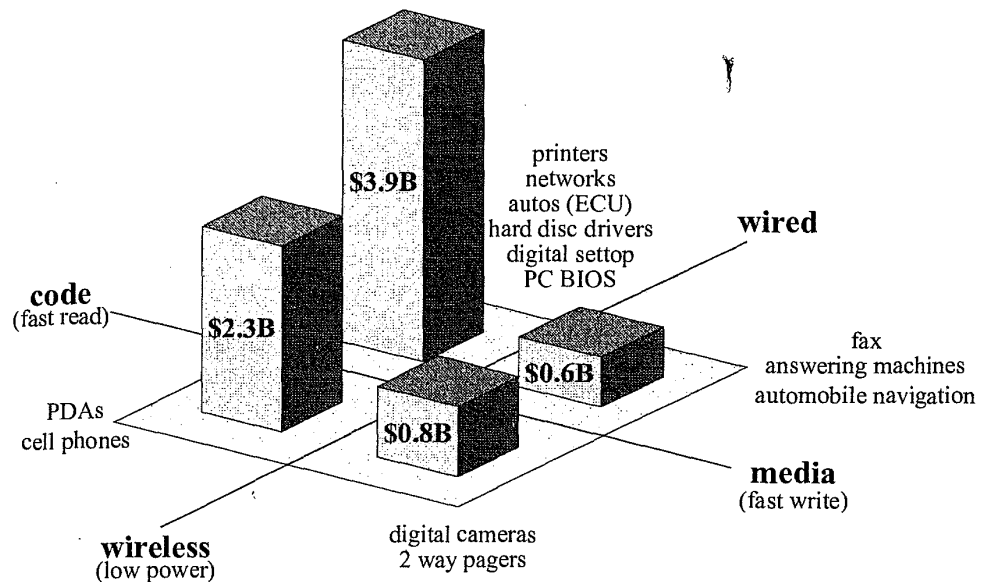


Figure 1.17 Flash market shares divided by application, 2002 forecast [24].

Cellular phones are a key Flash market driver, accounting for about the 30% of the total revenue and driving technical requirements such as low voltage and low energy consumption. Within less than 5 years, the Flash memory content in digital cellular phones will increase rapidly from 4 to 16 Mbit. Other important code and data storage applications are in the automotive electronics field: automotive industry has been one of the first Flash memory adopters. Vital functions such as Engine Control Units (ECUs) or automatic gear boxes, as well as Global Positioning Systems (GPS) and other driving assistance systems currently use Flash with densities up to 8 Mb.

The large majority of the above mentioned applications, and other emerging consumer applications, such as Set Top Boxes for digital TV, adopt NOR Flash memories. Mass storage applications, such as voice recorders and digital cameras, which require very high densities and low cost per bit, can take advantage of multilevel memories or NAND architectures.

The same applies to hard disk replacements, such as those required for data recording. A significant example is represented by portable medical monitors,

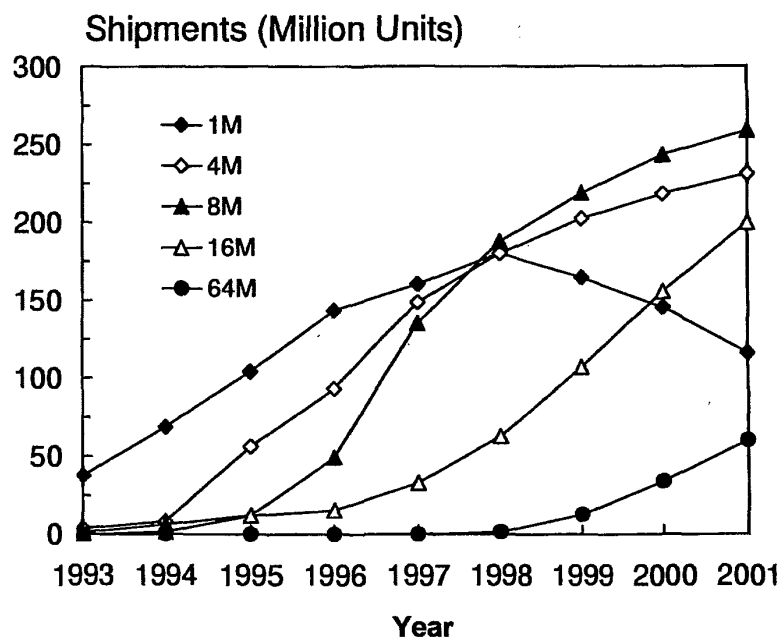


Figure 1.18 Growth of Flash memory shipments, 1993-2001 forecast

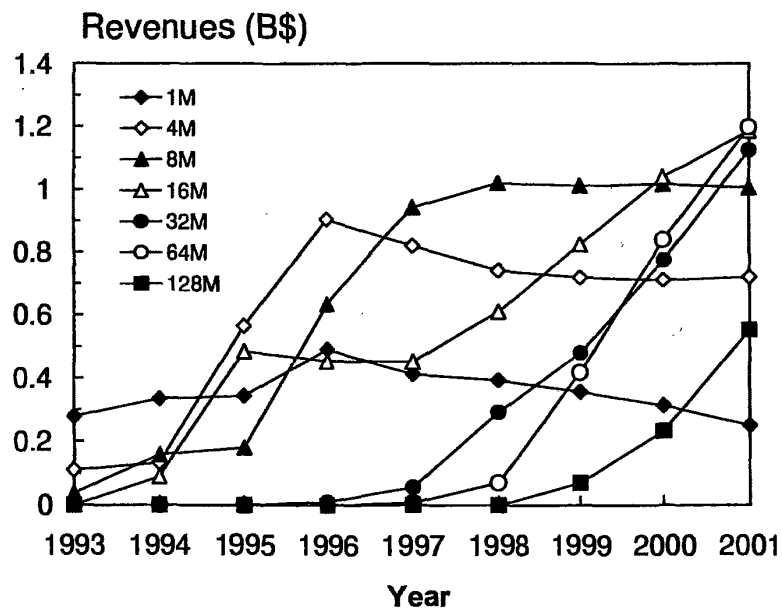


Figure 1.19 Flash market revenues divided by density, 1993-2001 forecast

which used to store data in bulky cassette tape recorders, substituted by Flash memory cards, with a substantial improvement in terms of weight, size and robustness. Flash memory cards are challenged by the evolution of small hard disk drives, which are still ten to one hundred times cheaper than Flash memories in terms of cost per bit. For those applications which require low power consumption and long battery life, however, Flash memory cards have no competitors [1].

The trend of Flash market can be also described in terms of single vs. dual supply voltage, or in terms of high vs low voltage. Since low power and low voltage are mandatory for portable applications, there is a growing demand for 3.3 V (or lower) single voltage devices, which can operate over a wide range of voltages without compromising access speed and functionality and can be switched into very low power standby mode.

The evolution of the Flash memory market has been accompanied by a continuous decrease in the Flash average selling price (8% decrease in 1996!) which has at the same time encouraged new applications of stand alone memory chips, and promoted the development of embedded Flash applications as a way to recover their added value.

The time evolution of the worldwide Flash memory market forecast for the 1993-2001 period is depicted in Figs. 1.18 and 1.19, which puts into evidence the continuous shift to higher densities, with the 8 Mb - 16 Mb dominating in these years, and the 64 Mb rapidly growing, both in terms of number of unit shipments and of revenues.

The improvement in multilevel storage solutions, and the diversification of the Flash memory products will make this area of the semiconductor market one of the most dynamic, both in terms of process and technology challenges and in terms of market evolution.

References

- [1] Lai S. (1998) "Flash memories: where we were and where we are going". *IEEE IEDM Tech. Dig.*, p. 971.
- [2] IEEE Standards Department (1998) "IEEE P1005 draft standard for definitions, symbols, and characteristics of floating gate memory arrays" (approved 1998). IEEE, 445 Hoes Lane, Piscataway, NJ (USA).
- [3] Kahng D. and Sze S.M. (1967) "A floating gate and its application to memory devices". *Bell Syst. Tech. J.*, **46**, p. 1288.
- [4] Frohman-Bentchkowsky D. (1971) "A fully decoded 2048-bit electrically programmable MOS-ROM". *IEEE ISSCC Tech. Dig.*, p. 80.

- [5] Wegener H.A.R. *et al.* (1967) "The variable threshold transistor, a newly electrically alterable nondestructive read-only storage device". *IEDM Tech. Dig.*.
- [6] Libsch F.R. and White M.H. (1998) "SONOS nonvolatile semiconductor memories". J.E. Brewer and W.D. Brown (Eds.) *Nonvolatile Semiconductor Memory Technology. A comprehensive guide to understanding and using NVSM devices*, IEEE Press, Chapter 5, p. 309.
- [7] Harari E., Schmitz L., Troutman B. and Wang S. (1978) "A 256 bit non-volatile static RAM". *IEEE ISSCC Tech. Dig.*, p. 108.
- [8] Johnson W.S., Perlegos G., Renninger A., Kuhn G. and Ranganath T. (1980) "A 16 Kbit electrically erasable nonvolatile memory". *ISSCC Tech. Dig.*, p. 152.
- [9] Hirano H., Honda T., Moriwaki N., Nakakuma T., Inoue A., Nakane G., Chaya S. and Sumi T. (1997) "2 V/100 ns 1T/1C nonvolatile ferroelectric memory architecture with bitline-driven read scheme and nonrelaxation reference cell". *IEEE Journal of Solid State Circuits*, **32**, p. 649.
- [10] Kunishima I. and Takashima D. (1998) "High-density chain ferroelectric random access memory (chain FRAM)". *IEEE Journal of Solid State Circuits*, **33**, p. 787.
- [11] Koike H. *et al.* (1996) "A 60 ns 1 Mb nonvolatile ferroelectric memory with a nondriven cell plate line write/read scheme". *IEEE Journal of Solid State Circuits*, **31**, p. 1625.
- [12] Pavan P., Bez R., Olivo P., and Zanoni E. (1997) "Flash memory cells - An overview". *Proc. of the IEEE*, **85**, p. 1248.
- [13] Guterman D.C., Rimawi I.H., Chiu T.L., Halvorson R.D. and McElroy D.J. (1979) "An electrically alterable nonvolatile memory cell using a floating-gate structure". *IEEE Trans. on Electron Devices*, **26**, p. 576.
- [14] Masuoka F., Asano M., Iwahashi H., Komuro T. and Tanaka S. (1984) "A new Flash E²PROM cell using triple polysilicon technology". *IEDM Tech. Dig.*, p. 464.
- [15] Verma G. and Mielke N. (1988) "Reliability performance of ETOX based Flash memories". *Proc. IRPS*, p. 158.
- [16] Kynett V.N., Baker A., Fandrich M., Hoekstra G., Jungroth O., Kreifels J. and Wells S. (1988) "An in-system reprogrammable 256 KCMOS Flash memory", *ISSCC Tech. Dig.*, p. 132.

- [17] Olivo P., Nguyen T. and Riccò B. (1988) "High-field-induced degradation in ultra thin SiO₂ films". *IEEE Trans. on Electron Devices*, **35**, p. 2259.
- [18] Camerlenghi E., Crisenza G., Annunziata R. and Cappelletti P. (1996) "Non volatile memories: issues, challenges and trends for the 2000's scenario". *Proc. of ESSDERC*, p. 121.
- [19] Rodjy N. (1992) "0.85 μ m double metal CMOS technology for 5 V Flash EPROM memories with sector erase". *Nonvolatile Semiconductor Memory Symposium*.
- [20] Bergemont A., Haggag H., Anderson L., Shacham E. and Woltsenholme G. (1993) "NOR virtual ground (NVG) - A new scaling concept for very high density FLASH EEPROM and its implementation in a 0.5 μ m process". *IEDM Tech. Dig.*, p. 15.
- [21] Bergemont A., Chi M.H. and Haggag H. (1995) "Low voltage NVG: a new high performance 3 V/5 V Flash technology for portable computing and telecommunication applications". *Proc. ESSDERC*, p. 543.
- [22] Bude J.D., Frommer A., Pinto M.R. and Weber G.R. (1995) "EEPROM/Flash sub 3.0 V drain-source bias hot-carrier writing". *IEDM Tech. Dig.*, p. 989.
- [23] Fischer B., Ghetti A., Selmi L., Bez R. and Sangiorgi E. (1997) "Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages". *IEEE Trans. on Electron Devices*, **44**, p. 288.
- [24] Rau W. (1998) "Quality & Testing: How Much Can We Afford? A Customer Perspective". *IEEE Non-Volatile Semiconductor Memory Workshop*, p. 7.

2 THE INDUSTRY STANDARD FLASH MEMORY CELL

Paolo Pavan¹, Roberto Bez²

¹ Dipartimento di Scienze dell'Ingegneria and INFM
Università di Modena e Reggio Emilia
Via Campi 213/B, 41100 Modena, Italy
paolo.pavan@unimo.it

² STMicroelectronics, Central R&D
Via Olivetti 2, 20041 Agrate Brianza (Milano), Italy
roberto.bez@st.com

Flash, a-ah
Savior of the Universe

—Queen, *Flash Gordon Soundtrack*, 1980

Abstract: This chapter gives a thorough overview of the Industry Standard Flash Memory Cell. More than 85% of today Flash memories rely on this concept. We will describe the basic structure of the floating gate device, and its operating conditions. We will highlight the main differences in the technology and process with respect to a standard CMOS process. Finally, a brief introduction on some of the more important yield and reliability issues will be given.

2.1 INTRODUCTION

A Flash memory is a Non Volatile Memory (NVM) whose "unit cells" are fabricated in CMOS technology and programmed and erased electrically.

In 1971, Frohman-Bentchkowsky developed a floating polysilicon gate transistor [1, 2], in which hot electrons were injected in the floating gate and removed by either Ultra-Violet (UV) internal photoemission or by Fowler-Nordheim tunneling. This is the "unit cell" of EPROM (Electrically Programmable Read Only Memory), which, consisting of a single transistor, can be very densely integrated. EPROM memories are electrically programmed and erased by UV exposure for 20-30 mins. In the late 1970s, there have been many efforts to develop an electrically erasable EPROM, which resulted in EEPROMs (Electrically Erasable Programmable ROMs). EEPROMs use hot electron tunneling for program and Fowler-Nordheim tunneling for erase. The EEPROM cell consists of two transistors and a tunnel oxide, thus it is two or three times the size of an EPROM. Successively, the combination of hot carrier programming and tunnel erase was rediscovered to achieve a single transistor EEPROM, called Flash EEPROM. The first cell based on this concept has been presented in 1979 [3]; the first commercial product, a 256 K memory chip, has been presented by Toshiba in 1984 [4]. The market did not take off until this technology was proven to be reliable and manufacturable [5].

The first Flash prototypes needed an external supply voltage for programming, and external management of the erasing procedure; they featured only bulk erase capability, and their endurance was very poor, less than 10,000 cycles. As an advantage versus EPROMs they just offered electrical erase capability. Modern Flash memories have an embedded microcontroller to manage the erase operation, they offer sector erase capability and single power supply.

The growing demand of high density nonvolatile memories for portable computing and telecommunications market has encouraged serious interest in Flash memory with capability of multi-level storage [6, 7, 8] and low voltage operation [9, 10, 11]. Multi-level storage implies the capability of storing two or more bits in a single cell.

Fig. 2.1 shows the cross section of a Flash cell. This cell structure has been presented for the first time by INTEL in 1988 and named ETOX™ (EPROM Tunnel OXide) [12].

In this chapter we will refer to what is called "Industry Standard Flash Memory Cell", which is a NOR, common ground, stacked gate double polysilicon device which is programmed using channel hot-electron injection, and erased via Fowler-Nordheim tunneling at the source junction. Many new cells and concepts have been presented in the recent past (for an overview see [13]):

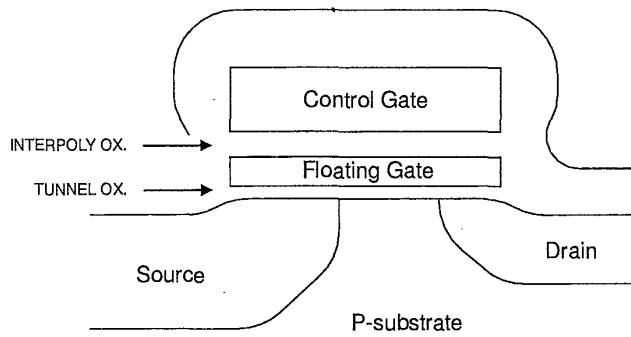


Figure 2.1 Schematic cross section of a Flash cell.

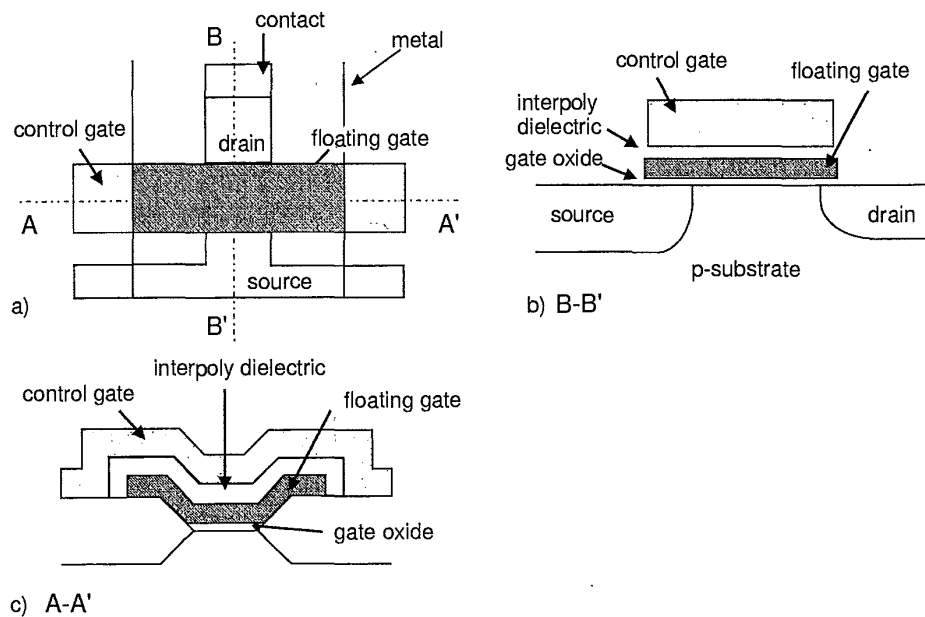


Figure 2.2 a) Layout of a T-shaped double polysilicon stacked gate Flash cell. The schematic cross sections along perpendicular directions are also shown in b) and c).

NAND, AND, DINOR, HiCR, etc. Today, almost 85% of Flash devices is based on NOR structures.

The cross section of an Industry Standard cell is shown again in Fig. 2.2b. It is composed by a n-channel MOSFET transistor with an additional isolated

gate which is floating in the dielectric material and called "floating gate". The floating gate is separated from the channel by a thin oxide layer (around 10 nm) which is called "gate oxide", and from the overlying gate (called "control gate") by a triple layer dielectric (oxide-nitride-oxide, ONO), called "interpoly dielectric", whose thickness is equivalent to a 20 nm silicon dioxide layer. Source junction is smoother and deeper than drain junction, to achieve higher source-substrate breakdown voltages. In fact, as we will see, the erase operation requires high voltages to be applied to the source.

Fig. 2.2a shows a typical layout of an Industry Standard Flash unit cell. Arrays of these cells will compose the memory bench. As it can be seen, there is only one contact to the drain junction (indeed half contact, since it is shared between two opposite cells), while the source junction is parallel to the control gate. The active area looks like a "T" (upside down in this picture): for this reason the cell is called T-shaped cell. The floating gate extends over the field oxide, to have a complete coverage of the channel region and to increase the gate coupling ratios. The part of the floating gate overlapping the field oxide, clearly sketched in the cross section in Fig. 2.2c, reminds of wings, due to the typical beak shape of the field oxide. The wing extension is one of the critical geometric parameters of the cell.

Reading an Industry Standard cell means to decode the information which is stored, while programming and erasing mean to change the stored information and have the cell in the "0" state and in the "1" state, respectively. Flash cell read, program, and erase bias configurations are summarized in Tab. 2.1.

Table 2.1 Source, Control Gate, and Drain biases during operations of a typical Flash cell. DV: dual voltage power supply; SV: single voltage power supply. Typical reference values can be: $V_{cc} = 5\text{ V}$, $V_{pp} = 12\text{ V}$, $V_{dd} = 5 - 7\text{ V}$, $V_{read} = 1\text{ V}$, $V_{NEG} = -8\text{ V}$.

	SOURCE	CONTROL GATE	DRAIN
READ	GND	V_{cc}	V_{read}
PROGRAM	GND	V_{pp}	V_{dd}
ERASE (DV)	V_{pp}	GND	FLOAT
ERASE (SV)	V_{cc}	V_{NEG}	FLOAT

In Tab. 2.1 we have reported two different erase configurations. DV (dual voltage) refers to devices which are operated with two different supply voltages, one for reading (3 V or 5 V), and one for programming (usually 12 V). SV (single voltage) refers to devices which are functioning with a single supply voltage (which can be 3 V or 5 V). In this case, the voltage values needed to program and erase the cell are internally generated with charge pumps.

Gate oxide is very thin. Therefore, if a high voltage is applied at the source when the control gate is grounded, a high electric field exists in the oxide enabling tunneling effects from the floating gate to the source. This bias condition is dangerously close to the breakdown of the source-substrate junction. Therefore, the source diffusion is processed differently from the drain one, which does not undergo to such bias conditions. To do so, a specific mask is used in the technological process to discriminate source and drain implants. The cell is not symmetrical, but this is the only difference with the standard EPROM process; this fact has been a great advantage, since all the accumulated experience in process development could be used to manufacture these devices. The same EPROM T-shaped cell layout has been used in Flash memory. In the late 80s-early 90s, EPROM technology was the leading technology for NVMs, and Flash Memories were derived from it, resulting in a 8 to 10% larger cell than an EPROM cell on the same technology, due to the different source diffusion. Today, Flash technology is the leading technology, and EPROMs or One Time Programmable memories (OTPs) are fabricated using the Flash process, thus using the same area per cell.

Oxide/nitride/oxide (ONO) interpoly dielectric thickness heavily influences program/erase speed and the magnitude of read current for a Industry Standard Flash cell [14]. Moreover, its good quality is essential for reliability issues of Flash cells, like low defect density and long mean time to failure, together with charge retention capability.

Most of the Flash devices today are based on the Industry Standard type of structure. A 16 Mbit, $0.35\text{ }\mu\text{m}$, 55 ns, 5 V-only Flash has been recently proposed, based on this type of cell [15]. Many new cell structures are continuously proposed. Most of the devices now announced are single power supply and all the program and erase algorithms are built-in the memory chip. Many are the differences among them, some of which can be listed as follows: cell size, program mechanism, process complexity, maximum voltage, application, array efficiency, redundancy efficiency. Higher levels of integration are required not only in Integrated Circuits, but also in systems. Non Volatile Memories are being integrated on the same chip with other circuits to reduce the number of ICs in a system, to decrease access time, power dissipation, area, and so on. Integration of logic and memory on the same chip is non trivial: the constraints imposed by the different fabrication processes lead to different solution according to the specific embedded application. Flash memories are now becoming ASP (Application-Specific-Products), requiring Application-Specific-Integrated-Circuit (ASIC) designs, and with many new cell structures and product applications new design techniques and architecture are evolving.

In this chapter, after the analysis of the basic structure of the Industry Standard Flash Memory cell, we will analyze its operating modes. After the

description of the main differences between a standard CMOS process and a Flash process, we will briefly address some reliability issues which are related mostly to the single cell structure, leaving the more detailed analysis of reliability issues to Chapter 7. Scaling issues will be also briefly addressed.

A detailed analysis of the physical mechanisms involved in Flash operations is carried out in Chapter 4.

2.2 BASIC STRUCTURE

The basic concepts and the functionality of an Industry Standard device are easily understood if it is possible to determine the Floating Gate (FG) potential. The schematic cross section of a generic FG device is shown in Fig. 2.3, where energy band diagrams are also drawn; the upper gate is the control gate and the lower gate, completely isolated within the gate dielectric, is the floating gate. The floating gate acts as a potential well, see again Fig. 2.3. If a charge is forced into the well, it cannot move from there without applying any external force: the floating gate stores charge.

The simple model shown in Fig. 2.4 helps in understanding the electrical behavior of a FG device. C_{FC} , C_S , C_D , and C_B are the capacitances between FG and control gate, source, drain and substrate regions, respectively.

Consider the case when no charge is stored in the FG, i.e. $\bar{Q} = 0$:

$$\begin{aligned} \bar{Q} = 0 &= C_{FC}(V_{FG} - V_{CG}) + C_S(V_{FG} - V_S) + \\ &+ C_D(V_{FG} - V_D) + C_B(V_{FG} - V_B) \end{aligned} \quad (2.1)$$

where V_{FG} is the potential on the floating gate, V_{CG} is the potential on the control gate, V_S , V_D , V_B are potentials on source, drain and bulk, respectively. If we name $C_T = C_{FC} + C_D + C_S + C_B$ the total capacitance of the floating gate, and we define $\alpha_J = C_J/C_T$ the coupling coefficient relative to the electrode J , where J can be one among G , D , S , and B , the potential on the floating gate due to capacitive coupling is given by:

$$V_{FG} = \alpha_G V_{GS} + \alpha_D V_{DS} + \alpha_S V_S + \alpha_B V_B \quad (2.2)$$

As an example, in a $0.4\mu\text{m}$ CMOS technology, $\alpha_G \approx 0.6$, $\alpha_D \approx 0.1$, $\alpha_S \approx 0.15$. It should be pointed out that (2.2) shows that the floating gate potential does not depend on the control gate voltage only, but also on source, drain and bulk potentials. If source and bulk are both grounded, Eq. (2.2) can be rearranged as:

$$V_{FG} = \alpha_G \left(V_{GS} + \frac{\alpha_D}{\alpha_G} \cdot V_{DS} \right) = \alpha_G (V_{GS} + f \cdot V_{DS}) \quad (2.3)$$

where:

$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_{FC}} \quad (2.4)$$

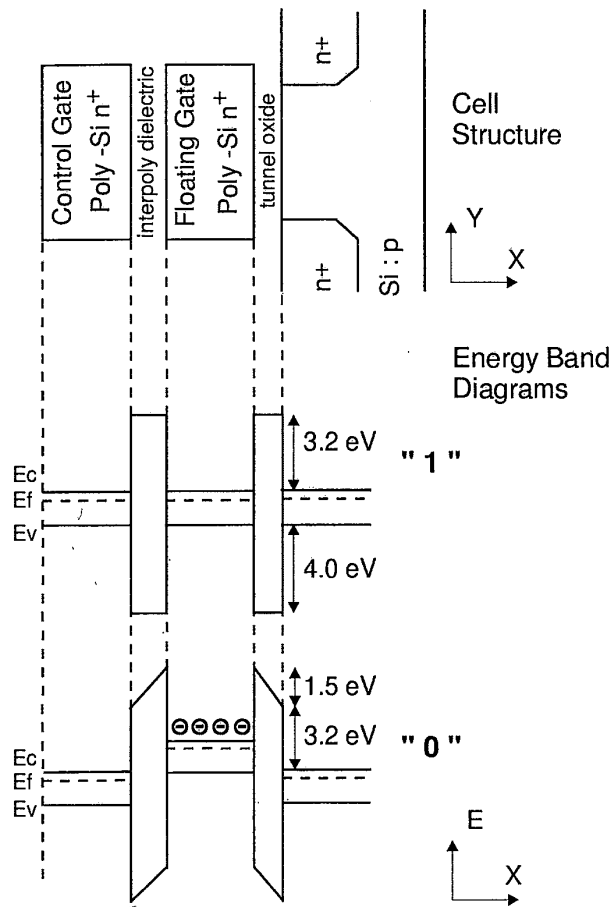


Figure 2.3 Cell structure and energy band diagrams of a Floating Gate transistor. The floating gate stores charge in state "0".

Device equations for the FG MOS transistor can be obtained from the conventional MOS transistor equations by replacing MOS gate voltage, V_{GS} , with FG voltage, V_{FG} , and transforming the device parameters, such as threshold voltage, V_T , and conductivity factor, β , to values measured with respect to the control gate. If we define, for $V_{DS} = 0$:

$$\begin{aligned} V_T^{FG} &= V_T(\text{floating} - \text{gate}) \\ &= \alpha_G V_T(\text{control} - \text{gate}) = \alpha_G V_T^{CG} \end{aligned} \quad (2.5)$$

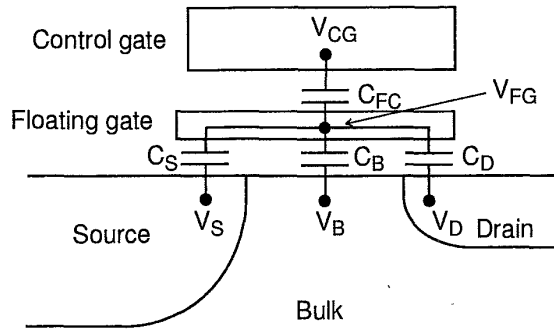


Figure 2.4 Schematic cross section of a Floating Gate transistor. The model using the capacitances between the Floating Gate and the other electrodes is described.

and

$$\begin{aligned}\beta^{\text{FG}} &= \beta(\text{floating} - \text{gate}) \\ &= \frac{1}{\alpha_G} \beta(\text{control} - \text{gate}) = \frac{1}{\alpha_G} \beta^{\text{CG}}\end{aligned}\quad (2.6)$$

it is possible to compare the current-voltage (I-V) equations of a conventional and a floating gate MOS transistor in the triode region (TR) and in the saturation region (SR) [16].

For a conventional MOS Transistor, in TR, when $|V_{DS}| < |V_{GS} - V_T|$

$$I_{DS} = \beta \left[(V_{GS} - V_T) V_{DS} - \frac{1}{2} V_{DS}^2 \right] \quad (2.7)$$

In SR, when $|V_{DS}| \geq |V_{GS} - V_T|$

$$I_{DS} = \frac{\beta}{2} (V_{GS} - V_T)^2 \quad (2.8)$$

For a Floating gate MOS transistor, in TR, when $|V_{DS}| < \alpha_G |V_{GS} + fV_{DS} - V_T|$,

$$I_{DS} = \beta \left[(V_{GS} - V_T) V_{DS} - \left(f - \frac{1}{2\alpha_G} \right) V_{DS}^2 \right] \quad (2.9)$$

In SR, when $|V_{DS}| \geq \alpha_G |V_{GS} + fV_{DS} - V_T|$

$$I_{DS} = \frac{\beta}{2} \alpha_G (V_{GS} + fV_{DS} - V_T)^2 \quad (2.10)$$

β and V_T of (2.9) and (2.10) are measured with respect to the control gate rather than with respect to the FG of the stacked gate structure, and then they are to be read as $\beta(\text{control} - \text{gate}) = \beta^{\text{CG}}$ and $V_T(\text{control} - \text{gate}) = V_T^{\text{CG}}$.

Several effects can be observed from these equations, many of them due to the capacitive coupling between the drain and the FG which modifies the I-V characteristics of floating gate MOS transistors with respect to conventional MOS transistors [16]:

- 1) The floating-gate transistor can go into depletion-mode operation and can conduct current even when $|V_{GS}| < |V_T|$. This is because the channel can be turned on by the drain voltage through the $f \cdot V_{DS}$ term in (2.9). This effect is usually referred to as "drain turn-on".
- 2) The saturation region for the conventional MOS transistor is where I_{DS} is essentially independent of the drain voltage. This is no longer true for the floating gate transistor in which the drain current will continue to rise as the drain voltage increases and saturation will not occur (see Fig. 2.5a).

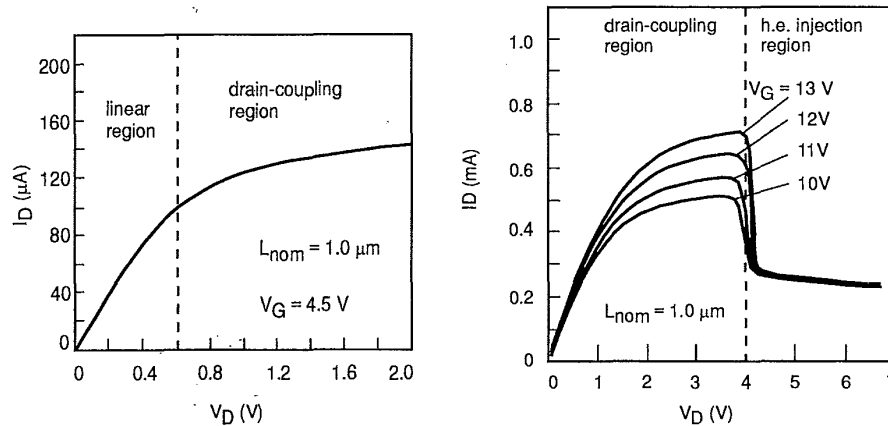


Figure 2.5 I-V characteristics of a floating gate transistor. In the right figure the effect of programming is also shown.

- 3) The boundary between triode and saturation regions for the floating gate transistor is expressed by the equation:

$$|V_{DS}| = \alpha_G |V_{GS} + f \cdot V_{DS} - V_T| \quad (2.11)$$

compared to the conditions valid for the conventional transistor $|V_{DS}| = |V_{GS} - V_T|$.

- 4) The transconductance in SR is given by:

$$\begin{aligned} g_m &= \frac{\partial I_{DS}}{\partial V_{GS} (V_{DS}=\text{constant})} \\ &= \alpha_G \beta (V_{GS} + f V_{DS} - V_T) \end{aligned} \quad (2.12)$$

g_m increases with V_{DS} in the floating-gate transistor in contrast to the conventional transistor where g_m is relatively independent of the drain voltage in the saturation region.

- 5) The capacitive coupling ratio, f , depends on C_D , and C_{FC} only ($f = \alpha_D/\alpha_G = C_D/C_{FC}$) and its value can be verified by:

$$f = -\frac{\partial V_{GS}}{\partial V_{DS} (I_{DS}=\text{constant})} \quad (2.13)$$

in saturation region.

Many techniques have been presented to simply extract the capacitive coupling ratios from DC measurements [17, 18, 19]. Most widely used methods are [20, 21]: i) linear threshold voltage technique; ii) subthreshold slope method; iii) transconductance technique. These methods require the measurement of the electrical parameter in both a memory cell and in a "dummy cell", i.e. a device identical to the memory cell, but with floating and control gates connected. By comparing the results, the coupling coefficient can be determined. Other methods have been proposed to extract coupling coefficients directly from the memory cell, without using a "dummy" one, but they need a more complex extraction procedure [22, 23, 24].

2.3 OPERATING CONDITIONS

Tab. 2.1 summarizes the bias conditions which identify the three operations of an Industry Standard cell, namely: Read, Program, Erase. In the following, these operations are described in detail, indicating the time ranges needed to perform them, the electrical characteristic variations which are induced, and highlighting the main unwanted effects that can arise by simply applying the configuration biases.

2.3.1 Read

To better understand the reading operation, we will refer again to Fig. 2.4. Let's consider the case when charge is stored in the FG, i.e. $\bar{Q} \neq 0$. All the hypotheses made in Section 2.2 hold true, and the following modifications need to be included.

Eqs. (2.3), (2.5), and (2.9) become, respectively:

$$V_{FG} = \alpha_G V_{GS} + \alpha_D V_{DS} + \frac{\bar{Q}}{C_T} \quad (2.14)$$

$$V_T^{CG} = \frac{1}{\alpha_G} V_T^{FG} - \frac{\bar{Q}}{C_T \alpha_G} = \frac{1}{\alpha_G} V_T^{FG} - \frac{\bar{Q}}{C_{FC}} \quad (2.15)$$

$$I_{DS} = \beta \left[\left(V_{GS} - V_T - \left(1 - \frac{1}{\alpha_G} \right) \frac{\bar{Q}}{C_T} \right) V_{DS} + \left(f - \frac{1}{2\alpha_G} \right) V_{DS}^2 \right] \quad (2.16)$$

Eq. (2.15) shows the V_T dependence on $\bar{Q}$. In particular, the threshold voltage shift ΔV_T is derived as:

$$\Delta V_T = V_T - V_{T0} = -\frac{\bar{Q}}{C_{FC}} \quad (2.17)$$

where V_{T0} is the threshold voltage when $\bar{Q} = 0$.

Eq. (2.16) shows that the role of injected charge is to shift the I-V curves of the cell. If the reading biases are fixed (usually $V_{GS} \simeq 5\text{ V}$, $V_{DS} \simeq 1\text{ V}$), the presence of charge greatly impacts the current level used to sense the cell state. Fig. 2.6 [25] shows two curves: curve A represents the "1" state, and curve B the same cell in the "0" state obtained with a 3 V threshold shift. In the defined reading condition, I_D ("1") is approximately $100\text{ }\mu\text{A}$ and I_D ("0") $\simeq 0$, respectively.

2.3.2 Program

Hot electron injection is used to move charge in the floating gate, thus changing the threshold voltage of the floating gate transistor. Programming occurs applying simultaneously pulses to the control gate and to the drain, when the source is grounded, see Fig. 2.7. This operation can be performed in an array by selectively applying the pulse to the Word Line (WL) which connects the control gates, and biasing the Bit Line (BL) which connects the drains. Hot electrons are injected in the floating gate and they change the floating gate potential, which becomes more negative. Therefore, injection tends to saturate, and the threshold voltage of a floating gate transistor follows the same trend, see also Fig. 2.5. The dependence of the threshold voltage shift on programming time is described by the following equation:

$$\Delta V_T(t) = V_T(t) - V_{T0} = -\frac{\bar{Q}(t)}{C_{FC}} \quad (2.18)$$

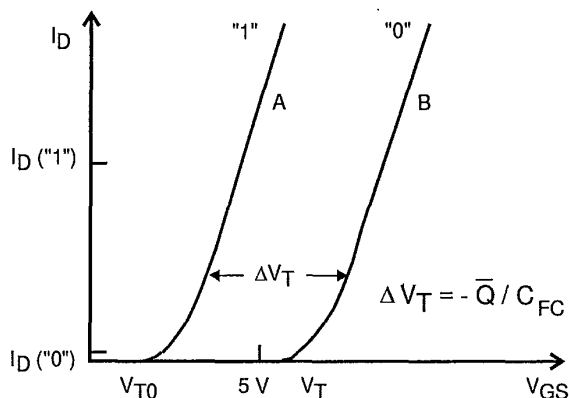


Figure 2.6 I-V curves of a floating gate device when there is no charge stored in the FG ("1" - curve A) and when a negative charge $\bar{Q}$ is stored in the FG ("0" - curve B) [25].

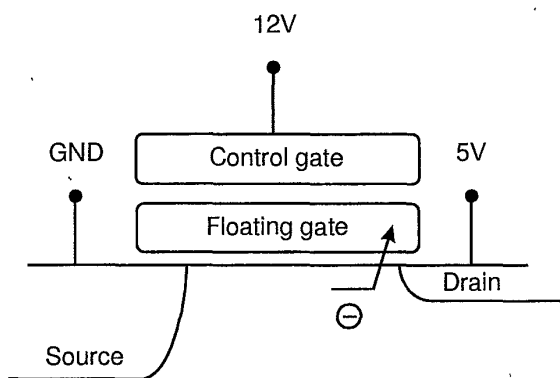


Figure 2.7 Schematic of the cell bias during program.

which relates the change in threshold voltage to the change in the charge stored in the floating gate. In its turn, the FG charge is related to the gate current by the equation:

$$\bar{Q}(t) = \int_0^t J_G(\tau) d\tau \quad (2.19)$$

Fig. 2.8 [26] shows the programming curves of Flash cells with different channel length. The programming curve is defined as the threshold voltage shift as a

function of programming time. To have a threshold voltage shift around 3, 3.5 V typical pulse width are in the range 1–10 μ s, see Fig. 2.8 with reference to the curve with $L_{\text{eff}} = 0.6 \mu\text{m}$.

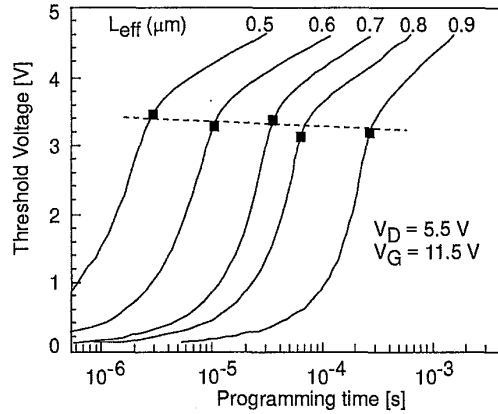


Figure 2.8 Programming curves of Flash cells with different channel lengths; t_{ox} is 12 nm. Black squares indicate ΔV_T^* [26].

A rapid change in cell V_T occurs initially and then, as V_{FG} drops below V_D , V_T tends to saturate. The same behavior can be observed in Fig. 2.9 [26], where the programming curves of a cell ($L_{\text{eff}} = 0.7 \mu\text{m}$, $V_G = 11.5 \text{ V}$) are plotted with V_D as a parameter.

It has to be reminded that the problem that worries every nonvolatile cell designer is to ensure a fixed threshold voltage shift of the cell, ΔV_T , in the shortest programming time, t_p , i.e. with the fastest programming speed $\Delta V_T/t_p$, with the drain voltage and the channel length of the cell as constraints. These two parameters are the ones with a direct influence on the lateral electric field, thus influencing hot-electron generation.

The programming characteristics of a cell can be better understood in terms of another parameter, called intrinsic threshold, ΔV_T^* , which is defined as the threshold voltage shift when $V_D = V_{\text{FG}}$, and can be expressed through Eq. (2.20) as a function of programming voltages:

$$\Delta V_T^* = V_{\text{CG}} - \frac{1 - \alpha_D}{\alpha_G} V_D \quad (2.20)$$

ΔV_T^* characterizes every cell and limits the programming speed performance. ΔV_T^* divides every programming curve in two parts, see Figs. 2.8 and 2.9.

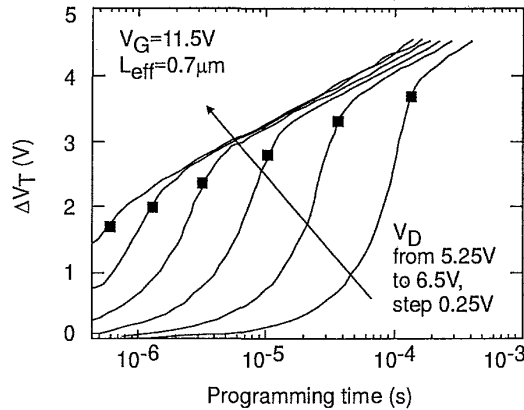


Figure 2.9 Programming curves of Flash cells for different drain voltages; t_{ox} is 12 nm. Black squares indicate ΔV_T^* [26].

Fig. 2.10 [26] shows the programming time to obtain $\Delta V_T = 3$ V as a function of V_D , and of ΔV_T^* . From the curves, one can see that there is a critical V_D above which the programming speed does not increase. This value does not depend on the channel length of the cell. In the same figure, the condition $\Delta V_T = \Delta V_T^* = 3$ V separates the t_p/V_D plane in two parts: for $\Delta V_T^* > \Delta V_T$ (left side of the figure), t_p decreases with exponential trend with V_D ; for $\Delta V_T^* < \Delta V_T$ (right side of the figure), t_p tends quickly to a nearly constant value, which depends only on channel length, once V_{CG} is set. From Eq. (2.20), a higher V_{CG} (or ΔV_T) shifts the critical V_D towards a higher (or lower) value. Moreover, ΔV_T^* is the upper limit for the achievable threshold voltage shift once programming time is fixed. Programming time becomes very large, i.e. programming speed very slow, when ΔV_T overtakes ΔV_T^* .

From these observations, the influence on the programming speed of different parameters, intrinsically related to the manufacturability or functionality of a Flash cell, can be analyzed.

Both channel length reduction and V_D increase can be used to increase the lateral electric field. Therefore, length variations result in programming time variations: the longer the cell, the longer the programming time. This relation depends also on V_D : the higher V_D is, the lower is the intrinsic threshold voltage shift, the shorter is the programming time. But once the intrinsic threshold voltage shift is reached, further increases in V_D do not improve the performance. Channel length shortening has to be correlated to an appropriate V_D choice.

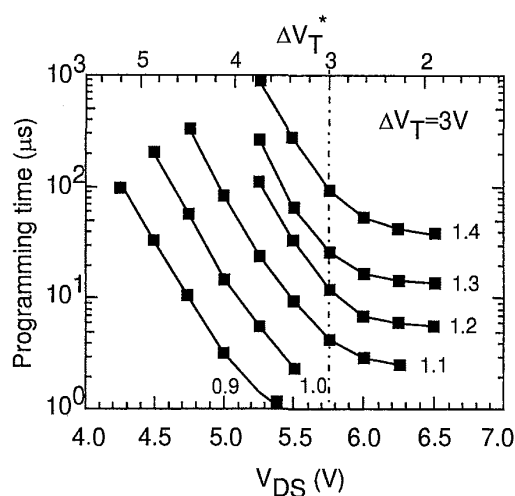


Figure 2.10 The programming time to obtain $\Delta V_T = 3V$ is plotted vs. V_{DS} and ΔV_T^* for $V_{CG} = 11.5V$. The number near each curve marks the nominal channel length (μm) [26].

For a fixed channel length, a V_D increase above the critical value does not improve performances.

The geometrical or process parameters of the cell do not have a direct impact on the programming speed, but only on the coupling ratios α_G and α_D , consequently on ΔV_T^* . Fig. 2.11 [26] shows that different ΔV_T^* are obtained for different wings (therefore different α_G and α_D), and that the programming speed is almost the same for cells with different α_G, α_D until $\Delta V_T \leq \Delta V_T^*$.

Other parameters which can influence the programming characteristics can be: i) source series resistance, its increase results in a lower effective V_{DS} , thus reducing the programming speed; ii) V_T dispersion after erase, which implies a longer t_p to get to the same final V_T . Temperature does also have an influence on programming speed (Fig. 2.12): a higher temperature reduces the number of hot electrons available for injection into the FG, hence retarding the programming characteristics [26].

In recent SV products, programming is achieved using charge pumps, therefore a key parameter is the “programming efficiency”, defined as the ratio between gate and drain programming currents, γ_{pr} :

$$\gamma_{pr} = \frac{I_G}{I_D} \quad (2.21)$$

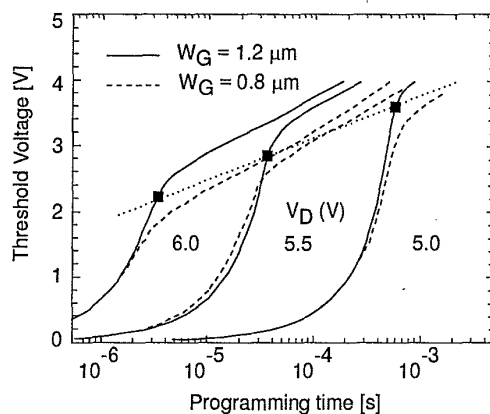


Figure 2.11 Programming curves of Flash cells with different coupling ratios and at different V_D ; $L_{\text{eff}} = 0.7 \mu\text{m}$, $t_{\text{ox}} = 12 \text{ nm}$ [26].

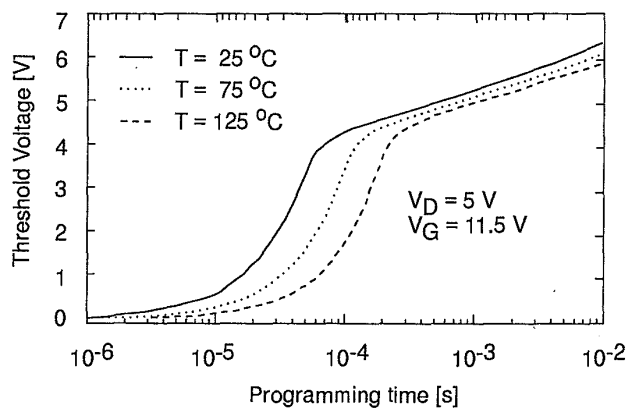


Figure 2.12 Programming curve of a Flash cell at different temperatures; $L_{\text{eff}} = 0.7 \mu\text{m}$, $t_{\text{ox}} = 12 \text{ nm}$.

To allow the most efficient programming, γ_{pr} has to be maximized. Being γ_{pr} intrinsically related to hot carrier generation mechanism, its value is $\approx 10^{-6}$, therefore a high drain current, $I_D \approx 1 \text{ mA}$, is required to have the gate current, $I_G \approx 1 \text{ nA}$, necessary to program the cell. γ_{pr} can be optimized either

introducing new technology solutions, or changing the electrical stimuli to be applied to the cell.

It can be interesting to notice that a floating gate cell can be used to measure very small currents. In fact, from Eq. (2.17) and from the definition of gate current as:

$$I_G = \frac{\Delta Q_{FG}}{\Delta t} = C_{FG} \frac{\Delta V_T}{\Delta t} \quad (2.22)$$

we can see that the derivative of the programming curve allows a precise evaluation of the gate current. Since ΔV_T can be around 1 mV–1 V, and Δt around 1 μ s–1 ms, and C_{FG} is around 1 fF, we can measure gate currents in the range 0.01 fA to 1 nA. An example of measurement of the derivative of ΔV_T is shown in Fig. 2.13 [22], where ΔV_T^* is also shown.

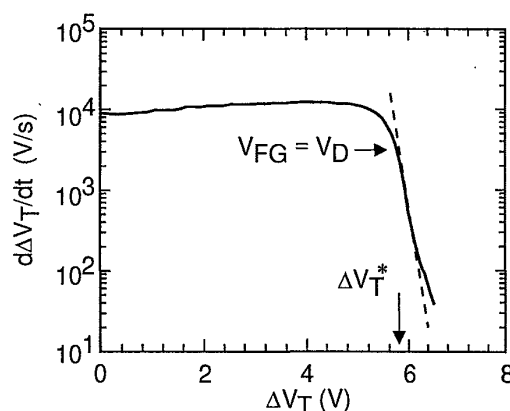


Figure 2.13 Time derivative of the programming curve, $d\Delta V_T/dt$ vs. ΔV_T . The point ΔV_T^* corresponds to the physical condition $V_{FG} = V_D$ [22].

2.3.3 Erase

Electrical erase is achieved by applying a high electric field through the tunnel oxide, see Fig. 2.14. The high electric field gives rise to a gate current I_G due to FN tunneling of charge from the floating gate to the source:

$$I_G = A_{FN} E_{ox}^2 \exp\left(-\frac{B_{FN}}{E_{ox}}\right) \quad (2.23)$$

where A_{FN} , B_{FN} are constant, and E_{ox} is the electric field in the tunnel oxide.

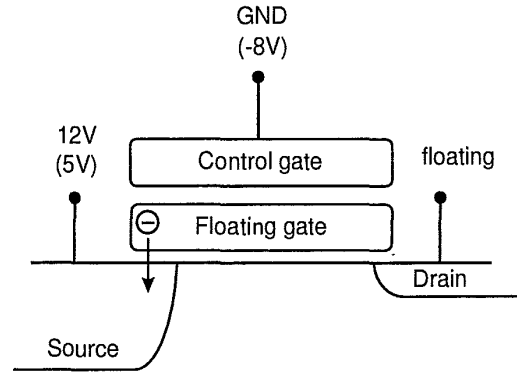


Figure 2.14 Schematic cross section of a Flash cell during erase. Voltages on electrodes refer to Dual Voltage operation mode, and in brackets to Single Voltage.

Continuity of the displacement vector at the surface gives:

$$\epsilon_{ox} E_{ox} = \epsilon_{Si} E_{Si} \quad (2.24)$$

where ϵ_{ox} , ϵ_{Si} are oxide and silicon permittivity, respectively, and E_{Si} is the electric field at the silicon surface underneath the gate. In particular, considering their values ($\epsilon_{ox} = 3.9$, $\epsilon_{Si} = 11.7$), $E_{ox} = 3E_{Si}$. The high electric field in the silicon is responsible for the source/substrate current I_S , due to band-to-band tunneling:

$$I_S = A_{BB} E_{Si}^2 \exp\left(-\frac{B_{BB}}{E_{Si}}\right) \quad (2.25)$$

where A_{BB} , B_{BB} are constants. These two currents are shown in Fig. 2.15 [27].

Band-to-band tunneling (BBT) occurs when band-bending is higher than the energy gap of the semiconductor, and the surface electric field is higher than 1 MV/cm. In this condition, tunneling of electrons from the valence band to the conduction band becomes significant, and holes are left in the valence band. Electrons are collected at the source terminal, while holes at the substrate contact, thus generating the leakage current. This substrate current depends only on the vertical electric field in the oxide, i.e. on the voltage drop between source and gate. Lateral electric field does not allow the inversion layer to be generated at the n^+ -Si/SiO₂ interface and leads the space charge region in deep depletion, sweeping all the free carriers. When source voltage is high enough, impact ionization becomes significant and contributes to the leakage

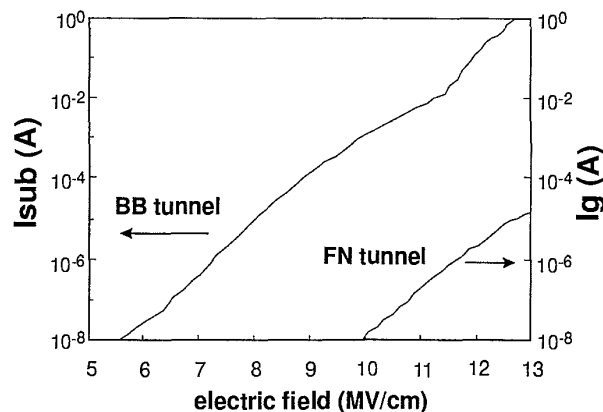


Figure 2.15 Relation between substrate and gate currents as a function of electric field measured on a large perimeter gated diode, with the p-n junction having the same profile of the source junction of a Flash cell [27].

current, thus starting the breakdown mechanism. The minimum voltage to start BBT decreases on decreasing the oxide thickness, and this is one of the major scaling limits. Generated holes can gain enough energy to be injected in the oxide where they are trapped at the Si/SiO₂ interface, thus becoming a concern from the reliability point of view.

To have a junction which can sustain the high applied voltages without breaking down, the source junction needs to be carefully designed.

Source breakdown is indeed one of the major limiting factors to erase time reduction, since the higher the voltage applied to the source is, the shorter the erasing time is. A solution to the problem is achieved by optimizing the source junction profile to a more gradual one, in order to reduce the electric field at the junction, and consequently the substrate current of some order of magnitude. A pictorial view of the source junction is depicted in Fig. 2.16. The source final doping profile is smoother and deeper; the profile of the high dose implant is also shown, and the details of the process are described in the following section.

In a conventional dual-voltage (DV) Flash, where besides V_{cc} (3 V or 5 V) a high voltage V_{pp} (about 12 V) is available, erasing is obtained by applying a high positive voltage to the source region (V_S), while the WL terminal (control gate of the memory cell) is grounded. In a single-voltage (SV) Flash, the lack of the high voltage V_{pp} implies the on-chip generation of a negative voltage by means of charge-pumps. In fact, in this case, the necessary voltage drop between the

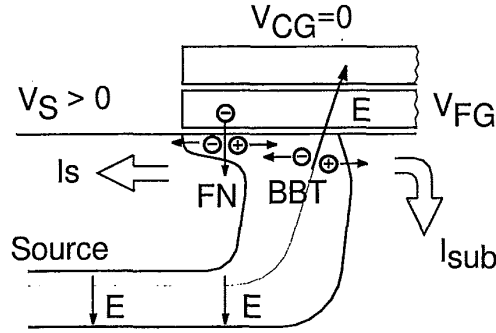


Figure 2.16 Pictorial description of Band-to-Band Tunneling (BBT) and Fowler-Nordheim (FN) currents ($V_S > 0$, and $V_{CG} = 0$ V) occurring in the n^+ source junction of a Flash cell [30].

source and the control gate is obtained applying V_{cc} to the source and a negative voltage V_{GN} to the control gate [15, 28, 29]. No matter how the voltage drop is obtained, in both cases the high electric field in the oxide between FG and source gives rise to a gate current due to FN tunneling, and simultaneously the high electric field in the silicon is responsible for the source/substrate current due to BBT tunneling.

To evaluate the voltages to be applied to the electrodes, we can recall that, neglecting the voltage drop in the FG and in the silicon:

$$E_{ox} \approx \frac{|V_{FG} - V_S|}{T_{ox}} \quad (2.26)$$

where T_{ox} is the gate oxide thickness, and express V_{FG} as a function of ΔV_T from Eqs. (2.14) (when $V_S \neq 0$ V), and (2.17). In a DV, $V_G = 0$ V, therefore:

$$|V_{FG} - V_S| = |\alpha_S V_S - \alpha_G \Delta V_T - V_S| = |(\alpha_S - 1) V_S - \alpha_G \Delta V_T| \quad (2.27)$$

while, in a SV Flash, $V_G < 0$ V, therefore:

$$\begin{aligned} |V_{FG} - V_S| &= |\alpha_S V_S + \alpha_G V_{CG} - \alpha_G \Delta V_T - V_S| = \\ &= |(\alpha_S - 1) V_S + \alpha_G (V_{CG} - \Delta V_T)| \end{aligned} \quad (2.28)$$

If a $\Delta V_T \approx 3$ V has to be achieved in a cell with $T_{ox} = 10$ nm, $\alpha_G \approx 0.6$, $\alpha_S \approx 0.15$, and $E_{ox} = 10$ MV/cm, we obtain V_S around 10 V in DV Flash, while in SV Flash V_{CG} is around -8 V, if V_S is 5 V. The DV erase operation requires a high voltage pulse to be applied to the source (common to all the cells in the

array/block) when control gates (WL) are grounded and drains (BL) floating. Before applying the erase pulse, all the cells in the array/block are programmed, to start with all the thresholds approximately at the same value. After that, an erase pulse having controlled width is applied. Similarly to programming curve, we can define the erasing curve as the curve that shows the threshold voltage shift as a function of erasing time. Erasing time is defined as the time needed to go to state "1", starting from state "0". Erasing time depends on the electric field in the tunnel oxide, that is on the voltage drop between source and floating gate, see Eq. (2.26), but it is independent of the starting threshold voltage value of the programmed cell, see Fig. 2.17. From the same figure, it is clear that cells with the same oxide thickness but different initial values of threshold voltage will reach the same threshold voltage at the end of the erase operation. The threshold shift depends on source voltage (Fig. 2.18) and, as a

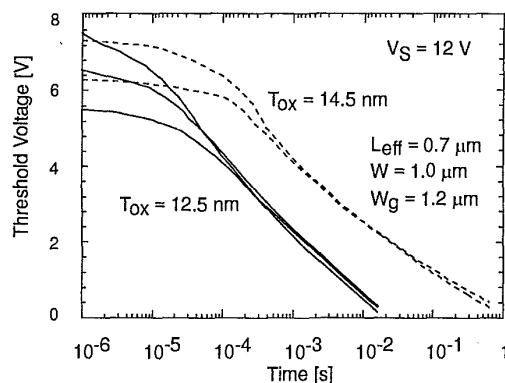


Figure 2.17 Erase curves of two Flash cells having different oxide thicknesses, and same L_{eff} .

rule of thumb, a one order of magnitude increase in erasing time occurs for each volt reduction in source voltage [30]. Typical erasing times are in the range 100ms–1 s.

Also in this case, like for program, it is possible to define the efficiency of the erase operation. We will evaluate the ratio γ_{er} between the current which is used to erase (gate current, I_G), and the unwanted current generated by the erase voltages (substrate current, I_{SUB}):

$$\gamma_{\text{er}} = \frac{I_G}{I_{\text{SUB}}} \quad (2.29)$$

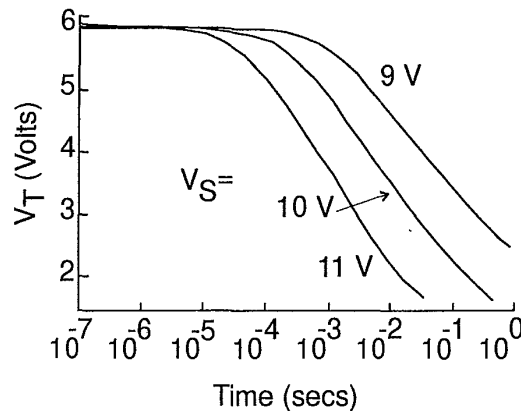


Figure 2.18 Erase curves of a cell with $T_{\text{ox}} = 12 \text{ nm}$, when different source voltages are applied [30].

Fig. 2.19 shows this ratio plotted on increasing the electric field applied at the tunnel oxide in the source region. Again, optimization of the erase operation means maximization of this ratio. Moreover, one can observe that this ratio is almost independent of the voltage applied to the control gate, which can be positive, grounded or negative.

2.4 TECHNOLOGY AND PROCESS

The evolution of the silicon planar technology for Flash memory follows the general trend of semiconductor industry. Up to now, Moore law has hold true also for Flash technology. Fig. 2.20 shows Flash evolution in terms of bit density versus production year: from the first product in the late eighties, a 256 K memory, to the up-to-date 16 Mb.

Flash process is basically a full CMOS process in which the building blocks to get an “ad hoc” floating gate device are incorporated. Hence, the basic process steps have been developed following the standard CMOS technology, but there is still a short delay with respect to microprocessor and Dynamic Random Access Memory (DRAM) technologies. In fact, the first product has been obtained with a $1.2 \mu\text{m}$ technology, while today production is peaked on the $0.35 \mu\text{m}$ technology (Fig. 2.20), comparing with the $0.25 \mu\text{m}$ DRAM in production. The technology road-map for Flash is given to follow the general semiconductor road-map: Flash device will be produced at the beginning of the new thousands in $0.25 \mu\text{m}$ and $0.18 \mu\text{m}$ technologies.

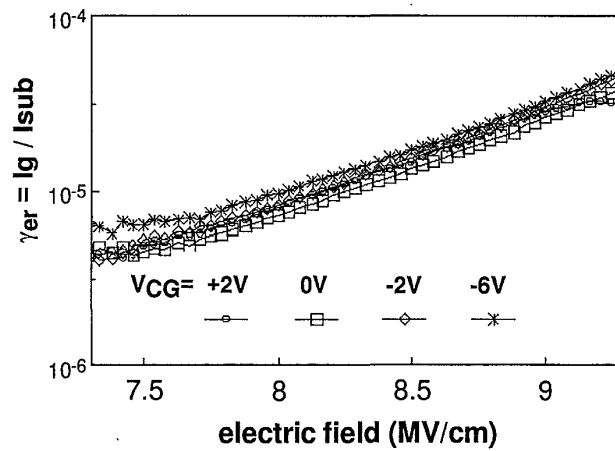


Figure 2.19 Erase efficiency as a function of electric field, when different gate voltages are applied.

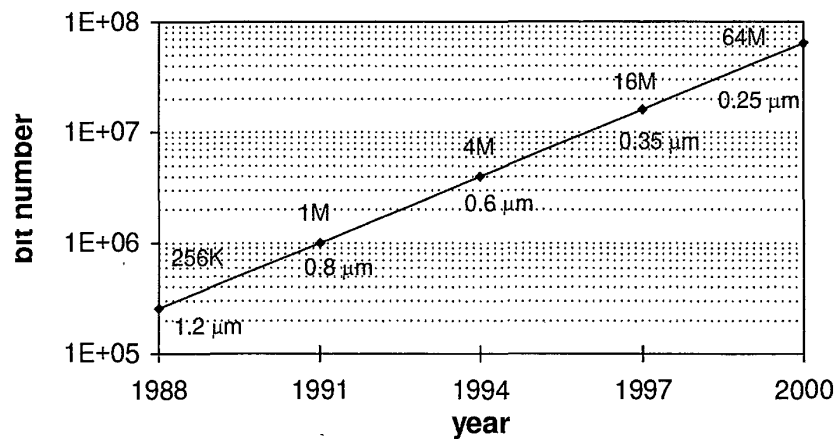


Figure 2.20 Evolution of Flash products in production.

The basic process flow for actual Flash device (16 Mb in $0.35\ \mu\text{m}$) is composed of the main blocks listed in Tab. 2.2, which can be grouped in Front-End and Back-End sets.

Table 2.2 Main blocks of a basic process flow for Flash devices.

Front-End	1) isolation
	2) well and channel doping
	3) cell structure definition
	4) transistor definition
Back-End	5) interlevel dielectric
	6) interconnections
	7) passivation

Flash technology is one of the most difficult to be mastered, since it requires a very accurate process optimization and a severe process control [31]. In fact, besides the usual requirements of a standard CMOS device in terms of timing (access time), low operating voltage and temperature range, a high voltage (externally forced or internally generated) is needed for writing operations, and this involves both Fowler-Nordheim tunneling and hot carriers; moreover, program/erase cycling must endure without degrading cell performance and data retention. Hence, in developing a process step, besides the issues related to a standard CMOS device, constraints strictly related to Flash cell must be taken into account.

2.4.1 Isolation

Two are the requirements for the field oxide isolation in Flash memory applications. The first one is to prevent parasitic leakage current between neighboring devices. The parasitic transistor must sustain the high voltage (greater than 10 V, in absolute value) necessary to program the cell. In particular this is mandatory where the circuitry is very dense, like the row decoding area, where the active-to-active area space become a constraint. The spacing reduction and, more in general the increased aspect ratio of the isolation geometry, enhances field oxide thinning effects. These effects have to be minimized as much as possible in order to achieve a more uniform and geometry-independent field oxidation.

The second is that the active area pitch in the cell array must be as small as possible, to increase the array efficiency in terms of cell size, obviously while

maintaining good quality of tunnel oxide and good crystallographic integrity with low junction leakage current.

To this aim different LOCOS (L) schemes, proposed for $0.5\ \mu\text{m}$ to $0.25\ \mu\text{m}$ technologies, can be used in Flash technology: modified L [32, 33], Poly Buffer L [34, 35], recessed L [36] (Figs. 2.21 and 2.22), recessed Poly Buffer L [37], NCLAD [38], BOX.

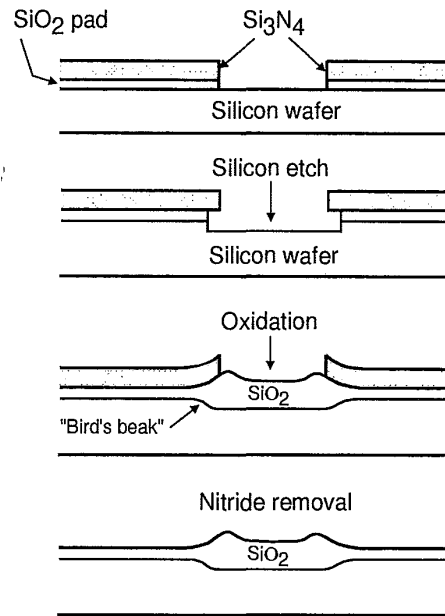


Figure 2.21 Process steps to obtain recessed LOCOS isolation.

2.4.2 Well and Channel Doping

The introduction of high energy implantation [39] has allowed the optimization of the well formation. Fig. 2.23 shows the different process flows for isolation and well fabrication in a retrograde well (Fig. 2.23a) obtained by means of high energy implantation, and a diffused well (Fig. 2.23b). Tab. 2.3 reports typical dose and energy ranges of the implantation steps for well formations, considering junction depth (retrograde well), channel stopper (isolation), anti punch-through and channel doping (threshold voltage shift) implants. The use of the high energy implantation and then of the retrograde well will play a special role in the Flash development since it allows the formation of triple well

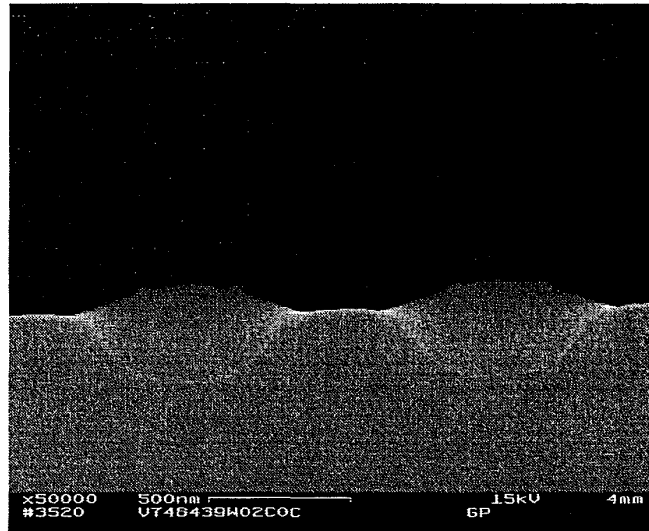


Figure 2.22 SEM micrograph of recessed LOCOS isolated devices.

structures (Figs. 2.24 and 2.25) [40, 41], by means of a very high energy n-buried channel (Tab. 2.3). The triple well structure is of particular interest for Flash applications, since it can be used in the decoding circuitry for managing the negative voltage (Fig. 2.26) for the erase operation in single voltage devices [42].

Table 2.3 Dose and Energy ranges of implantation steps in 16 Mbit Flash Memory.

IMPLANTATION STEP	DOSE (Atoms/cm ³)	ENERGY (KeV)	TILT ANGLE (deg)
Retrograde n/p well (n/p)	5×10^{12} – 1×10^{14}	500–1500	0–7
Isolation	1×10^{12} – 1×10^{13}	250–750	0–7
Anti Punch-Through	1×10^{12} – 1×10^{13}	100–200	0–7
Threshold Voltage Shift	5×10^{11} – 5×10^{12}	20–100	0–7
N-buried for triple well	5×10^{12} – 5×10^{13}	2000–3000	0–7

Another application of triple well process is strictly connected with the specific Flash cell architecture and write/erase schemes. For example, the whole array can be put in a triple well to allow a different programming or erasing

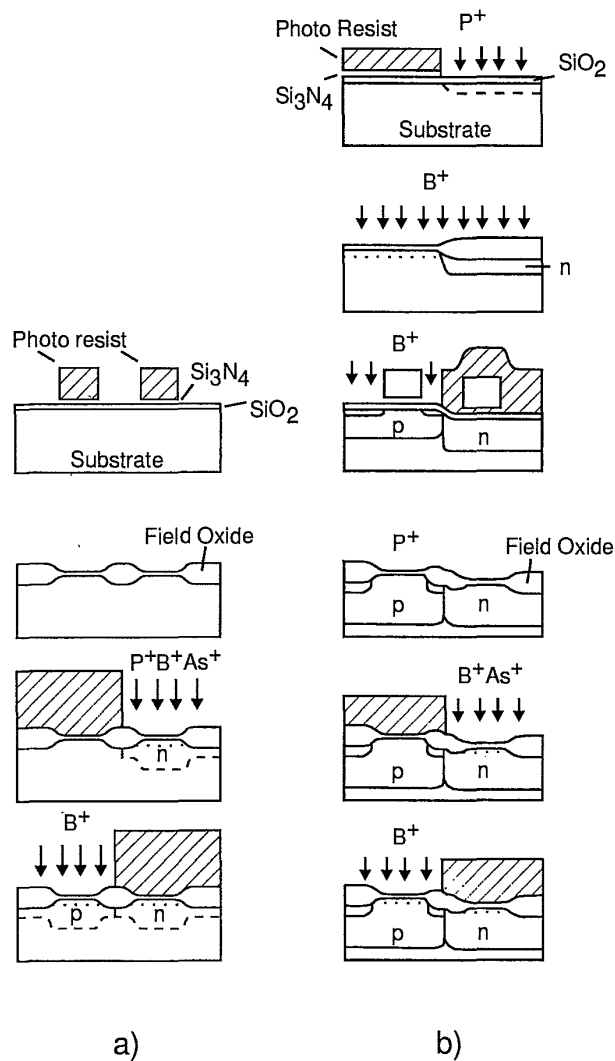


Figure 2.23 Isolation techniques: a) retrograde well, and b) conventional process flows.

operation, as proposed for the conventional NOR type cell [43]. Moreover, the triple-well has become a basic structure for alternative cell approaches, as NAND [44] or DINOR cells [45].

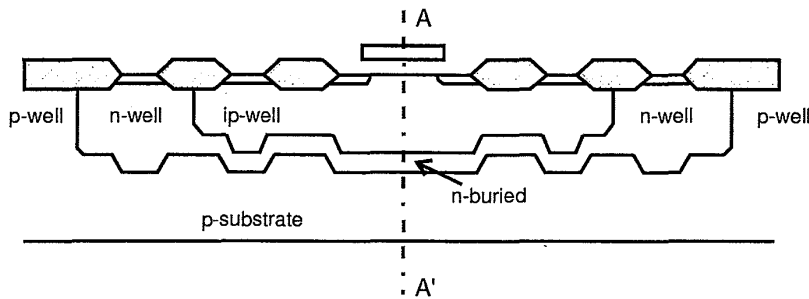


Figure 2.24 Triple well structure [41].

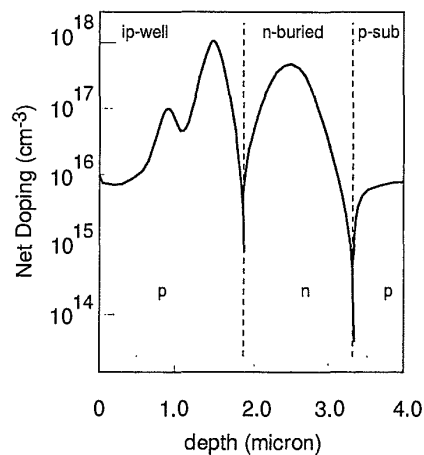


Figure 2.25 Doping profile of a triple well structure along A-A' section of the previous figure [40].

2.4.3 Cell Structure Definition

This module obviously differentiates a Flash process from a standard CMOS process. The goal is clear: obtaining millions of bits which can be heavily stressed, both in terms of temperature (between -40°C and 125°C) and in terms of electric fields and hot carriers, and nevertheless able to retain charge for a long time. In developing a technology for Flash, all functionality and reliability constraints must be taken into account. The proper Flash cell module is basically composed by the following steps.

1. *Tunnel oxide growth.* The quality of the thin gate oxide, thickness around 12–8 nm, as a function of the different memory device generation, is the

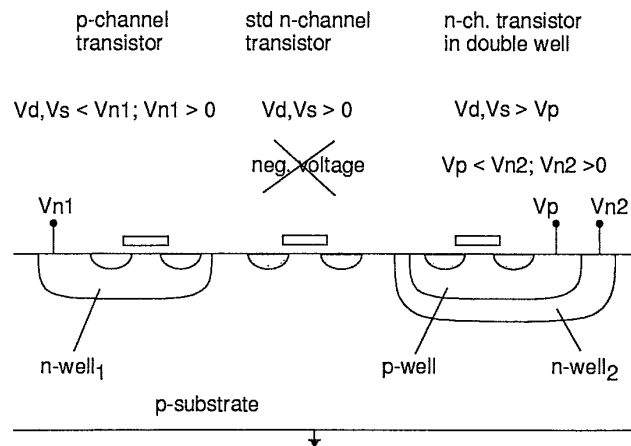


Figure 2.26 Schematic cross section of transistors in a triple well process [40].

most important factor contributing to a robust Flash technology, both in terms of yield and reliability.

The first requirements of the oxidation process are the uniformity and repeatability at single wafer, batch and batch-to-batch level. Oxide thickness disuniformity will heavily impact Flash device characteristics, in particular programming and erasing performances (see Fig. 2.17). The second requirement is the control of oxide defect density, which will influence memory functionality and reliability. Defectivity is mainly due to macroscopic defects generated during the silicon oxidation process. Other defects come from chemical (heavy metals) and physical (particles) contaminants, coming from wafer management and from the oxidation (furnaces) and cleaning (chemical bench) equipments.

From a point of view of film formation, besides the optimization of the well-known wet and dry oxidation process, starting from 1990 a nitridation of the film has been introduced [45, 46, 47] to further improve the thin dielectric properties. Nitridation occurs by a rapid thermal annealing in N_2O ambient after the oxidation process. In particular, nitride-oxide shows either smaller flat band voltage shift and Fowler-Nordheim voltage variation versus injected charge under both constant current and exponential ramp current stresses [48]. This result is associated with an increase of the hardness to the injected charge. Moreover, a higher immunity against hot carriers has been also shown [49]. The explanation of this improved oxide quality is due to the nitrogen atoms which accumulate at the silicon silicon-dioxide interface releasing mechanical stress, saturating

the dangling bonds and giving rise to a higher energy of the Si-N bond with respect to the Si-O one.

2. *First polysilicon deposition.* The floating-gate of the cell is formed by a polycrystalline silicon film, the first deposition of this type, and usually called poly1. The second one will form the control-gate. Polycrystalline silicon is usually formed by CVD technique, using pyrolysis of SiH_4 . Typical pressure deposition and times are around hundreds of mTorr and hours to form 100–150 nm thick films. The technology issues involved with this step are mainly related to two problems. First, the possibility to damage the underneath tunnel oxide during poly1 deposition, doping, and post thermal treatments. Second, the impact of the grain size on the erasing characteristics of a memory array. An alternative to the polycrystalline silicon is represented by the amorphous silicon (a-Si), due to its very low surface roughness and excellent structure homogeneity. a-Si deposition generates a smoother interface with the active dielectric, thus increasing the electrical properties of devices. Furthermore an a-Si film will improve the uniformity of the electric field across the tunnel oxide during the erasing operation. This is essentially due to the reduction of “point effects” of big grain size in polysilicon. As a result a tighter erasing threshold voltage distribution can be obtained.
3. *Interpoly dielectric.* It is formed on top of poly1 and it must have good dielectric performance to guarantee retention requirements. In fact, it separates the floating-gate from the control-gate and it acts as floating gate sealing film in the word-line direction. Moreover, it must be as thin as possible to increase the gate coupling ratio, thus improving cell performance. Good oxides over polysilicon are obtained by adopting high growth temperature ($> 1100^\circ\text{C}$), but this has the drawback that tunnel oxide quality is greatly reduced by high temperature post-annealing. A trade-off has been reached using triple dielectric films composed by Oxide-Nitride-Oxide (ONO). The thicknesses of the three different layers have been optimized to obtain the best retention performances with the thinnest equivalent thickness [14, 50, 51, 52]. Actual equivalent thickness are in the range of 15–30 nm.
4. *Second polysilicon deposition.* The control gate is formed by this second polycrystalline silicon deposition, commonly called poly2. The issue related to this film is the resistivity of long and narrow strips. In fact the control gate connects many cells (hundreds) on the same line (row), the so-called word line. Usually, to reduce word-line resistivity, a thick film of tungsten silicide is deposited on top of poly2.

5. *Drain and source junction architecture.* The different optimization of drain and source junctions of a Flash cell, respectively to enhance programming and to reduce substrate current in erasing, leads to the asymmetric structure shown in the scheme in Fig. 2.1 and in the TEM picture in Fig. 2.27. Drain junction must be optimized to improve the program-



Figure 2.27 TEM picture of a unit cell for 16 Mbit Flash Memory cell in 0.5 μm technology.

ming characteristics of the cell. Since program is obtained by channel hot electron injection at the drain, the junction must be as efficient as possible in generating hot electron. As known, this is achieved with abrupt junctions, since the longitudinal electric field responsible of the channel electron energy distribution is increased. In Flash cells, the drain junction has been studied for the opposite reasons of what happened in MOS transistors, where hot electron effects must be minimized to reduce charge injection into oxide and interface traps generation, both causing transistor aging. Therefore, in MOS transistors light doped diffusion (LDD) junctions are used. In Flash cells, the drain junction is obtained by a high dose arsenic implant. Tab. 2.4 compares the ion implantation parameters used to form drain junctions in a memory cell and MOS transistor.

Source junction must be optimized to minimize the parasitic substrate current due to band-to-band tunneling during erase. To this purpose, a phosphorus diffusion is added to the arsenic diffusion to obtain a very deep junction [53]. This greatly helps in controlling the substrate current and in avoiding hot hole generation, detrimental for the tunnel oxide reliability. Nevertheless, the n^+ region must be coupled with the floating-gate

Table 2.4 Dose and Energy ranges of implantation steps in 16 Mbit Flash Memory.

IMPLANTATION STEP	DOSE (Atoms/cm ³)	ENERGY (KeV)	TILT ANGLE (deg)
Memory Drain P-Pocket	1×10^{13} – 1×10^{14}	30–100	30–60
Flash memory cell Drain	5×10^{14} – 5×10^{15}	30–100	0
Flash memory cell Source	1×10^{15} – 5×10^{15}	30–100	0
Transistor LDD (n/p)	1×10^{13} – 1×10^{14}	30–100	30–60
Transistor S/D (n/p)	1×10^{14} – 1×10^{15}	30–100	0

to avoid the complete deep-depletion of the surface region underneath the tunnel oxide. If this happens, it gives rise to a decrease of the surface potential and, as a consequence, to a variation of the oxide electric field, responsible for tunneling. As a result, a slower erasing time and broader erase distribution are obtained.

6. The *channel doping* must be carefully adjusted to trade off the conflicting requirements for source and drain junctions. In fact, a high channel doping (surface acceptor concentration in the order of 10^{17} – 10^{18}) helps to further optimize the programming efficiency, giving rise to a higher lateral electric field. On the contrary, a high channel doping decreases the source junction breakdown and worsens the leakage current performances. To completely decouple the drain from the source junction formation a large tilt angle implantation is used in order to increase the boron dopant concentration below the gate after the gate formation only at the drain side (Tab. 2.4). It allows to surround the drain junction with a p-pocket (Fig. 2.28) and to optimize the programming efficiency without increasing the overall channel doping [31, 54]. In this way it is possible to increase the electric field at the drain side without affecting breakdown and BBT characteristics of the source junction.

2.4.4 Interlevel Dielectrics

The usual role played by interlevel dielectric in a standard CMOS process is, in a Flash process, coupled with the non-volatile memory charge retention problem. Intrinsic charge loss is related to ionic process [55, 56]: a field assisted, thermally activated release of mobile ions from the interlevel dielectric. This problem can be faced with contamination gettering in the interlevel film. For example data retention strongly depends on the phosphorus content of a boron-

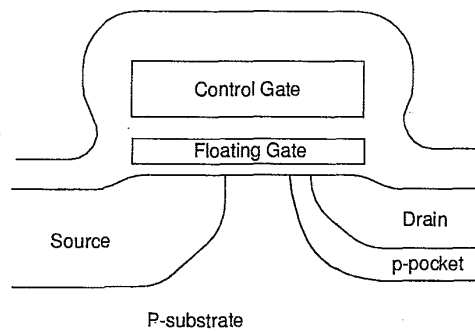


Figure 2.28 Cross section of a generic Flash device with p-pocket.

phosphosilicate (BPSG) film used as interlevel dielectrics. Cells with the same interpoly dielectric and gate oxide thicknesses, but with different BPSG show an increasing charge loss on decreasing the phosphorus concentration of the BPSG film (Fig. 2.29) [55].

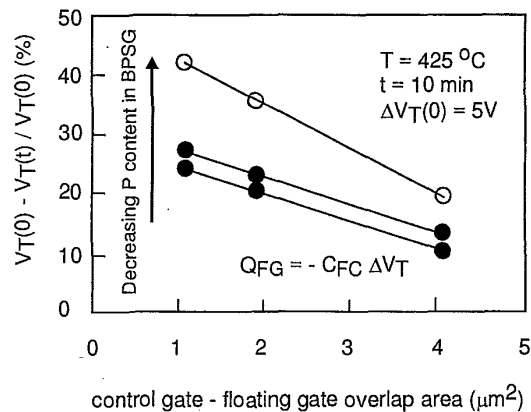


Figure 2.29 Charge loss as a function of control gate/floating gate overlapping area in cells with the same gate oxide thickness but different P concentration in BPSG [55].

2.4.5 Interconnections

Interconnection structure plays a fundamental role in the building process of an integrated circuit. Interconnection technology has a major impact both on the manufacturing and the electrical performances of a circuit. The increasing complexity (Fig. 2.20) of the integrated circuit requires a continuous increase of interconnection levels: Tab. 2.5 reports the Semiconductor Industry Association (SIA) roadmap.

Table 2.5 SIA (Semiconductor Industry Association) forecasted roadmap for interconnections.

Lithography generation (μm)	0.35	0.25	0.18	0.13	0.10
Year	1995	1998	2001	2004	2007
min poly dimensions (nm)	350	250	180	130	100
metal levels (logic)	4-5	5	5-6	6	6-7
metal levels (memory)	2	2-3	3	3	3
max length (km) of interconnections in a chip	0.38	0.84	2.1	4.1	6.3
metal spacing (μm)	1	0.75	0.55	0.4	0.27
contact dimensions (μm)	0.4	0.28	0.2	0.14	0.11

1. *Contacts.* As clearly shown in Tab. 2.5, one of the most critical steps of the interconnection technology is represented by contacts between metal and silicon and by vias, that are contacts between different metal levels. The contact dimensions are very close to the minimum lithography dimension. Therefore, contact technology becomes particularly important in dense Flash memory since every two bits there is one contact. For example, this means that in a 16 Mb Flash there must be at least 8 million "good" contacts. The failure of one contact results in the failure of a couple of bits, then redundancy cells have to be used. It is clear that a low level of contact defectivity means a high manufacturing yield. From the electrical point of view the contact integrity depends on two main aspects: correct definition (lithography and etching), and hole filling by the metal layer. To optimize parasitic capacitance, the pre-metal dielectric does not scale with the same factor of the contact dimension; therefore, the aspect ratio (diameter/depth) of the contacts is becoming more and more critical. The problem of the filling of a hole with a bad aspect ratio ($< 1/2$) is not trivial. Up to now, in 0.35 μm technology, different solu-

tions are still under evaluation: tungsten plugs with CVD technique [57], collimated sputtering or, more recently, ion metal plasma sputtering.

2. *Planarization.* Another key solution to allow interconnection scaling has been surface planarization. In Flash memory this becomes a key issue due to the critical morphology of the field oxide resulting from the specific active area in the cell array. Moreover, it can be added that interconnection process uses films with thickness of the same order of the stepper focus depth. This problem is solved using special planarization techniques. One consists in a sacrificial use of a liquid phase deposited glass, so called Spin On Glass (SOG). As shown in Fig. 2.30, SOG deposition is able to fill very small gaps, resulting in a planar surface. The following anisotropic etch moves the planarity level to the underlying dielectric. This procedure conducts to a good planarization level, but only on localized regions inside the circuit. A complete planarization is obtained with Chemical Mechan-

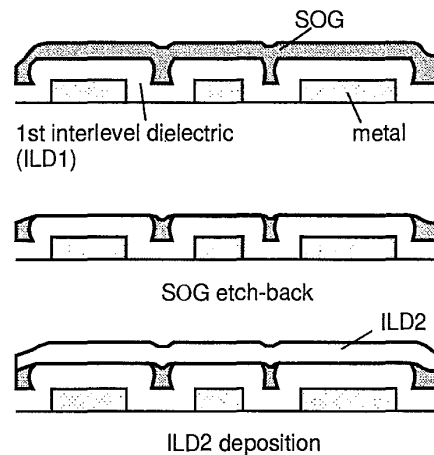


Figure 2.30 Planarization techniques: Spin on Glass (SOG).

ical Polishing (CMP), Fig. 2.31 [58]. This is very similar to the wafer finishing technique. The process steps before CMP are the high density plasma deposition followed by a thick sacrificial dielectric by CVD. The CMP process is controlled by the chemical composition, the slurry, the head pressure and the wafer surface effective speed. Although this process seems to be very heavy to be applied in the silicon technology – it is hard to find the process end point, difficulties due to contaminant and solid particles cleaning – this technique is today mature and it has a diffused industrial application even with most advanced Flash technologies.

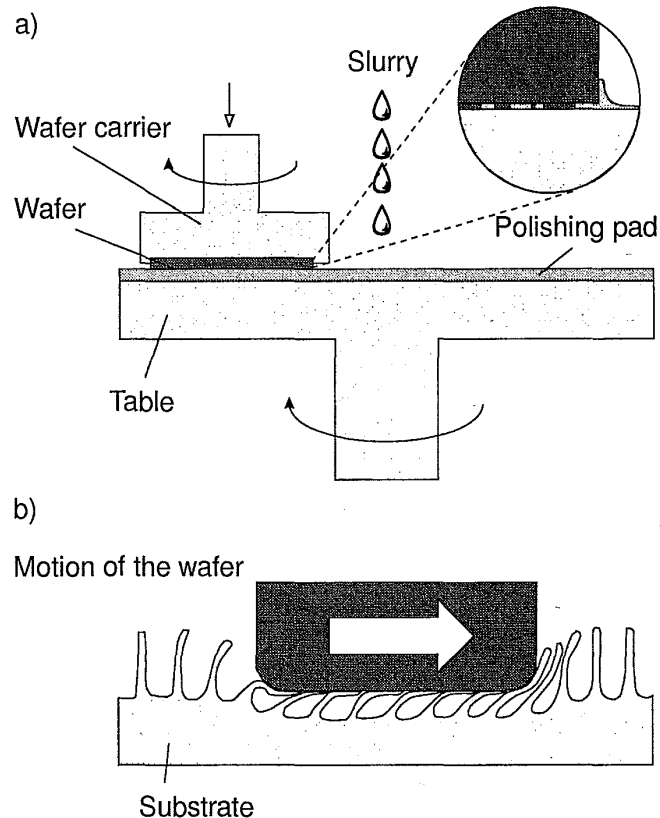


Figure 2.31 Planarization techniques: Chemical Mechanical Polishing (CMP). a) Schematic of the machinery; b) detail of the wafer surface during the procedure [58].

3. *Metallization.* Coming back to Tab. 2.5, it can be observed that complexity constraints for memories in terms of interconnection levels are relaxed compared to logic ones, implying an obviously different cost/performance ratio. The interconnection roadmap shows a metal pitch reduction with constant or reduced metal thickness. The constant thickness pitch scaling implies a linear increase of current density, parasitic capacitance and resistance proportional to the scaling factor. In this case, the aspect ratio thickness/distance becomes critical. From the morphology point of view, the very short distance between two metal stripes and the high metal thickness to be etched worsen the definition characteristics, in particular by plasma etch, and could result in short circuits. On the contrary

the contemporary scaling of pitch and thickness implies the increase of resistance and current density with a square law of the scaling factor. In this case the critical issue is represented by metal layer composition and this can be dealt with by using a low resistivity and low electromigration material. To this purpose, recently, copper has been proposed as an ideal candidate [59], due to its $1.7\mu\Omega\text{cm}$ resistivity and its good electromigration properties. Different deposition techniques have been proposed: sputtering, CVD, electrochemical plating or electroless. Common to all of them is the need of a good barrier to prevent copper diffusion towards the junctions. Also in this case a TiN film is a good solution. Nevertheless, the key issue for copper integration in silicon process is plasma etch to define copper stripes. The presence of volatile chlorine compounds produced by reactions and corrosion phenomena has pushed towards solutions that do not imply copper etch, like for example damascene technique (Fig. 2.32) [60].

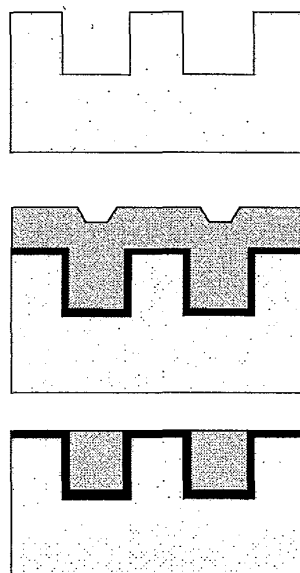


Figure 2.32 Sequence of damascene technique; top: trench definition and etch; middle: copper deposition; bottom: interconnections after chemical-mechanical-polishing [60].

A SEM picture of an array is shown in Fig. 2.33, where a cross section along BL is shown: single cells, source and drain diffusion lines, contact plugs can be observed.

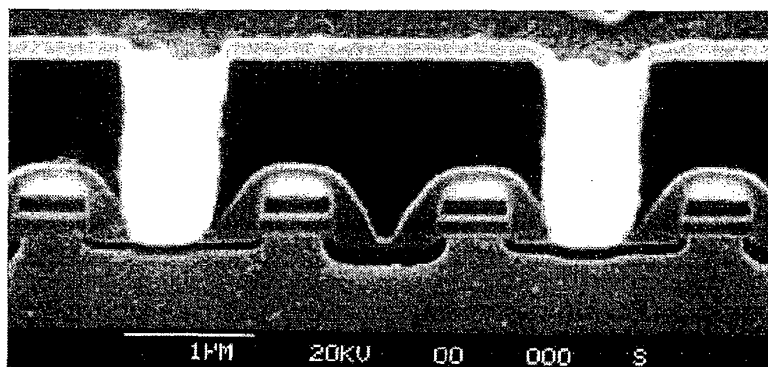


Figure 2.33 SEM micrograph of a few cells in an array.

2.4.6 Final Passivation

As in all integrated circuits, also in Flash memories the final passivation is the “protective shield” towards the outside aggressive world. It must protect against contamination compounds, since retention can be greatly affected by them. Moreover, the functionality temperature range of a Flash commodity spans between -40°C and 125°C . Finally, a passivation specific, which is necessary in EPROM, but typical to Flash too, is to be UV light transparent. In fact, since different process steps can induce charge into the floating gate, before testing all wafers are submitted to UV exposure, to reset the floating gate to its virgin state.

2.5 YIELD AND RELIABILITY

Many issues have to be addressed when, from the theoretical model of a single cell, a “real” product has to be designed, integrating millions of devices in an array. Nonvolatility implies at least 10 years of charge retention, and the cell has to store information also after many read/program/erase cycles. Cycling and retention experiments are performed to investigate Flash cell reliability. The confidence on Flash memory reliability has grown together with the understanding of the single memory cell failure mechanisms.

Other issues are specific to the organization which is used to access the array of cells. These issues can give rise to malfunctioning of the product either when it comes out of the production line, or during its operating life. Flash arrays are verified analyzing array disturbs and erase threshold distribution. New architecture solutions, however, open new issues on Flash array reliability.

The high degree of testability allows the detection at wafer level of latent defects which may cause single bit failures related to programming disturbs, data retention and premature oxide breakdown, thus making Flash memories more reliable than full featured EEPROMs, at equivalent density [61].

2.5.1 Retention

Fast program and erase operations require high voltages and currents through thin oxides, which in turn are easily degraded. In modern Flash cells, for example a 16 Mbit fabricated in a $0.4\mu\text{m}$ technology, floating gate capacitance is approximately 1 fF, a threshold voltage shift of 3 V is requested, and a programmed cell stores around 10,000 electrons in its floating gate. A loss of only 10% in this number can lead to a wrong read of the cell, therefore a loss of less than 2 electrons per week can be tolerated.

Mechanisms which lead to charge loss or charge gain can be divided into extrinsic and intrinsic ones: the former are due to defects in the device structure, the latter to the physical mechanisms which are used for program and erase operations.

Retention capability of Flash memories has to be checked by using accelerated tests which usually adopt high electric fields and hostile environments at high temperatures.

2.5.2 Endurance

Today Flash cells are requested to guarantee 100,000 erase/program cycles. Cycling is known to cause a fairly uniform wear-out of cell performance [62], due to the degradation of the tunnel oxide, which eventually limits Flash memory endurance. A typical result of an endurance test on a single cell is shown in Fig. 2.34 [61]; as the experiment was performed applying constant pulses, the variations of program and erase threshold levels are described as "program/erase threshold window closure", and give a measure of oxide aging. In real devices this corresponds to longer program/erase times.

In particular, the reduction of the programmed threshold with cycling is due to trap generation in the oxide and to interface state generation at the drain side of the channel [63], which are mechanisms specific to hot electron degradation (see Fig. 2.34). The evolution of erase threshold voltage reflects the dynamics of net fixed charge in the tunnel oxide as a function of the injected charge [64]: the initial lowering of the erase V_T is due to a pile-up of positive charge which enhances tunneling efficiency, while the long term increase of the erase V_T is due to generation of negative traps (see Fig. 2.34).

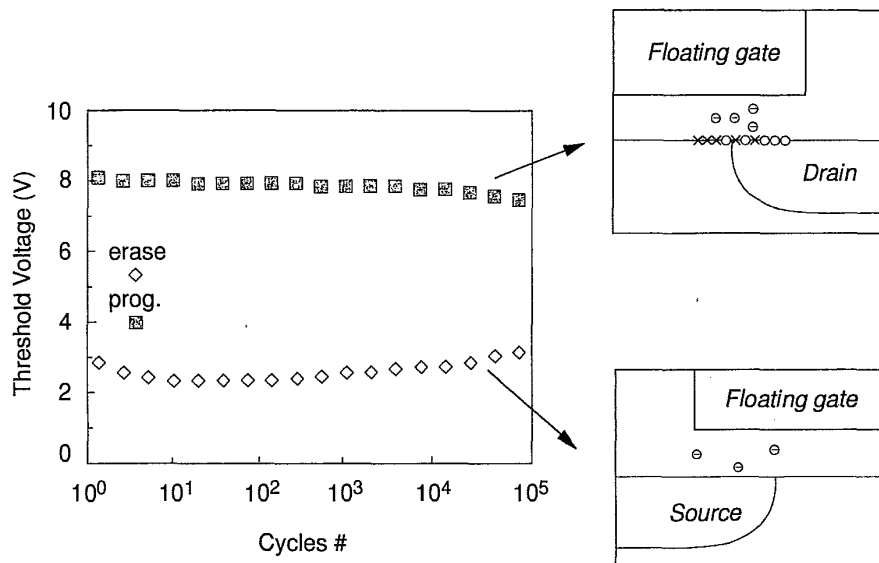


Figure 2.34 Threshold voltage window closure as a function of program/erase cycles on a single cell [61].

2.5.3 Reading Disturbs

During the Read operation, biases applied to the cell are of the same kind of biases applied during programming, only lower in magnitude. Flash cell scaling in the lateral dimension requires careful optimization of the doping profiles in order to enhance programming and erasing performances, while keeping under control short channel effects and preserving reliability [65]. The drain implant in asymmetrical cells with p-pocket, introduced in the process to increase programming efficiency and separately optimize drain and source junctions, increases also the spurious hot carrier generation at low V_{DS} used for reading. This leads to the read disturb called "soft-programming" [66], by which a "one" (no charge stored in the FG) can in principle be slowly converted into a "zero" by the cumulative low level injections caused by frequently repeated read cycles, situation which is particularly harmful in reference cells of the sense amplifier.

2.5.4 Programming Disturbs

Considering an array as in Fig. 2.35, if we want to program the highlighted transistor, a high voltage ($V_{pp} = 12\text{ V}$) is applied to the WL, and a sufficiently high voltage ($V_{dd} = 5 - 7\text{ V}$) is applied to the BL, in order to generate hot-electrons to program the cell. In this bias condition, though, there are two major disturbs, one due to the high voltage applied to the WL and to the transistors on the same line, the second to the medium-high voltage applied to the BL and to the transistors on the same column.

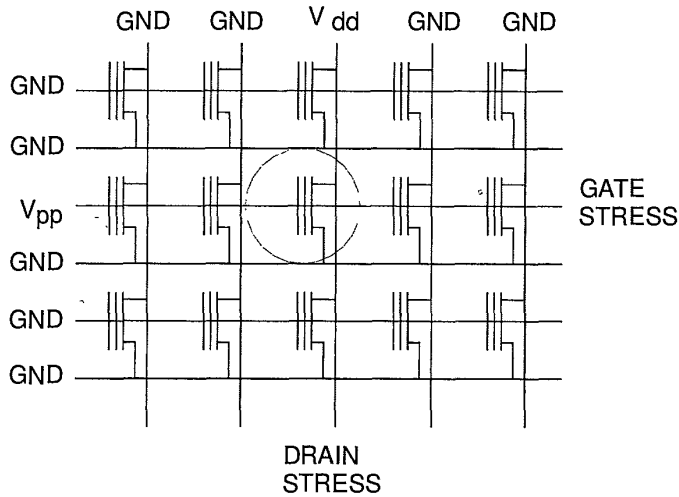


Figure 2.35 Programming disturbs.

High voltages applied to the WL can stress the gate of transistors which have their gate connected to the WL but are not selected. There might be tunneling of electrons from floating to control gate through the interpoly oxide in all the programmed cells, i.e. in those cells where the FG is filled with electrons, since they have 12 V applied to the gate and 0 V on both source and drain. This is the “DC-erasing” disturb (Fig. 2.36), which induces charge-loss, and therefore reduces the margin for the high level of threshold voltage.

There might be also tunneling of electrons from the substrate to the floating gate in all the non-programmed cells, i.e. in those cells where the FG is “empty”. This is the “DC-programming” disturb (Fig. 2.37), which induces charge-gain and reduces the margin for the low level of threshold voltage.

Both these disturbs are called “gate-disturbs” and are present even during reading operations. They are triggered to test gate-oxide quality.

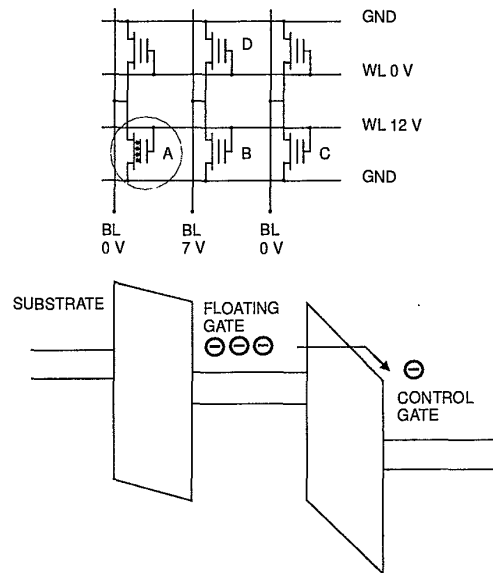


Figure 2.36 Programming disturbs - DC erasing of a programmed cell. Cell A is programmed, cell C is non-programmed.

A relatively high voltage ($V_{dd} = 5 - 7 \text{ V}$) applied to the BL can stress the drains of all FG transistors in the same column. Namely, in cells which share the BL with cells which are to be programmed, electrons tunnel from the floating gate through the gate oxide to the drain [67]; moreover, holes can be generated via impact-ionization in the substrate and then injected in the floating gate. This disturb, called “drain-disturb” (Fig. 2.38), causes charge-loss and, consequently, a decrease in the high value of the threshold voltage. The same disturb can result from extensive reading cycles, and can be used as a gate oxide quality monitor.

These disturbs become important when the same reading or programming operation needs to be repeated continuously many times, for example when a complete row, or column, is to be programmed in an array; this takes, in a 1Mbit array, a thousand repetitions. Disturb influence becomes more and more important on increasing the number of reading-programming cycles, or programming-erasing cycles.

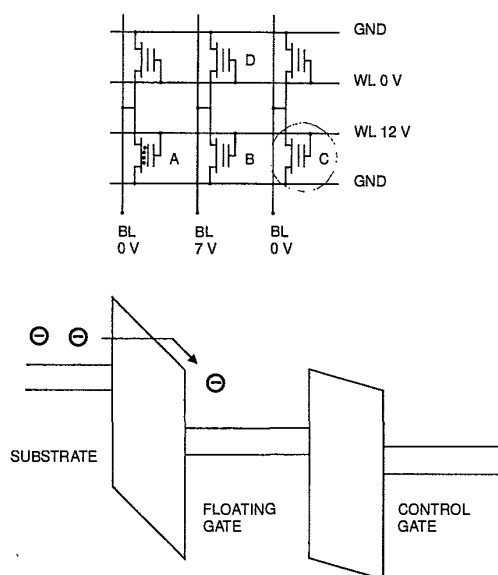


Figure 2.37 Programming disturbs - DC programming of a non-programmed cell. Cell A is programmed, cell C is non-programmed.

2.5.5 Erasing Disturbs

The electrical erase operation is performed on a block or sector, or on the whole array. This results in a distribution of erase threshold voltages of the single cells which is shown in Fig. 2.39 (dotted line). If the erase operation is performed by UV exposure, the distribution of the erase threshold voltages is almost a Gaussian around a mean value, see Fig. 2.39 (solid line).

The exponential tail in threshold voltage distribution represents a large population of cells which erase faster than typical bits. This population is too large to be attributed to extrinsic defects, and it is believed to be related to statistical fluctuations of oxide charge and to the structure of the injecting electrode [61]. Positive charges in the tunnel oxide and irregular polycrystal grains may induce a local increase of the electric field, thus enhancing current injection locally, and making individual cells to erase faster than average.

A relevant mechanism of single bit failure during program/erase cycling is the occurrence of an "erratic bit" [61, 68, 69]. An erratic bit shows an unstable and unpredictable behavior in erasing, since its erase threshold voltage changes randomly from cycle to cycle, from the bulk Gaussian distribution to the lower part of the tail. This behavior is expected to be due to hole trapping

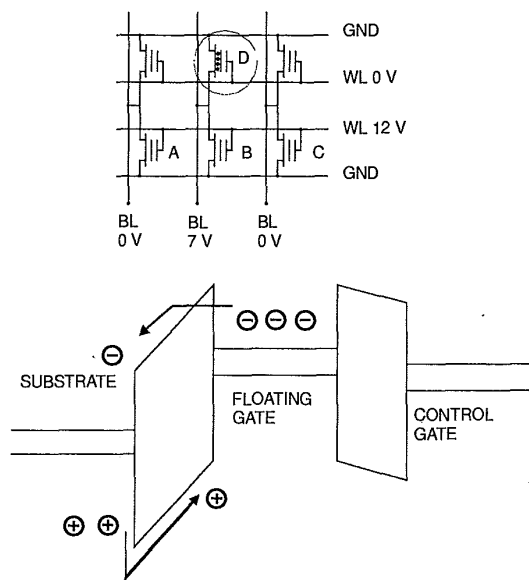


Figure 2.38 Programming disturbs - Drain disturb. Cell D is programmed.

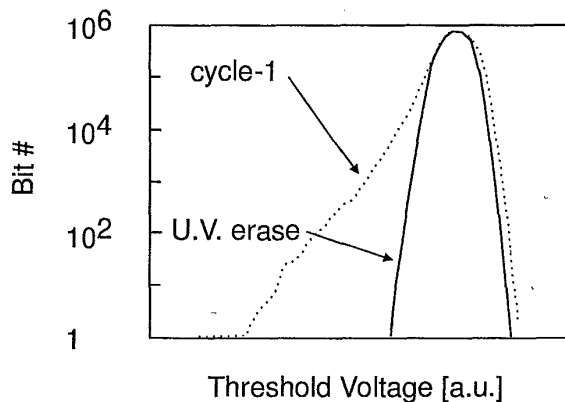


Figure 2.39 Threshold voltage distribution after different erase procedures: UV erase (solid curve), after the first cycle (dotted line) [61].

in the tunnel oxide. Wentzel-Kramers-Brillouin (WKB) calculations [68, 69] have shown that the statistical distribution of hole traps gives a low but finite probability of having clusters of three or more positive charges, whose com-

bined electric field effect induces a huge local increase in the tunnel current. In this condition, trapping/detrapping of an individual positive charge causes a detectable change in erasing speed and threshold level. Since this behavior is due to statistical fluctuations of intrinsic oxide defects, erratic bit occurrence can be reduced by process optimization, but cannot be completely eliminated; design solutions have been developed to cope with this problem at circuit level.

Other failures are related to the erase mechanism. Since FN tunneling is not self-limiting, it can lead to over-erasing of the cells, i.e. more electrons than those which have been stored are removed from the floating gate; the device has less negative charges than in the non-programmed case and a net positive charge is now present, thus transforming the device from an enhancement to a depletion one. In this case, the read operation will always give a wrong result.

2.6 SCALING ISSUES

The architecture of an Industry Standard Flash cell array is typical of a NOR gate array, where every single cell is addressed by two signals, one for the BL and one for the WL; the source line and body are common to the whole array. Moreover, in standard arrays a contact is shared between two cells, thus consuming cell area. The common issue among the different solutions and applications is the cost-per-bit reduction, which will be mainly provided by technology scaling. No consolidated theory has been developed for Flash cell scaling [70].

Scaling issues deal then with the single cell layout. The goal is to reduce the area used for contacts, and layout issues are contact placement issues. To improve integration, many new solutions have been proposed, mainly new array architectures.

A reduction of the area occupied by a Flash memory cell when fabricated in a double poly stacked gate structure, particularly the reduction of the effective channel length, L_{eff} , gives many advantages, not only from the density point of view, but also for the performances. In fact, the efficiency of the carrier injection into the floating gate increases on decreasing L_{eff} , thus speeding up the program operation. On the contrary, decreasing L_{eff} enhances punch-through and drain turn on, since the capacitive coupling between drain and floating gate increases. The final value of L_{eff} comes from a trade-off between performances and disturbs.

Another relevant issue in Flash memories is the need of high voltages for program and erase. While CMOS technology scaling requires the reduction of the operating voltages, the actual program/erase operations are based on physical mechanisms whose major parameters do not scale (3.2 eV energy barrier for channel hot electrons and at least 8–9 MV/cm for FN data alteration in

0.1–1 s). Moreover, the trend towards increasing the programming throughput will even force the internal voltage to rise. Double voltage supply simplifies memory design and minimizes the area, since there is no need to generate the high voltage internally. Therefore, they are preferably used when present in the circuit, even though internal generation is sometimes preferred for the correct operation of the memories in the chip. If internal generation is to be done, many issues need to be discussed. For example, hot electron programming is not an efficient mechanism, and if only a 5 V voltage supply is used and current levels are not increased, programming times can become unacceptably slow. On the other hand, internal charge pumping circuits can be used only when small currents flow in the channel. Erasing opens similar issues.

Nonvolatility implies at least 10 years of charge retention. Nonvolatility issues affect the scalability of thin active dielectrics (tunnel and interpoly). Direct tunneling mechanism fixes the tunnel oxide limit to 6 nm, which needs to be increased more realistically up to 7–8 nm, due to trap assisted electron tunneling caused by oxide aging [71]. The scalability limit of the interpoly dielectric (ONO) has been reported to be around 12–13 nm [14]. These thicknesses can be combined to give an equivalent memory cell oxide (defined as tunnel oxide thickness divided by the coupling coefficient α_G), which sets the limit for the memory cell poly length. Other constraints limit the minimum poly length, namely: channel hot electron injection requires some minimum drain-gate overlap and abrupt junction to maximize injection efficiency; FN tunneling to the source requires an overlap with the n^+ region below the gate; FN tunneling to the channel requires small gate/diffusions overlaps. Moreover, when charge is injected from the polysilicon floating gate the number of poly grains in the tunneling area plays an important role in determining the V_T distribution width [72].

In this scenario, the search for higher integration goes towards new architectural solutions, towards the reduction of the number of contacts, and towards the reduction of alignment tolerances. Contactless (virtual ground) configurations have been proposed and used [73]. Fig. 2.40 shows the layouts of a T-shaped staked-gate Flash cell and of a contactless one. In the second case, a 50% area reduction can be achieved, only by self-aligning every single device, but this induces a higher complexity.

Other structures alternative to Industry Standard Flash cells have been considered for Flash memories, see again [13]. Differences are mainly due to the array organization, program/erase mechanisms, and approaches to overerasing (which is solved algorithmically in standard structures). Many new cells use FN tunneling to the channel for both erase and programming, to enable an easier supply scaling and to reduce cell size. These new cells are used in NOR or AND architectures according to their specific details. A completely different

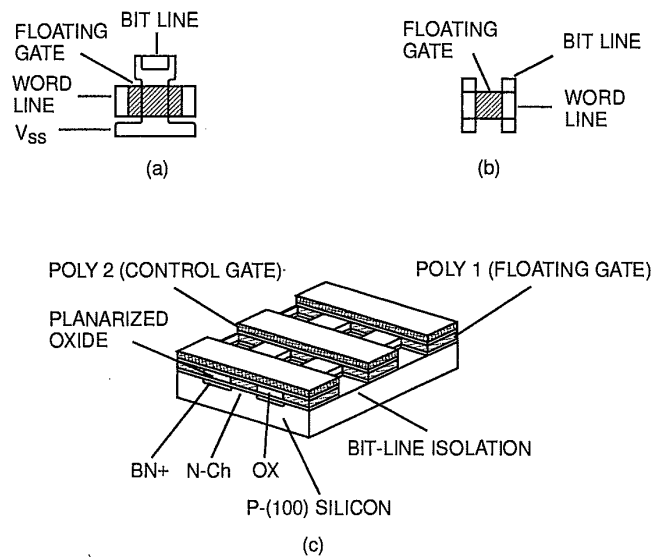


Figure 2.40 Layout of a T-shaped staked gate Flash cell (a). A contactless (Virtual Ground) cell is also shown in (b) and (c) [73].

approach in array organization can be followed by using a NAND architecture, in which the elementary unit is not composed by the single three-terminal cell, which stores one single bit, but by more FG transistors connected in a series of 8 or 16, realizing a chain connected to the bit line and ground through two select transistors. This kind of memory organization with a unit element with a dimension of one byte (or one word) is closer to the ideal memory with parallel access. It allows even page (256-byte) programming, resulting in a greatly improved versatility. Many efforts are currently being done to produce this kind of memory for mass storage applications.

Acknowledgments

The authors wish to thank Roberto Formentini and Sabina Gheduzzi (University of Modena) for their help in figure editing.

References

- [1] Frohman-Bentchkowsky D. (1971) "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure". *Applied Physics Letters*,

18, pp. 332–334.

- [2] Frohman-Bentchkowsky D. (1974) "FAMOS - a new semiconductor charge storage device". *Solid-State Electronics*, **17**, p. 517.
- [3] Guterman D.C., Rimawi I.H., Chiu T.L., Halvorson R.D. and McElroy D.J. (1979) "An electrically alterable nonvolatile memory cell using a floating-gate structure". *IEEE Trans. on Electron Devices*, **26**, 4, pp. 576–586.
- [4] Masuoka F., Asano M., Iwahashi H., Komuro T. and Tanaka S. (1984) "A new Flash E²PROM cell using triple polysilicon technology". *IEDM Tech. Dig.*, pp. 464–467.
- [5] Verma G. and Mielke N. (1988) "Reliability performance of ETOX based Flash memories". *Proc. IRPS*, p. 158.
- [6] Bauer M. *et al.* (1995) "A multi-level 32 Mb Flash memory". *ISSCC Conf. Proc.*, p. 132.
- [7] Hemink G.J., Tanaka T., Endoh T., Aritome S. and Shirota R. (1995) "Fast and accurate programming method for multi-level NAND EEPROMs". *Digest of VLSI Symposium on VLSI Technology*, no. 10B-4, pp. 129–130.
- [8] Eitan B., Kazerounian R., Roy A., Crisenza G., Cappelletti P. and Modelli A. (1996) "Multilevel flash cells and their trade-offs". *IEDM Tech. Dig.*, pp. 169–172.
- [9] Rodjy N. (1992) "0.85 μ m double metal CMOS technology for 5 V Flash EPROM memories with sector erase". *12th Nonvolatile Semiconductor Memory Symposium*, Monterey, California (USA).
- [10] Bergemont A., Haggag H., Anderson L., Shacham E. and Woltsenholme G. (1993) "NOR virtual ground (NVG) - a new scaling concept for very high density FLASH EEPROM and its implementation in a 0.5 μ m process". *IEDM Tech. Dig.*, pp. 15–18.
- [11] Bergemont A., Chi M. and Haggag H. (1996) "Low voltage NVG: a new high performance 3 V/5 V Flash technology for portable computing and telecommunications applications". *IEEE Trans. on Electron Devices*, **43**, 9, pp. 1510–1517.
- [12] Kynett V.N., Baker A., Fandrich M., Hoekstra G., Jungroth O., Kreifels J. and Wells S. (1988) "An in-system reprogrammable 256 K CMOS flash memory". *ISSCC Conf. Proc.*, p. 132.

- [13] Pavan P., Bez R., Olivo P. and Zanoni E. (1997) "Flash memory cells - an overview". *Proc. of the IEEE*, **85**, 8, pp. 1248-1271.
- [14] Mori S., Yamaguchi Y., Sato M., Meguro H., Tsunoda H., Kamiya E., Yoshikawa K., Arai N. and Sakagami E. (1996) "Thickness scaling limitation factors of ONO interpoly dielectric for nonvolatile memory devices". *IEEE Trans. on Electron Devices*, **43**, 1, pp. 47-53.
- [15] Venkatesh B., Chung M., Govindachar S., Santurkar V., Bill C., Gutala R., Zhou D., Yu J., Van Busirik M., Kawamura S., Kurihara K., Kawashima H. and Watanabe H. (1996) "A 55 ns 0.35 μ m 5 V-only 16 M flash memory with deep-power-down". *ISSCC Conf. Proc.*, no. TP 2.7, pp. 44-45.
- [16] Wang S.T. (1979) "On the I-V characteristics of floating-gate MOS transistors". *IEEE Trans. on Electron Devices*, **26**, 9, pp. 1292-1294.
- [17] Wada M., Mimura S., Nihira H. and Iizuka H. (1980) "Limiting factors for programming EPROM of reduced dimensions". *IEDM Tech. Dig.*, pp. 38-41.
- [18] Kolodny A., Nieh S.T.K., Eitan B. and Shappir J. (1986) "Analysis and modeling of floating gate EEPROM cells". *IEEE Trans. on Electron Devices*, **33**, 6, pp. 835-844.
- [19] Prall K., Kinney W.I. and Marco J. (1987) "Characterization and suppression of drain coupling in submicrometer EPROM cells". *IEEE Trans. on Electron Devices*, **34**, 12, p. 2463.
- [20] Wong M., Liu D.K.-Y. and Huang S.S.-W. (1992) "Analysis of the sub-threshold slope and the linear transconductance techniques for the extraction of the capacitive coupling coefficients of floating-gate devices". *IEEE Electron Device Letters*, **13**, 11, pp. 566-568.
- [21] Choi W.L. and Kim D.M. (1994) "A new technique for measuring coupling coefficients and 3-D capacitance characterization of floating gate devices". *IEEE Trans. on Electron Devices*, **41**, 12, pp. 2337-2342.
- [22] Bez R., Camerlenghi E., Cantarelli D., Ravazzi L. and Crisenza G. (1990) "A novel method for the experimental determination of the coupling ratios in submicron EPROM and Flash EEPROM cells". *IEDM Tech. Dig.*, pp. 99-102.
- [23] San K.T., Kaya C., Liu D.K.Y., Ma T.P. and Shah P. (1992) "A new technique for determining the capacitive coupling coefficients in FLASH EPROMs". *IEEE Electron Device Letters*, **13**, 6, pp. 328-331.

- [24] Moison B., Papadas C., Ghibaudo G., Mortini P. and Pananakakis G. (1993) "New method for the extraction of the coupling ratios in FLOTOX EPROM cells". *IEEE Trans. on Electron Devices*, **40**, 10, pp. 1870-1872.
- [25] Woods M. (1991) "An E-PROM's integrity starts with its cell structure". C. Hu (Ed.), *Nonvolatile semiconductor memories: technologies, design, and application*, IEEE Press, Chapter 3, pp. 59-62.
- [26] Bez R., Cantarelli D. and Serra S. (1992) "The channel hot electron programming of a floating gate MOSFET: an analytical study". *12th Nonvolatile Semiconductor Memory Workshop*, Monterey, California (USA).
- [27] Bez R., Cantarelli D., Moioli L., Ortolani G., Villa C. and Dallabora M. (1998) "A new erasing method for a single-voltage long-endurance Flash memory". *IEEE Electron Device Letters*, **19**, 2, pp. 37-39.
- [28] Cagnina S., Chang C., Haddad S., Lien J., Radjy N., Sun Y., Tang Y., Van Buskirk M. and Wang A. (1992) "A 0.85 μm double metal CMOS technology for 5 V Flash memories with sector erase". *12th Nonvolatile Semiconductor Memory Workshop*, Monterey, California (USA).
- [29] Yoshikawa K., Yamada S., Miyamoto J., Suzuki T., Oshikiri M., Obi E., Hiura Y., Yamada K., Ohshima Y. and Atsumi S. (1992) "Comparison of current Flash EEPROM erasing methods: stability and how to control". *IEDM Tech. Dig.*, pp. 595-598.
- [30] Keenney S., Bez R., Cantarelli D., Piccinini F., Mathewson A. and Lombardi C. (1992) "Complete transient simulation of Flash EEPROM devices". *IEEE Trans. on Electron Devices*, **39**, 12, pp. 2750-2757.
- [31] Cappelletti P., Panchieri A. and Ravazzi L. (1993) "Mastering key factor which affect flash memory reliability". *ESREF 93*, Bordeaux (France), pp. 77-82.
- [32] Tsai H., Yu C.L. and Wu C.Y. (1986) "A bird's beak technique for LOCOS in VLSI fabrication". *IEEE Electron Device Letters*, **7**, 2, pp. 122-123.
- [33] Wils N.A.H., van der Plas P.A. and Montree A.H. (1990) "Dimensional characterization of poly buffer LOCOS in comparison with suppressed LOCOS". *ESSDERC 90*, pp. 535-538.
- [34] Guldi R.L., McKee B., Damminga G.M., Young C.Y. and Beals M.A. (1989) "Characterization of poly buffer locos in manufacturing environment". *Journal of Electrochemical Society*, **136**, p. 3815.

- [35] Miéville J.P., Rooyackers R. and Deferm L. (1994) "An optimized poly-buffered LOCOS process for a 0.35 μm CMOS technology". C. Hill and P. Asburn (Eds.), *Proc. 24th ESSDERC 94*, pp. 99–202.
- [36] Burton G., Tuntasood P., Chien F., Kovacs R. and Vora M. (1984) "New techniques for elimination of the bird's beak". *IEDM Tech. Dig.*, pp. 582–585.
- [37] Shimizu N., Naito Y., Itoh Y., Shibata Y., Hashimoto K., Nishio M., Asai A., Ohe K., Umimoto H. and Hirofiji Y. (1992) "A poly-buffer recessed LOCOS process for 256Mbit DRAM cells". *IEDM Tech. Dig.*, pp. 279–282.
- [38] Pfister J.R., Kenkare P.U., Subrahmanyam R., Lin J.H. and Crabtree P. (1993) "Nitride-clad LOCOS isolation for 0.25 μm CMOS". *VLSI Tech. Symp.*, no. 11-2, pp. 139–140.
- [39] Borland J.O. and Koelsch R. (1993) "MeV implantation technology: next generation manufacturing with current generation equipment". *Solid State Technology*, p. 1.
- [40] Cappelletti P., Fratin L. and Ravazzi L. (1995) "Application of advanced ion implantation techniques to Flash memories". *Nuclear Instruments and Methods in Physics Research, B*, **96**, pp. 405–410.
- [41] Auricchio C., Bez R., Losavio A., Maurelli A., Sala C. and Zabberoni P. (1996) "A triple-well architecture for low voltage operation in submicron CMOS devices". G. Baccarani and M. Rudan (Eds.), *Proc. ESSDERC 96*, Bologna (Italy), p. 613.
- [42] Umezawa A., Atsumi S., Kuryiama M., Banba H., Imamiya K., Naruke K., Yamada S., Obi E., Oshikiri M., Suzuki T. and Tanaka S. (1992) "A 5 V-only operation 0.6 μm Flash EEPROM with row decoder scheme in triple-well structure". *IEEE Journal of Solid State Circuits*, **27**, 11, p. 1540.
- [43] Momodomi M., Tanaka T., Iwata Y., Tanaka Y., Oodaira H., Itoh Y., Shirota R., Huchi K.O. and Masuoka F. (1991) "A 4Mb NAND EEPROM with tight programmed V_t distribution". *IEEE Journal of Solid State Circuits*, **26**, 4, p. 492.
- [44] Kobayashi K., Nakai H., Kunori Y., Nakayama T., Miyawakiand Y., Terada Y., Onoda H., Ajika N., Hatanaka M., Miyoshi H. and Yoshihara T. (1993) "Memory array architecture and decoding scheme for 3 V-only sector erasable DINOR flash memory". *VLSI Circuit Digest Technical Papers*, p. 93.

- [45] Uchiyama A., Fukuda H., Hayashi T., Iwabuchi T. and Ohno S. (1990) "High performance dual-gate sub-half micron CMOS with 6 nm thick nitrided SiO₂ films in N₂O ambient". *IEDM Tech. Dig.*, pp. 425–428.
- [46] Fukuda H., Arakawa T. and Ohno S. (1990) "High reliable thin nitride films formed by rapid thermal processing in N₂O ambient". *Journal of Applied Physics*, **29**, 12, p. L2333.
- [47] Ting W., Hwang H., Lee J. and Kwong D. (1990) "Composition and growth kinetics of ultrathin SiO₂ by oxidizing Si substrate in N₂O". *Applied Physics Letters*, **57**, p. 26.
- [48] Bellafore N., Pio F. and Riva C. (1993) "Thin oxide nitridation in N₂O by RTP for non-volatile memories". *Microelectronics Journal*, **24**, p. 453.
- [49] Dunn G. and Scott S.A. (1990) "Channel hot-carrier stressing of reoxidized nitrided silicon dioxide". *IEEE Trans. on Electron Devices*, **37**, 7, p. 1719.
- [50] Pan C., Wu K.J., Freiberger P.P., Chatterjee A. and Sery G. (1990) "A scaling methodology for oxide-nitride-oxide interpoly dielectric for EPROM applications". *IEEE Trans. on Electron Devices*, **37**, 6, p. 1439.
- [51] Pan C., Yu K.W. and Sery G. (1991) "Physical origin of long-term charge loss in floating-gate EPROM with interpoly oxide-nitride-oxide stacked dielectric". *IEEE Electron Device Letters*, **12**, 2, p. 51.
- [52] Mori S., Sakagami E., Araki H., Kaneko Y., Narita K., Ohshima Y., Arai N. and Yoshikawa K. (1991) "ONO interpoly dielectric scaling for nonvolatile memory applications". *IEEE Trans. on Electron Devices*, **38**, 2, pp. 386–391.
- [53] Tang Y., Chen J., Chang C., Liu D., Haddad S., Sun Y., Wang A., Ramskey M., Kwong M., Kinoshita H., Chan W. and Lien J. (1996) "Different dependence of band-to-band and Fowler-Nordheim tunneling on source doping concentration of an n-MOSFET". *IEEE Electron Device Letters*, **17**, 11, p. 525.
- [54] Cappelletti P., Cutolo A., Fratin L., Ravazzi L. and Riva C. (1995) "The Flash E²PROM cell with Boron p-pocket architecture: advantages and limitations". *Proc. Nonvolatile Semiconductor Memory Workshop*.
- [55] Crisenza G., Ghidini G., Manzini S., Modelli A. and Tosi M. (1990) "Charge loss in EPROM due to ion generation and transport in interlevel dielectrics". *IEDM Tech. Dig.*, p. 107.

- [56] Crisenza G., Clementi C., Ghidini G. and Tosi M. (1992) "Floating gate memories reliability". *Quality and reliability engineering international*, **8**, 177.
- [57] Ohmi T., Miyashita M. and Imaoka T. (1991) *Proc. Microcontamination 91*, pp. 491-510, 1991.
- [58] Tissier A. *et al.* (1994) "Planarization of pre-metal and metal levels for 0.5 μm and 0.35 μm logic CMOS processes". *Proc. Conferences on Advanced Metallization for ULSI Application MRS*.
- [59] Bai G., Chiang C., Cox N., Fang S., Gardner D.S., Mack A. and Marieb T. (1996) "Copper interconnection deposition techniques and integration". *Proc. 1996 Symposium on VLSI Technology*, Honolulu, pp. 48-49.
- [60] Morand Y., Lerme M., Palleau J., Torres J., Vinet F., Demolliens O., Ulmer L., Gobil G., Fayolle M., Romagna F. and Bihan R.L. (1997) "Copper integration in self aligned dual damascene architecture". *Proc. 1997 Symposium on VLSI Technology*, Kyoto (Japan), pp. 31-32.
- [61] Cappelletti P., Bez R., Cantarelli D. and Fratin L. (1994) "Failure mechanisms of Flash cell in program/erase cycling". *IEDM Tech. Dig.*, pp. 291-294.
- [62] Haddad S., Chang C., Swaminathan B. and Lien J. (1989) "Degradation due to hole trapping in Flash memory cells". *IEEE Electron Device Letters*, **10**, 3, pp. 117-119.
- [63] Yamada S., Hiura Y., Yamane T., Amemiya K., Oshima Y. and Yoshikawa K. (1993) "Degradation mechanism of Flash EEPROM programming after program/erase cycles". *IEDM Tech. Dig.*, pp. 23-26.
- [64] Olivo P., Riccò B. and Sangiorgi E. (1986) "High field induced voltage dependent oxide charge". *Applied Physics Letters*, **48**, pp. 1135-1137.
- [65] Esseni D., Selmi L., Bez R., Ravazzi L. and Sangiorgi E. (1995) "Soft-programming in scaled flash memory cells". H.C. de Graaf and H.V. Kranenburg (Eds.), *Proc. ESSDERC 95*, The Hague (The Netherlands), pp. 549-552.
- [66] Van Houdt J.F., Wellekens D., Groeseneken G. and Maes H.E. (1995) "Investigation of the soft-write mechanism in source side injection flash EEPROM devices". *IEEE Electron Device Letters*, **16**, p. 181.

- [67] Aritome S., Shirota R., Hemnik G., Endoh T. and Masuoka F. (1993) "Reliability issues of Flash memory cells". *Proc. of the IEEE*, **81**, 5, pp. 776-788.
- [68] Ong T.C., Fazio A., Mielke N., Pan S., Righos N., Atwood G. and Lai S. (1993) "Erratic erase in ETOX Flash memory array". *VLSI Symp. on Tech.*, pp. 82-83.
- [69] Dunn C., Kaya C., Lewis T., Strauss T., Schreck J., Hefley P., Middendorf M. and San T. (1994) "Flash EPROM disturb mechanism". *Proc. IRPS*, pp. 299-308.
- [70] Crisenza G., Annunziata R., Camerlenghi E. and Cappelletti P. (1996) "Non volatile memories: issues, challenges and trends for the 2000's scenario". *Proc. ESSDERC '96*, Bologna (Italy), pp. 121-130.
- [71] Moazzami R. and Hu C. (1992) "Stress-induced current in thin silicon dioxide films". *IEDM Tech. Dig.*, pp. 139-142.
- [72] Maramatsu S., Kubota T., Nishio N., Shirai H., Matsuo M., Kodama N., Horikawa M., Saito S., Arai K. and Okazawa T. (1994) "The solution of over-erase problem controlling poly-Si grain size modified principles for Flash memories". *IEDM Tech. Dig.*, pp. 847-850.
- [73] Mitchell A.T., Huffman C. and Esquivel A.L. (1987) "A new self-aligned planar array cell for ultra high density EPROMs". *IEDM Tech. Dig.*, pp. 548-553.

3 BINARY AND MULTILEVEL FLASH CELLS

Boaz Eitan¹, Anirban Roy²

¹ WSI Ltd.
Netanya, Israel

² WSI Inc.
Fremont, CA

3.1 INTRODUCTION TO FLASH CELL DESIGN

The selection of a Flash cell approach is a reflection of the market and product features that a company decides to pursue. There are two major markets for Flash memories: one is the traditional embedded memory, and the other is the new emerging market of mass storage.

The embedded memory market includes stand alone memory products and the integration of memory and logic on the same product (microcomputers with memory on board, as an example). In this market, the Flash functionality follows the same trend as the traditional NVM products, ROM, EPROM and EEPROM. These applications require a complete memory array, and not even a single defective bit is acceptable in an array during product operation, as

requested by the customer. Product access times are in the range of 25–100 nS, programming time of 10 μ s/byte and erase time of a few hundred milliseconds to few seconds. The supply voltage in operation is typically $5\text{ V} \pm 10\%$, with many emerging requirements for $3\text{ V} \pm 10\%$ and even $2\text{ V} \pm 10\%$. This supply voltage is the only source of power in Read, Programming and Erase (dual power supply Flash products are loosing appeal). Operations must be guaranteed across all the temperature ranges dictated by commercial, industrial or military standards. Redundancy concepts to enhance yield are required. As will be explained in this chapter, Flash redundancy is much more complex to implement compared to EPROM or volatile memories like SRAM and DRAM.

The size of the embedded market is by far larger than the mass storage one, despite early predictions of a major storage market. The key element in this last market is cost per bit. So far, the complexity of solid-state Flash memories, and the continuous and consistent reductions in the cost-per-bit of magnetic disks have hindered the wide spread acceptance of Flash as a storage replacement. Nevertheless, the Flash memory technology is widely penetrating into markets like digital cameras, palm size computers and medium density mass storage. Moreover, the system support of Flash memory helps in reducing costs. System management allows replacing defective portions of the array before and during the product operation, sharing the control functions between many dies on a board, implementing multilevel cell concepts. These are only some examples of system related cost reductions.

The product features in mass storage applications are different from the ones required for embedded memory products. Access times in ms are acceptable, with much more emphasis on the programming speed. Parallel programming is required for digital video cameras. Array segmentation that reflects a 512 byte magnetic disk segmentation is a key array architecture for this market.

Market requirements are clearly a key factor in the selection of cell and array concept. The company past experience plays a significant role, as well. The complexity of Flash is so overwhelming that fall backs on past NVM experience may significantly reduce the time to market of products. In this chapter, we will attempt to evaluate Flash cell and array architectures from the point of view of device functionality, process complexity and applications. To simplify the discussion, we will divide the chapter into two sections. In the first section, we will discuss Flash concepts in a single bit per transistor implementation. This will allow us to focus on array functionality, process and product implementation. In the second part, we will focus on the multilevel compatibility of the same concepts as an add on feature.

3.2 BINARY FLASH CELLS

3.2.1 *Figures of Merit*

In describing Flash cells and their operations, we will refer to a set of figures of merit [1-4]. This will simplify the discussion and will serve as a mean of comparison. In Chapter 2, a very thorough discussion on the Industry standard cell and array (the T-shaped cell) has been presented. We will use this architecture as a reference for our analysis.

The figures of merit which will be used in the following to evaluate different solutions are:

1. Cell size;
2. Process complexity;
3. High Voltage requirements;
4. Array efficiency;
5. Redundancy implementation;
6. Low Voltage simplicity;
7. General purpose or dedicated application.

To generalize the terms in which we compare the *cell size*, we will refer to absolute size in F^2 (Feature size squared). This concept has the advantage of establishing a common term of comparison. The disadvantage is in the assumption that this normalized cell size is invariant under design rule scaling. Once the cell size is established the next question is *how complex is the process*. This figure of merit focuses on the manufacturing of the cell. Another important figure of merit is the value of the *maximum voltages* required to operate the cell. This figure of merit is very important, since it reflects the level of complexity of the CMOS process needed to sustain the high voltages. It also has implications on the circuitry surrounding the array and on the reliability of the memory itself. The high voltage requirements are orders of magnitude more stringent than for an EPROM. Flash technology requires reliable oxide with an endurance of 100,000 cycles or more with high voltage stress [5].

The *array efficiency* and the *redundancy implementation* are two figures of merit that correlate the cell and array complexity to the product implementation and yield. In case the cell size reduction is accompanied by an enhanced array complexity, the array efficiency will be poor. In Flash memories, the redundancy implementation is critical: many single cell failures need great

flexibility in the redundancy. Many concepts with limited natural bitline segmentation suffer from lack of row redundancy.

The *low voltage implementation* depends on charge pumping techniques for voltage boosting during programming and read. Issues like maximum voltage, maximum currents and read speeds determine the ease of applying a given approach to lower supply voltage. In logic products, low supply voltage is a reflection of the internal voltages of operation. In Flash this is true only for the logic part of the chip, while the array will need several internally generated voltages for programming, erase and read. No Flash concept to-date is a true low voltage for all operating functions.

Based on Section 3.1, it is clear that the Flash market is not yet settled. We think it is advantageous to have a *general application Flash concept* rather than a dedicated market one. It is clear that, if in the future partial markets will become big enough, this benefit may no longer be relevant. Finally, scalability has not been identified here as an independent figure of merit, since it results from the sum of the previously listed figures of merit.

3.2.2 Cell Design Complication Hierarchy from ROM to Flash

The product functionality increases from ROM to EEPROM, from a read only memory to a byte programmable, erasable read only memory. The impact of the enhanced functionality leads to increased cell and array design complications as we migrate up the functionality ladder. In order to fully appreciate Flash cell design, it is imperative to compare its cell design constraints to other NVM cell constraints. Fig. 3.1 shows the device schematic of ROM, EPROM and Flash cells.

As the most significant difference, the ROM cell/array never "sees" high voltage. It is significantly simpler than EPROM or Flash. ROM is the only NVM concept where all internal voltages can be scaled to the external supply levels. In ROMs, there is no floating gate for charge storage, and this significantly simplifies the process: no charge leakage considerations, no disturbs, and no drain turn-on issues are used to limit channel length scaling.

The EPROM, incorporating the electric programming capability, is a big step up from ROM. The issues of high voltage, floating gate reliability and process complexity are present but in a simpler form than in Flash. A major complication as we migrate from EPROM to Flash is electrical erase. The EPROM cell virgin V_t is a fixed parameter based on cell coupling ratios and floating gate threshold voltage. The electrical erased V_t of the Flash cell, even though controllable, has a distribution on the array since the erase operation is carried out on an entire memory block or sector at a time ("bulk" erase). The cell design has to be carried out with the goal of minimizing coupling variations.

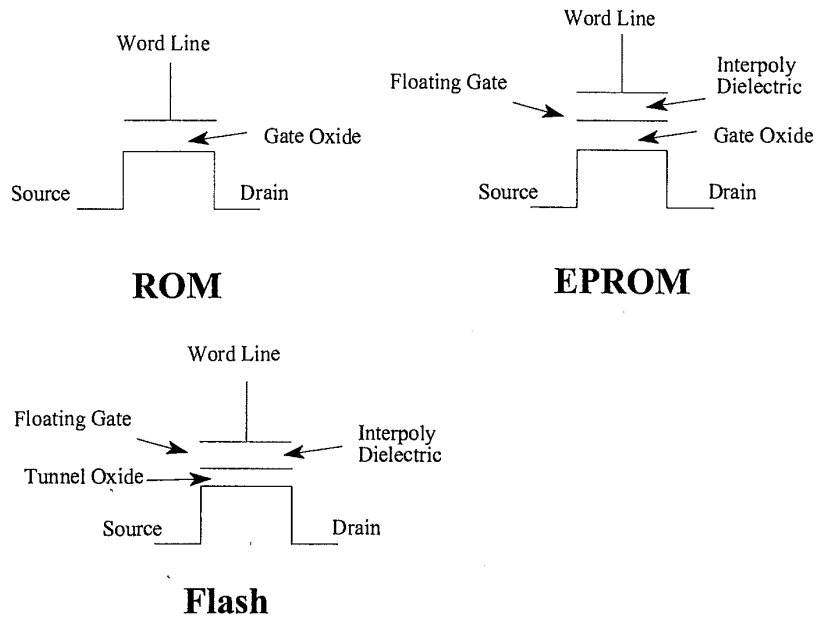


Figure 3.1 Device schematics for ROM, EPROM and Flash cells.

This is the main motivation to move to self aligned source oxide etch as we scale down source erase Flash cells, for example. The electrically erased V_t and its distribution complicate the circuit and product design of Flash with respect to EPROM.

The electrical erase of Flash cell with a thin tunnel oxide needs an extended overlap of the floating gate on the field oxide to obtain a coupling ratio capable of providing good programming, erase and read performance. If the diffusion pitch is the limiting pitch, this becomes a challenge as we scale down Flash cell size to approach EPROM cell size. To minimize the total program-erase cycle time, Flash cells must also exhibit superior programming behavior with respect to EPROM cells (for which programming time during wafer sort is the major consideration). The Flash cell operates with high voltages on both source and drain, and this imposes tough junction engineering requirements.

A large variety of possible "disturb" effects adds on the complexity of Flash over EPROM. The Flash array has to be designed in view of the Flash cell capability to achieve immunity to such disturbs. This dictates the need for array segmentation into sectors both during programming and electrical erase.

3.2.3 Basis for Flash Cells/Array Classification

There are different ways one can classify cells, viz., by the kind of system application of the memory chip or by the physical mechanisms for memory operations, like tunneling and hot electron injection. In fact, different Flash memory vendors often quote subtle variations for the same generic Flash concept, which we will refrain from discussing in this chapter.

According to the classification introduced at the beginning of this chapter [6], there are two distinct classes of Flash erase, one for general purpose stand-alone applications, and the other for specific applications like mass storage or logic with embedded memory. Stand alone Flash memories have to possess all-around good characteristics for cell size, program and erase speed and reliability, read current and peripheral overhead. The mainstream Flash memory put special emphasis upon one attribute or another while compromising on the rest, e.g., Flash memory for mass storage application puts big emphasis on cell size and redundancy, while less emphasis on program and erase speed or read current. On the other hand, for logic with embedded Flash the emphasis is put on process and device integration [7]: the relative size of the embedded Flash memory with respect to logic defines how much effort has to be spent on cell size reduction and which will be the cost increase due to the integration of Flash technology with logic or other technologies. Nevertheless, the present trend towards higher Flash densities also for embedded applications emphasizes strongly smaller array concepts.

As far as a classification according to memory function is concerned, programming and erase mechanisms of Flash cells are a good discriminating benchmark. The programming mechanisms most widely adopted are channel hot electron injection and Fowler-Nordheim tunneling, while the most widely utilized erase mechanism is Fowler-Nordheim tunneling to either the silicon channel (or to a junction) or to another polysilicon layer. The details of programming and erase mechanisms are discussed in details in Chapter 2. This approach to Flash cell classification is partly historical in nature. Flash cells have evolved along two different paths, one coming from the UV-EEPROM, which is channel hot electron programmable and UV erasable, while the other is coming from EEPROMs, which traditionally have relied on tunneling for both programming and erase. Another attribute of the evolution process from EPROM to EEPROM that becomes apparent in Flash cells is the choice of a one transistor cell or of a two transistor (or split gate) cell.

As the drive towards lower supply voltages continues, low current programming mechanisms are being favored. Thus, as we scale down the power supply, the channel hot electron mechanism with highest injection efficiency (e.g., source side injection) and the Fowler-Nordheim (FN) tunnel programming

mechanism become the two dominant competing mechanisms driving Flash development. Since tunneling is a relatively slower process, many bytes (or pages) must be written in parallel to attain competitive programming times relative to channel hot-electrons programming. The two main disadvantages of FN-tunnel programming are the need of high electric fields, and the wider final distribution of programmed V_t due to the multibyte programming algorithm. These two factors are crucial to determine device reliability and endurance characteristics.

The disadvantage of channel hot-electrons programming is that the programming current is still many times higher than the one in FN tunnel programming. Today, we are seeing a renewed interest in other hot electron programming mechanisms which could allow to reduce the programming current, like substrate hot electron injection. The choice of the programming mechanism becomes very important in implementing multilevel Flash operation (to be discussed in detail in Section 3.3.2). In the choice of the programming/erasing mechanism another constraint for Flash cells has to be considered, which is not required for EPROMs: for Flash, program and erase operations have to be guaranteed at any temperature in the operating range, while there are only room temperature programming requirements for the EPROM array.

The number of EPROM array program/erase cycles is limited to 1,000, while Flash array cycles can be as high as 1,000,000; this is the biggest incremental difference from EPROM to Flash, which is reflected into much tighter requirements concerning the integrity of the cell after program/erase cycles. Dielectric reliability under high field stress and charge injection must be guaranteed for Flash cells. The impact of cycling requirements on Flash cell design is evident, for instance, in the tailoring of the graded source junction of the industry standard Flash cell.

For Flash memories, the possible array configurations can be classified as having a *common ground* or a *virtual ground* architecture. *Common ground* architectures implement a bit-line contact and a separate diffusion source line every two cells. The diffusion source line is connected to ground through a dedicated metal line every sixteen bit lines. *Virtual ground architectures* adopt one contact every 64 cells with the diffusion lines serving as drain of one column of cells or as the source of an adjacent column of cells (like in the split gate, triple poly, Virtual Ground architecture). More advanced Virtual Ground arrays adopt one metal line every two bit lines like in the Alternate Metal Ground, AMG architecture. The common ground array is traditionally the most popular array architecture, but with technology (density) scaling virtual ground array architecture are becoming increasingly favorable. The main motivation for virtual ground array architecture is the smaller effective size.

Virtual ground arrays have diffusion bitlines which are buried beneath the control gate polysilicon layer. The major contribution to smaller effective cell size reduction is the dramatic reduction in the number of bitline contacts in the array by a factor typically ranging from 8 to 64. The channel length of virtual ground cells is defined primarily or in conjunction with the floating gate polysilicon layer. This is a difference from the common ground cell, where the channel length is defined by the self aligned double poly etch. The presence of bitline contacts shared by 16–128 cells leads to an increased bitline resistance for the virtual ground arrays. In comparison, the common ground array cells have negligible drain resistance but considerable source resistance due to the source contact which is shared by 8–16 cells. The bitline resistance and its associated impact on read current and programmed V_t are an attribute for array performance.

3.2.4 Detailed Description of Flash Cells and Architectures

Following the classification reported in Fig. 3.2, we see that the memory array can be implemented in NOR, NAND or AND configurations. The industry standard has a NOR, common ground array configuration. We will consider in the following other NOR common ground implementations of the array: Source Injection Flash memories, DINOR (*DIVided bit line NOR*) and EEPROM-like Flash, as well as three variants of NOR virtual ground arrays, namely AMG, Split Gate and *Asymmetrical Contactless Transistor* (ACT). As it will be shown in the following paragraphs, the virtual ground concept is present also in the NAND configuration and in the AND arrays.

The “programmed” state of the cells is identified here as the result of the operation which can be carried out byte by byte, since this is the convention usually adopted by the developers of the various types of architectures. According to the array organization and operation, this can correspond to either a high V_t or a low V_t state of the cell. It should be noted, however, that the JESD 100 “Terms, Definitions, and Letters Symbols for Microcomputers and Memory Integrated Circuits” and the IEEE Floating Gate Array Standard P1005 (draft version) always identify “program” as the operation of injecting electrons onto the floating gate of the memory cell (thus increasing V_t) and “erasing” as the opposite operation. In the following, we will alert the reader when the terminology adopted deviates from the IEEE standard above quoted. According to the terminology we have chosen, the NOR arrays have a “programmed” state which is high V_t (same as IEEE standard), while the DINOR and AND arrays have “programmed” state which is low V_t (so the IEEE standard calls this state “erased”).

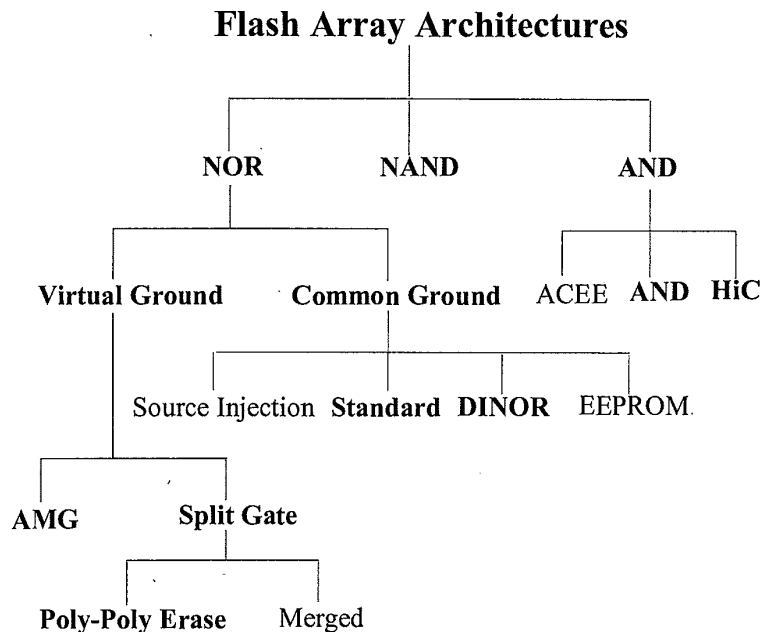


Figure 3.2 The family tree of Flash array architectures. The NOR architecture variations are the majority of the concepts. The Common Ground is the predominant concept in production today. The ACEE (Array Contactless EEPROM) and HiC (High Coupling ratio) are among the many variants of the Flash cells developed.

The choice of the cells which will be discussed in detail in the following has been based on their relative importance, judged by the number of companies that adopted it, or on their significance as a basic concept. We will start with a description of the standard array focusing on the cell and array features. This will assess the terminology adopted and will be used as a reference throughout the chapter.

Industry Standard Common Ground NOR

This concept was extensively described in Chapter 2. The cell is based on the standard EPROM cell and array architecture, and adopts the same channel hot-electrons programming mechanism [8–10]. Erasing is accomplished through Fowler-Nordheim tunnelling of electrons from the floating gate to the source. The cell design relies on many years of experience accumulated on EPROM production, which is an important part of its appeal.

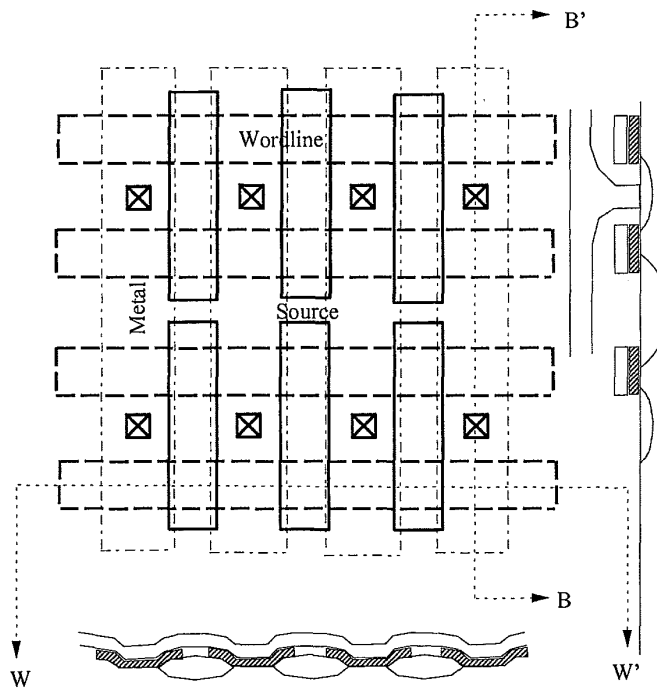


Figure 3.3 The schematic layout of the industry standard common ground NOR array [3]. The polycide wordlines traverse left to right (dashed borders in the figure), while the metal bitlines (dash-and-dot borders) traverse top to bottom. Notice the sharing of bitline contact between two cells and the definition of the poly1 layer which is etched twice; the first etch defines the width while the second etch (self aligned to the wordline) defines the channel length.

The array layout and the corresponding cross-sections along the bit-line B-B' and along the word line W-W' are shown in Fig. 3.3. In the layout of Fig. 3.3, the metal bit lines (BL) are the dashed vertical lines and the horizontal lines are the polycide word lines (WL). There is a BL contact every two cells and there is a diffusion source line every two cells. The diffusion source line, which is parallel to the wordline, is connected through a dedicated common source metal line every sixteen bitlines.

Source and drain regions are self-aligned to the 2nd-poly/ONO/1st-poly stacked gate structure. Among the critical design rules, the metal bit line pitch with a contact ranks high. Another critical feature is the distance of the contact to the double poly gate stack. The diffusion pitch is also a critical design rule. Obviously there are many other design rules that affect this cell

size. The $0.6\text{ }\mu\text{m}$ cell size is $3.6\text{ }\mu\text{m}^2$, the $0.5\text{ }\mu\text{m}$ generation is $\approx 2.8\text{ }\mu\text{m}^2$ and the $0.35\text{ }\mu\text{m}$ generation is $1.2\text{ }\mu\text{m}^2$. All the above sizes can be translated to 10 F^2 as the cell's figure of merit. Fig. 3.4 shows a sketch of the array organization.

The schematic cross-section in Fig. 3.5 refers to the n^+ source region between two wordlines and allows the identification of the key critical process steps which characterize this cell. The first critical point is the definition of the polycide word lines, which consists in a *self align* step needed for defining the

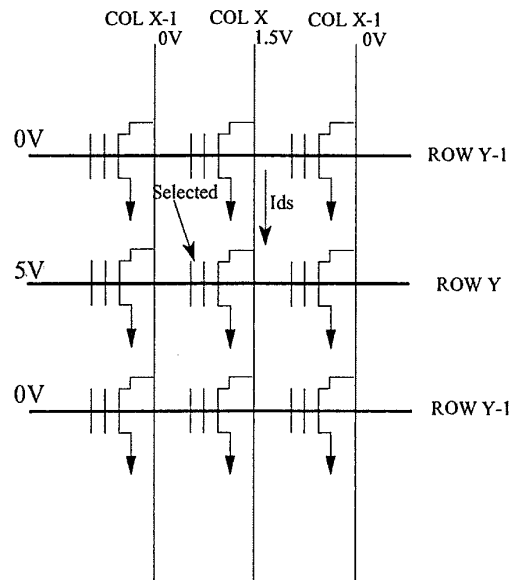


Figure 3.4 The device schematic of the layout in Fig. 3.3 [3]. The cell during read/programming is selected with a combination of wordline and bitline, while the erase is done for the whole block of the array.

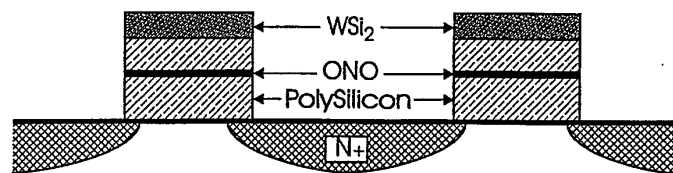


Figure 3.5 The basic process steps involved in fabricating the standard common ground NOR array [3]. The figure illustrates the misalignment critical masking required for the graded source junction.

floating gates. The difficulty of this step is in maintaining a vertical profile over a very thick double poly feature that crosses over the steep slopes and thick field area. The control on the vertical profile of the stacked gate defines the channel length. In order to minimize the cell size, the source side is defined through a self aligned etch of the field oxide to the word line. The difficulty in this process step is that the selectivity of the oxide etch to the silicon etch is very critical since the source line is defined through diffusion and field areas. Another criticality of this source definition step is the mask definition in the middle of the narrow word lines (Fig. 3.5 bottom).

To summarize, the critical technology steps in defining the cells in and "industry standard" array are: 1) Contact between the two cells, 2) Double poly etch, 3) Self aligned source etch and mask, 4) Tunnel oxide growth and quality, and 5) Tight metal pitch. The combination of so many critical process steps and design rules results in the relatively large cell size.

Functionally this array is very simple. Its schematic diagram is shown in Fig. 3.4 and described in Tab. 3.1. The read and programming operations differ only by the voltage levels. The erase operation is on the source side. Separating the programming and erase junctions helps in the optimization of both mechanisms and improves the endurance reliability. Negative gate voltage is a common feature in the erase operation.

Table 3.1 Different voltage levels required to perform the various array operations, viz., read, programming and erase in the standard common ground NOR array.

MODE	BL	WL	SL	COMMON
READ	1V	Vcc	Gnd	
PROGRAMMING	6V	10V	Gnd	
ERASE	Float	-10V	Vcc	Block

Two types of array segmentations are being implemented, generating either BL or WL blocks. In the BL block, a group of full BL's serve as a segment [9]. This solution is very natural to the architecture and metal 2 can serve for WL strapping providing speed enhancement. The WL block is achieved by segmenting the BL into 256 bits local BL's (metal 1), connected through select transistors to metal 2 Global BL's. This segmentation adds about ten percent of the array size [10].

The industry standard array can be used for both embedded and mass storage applications. However, it is a better fit for the embedded market. Redundancy effectiveness is a key parameter in product cost structure. For embedded memory there can be two types of redundancies, viz., column and row redundancy [11]. In the case of mass storage products, redundancy can be achieved

via system level concepts that do not require a perfect die. While discussing redundancy we will always refer to the first concept which is the perfect die rather than the concept of the mass storage market. For the industry standard cell, the column redundancy is very effective. Disturb problems and lack of natural segmentation limits implementation of effective row redundancy.

As is explained in Chapter 2, the erase V_t distribution control is one of the most significant problems in implementing this Flash concept. In order to enlarge the tolerance to the V_t distribution, the maximum V_t of the cell can reach a value as high as 3 V, hence charge pumping the WL in low supply levels is essential. The methods to implement this feature, though well established, result in speed and power penalty [9]. The array efficiency, defined as the percent of the array size relative to the die size, for a 16 Mbit product is about 51%. This is an important figure of merit that we will review as we analyze the different array architectures. This relatively low array efficiency is the consequence of the complexity in Flash product implementation. A complex controller is needed on board to guarantee programming and erase across specified temperature and supply voltage variations, without affecting the product performance.

In Tab. 3.2 the common ground figures of merit are summarized. The cell size is large since it is 2.5X larger than the minimum geometry (for a process with minimum feature size F , the minimum geometry is $4F^2$). This is not the most difficult concept, even-though the process complexity is high. Low voltage implementation is suffering from the high current requirement in programming and WL voltage pumping in read. The array efficiency for this array is better than for most other, but lags behind the pure mass storage products.

Table 3.2 A figures of merit summary for the standard common ground array. A higher grade means worse performance. This is a robust technology, widely adopted in production.

FEATURE	GRADE
Cell Size	3
Process Complexity	2
High Voltage (On-Chip)	2
Array Efficiency	2
Redundancy	2
General Purpose	1
MultiLevel	1
Low Voltage	2
TOTAL	15

It is very important to realize that currently the common ground architecture is by far the highest volume Flash memory in production.

NAND Cell and Array

The NAND array reduces the cell size with respect to the NOR, *parallel*, common ground architecture, by connecting the cells *in series* between a bit line and the source line. The NAND array architecture is a mass storage concept where scaling is achieved by word line pitch scaling [12–16]. The word line pitch scaling can be better understood with the help of Fig. 3.6. Scaling is achieved by the positive effect of two factors: 1) a significant reduction of the total number of contacts, from one contact every two cells to one contact every sixteen cells; 2) the reduction of the requirement on the punch-through immunity of the cell. The outcome is a significant reduction of the cell size from $10F^2$ to $6F^2$. In order to understand the dramatic improvement on the reduction of the number of contacts, we will use the schematic diagram of the array shown in Fig. 3.7. The unit block is formed by sixteen elementary cells in series with two select transistors, i.e., the ground select transistor (GSL) and the bit line select transistor (SSL). The read is accomplished using the “Read Through” concept: the fifteen cells neighbor to the selected cell serve as pass gates. This can be viewed as if every cell has two functions, viz., to store the information and as well as to be a pass gate when reading its neighbors.

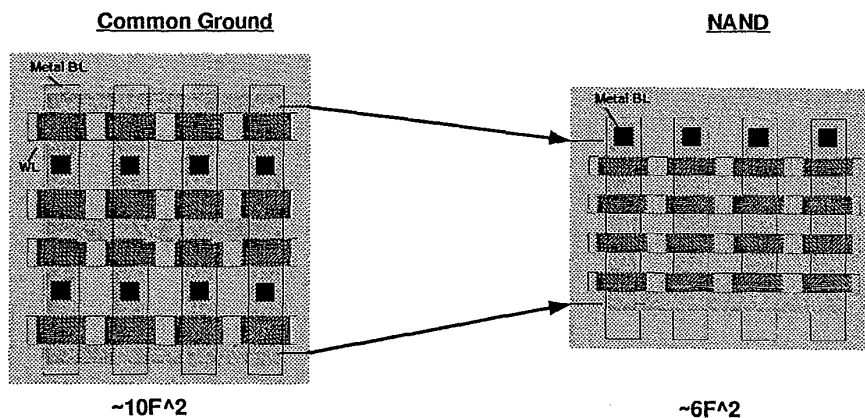


Figure 3.6 Illustration of the wordline pitch scaling in the NAND array. The cell size scales down from a $10F^2$ for the standard common ground to $6F^2$ for the NAND [1]. Note the reduced number of contacts.

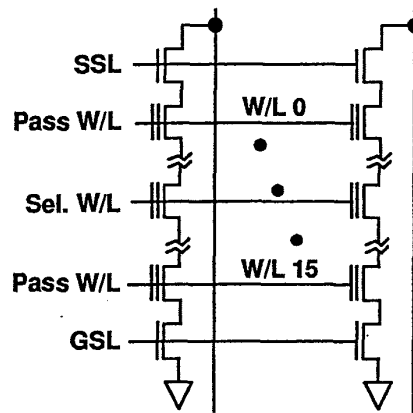


Figure 3.7 The device schematic diagram of the NAND array. The group of 16 cells are connected to the global bitline with select transistors and the selected cell access requires the neighboring cells to provide a read through [14].

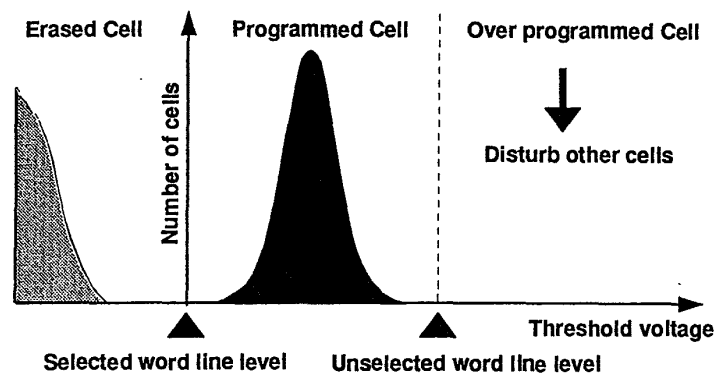


Figure 3.8 The programmed and erased array V_t distribution for the NAND array. Over-programmed cells can potentially affect the read access of the other cells [13].

Fig. 3.8 illustrates the above point with the different wordline voltage levels. The cell to be read has a grounded WL (selected wordline level), the other cells (the pass gates) have the unselected level of 4.5 V. In the "0" state the cell has a threshold voltage higher than 0 V (programmed cell), and in the "1" state the cell has a negative threshold (erased cell). The programmed level has to be lower than the pass gate voltage, to enable the neighbor read. The logic level of

the cell is detected by a sense amplifier connected to the bit line. It is clear that "Read Through" is a slow read concept, and the control of the programmed threshold distribution is a key issue. The random access speed is slow, around $10\ \mu\text{s}$. This is the main reason why the NAND concept is dedicated to the slow mass storage applications.

Programming is achieved through FN tunneling of electrons from the channel to the floating gate and results in a positive threshold voltage. Erasing is obtained by FN tunneling of electrons from the floating gate to the channel resulting in a negative threshold voltage V_t . Erasing requires biasing the p-well and the n-substrate with a high positive voltage ($\approx 20\text{ V}$). The consequences of this mode of operation are: a) the cell does not employ hot electron programming, the bias voltage applied to the drain is relatively low, and punch-through does not represent a problem; b) since channel tunneling instead of source tunneling is adopted, standard (not-graded), lightly-doped, source (and drain) junctions can be used; both a) and b) allow shorter gate lengths with respect to the standard cell, and improved scalability. Other relevant factors are: c) as tunneling is the program/erase mechanism, low currents are involved, and high voltages can be generated by charge-pump techniques; d) the erase procedure requires high voltages applied to the p-well; as a consequence, the control CMOS circuits and the NAND cell array must be located in different p-wells. The p-well which hosts the logic circuits is always at ground.

The "Read Through" concept, however, complicates the conceptually simple erase/program procedures based on F-N tunneling from/to the channel for both programming and erase. "Channel" F-N involves low coupling ratio and very high voltages. In the following, we describe program and erase operations and their drawbacks.

Programming is achieved by selective F-N tunneling from the channel to the floating gate. In Fig. 3.9, the programming of one bit and the inhibit of the others are shown. The programming is done by applying a high voltage (15–20 V) on the word line and grounding the channel. The grounding of the channel is obtained through the neighbor cells by applying an intermediate voltage level (10 V) to their gates and grounding the metal BL. To inhibit the devices which share the same word line but should not be programmed, the bit line of these devices is supplied with V_{cc} , thus turning off the corresponding select device (see Fig. 3.9). Through coupling from the gate to the channel, the channel potential of the program-inhibited cell is increased to 8 V and hence the potential drop across the tunnel oxide is less than the value which could initiate tunneling. This type of dynamic inhibit puts severe limitations on the cell disturbs: all the pass gates see 10 V on their gates with possible ground potential on their channel; also the unselected cells sharing the same word line see a similar voltage difference, with possible weak programming of the cell.

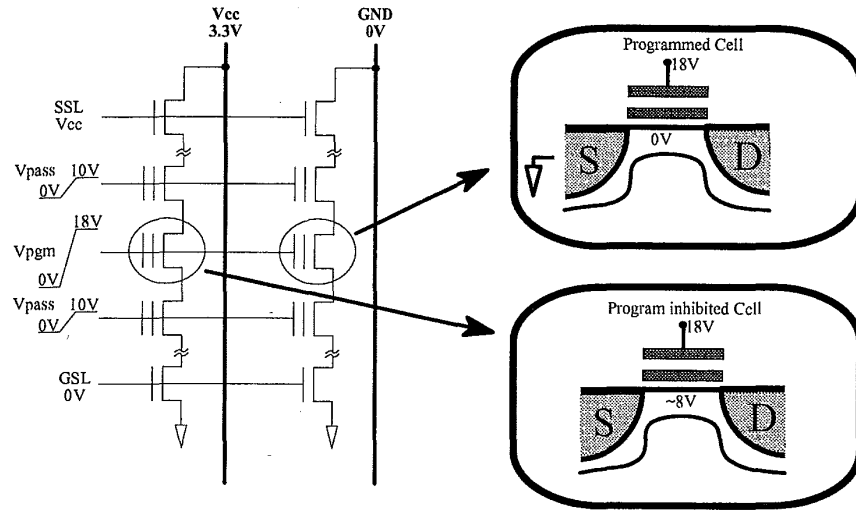


Figure 3.9 Dynamic program inhibit concept for the NAND array. Deep-depletion under the floating gate of the cells belonging to the wordline being programmed reduces the voltage stress for the gate disturb [14].

The latter inhibit margin is based on the dynamic charging of the channel. These severe disturb conditions and the tight requirement on the programming threshold distribution make this technology very sensitive to the coupling and tunnel oxide thickness variations.

During erase, a very high voltage is applied to the array well, with all the word lines grounded. The very high voltage is required because of the relatively low gate to substrate coupling. The functionality of the NAND architecture is illustrated in Tab. 3.3.

Table 3.3 Truth table for the NAND array summarizing the different voltage levels required to perform the various array operations, viz., read, programming and erase.

MODE	Sel. WL	Pass WL	SSL	GSL	"0" BL	"1" BL	Bulk
READ	Gnd	4.5V	4.5V	4.5V	1.8V	0.7V	Gnd
PROGRAMMING	15.5-20V	10V	Vcc	Gnd	Gnd	Vcc	Gnd
ERASE	Gnd	Gnd	Float	Float	Float	Float	21V

The process is relatively simpler than the common ground architecture. It is a single metal, triple well and double poly process. By reducing the number of contacts in the array, the design rule of contact to double poly is less critical. There is no need to have a self aligned source. The process complexity arises

from the requirements on a good and tight control of the coupling and on the use of very high voltages (above 20 volts) in the CMOS periphery. Product implementation for mass storage applications is simplified by the array natural segmentation. The small basic block makes this architecture very effective in incorporating both column and row redundancies.

A single power supply can be used for low voltage applications, but it requires WL charge pumping to voltages higher than in any other array architecture. This is due to the need of reading through the cells without introducing too high series resistance. Moreover, channel F-N tunneling requires extremely high voltages for program and erase, which add on other severe requirements on voltage pumping. This makes the charge-pump circuitry a significant part of the periphery.

The "Read Through" mechanism is not compatible with high speed applications. The array efficiency is good, with a value of 55% for the 16Mbit mass storage product. An evaluation of the performances of this architecture is shown in Tab. 3.4. The key difference between this architecture and the common ground array is a much smaller cell size with a more complicated read operation. This limits the application of the NAND architecture to mass storage only. Its advantage in redundancy implementation is important for this kind of application; the disturb margin, however, is very narrow and may cause significant yield problems.

Table 3.4 A figures of merit summary for the NAND array. This memory concept is well suited for mass storage applications, but suffers by high voltages and slow speed for read access. A higher grade means worse performance.

FEATURE	GRADE
Cell Size	1
Process Complexity	2
High Voltage (On-Chip)	3
Array Efficiency	2
Redundancy	1
General Purpose	3
MultiLevel	3
Low Voltage	2
TOTAL	17

DINOR (DlVided bit line NOR) Cell and Array

The DINOR array allows the scaling of the industry standard array without changing the array architecture [17–21]. The scaling is both in the word line

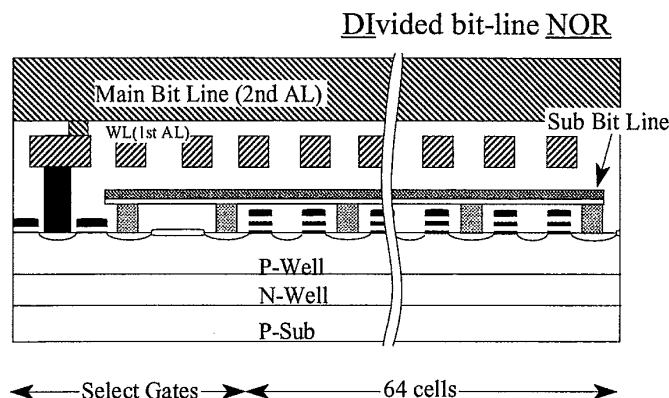


Figure 3.10 DINOR array device cross-section showing the triple well, triple poly and double metal process required to realize this array [17]. There are salicided local bitlines and the first metal is used for wordline strapping.

pitch and the bit line pitch. To achieve scaling in both dimensions, both the process and the cell programming and erase mechanisms are changed with respect to the standard NOR array. This is a quite different approach to scaling than the NAND, where a dramatic reduction of the word line pitch is achieved by eliminating the contacts between two bits.

A cross-section of the basic block is shown in Fig. 3.10. The global bit lines in metal 2 are connected through a select device to the local bit lines. When switched on, the local bit lines are connected to the drains of a sub-block of 64 transistors by means of a polycide line. There is one global metal BL for every two local polycide BLs (Fig. 3.11). The local bit lines are connected in the standard common ground array configuration: one contact between two word lines. There are diffusion source lines connected every sixteen cells through another global source metal line.

The BL pitch scaling is due to the smaller area required for the buried contact between the polycide local BLs and the drain diffusion layers, obtained through poly plug contacts, with respect to the BL metal/diffusion contact in the standard architecture. The overall memory array is smaller than the NOR array despite the select gate has been added to each sub bit-line.

The first metal is used to reduce the RC delay of the word lines. The poly WL resistance is high due to the limitation imposed on it by the polycide local BL. This is one of the disadvantages of the DINOR architecture.

Fowler-Nordheim tunneling of electrons from the floating gate to the drain junction, which is accomplished selectively and results in a low V_t , is usually called programming by DINOR developers. Erasing is the opposite operation,

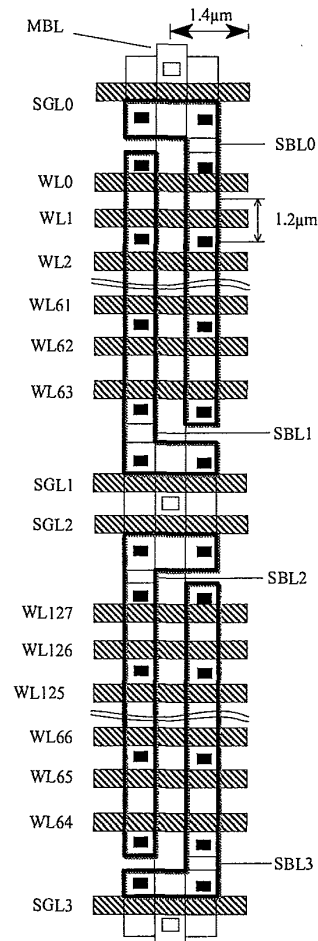


Figure 3.11 Schematic layout of the DINOR array [17]. The select transistors connect the global second metal lines to the salicided local bitlines. There is no double-diffused junction, and this helps in scaling the array pitch.

carried out via FN tunneling from the channel to the floating gate; it results in high V_t voltage and is accomplished block by block. These definitions are the opposite with respect to the industry standard NOR array and IEEE P1005 draft standard.

FN tunneling and drain programming and erase allow further scaling of the cell. Low current mechanisms are required due to the high resistivity of the polycide, so the choice of a low current programming mechanism is a consistent with the local polycide BL concept. In general, the length of the polycide is

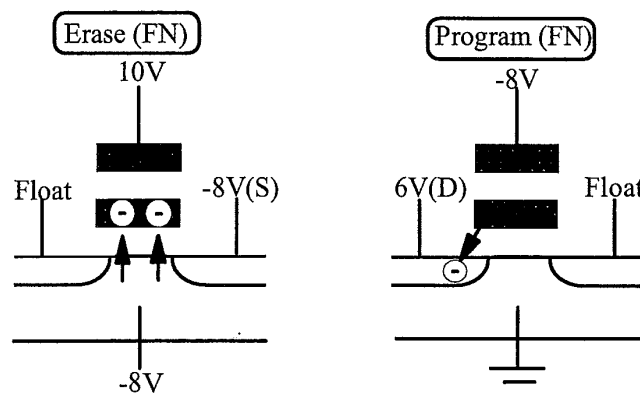


Figure 3.12 The erase and programming tunneling operations in the DINOR array [17]. Notice the programmed V_t is the low V_t state, a major departure from the standard common ground.

reduced, so its sheet resistance has a negligible effect on any of the cell electrical operations. F-N programming improves cell immunity to drain turn-on, helping in WL pitch scaling.

The array is implemented with a triple well process (Fig. 3.10). This provides the option to apply negative voltages to the p-well during erase. The triple well technology reduces the maximum voltage applied to any node, by dividing it into positive and negative voltages. The cross-section (Fig. 3.10) can be described as a triple poly concept with the third poly as the local bit line to reduce the bit line pitch. The metal 2 serves as a global bit line with one metal line for every two local bit lines and metal 1 provides RC reduction with word line strapping. The cell size in this array architecture is $1.68 \mu\text{m}^2$ in a $0.5 \mu\text{m}$ technology, or $6.7 F^2$, relative to the $10 F^2$ of the standard Flash architecture. It is a very significant improvement in scaling.

In order to understand the scaling of the word line pitch, we will explore the programming and erase mechanisms. During programming, electrons are ejected from the floating gate to the drain junction via FN-tunneling through the gate-drain overlapped area. The selection of the programmed bit is based on the combination of negative WL with positive drain voltage. This is shown in Fig. 3.12 (programming); the typical program time is $\approx 20 \mu\text{s}$ (see Fig. 3.13a). Since programming to a low V_t state is the selective operation, byte-by-byte program and verify algorithm can be applied, with two advantages: negative threshold values or “depleted” cells are avoided and the V_t distribution is tightened. Another advantage of this programming concept is the low current mechanism. Up to 256 bits can be programmed in parallel. The only disadvantage

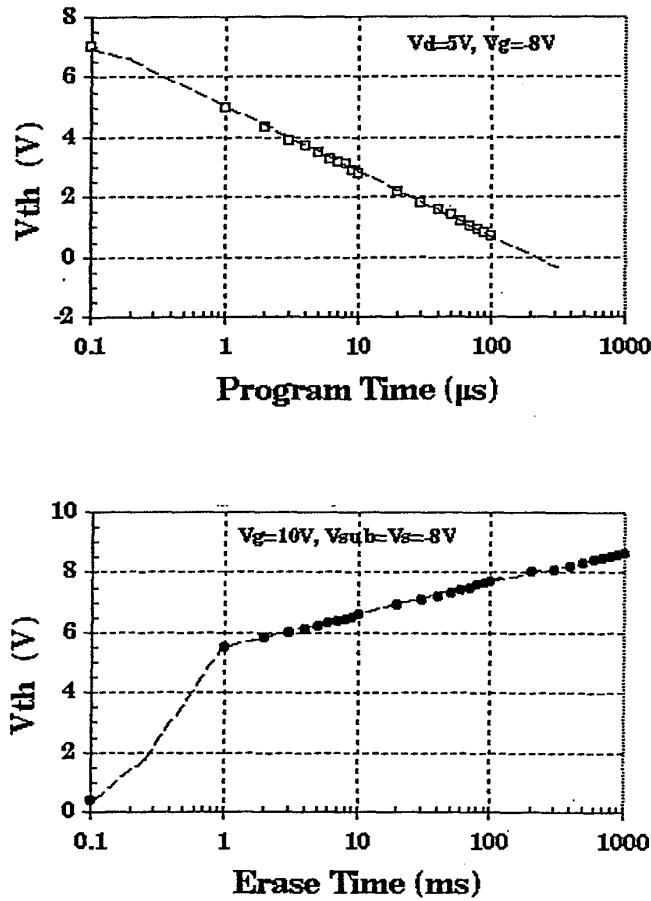


Figure 3.13 a) The programming time, and b) erase time characteristics for the DINOR array. To reduce the programming time as many as 256 bytes are programmed in parallel [18].

of this concept is the drain F-N programming, which requires very high electric fields.

Erasing – the block operation – is obtained by FN electron tunneling from the channel to the floating gate, which results in a high V_t ; the typical erase time is 200 ms (see Fig. 3.13b). To reduce the maximum applied voltage during erase, the 18 V needed to reach the required electric field are split into a positive 10 V voltage on the gate word line and a negative 8 V voltage on the source and p-well. All these voltages can be generated internally by charge-pump circuits.

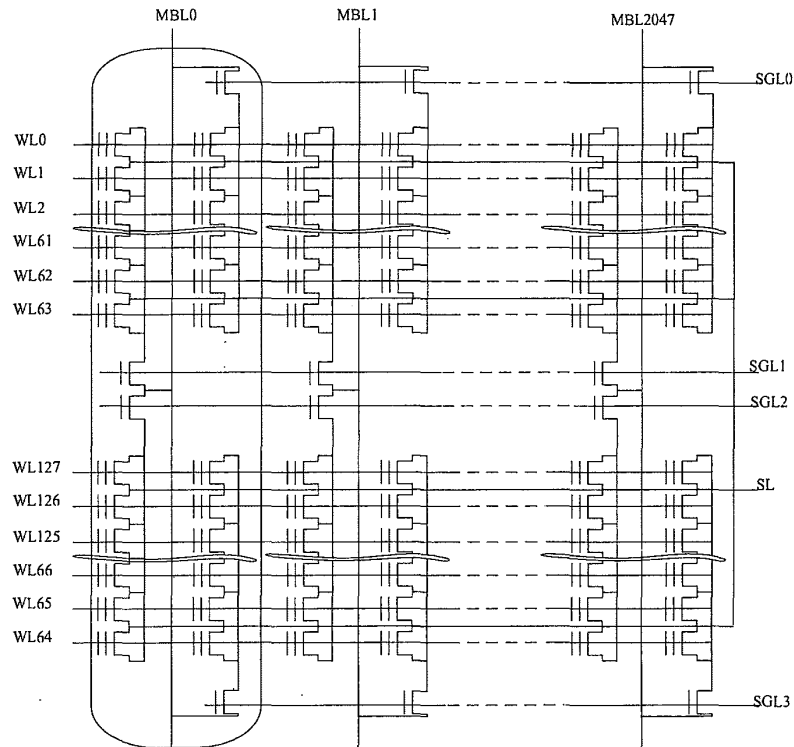


Figure 3.14 The DINOR array device schematic diagram showing blocks of 64 cells, accessible via select transistors [17].

A schematic diagram of the array architecture is shown in Fig. 3.14. Each global metal bit line serves two local bit lines. The source line connection can also be seen. Select devices add about 13% to the array size. The functionality truth table of the array is shown in Tab. 3.5. The product implementing a DINOR array can cater to both embedded and mass storage markets. Implementing redundancy in this architecture results in better solutions than in the case of the standard array since the column redundancy is very effective due to the 64 word lines segment size and the erase into the high V_t . Low voltage applications are possible with no obvious advantage over the industry standard array. During programming, the sub-block select transistors apply the positive drain voltage only to the selected sub-block. All the cells which belong to the sub-bit line of this block experience the program drain voltage; only the cell to be programmed has a negative control gate voltage; as a result, the unselected cells of the same sub-block experience a weak programming. Moreover, band-

Table 3.5 Truth table for the DINOR array summarizing the different voltage levels required to perform the various array operations, viz., read, programming and erase.

MODE	BL	WL	SL	SEL	WELL
READ	1V	Vcc	Gnd	Vcc	Gnd
PROGRAMMING	6V	-8V	Float	10V	Gnd
ERASE	Float	10V	-8V	-8V	-8V *

Table 3.6 A figures of merit summary for the DINOR array. The array size advantage is overshadowed by a large periphery. A higher grade means worse performance.

FEATURE	GRADE
Cell Size	1
Process Complexity	3
High Voltage (On-Chip)	3
Array Efficiency	3
Redundancy	1
General Purpose	1
MultiLevel	3
Low Voltage	2
TOTAL	17

to-band tunneling at the drain can lead to enhanced stress induced leakage current.

Tab. 3.6 summarizes the rating of this technology. The key advantages are clearly the small cell size, the better redundancy scheme and the fact that it is a general purpose architecture. The key disadvantages of this architecture are the small margin to disturbs, inherent to the F-N drain programming, and the extra process complexity due to the added number of polysilicon layers and the use of metal 1 in the array. The array efficiency is low at a mere 42% for 16 Mbit. This is due to the segmentation overhead and to the triple well technology required for the array.

AND Cell and Array

In the AND array architecture, the cells are connected in parallel between local bit and source lines (Fig. 3.15). This architecture achieves scaling of the word line pitch by reducing the number of BL contacts [22–26]. Compared to the common ground array, it eliminates the need for contacts every two word lines by using localized virtual ground. The typical AND array architecture has

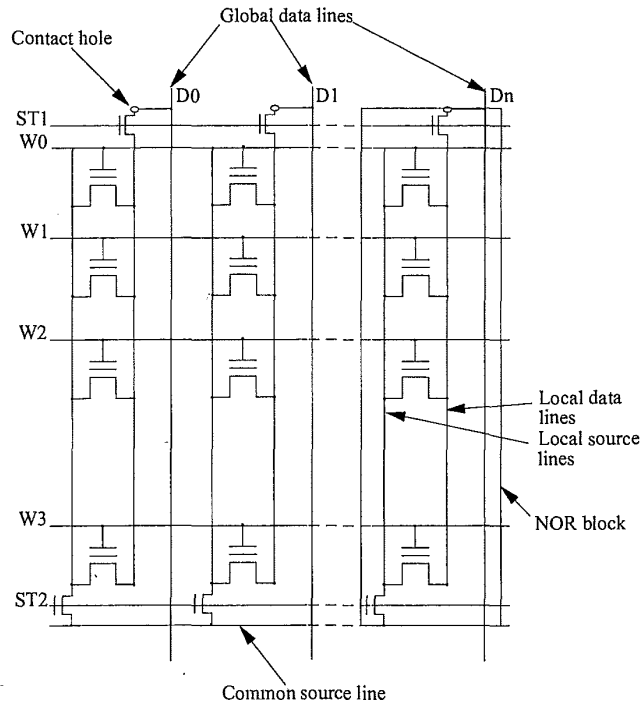


Figure 3.15 AND array device schematic showing 32 cell blocks accessible with select transistors. There are local diffusion bit lines and source lines [22]. The sector overhead for the 64 cell sector is 30%.

dedicated diffusion bit lines and source lines for every block of cells which are connected through select transistors to the global data lines and to the common source lines, respectively, as shown in Fig. 3.15. The local diffusion bit lines and source lines are implemented before the WL deposition, eliminating the need for a dedicated contact every two word lines. The size of the sector depends upon the series resistance of the diffusion lines. The select transistors and the metal contact to the global BL approximately adds a 30% overhead for a 64 cell sector.

Like the DINOR array architecture, the AND array architecture uses FN tunneling drain programming (which can be carried out byte-by-byte and corresponds to removal of electrons from the floating gate to the drain, thus decreasing V_t) and FN channel erase, i.e., injection of electrons from the channel into the floating gate, which is carried out on the whole block, and corresponds to increase of V_t . This definitions are the same as DINOR, and the opposite of the industry standard NOR array. Typically, programming is accomplished

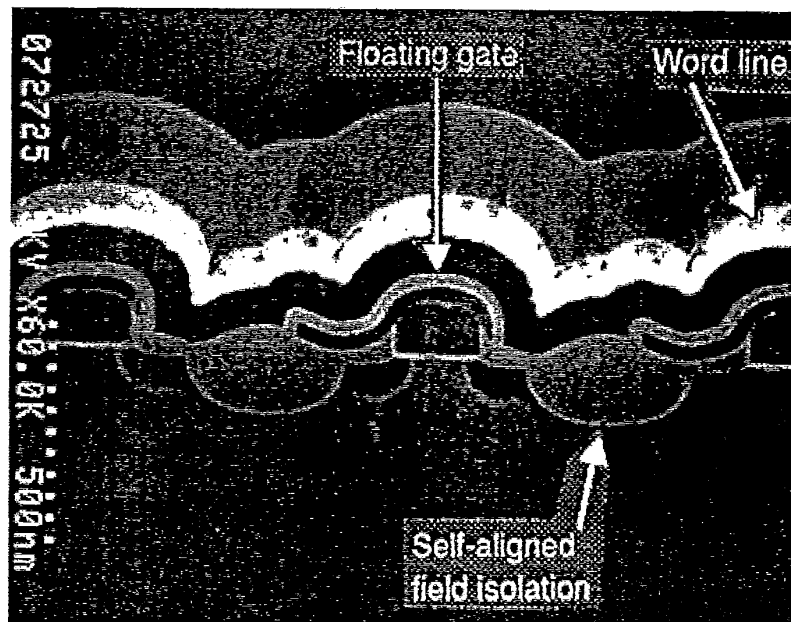


Figure 3.16 AND array device cross-section showing the self aligned field isolation, and the increased gate coupling with triple poly technology [22]. The wordline RC of the array is about 4 times than that of the common ground array.

with -9 V on the word line and $+3\text{ V}$ on the local bit line (source floating), while erasing is carried out with bitline and source line grounded, and $+13\text{ V}$ applied to the wordline. All voltages applied to the wordline are internally generated starting from $V_{dd} = 3\text{ V}$. Using program and verify, the low V_t distribution of the programmed state can be tightened. This allows read operation at 3 V , without the need for wordline boost.

This architecture has isolated segments, an excellent feature in implementing both row and column redundancies. However, to achieve the full separation between the segments, local array field isolation and dedicated diffusion BL and SL are required. The complicated morphology of the silicon area in between the bit lines, with a large amount of different features, creates high stress in the Si and limits the cell size scaling. The main advantage of the technology is in the reduction of the word line pitch, but the BL pitch is the limiting pitch. Moreover, the implementation of the gate coupling capacitor in the BL pitch helps in maintaining the cell size scaling advantage. A cell cross-section is shown in Fig. 3.16, where the self aligned field isolation is shown with the diffusion bit lines and the floating gate above it. A third poly extension of the

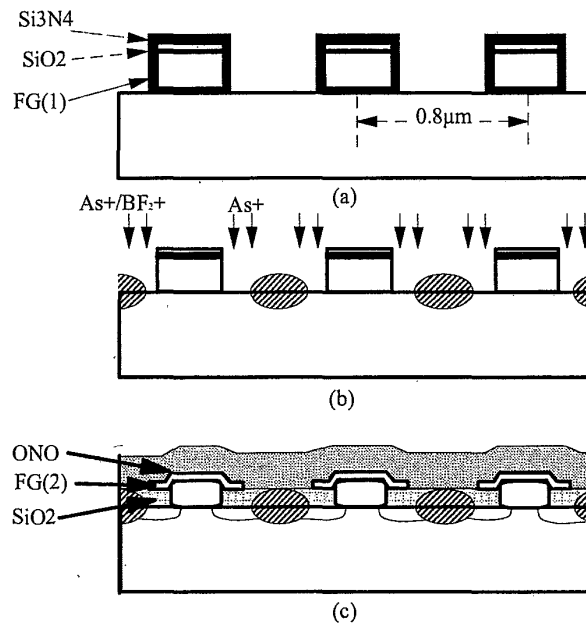


Figure 3.17 Few of the process steps in the formation of the AND array [22]. The process steps worthy of note are: a) double poly floating gate, b) tunnel oxide formation before the array field oxide, c) very thin poly-poly dielectric.

floating gate along the WL is the method used to increase the gate coupling. The typical cell size for this type of concept is $1.28 \mu\text{m}^2$ for $0.4 \mu\text{m}$ technology at 8 F^2 , and a $0.4 \mu\text{m}^2$ cell for $0.25 \mu\text{m}$ technology at 6.4 F^2 .

The floating gate consists of two layer of polysilicon, which are electrically connected. The lower layer, FG1, defines the channel length; the upper, FG2, realizes a large capacitive coupling between the control gate and the floating gate. Some of the most relevant processing steps of the AND technology are shown in the cross-sections of Fig. 3.17, which schematically illustrate the process of implementing self aligned field isolation and double floating gate. The critical steps in generating this array are the self aligned array field oxide, the double poly floating gates and the tunnel oxide growth and quality (especially since the tunnel dielectric is grown before the array local isolation). The triple poly self aligned etch to define the word line width is another critical process step. In this process, in order to eliminate the need for a triple well for the array, a very thin poly-poly dielectric is implemented, eliminating need for substrate biasing during tunneling.

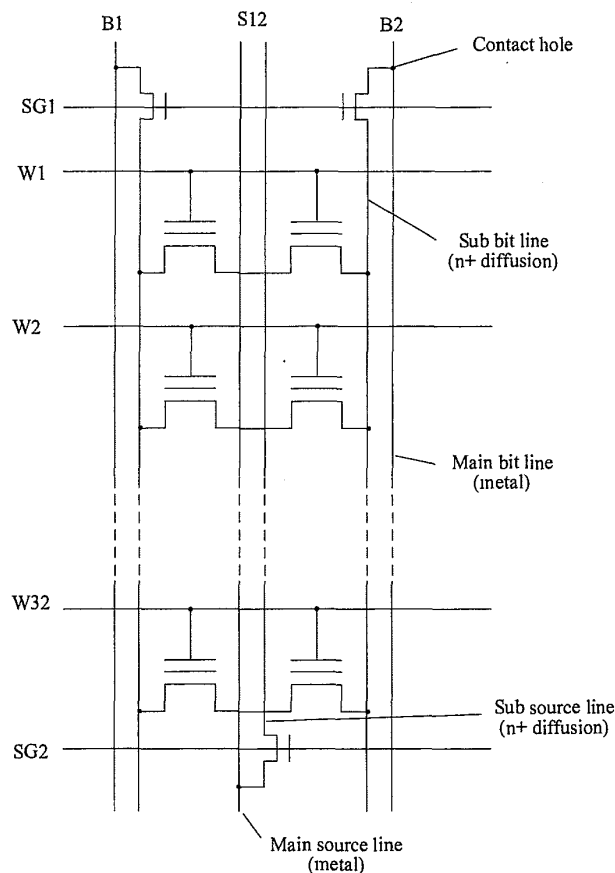


Figure 3.18 Device schematic for a modified AND array [23]. There is a dedicated source and contacts to bit lines and source lines are staggered to provide better scaling.

A modification of the basic AND array architecture is shown in Fig. 3.18. In this case, one source line is shared between two bit lines. This reduces the number of diffusion lines to one every two cells, thus improving the requirement on the tight isolation pitch. The rest of the functionality is very similar to the standard AND architecture.

The implementation of the AND architecture into a product is very similar to the DINOR architecture one, with a better efficiency in terms of column and row redundancies. The array efficiency is very similar to the DINOR case: 41% for 64Mbit product. The figures of merit of the AND array are shown in Tab. 3.7. The key advantages can be seen as the slight cell size improvement over the common ground array and its redundancy efficiency.

Table 3.7 A figures of merit summary for the AND array. This concept is well suited for mass storage applications without speed penalty.

FEATURE	GRADE
Cell Size	2
Process Complexity	3
High Voltage (On-Chip)	3
Array Efficiency	3
Redundancy	1
General Purpose	1
MultiLevel	3
Low Voltage	1
TOTAL	17

Split Gate Virtual Ground Cell and Array

The virtual ground is a concept to improve cell size by reducing the number of BL contacts [27]. It is possible to place a contact every 64 WL, or more, thanks to diffusion bit lines. The metal line with the BL contacts reduces the diffusion BL series resistance. Further scaling is achieved by eliminating the dedicated source line. In the split gate virtual ground cell, the name “virtual ground” reflects the dual functionality of the BL, both as a BL and as a SL of the neighbor cell. The bit line pitch is still limited by the pitch between metal and contact. The key disadvantage of this array is that the word lines are crossing the bit lines and there are no isolations among different cells. This complicates both read and programming conditions.

The split gate triple poly architecture [28–31] adopts programming by channel hot electron injection and erasing via poly-poly F-N tunneling. Two cross-sections of this cell are shown in Fig. 3.19. The top cross-section shows that the cell is actually a merged two transistor “split gate” concept. On the right side there is a floating gate device and on the left side there is a pass transistor. As long as the two transistors in series fit into the metal pitch, there is no cell size penalty.

The second cross-section, across the channel width, shows a third terminal which is shared between two adjacent cells. This is the erase gate. In this architecture, poly-poly erase from the floating poly to the erase gate replaces the floating gate to substrate erase. An added process complexity of this architecture is the control of a good quality active poly-poly oxide between the floating gate and erase gate, while maintaining the quality of the poly-poly oxide between the WL and the erase gate. The third poly as an erase gate prevents from using a polycide WL, which introduces speed disadvantages.

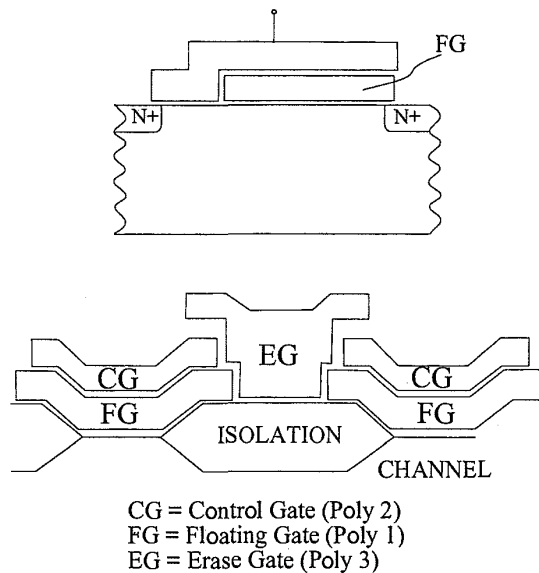


Figure 3.19 Triple poly split gate Flash cell device cross-sections: a) along the channel length, and b) along the channel width directions [28]. The first poly layer forms the floating gate, the second poly layer forms the wordlines and the third poly layer forms the erase line.

The layout of the cell is shown in Fig. 3.20. The word lines and the erase lines cross the diffusion lines, with the overlap area over the edges of the poly1 floating gate. Moreover, the extra feature and the coupling capacitor over the field isolation increase the width of the cell over the minimum word line pitch. Thus, part of the virtual ground advantage in cell area is compromised. For a $0.55\text{ }\mu\text{m}$ technology the cell size is $2.1\text{ }\mu\text{m}^2$ or 6.9 F^2 .

Programming is obtained by channel hot electrons, i.e., by raising the wordline of a selected row to 12 V and the drain bit lines of the cell to be programmed to 7 V, while the source lines are held at 0 V. Low channel currents are used, so that programming 64 bits in parallel is possible. Erasing is accomplished by tunneling, transferring the electrons from the floating gate to the erase gate, through the interpoly silicon oxide; the erase voltage ranges from 12 V to 22 V. "Program" and "erase" are identified as in the NOR standard array. The split gate architecture completely solves the overerase problem within the cell itself. The effect of the split gate architecture on the erase is shown in Fig. 3.21. Once the floating gate threshold is reduced below the pass transistor threshold, the threshold voltage V_t of the cell tends asymptotically to the V_t value of the transfer gate. While the floating gate V_t can be negative (overerased), the device as a whole is immune from this over erased state.

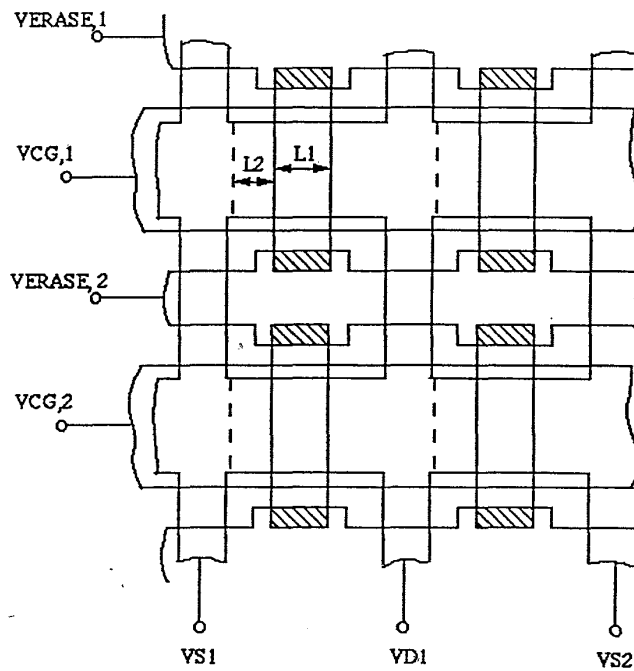


Figure 3.20 Cell layout of the triple poly split gate cell [29]. The erase line overlap on the floating gate is very misalignment sensitive.

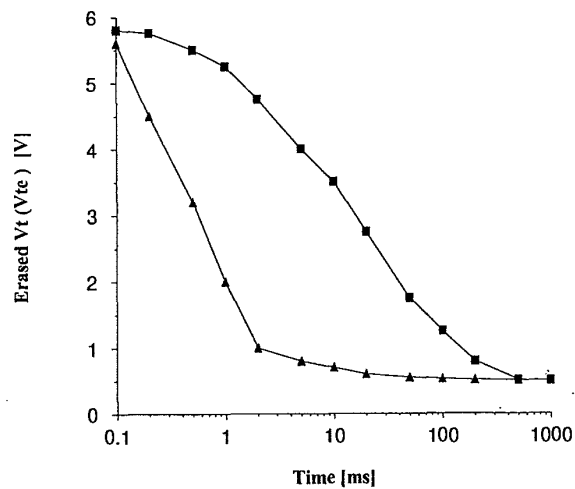


Figure 3.21 The saturated erase characteristics of the split gate cell [32]. The over-erase complications during read are avoided.

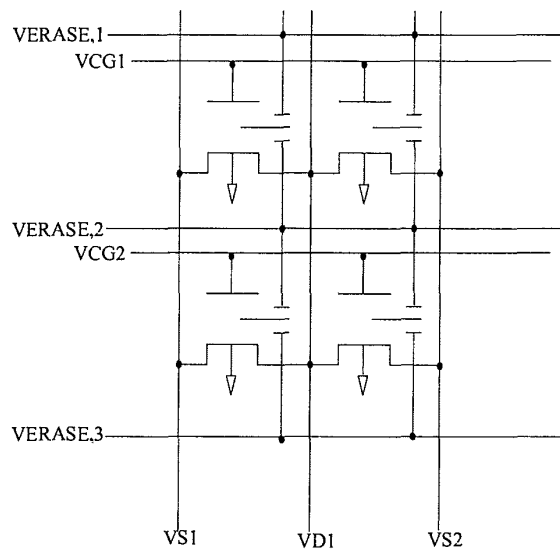


Figure 3.22 The array architecture schematic diagram of the split gate concept [29]. The array only requires positive voltages. The array also offers excellent natural segmentation, a feature useful for disk replacement products.

Another advantage of the split gate virtual ground (VG) architecture is that the “split gate” solves the program disturb problem; in particular there is no undesired programming of cells when the source of the cell is being used as the drain of another cell. Program disturb is a major issue for a symmetrical cell with a virtual ground architecture, and adds a major circuit overhead in case the cell is not inherently immune to it.

The functionality truth table of the split gate VG is shown in Tab. 3.8. The read and programming conditions are very similar to a common ground case, while the erase is unique. The wide range of erasing voltage, and its variation, is very typical of a poly-poly erase mechanism. Sensitivity of poly-poly F-N tunneling to poly grain size and trapping in the poly-poly oxide constitute the major sources of erase voltages variation. Erase voltage is automatically adapted to take into account these aging phenomena after memory cycling. The array schematic is shown in Fig. 3.22, where erase gate is the new feature. Both program-verify and erase-verify schemes can be implemented.

Difficulty in yielding a perfect die, the slow speed and very high voltages make this concept very suitable to mass storage applications. The presence of positive voltages only during programming and erase, the natural array segmentation and a reasonable cell size result in a good array efficiency. A 56% efficiency for a 18 Mbit and 65% for a 34 Mbit can be achieved. Tab. 3.9 shows

Table 3.8 Truth table for the split gate array summarizing the different voltage levels required to perform the various array operations, viz., read, programming and erase. The erase voltage can reach an excess of 20 V.

MODE	BL	WL	SL	EL
READ	1V	Vcc	Gnd	Gnd
PROGRAMMING	7V	10V	Gnd	Gnd
ERASE	Gnd	Gnd	Gnd	10-22V

Table 3.9 A figures of merit summary for the split gate array. The array requires only positive voltages and is immune to over-erase bits during read.

FEATURE	GRADE
Cell Size	2
Process Complexity	2
High Voltage (On-Chip)	3
Array Efficiency	1
Redundancy	2
General Purpose	3
MultiLevel	2
Low Voltage	3
TOTAL	18

that cell size, array efficiency, and process simplicity are some of the advantages. The main disadvantage is due to the limitations of the poly-poly erase concept, which allow only dedicated mass storage applications.

Alternate Metal Ground (AMG) Cell and Array

The AMG cell architecture is another virtual ground concept that scales the cell very effectively [33–37]. It relies on the same programming and erase mechanisms as the industry standard common ground cell does. The AMG concept was developed as an EPROM and ROM cell and array architecture, to achieve the cell size equal to the minimum geometry on the given technology ($4F^2$). This is accomplished by the combination of a virtual ground architecture which consists in a single metal bit line for every two diffusion bit lines. Two segment select transistors (2 per segment of 128 cells) replace the split gate concept (dedicated select per cell). The 10% overhead is a very reasonable penalty to pay for such a dramatic scaling. In the EPROM and ROM erase, a fieldless array is implemented to get the full scaling benefit. In the Flash implemen-

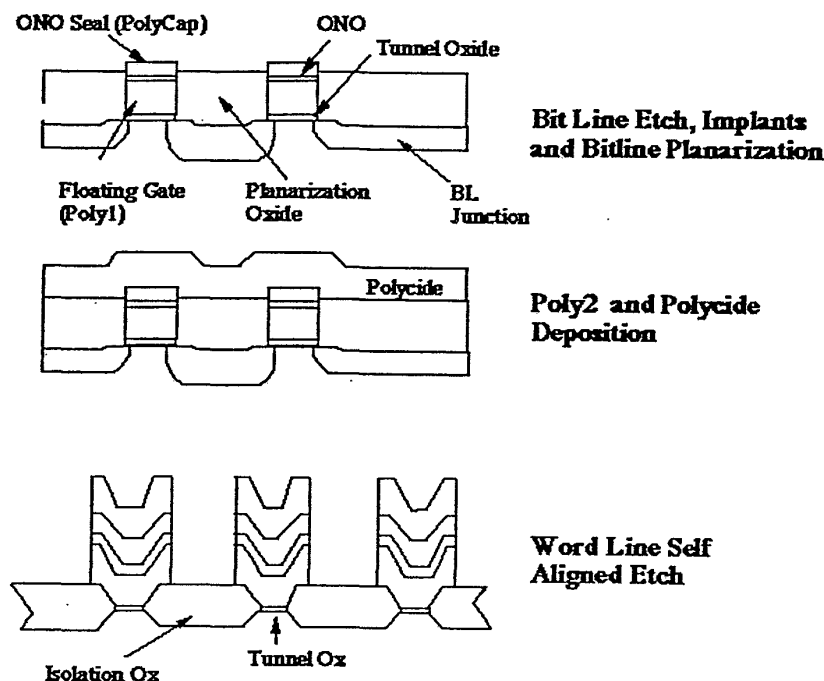


Figure 3.23 AMG Flash cell device cross-sections: a) along the channel length, and b) along the channel width directions [2]. The poly1 pitch is the limiting pitch along the channel length while the poly2 pitch (with gate coupling requirement) determines the pitch along the channel width.

tation, the need for increased gate coupling for the erase operation adds 50% area to the cell size. A cross-section of the cell is shown in Fig. 3.23. The cross-section along the WL (Fig. 3.23a) is very similar to the common ground cell, without the drain contact. In Fig. 3.23b, a cross-section across the word line is shown. The isolation oxide, for erase coupling enhancement, is the cause for the increase in cell size relative to the EPROM case.

The cell layout is shown in Fig. 3.24. Polycide wordlines cross over vertical bitlines in a simple geometry. Cell size of $2.1 \mu\text{m}^2$ in $0.6 \mu\text{m}$ technology is equivalent to a $5.8 F^2$ cell size. A schematic of the array architecture is shown in Fig. 3.25. Every block of 128 word lines has block select transistors (top and bottom) and AMG selects that connect the internal segment to the left or right BL. All the select transistors are integrated into the same pitch as the cell by a simple layout. The AMG select transistors serve to select a source line or a drain line into the segment area during read and programming, respectively.

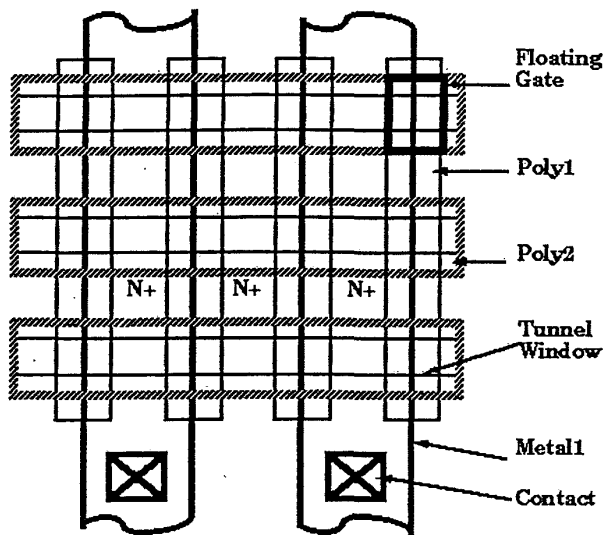


Figure 3.24 AMG Flash array layout [2]. Notice that each metal 1 bitline accommodates two n^+ diffusion bitlines.

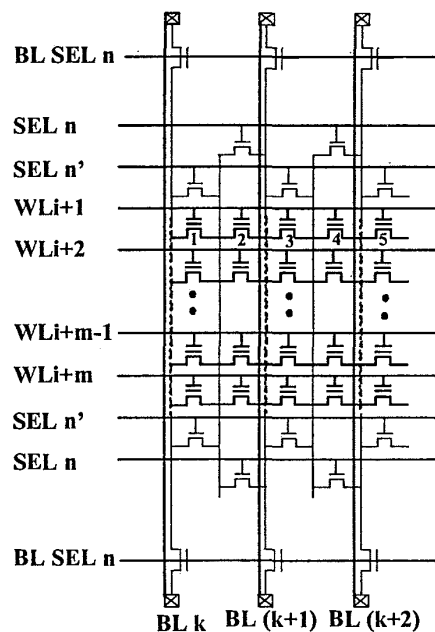


Figure 3.25 Schematic diagram of the AMG array architecture [35]. The natural segmentation of the array is provided with block select transistors BL SEL n , the AMG select transistors SEL n and SEL n' provide connection to the drain junction.

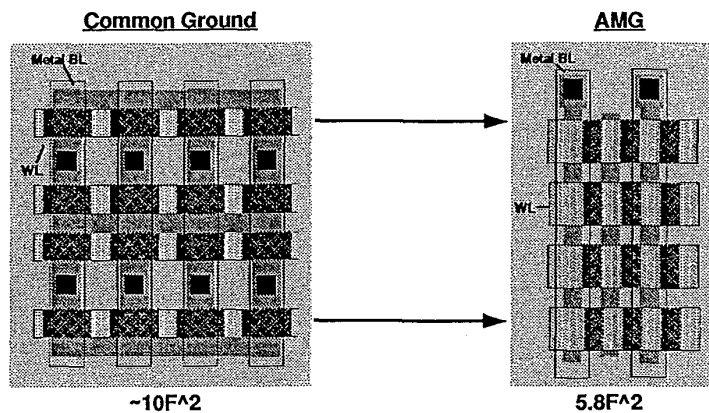


Figure 3.26 Comparison of the common ground and AMG arrays to illustrate the bitline scaling of the latter concept [1]. In the AMG concept, the number of contacts is significantly reduced and only one metal line is required for two bitlines.

As we can see, there is a metal bit line with a contact every two diffusion bit lines, which enables the aggressive scaling of the cell in the bit line dimension. The AMG array is a virtual ground array with respect to the metal bit lines. The natural segmentation is advantageous for any Flash application and the select transistors add $\approx 12\%$ overhead to the cell area.

The functional truth table is shown in Tab. 3.10. The erase source diffusion can be directly accessed via the block select. The drain during programming is the internal segment. The erase junction also serves as drain during the read operation. To better appreciate the AMG scaling, a comparison to the common ground array is shown in Fig. 3.26. The dramatic scaling is a combination of WL pitch scaling, eliminating the contacts, and BL pitch scaling, having one metal BL every two diffusion BLs.

Table 3.10 Truth table for the AMG array summarizing the different voltage levels required to perform the various array operations, viz., read, programming and erase.

MODE	BL k	WL	BL(k+1)	SEL n	SEL n'	BL SELn
READ	1V	Vcc	Gnd	Vcc	Gnd	Vcc
PROGRAMMING	Gnd	10V	6V	11V	Gnd	11V
ERASE	Vcc	-8V	Vcc	Gnd	Gnd	11V

The product implementation can take advantage of the natural segmentation to offer very effective column and row redundancy schemes. The small cell

size combined with a simple NOR architecture make AMG an effective concept for both mass storage and embedded applications. The array efficiency for a 16 Mbit product is reasonable at 52%. The summary table for the AMG array architecture is shown in Tab. 3.11. Small cell size, general purpose and effective redundancy implementation are the highlights of the AMG architecture. This array concept rates average on other figures of merit, but is an attractive architecture overall.

Table 3.11 A figures of merit summary for the AMG array. This concept is well suited for small cell size with common ground array like operations.

FEATURE	GRADE
Cell Size	1
Process Complexity	2
High Voltage (On-Chip)	2
Array Efficiency	2
Redundancy	2
General Purpose	1
MultiLevel	2
Low Voltage	2
TOTAL	14

Other Flash Cell and Array Concepts

Six different array architectures have been described in the last paragraphs. These architectures have been chosen because of their importance, measured by the production level reached, or by their significance as original and general concepts. There are many other array and cell operation concepts. In this section other interesting concepts will be described in a very short format.

Source Injection Concepts. Source injection is a hot electron mechanism, where electrons are accelerated in a region closed to the source. The key advantage of this concept is the high injection efficiency with a very low programming current. The cross-section of a device with source side injection is shown in Fig. 3.27. Note the lateral field along the channel. The lateral field enhancement next to the source is a combination of a low conductivity source transistor with a high conductivity floating gate device.

Another source injection Flash cell is shown in Fig. 3.28, where a split gate architecture is implemented. For some applications, like low voltage and low densities, this concept may be advantageous. There is no scaling benefit in source side injection concepts, since they need another geometry in the cell to

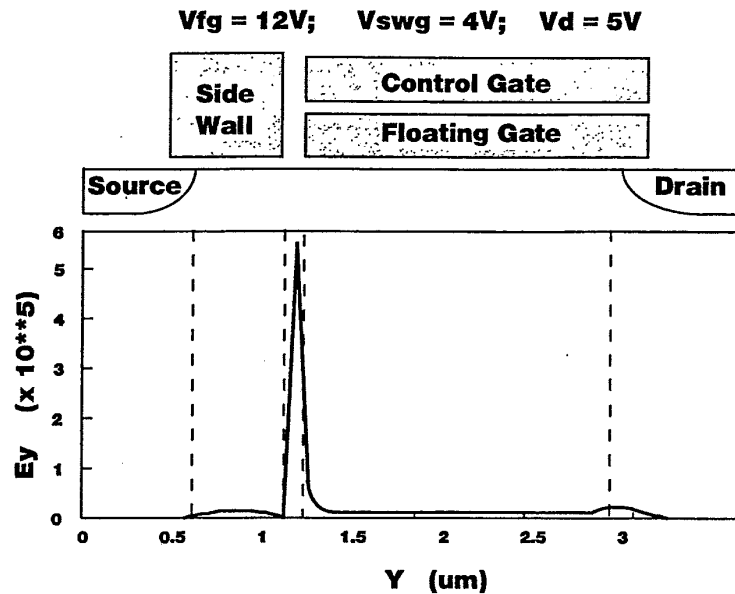


Figure 3.27 Split gate source injection concept is depicted [2, 39]. The electric field concentration on the source side of the floating gate is responsible for increased efficiency channel hot electron injection.

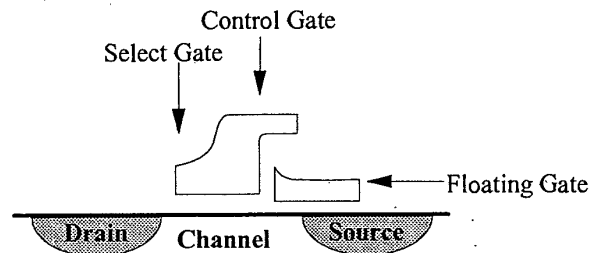


Figure 3.28 Device cross-section of the source injection common ground array. The programming is via source side injection while the erase is via poly-poly FN-tunneling from the sharply defined floating gate feature [39].

create the source lateral field enhancement. The cell in Fig. 3.28 is a $13.6 F^2$ in $0.5 \mu\text{m}$ technology ($3.4 \mu\text{m}^2$).

A merged split gate source injection concept is shown in Fig. 3.29 [42, 43]. These are two double poly cells connected back to back, with a WL as a select line. The read concept resembles the NAND “read through” concept. During

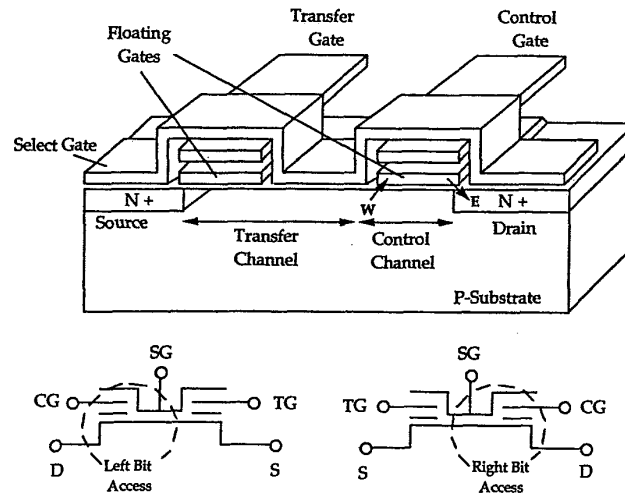


Figure 3.29 Device cross-section of the merged split gate source injection array [42]. Back to back split gates are constructed to share the third poly select gate for better cell size. Only one metal line is required for every two bitlines. The select gate controls the source injection.

programming, the select line is maintained at a low voltage to create the source side injection conditions. This array has some of the AMG advantages, in particular it needs only one metal BL for two cells. The read through concept replaces the AMG select transistors. The cell size is small at $5.4 F^2$, but the process and disturb modes are very complex.

EEPROM Based Flash Architecture. The architecture of the EEPROM based Flash cell is identical to that of the standard EEPROM, but has no WL segmentation [44, 45]. The cell size in this architecture is very large ($\approx 40 F^2$). The additional select transistor per bit and the high gate coupling needed for erase are the two main reasons for such a large cell.

The main advantage as a Flash is the EEPROM experience and very low voltage applications. This is a reasonable approach for a short term foot hold in the market, but remains a very unattractive long term solution.

Why so Many Concepts? Trying to analyze the reasons for the large number of Flash concepts, one has to understand the decision making process behind the choice of an array architecture.

There are three main considerations in selecting a Flash concept. The cell device physics or process technology comes first, followed by the company past

experience, and by the marketing beliefs. Hot electron programming, source side injection, F-N drain programming, channel or source side F-N erase are all quite complex mechanisms. Cell optimization for each one of them, including the disturb modes, is far from being a simple straightforward task. There is no obvious ideal concept that solves all possible problems in programming, erase and read. Based on their past experience, companies put different emphasis to a given device/process problem, thus further complicating the decision making process.

A company's past NVM experience may force a decision to a more "familiar" territory. For example, EPROM product line may favor Flash as the EPROM extension. This simplifies the programming understanding, circuit design concepts in read and programming, and test concepts. From the marketing point of view, the choice of the key application is significant, either embedded or mass storage. Anyone of these applications has different requirements on the Flash features. For example, low voltage, parallel programming, number of program erase cycles, all of these are marketing parameters that may affect the array concept selection.

Any selection requires years of investment, after which only very few concepts survive the test of time. The convergence into a smaller field of concepts will occur in the next few years. The understanding of more concepts helps in the development of each one of them. New and better concepts may emerge from this type of knowledge.

Summary of Array Architectures

Tab. 3.12 summarizes the figures of merit of all the Flash memory concepts that have been analyzed in this chapter. In this simplistic rating system, we have attributed an average score to all concepts. Obviously a 1 or 2 points differential is not enough to declare the best concept! This table does give a way to identify the best concept, once key assumptions have been made. There are three basic differences: cell size, programming mechanism, and application.

In case the application is general, and F-N programming is unacceptable, the list is reduced dramatically. This happens also when parallel programming is a must for the application: there the channel hot-electrons concepts are eliminated. The above examples show that the simplistic rating system can be improved substantially by putting a special weight on a specific requirement. Scaling of the single bit concepts and multilevel implementation, significantly enhances the differences between the various concepts.

Table 3.12 Comparison of different Flash concepts, based on the various figures of merit. Higher values correspond to worse performance.

FEATURE	Common Ground	NAND	DINOR	AND	Split Gate	AMG
Cell Size	3	1	1	2	2	1
Process Complexity	2	2	3	3	2	2
High Voltage (On-Chip)	2	3	3	3	3	2
Array Efficiency	2	2	3	3	1	2
Redundancy	2	1	1	1	2	2
General Purpose	1	3	1	1	3	1
MultiLevel	1	3	3	3	2	2
Low Voltage	2	2	2	1	3	2
TOTAL	15	17	17	17	18	14

3.2.5 Scaling and Conclusions

In analyzing the stability of the various Flash concepts, it is useful to review the past trends. The memory density as a function of time for EPROM and Flash over the last two and a half decades is shown in Fig. 3.30. The density has increased in an exponential way, a factor of two every 1.8 years. The question is: will this trend continue or will it slow down?

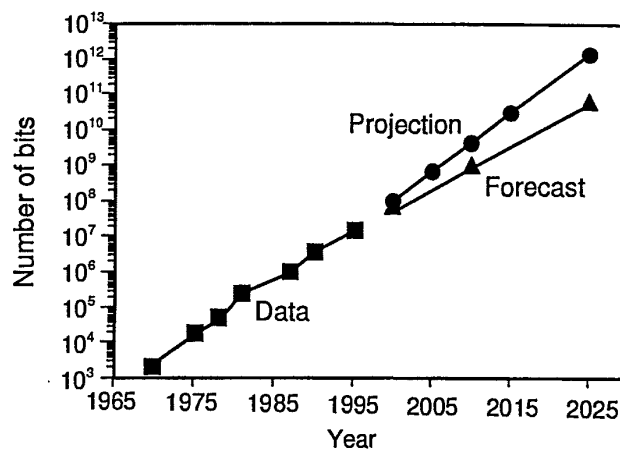


Figure 3.30 Flash density evolution curve. The historical growth rate is at about a doubling every 1.8 years.

Before we answer this question, the cell size scaling over the same period of time is shown in Fig. 3.31. The past trend showed a factor of two reduction of the cell size every four years! This has led to an exponential increase in array size, a factor of two every four years. Extrapolating all these trends into

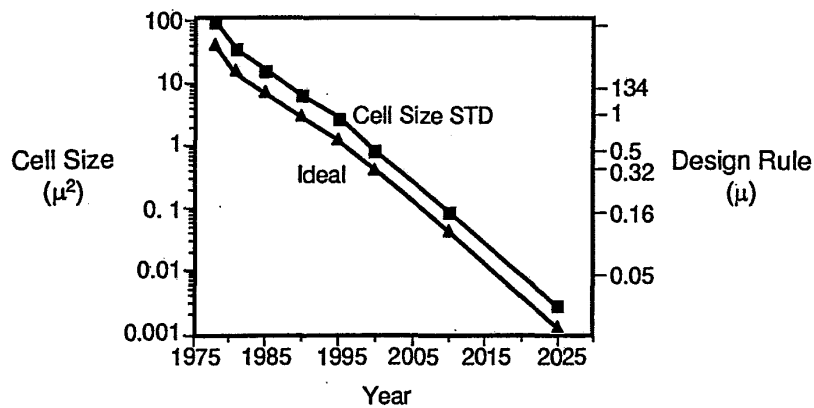


Figure 3.31 The history and future projection of cell size scaling for the common ground and ideal Flash arrays. The year 2025 is chosen to illustrate the demands on the process and device technology, if the present trend continues.

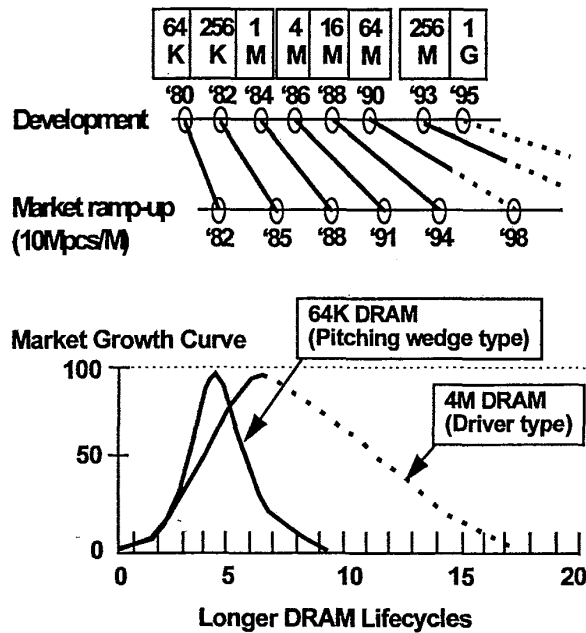


Figure 3.32 DRAM development and market ramp up time schedules and its impact on product lifecycles [46]. The product lifecycles in the future are expected to stretch longer.

the year 2025 results in density of 2×10^{12} , a cell size of $500 \text{ \AA} \times 500 \text{ \AA}$, with a $180 \text{ \AA}$ feature size and a die size of 64 cm^2 . As a reference, there are 18 dice on

an 8 inch wafer. From another point of view, we can note that a 10 V internal voltage across the cell effective channel of 180 Å results in an electrical field of 5.5 MV/cm in the silicon. This is almost ten times higher than the silicon breakdown field!

Two key assumptions are embedded in this extrapolation, first the scaling of design rules will continue at the same rate as before, second that voltage scaling for the above Flash concepts can continue in an accelerated way. The first assumption (design rule scaling) has shown signs of weakness. The time from development to market introduction for DRAM is shown in Fig. 3.32 (upper part) and the life cycle (bottom). Both trends are very clear: it takes a much longer time from the development phase to the introduction into the market for any new generation of DRAM memory, and the life cycle of the product is much longer. Both are a reflection of the huge investments in the development and production facilities. This trend by itself predicts that there will be a slow down in the introduction of scaled processes in the future.

For Flash, the issue of voltage scaling makes the above picture more dramatic. While CMOS voltage scaling is very difficult, today processes are 3.3 volts internally. For Flash, all concepts have internal voltages in the range of ten to twenty volts. Can these level be scaled significantly?

Internal Voltage Scaling

Hot electron channel injection is a relatively low field injection mechanism. The scaling of the programming voltages over the years has been very significant. From 28 V in the seventies to 12 V in the eighties and 10 V in the nineties. Key developments in voltage scaling, like source side injection, substrate injection, or the newest concept, secondary impact ionization enhanced programming, do not show a promising path to very low voltages. The accelerating voltage in the channel of the transistor can be reduced to less than 5 V. The best results so far are based on the secondary impact ionization enhanced programming, and are shown in Fig. 3.33 [47]. Both $V_d - V_{sub}$ and $V_g - V_{sub}$ are 6 V. Much lower voltages in channel hot electrons programming, as low as 1.5 V, have been demonstrated (Fig. 3.34 [48]). The programming rate is so slow that it is more like a disturb mechanism, than a programming technique. In conclusion, for channel hot electron programming, there is no clear path to a dramatic voltage scaling below 5 V.

Fowler-Nordheim tunneling, which is used for both programming and erase, is a high field mechanism. Voltage scaling can be achieved by scaling the dielectric thickness or by choosing a lower barrier gate dielectric. Both options are not very realistic, as long as the retention requirement is not reduced substantially. The 10 years retention lifetime limits the floating gate oxide thicknesses

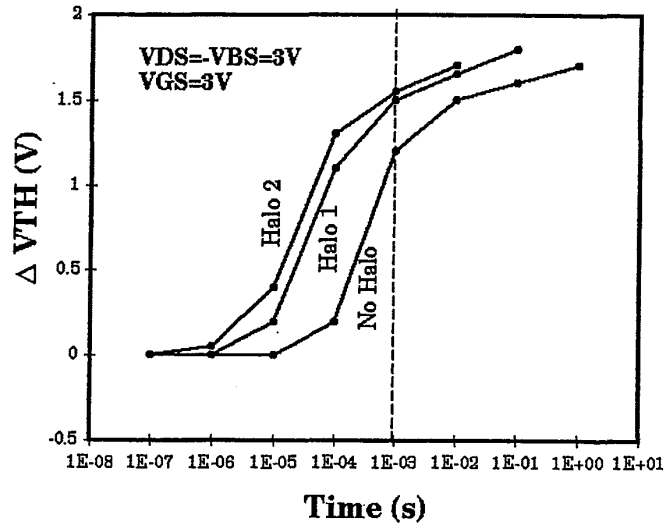


Figure 3.33 Secondary electron injection concept to provide ± 3 V programming in the millisecond time scale [47]. "Halo" refers to different boron halo doses ($5 \cdot 10^{13}$ for halo 1 and 10^{14} for halo 2, respectively) added to increase junction doping.

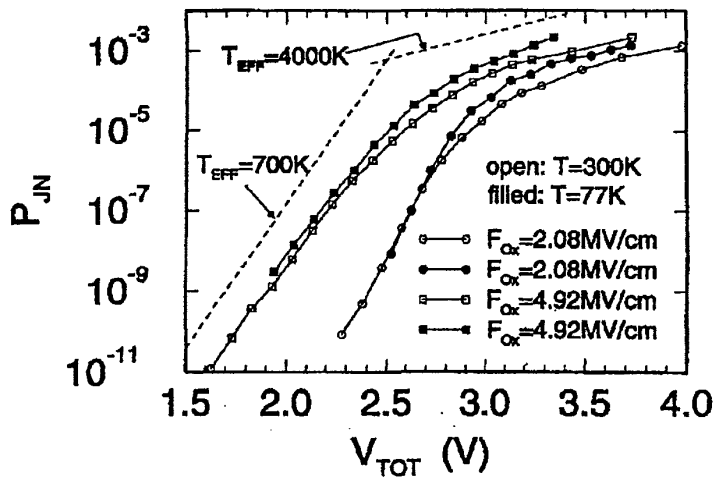


Figure 3.34 Very low accelerating voltages (V_d s) Hot Electron programming [48]. P_{JN} represents the measured electron injection probability as a function of the total accelerating voltage. F_{Ox} is the oxide field.

to 60–70 Å. Any further scaling will lead to the direct tunneling regime. Also stress induced leakage current becomes dominant for thinner tunnel oxide. The

retention time will be reduced to seconds once the oxide thickness is in the range of 30–50 Å.

In conclusion, in today programming mechanisms, both channel hot electron and F-N tunneling, no clear path is available to reduce the internal voltages. The assumptions are: floating gate cells and at least 10 years of retention. Obviously the dramatic effective gate length scaling to the range of 180 Å, if we ever achieve it, is not going to be a direct extension of today concepts!

Conclusions

The single bit per transistor Flash devices scaling is predicted to show a significant slow down. This is based on the current cell concepts with their limitation on voltage scaling. To continue the fast scaling, two possible scenarios can take place: first, the Flash concept altogether, including its physics, has to be changed; second, much less retention has to be accepted, leading to an acceptance of “imperfect Flash” concept for the future Flash memories development.

An example of a concept that is already demonstrated to require much lower voltages is the Ferroelectric NVM Secondary electron injection concept, to provide ± 3 V programming in the millisecond time scale [49–51]. A ferroelectric cell is shown in Fig. 3.35. The cell is based on storing the information in a magnetic domain within the ferroelectric capacitor. This capacitor is integrated with a pass transistor into a memory cell. While it is a very interesting concept with many benefits, there are two key disadvantages: a very large cell size and the problem of integrating ferroelectric material into silicon technology.

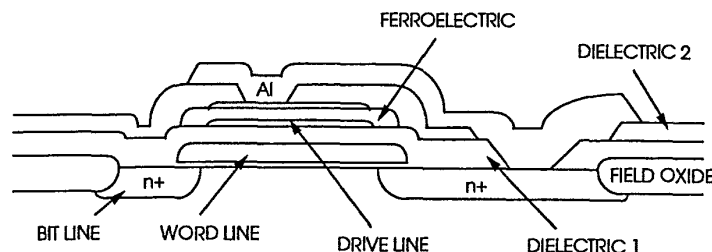


Figure 3.35 The ferroelectric array device cross-section secondary electron injection concept to provide ± 3 V programming in the millisecond time scale [3]. The ferroelectric capacitor is integrated above the CMOS array and promises very high endurance. Ferroelectric memory read operation being destructive to the stored information requires a restore operation after every read operation.

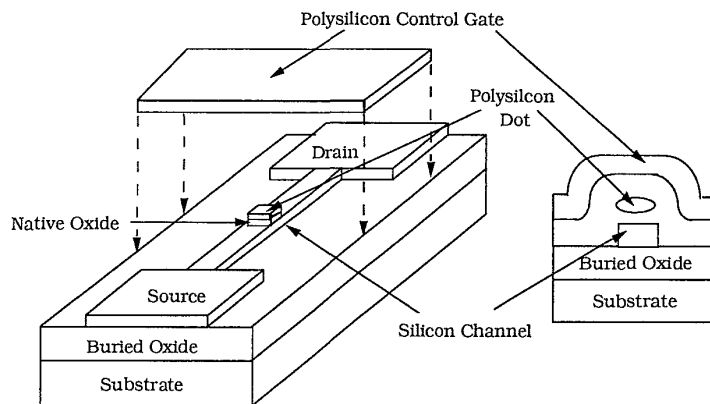


Figure 3.36 Demonstration of the single electron concept with a very narrow polysilicon floating gate structure [52]. The spatial localization of the electron wavefunction leads to quantized energy levels (and subsequently threshold voltage steps).

Another new concept that has emerged lately is the single electron Flash (Fig. 3.36) [52, 53]. This device is based on the scaling of the floating gate down to several hundred angstroms. Electron confinement in the potential well of the floating gate leads to discrete states in its conduction band. Every state can store one electron, which alters the transistor V_t . The cell and applied voltages in this concept are the same as in traditional floating gate transistors. Hence its scaling is not any better. For single electron Flash, to become a dramatic improvement over the other approaches, the advantage is in the multilevel potentials. This will be discussed in the second half of this chapter.

A third example of a concept change is reviving the MNOS concept. A true 5 V only (internal) MNOS device is shown in Fig. 3.37 [54–56]. We can take advantage of this concept to scale voltages and simplify the process, provided that the retention is much shorter (Fig. 3.37 below). Such a non-perfect memory concept is obviously a rather dramatic change in system architecture.

A final example of cell size scaling, that can help in keeping the cell size scaling per bit with the past trends, is storing several bits in a single cell. This concept has been used extensively in the last few years for voice storage, where the data integrity is not very critical. For data storage, multilevel cells products are being introduced presently. This is the next main topic of discussion. All cell concepts that have been introduced in the first portion of the chapter are applicable for multilevel applications, however, their differences will loom brighter.

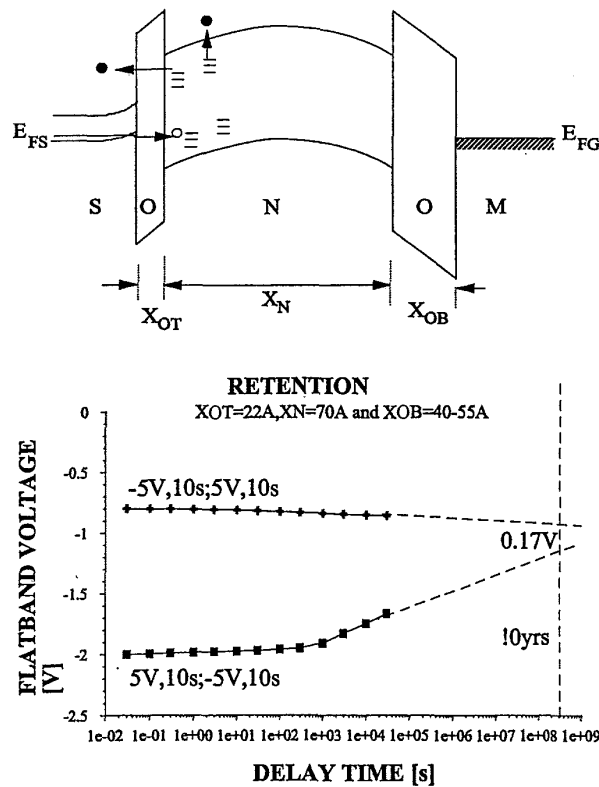


Figure 3.37 Scaled MNOS concept to achieve true 5V only Flash concept [54]. The charges are trapped in the nitride layer to provide threshold voltage shifts. The bottom figure provides the results of a retention measurement extrapolated to ten years. X_{OT} stands for tunnel oxide, X_N is the nitride layer thickness, X_{OB} is the blocking oxide thickness.

3.3 MULTILEVEL FLASH CELLS

3.3.1 Introduction to the Concept of Multilevel Flash

The multilevel Flash concept is a method to increase the number of stored bits in a memory cell [1, 26, 31, 57, 58]. This is the most efficient way to scale the effective cell size for any technology. In describing a single bit Flash, we have assumed two threshold states: an erased V_t and a programmed V_t . For the NOR array, the read V_t window is defined as the difference of the maximum erased V_t and minimum programmed V_t . The read V_t window has to accommodate the reliability requirements to ensure a sufficient signal for proper sensing (see Fig. 3.38a). In Fig. 3.38b, the two bit window with its four V_t states is shown. To incorporate four states, the total window is larger, while

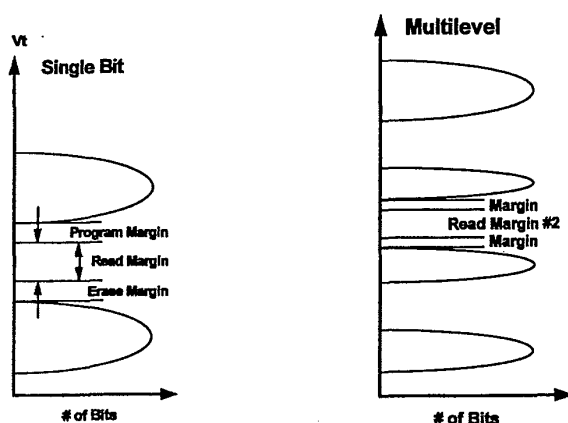


Figure 3.38 The program read and erase margins for a) single bit Flash cell, and b) multilevel Flash cell with 4 states [1].

the read window between neighboring states becomes smaller. The reliability margins are compromised, too. One of the consequences of this concept is a requirement for a smaller V_t distribution per state. This is due to the need to minimize the total magnitude of the window. The requirement for a smaller V_t distribution combined with the wider read window enhances the difficulty in implementing multilevel, due to disturbs.

To store N bits, 2^N different threshold states are required. There are many trade offs to face in multilevel Flash memories. Compromising on the read current by dividing it between the large number of bits, reduces the effective read signal per bit. This reduced signal results in a slower access time. A wider V_t window is associated with higher programming voltages. This implies higher electric fields, hence more disturb and wearout. The choice is between smaller required margins for reliability with an improvement on the process control, or reduced reliability levels. The advantage of increasing the density by multilevel is big enough to justify the extra complexity in different aspects of the cell and product. Its effectiveness as a cost reduction path is yet to be demonstrated. For mass storage applications, where some of the requirements are relaxed and the system support is extensive, multilevel is of great benefit.

3.3.2 Multilevel Programming Mechanisms

In our previous discussion (Section 3.2), the programming mechanism has been used as a key parameter in comparing the various array architectures. For multilevel Flash cells the programming mechanism is even more critical. The

two different mechanisms, channel hot electron programming and F-N tunneling, will be compared with reference to their specific application in multilevel Flash. From the previous discussion, the main differences between them are the higher fields in F-N tunneling programming and the longer time it takes to program a bit. These two basic differences have a strong impact on the multilevel applications of both techniques.

First, let us consider the impact of electric field and programming time on disturbs. In Fig. 3.39a, the disturb margin for channel hot electron programming is shown. The programming curve shows that programming is achieved

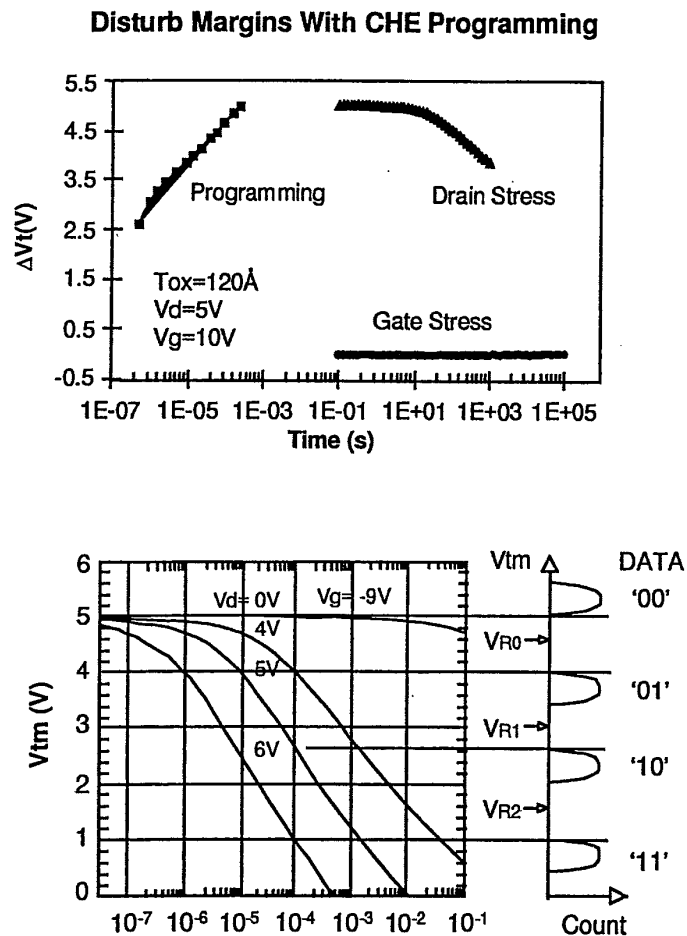


Figure 3.39 a) Drain and gate disturb margins for CHE programming mechanism [1], and b) the drain voltage enhanced gate disturb for F-N tunnel programming mechanism [26].

in the range of $10\ \mu\text{s}$ or less. The programming time is defined as the time to obtain a delta V_t of 2 V. There are two disturb conditions. Drain stress is for the bits on the same column as of the programmed bit that are not being programmed. Gate stress is for the bits that are on the same word line, but different columns. As you can see, the gate stress is not an issue since a margin of more than nine orders of magnitude is available. Disturb margin is defined as less than 10% reduction in V_t . The drain stress is more severe, with only six orders of magnitude of available margin relative to programming. In Fig. 3.39b the gate disturb for F-N drain erase mechanism is shown. In this case less than one order of magnitude margin relative to the programming time is available. Not shown is the drain stress, which is as severe as the gate stress. Such a small margin is actually considered unacceptable since it enlarges the V_t distribution of each of the V_t states. The reason why this single transistor margin is considered too small is found in the process variations within the many millions of bits. In a large array, the programming and disturb rates can be varying due to process variation by a factor of ten or more. The impact of cycling on disturb can amount to one or two orders of magnitude. These variations require that the margin between the function and its disturb will have more than 3 orders of magnitude. This permits safe operation and simple production screening. The sensitivity of the F-N programming to some of the cell parameters like gate and drain coupling and drain overlap is very high and this further enhances the sensitivity of F-N tunneling to the disturbs. Moreover, because of the long programming time in F-N tunneling one would prefer to do parallel programming, which typically results in less control of the threshold of each and every one bit.

Another disadvantage of F-N tunneling programming is the "erratic" bit phenomenon. This is a well known problem (see Chapter 7), whose effects are somewhat alleviated in the case of F-N erase. For programming, it is much more severe, since it cannot be screened and requires a new erase cycle while programming! An erase cycle within the programming cycle is not an acceptable flow for the programming of any product.

One of the advantages of channel hot electron programming for multilevel applications is the minimization of disturbs as a function of endurance. In Fig. 3.40 the read disturb as a function of program erase cycles is shown. The read disturb is defined as a V_t change due to drain stress. Both programming mechanisms were invoked on the same device. This eliminates the processing conditions and circuit related issues from affecting the comparison. As shown, the channel hot electron programming induces less disturb, as indicated by the shallower slope as a function of cycling. This result further emphasizes the disadvantage of F-N programming mechanism.

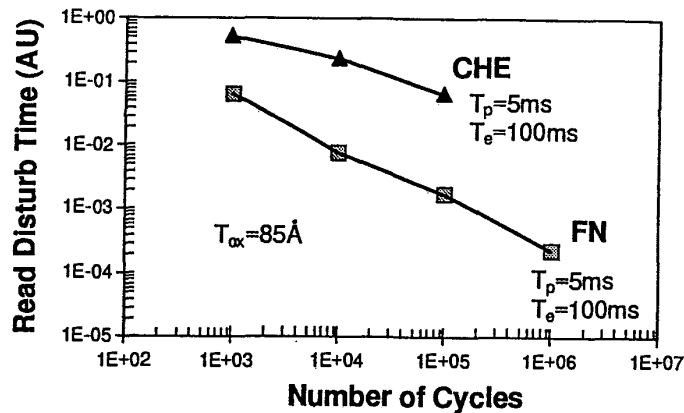


Figure 3.40 A comparison between the disturb due to channel hot electron and Fowler-Nordheim tunneling programming, as a function of program erase cycles [1]. This experiment is performed on the same product for both mechanisms.

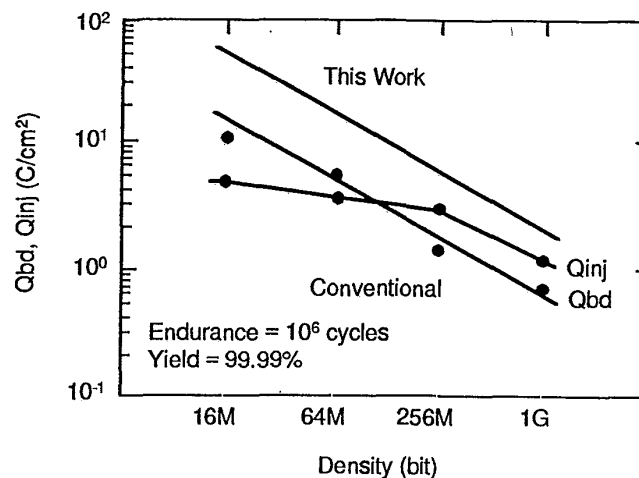


Figure 3.41 Q_{bd} and Q_{inj} as a function of memory density [5]. For very high density, the Q_{bd} limit is reached.

The higher field during programming for the F-N programming enhances the charge to breakdown (Q_{bd}) limitation for endurance. In Fig. 3.41 the Q_{bd} and injected charge (Q_{inj}) as a function of memory density are shown [5]. The margin built in "this work" is through a concept of maximum field reduction. For channel hot electron programming concepts, the equivalent to the above

F-N programming is the erase mechanism. As a block-erase mechanism, the available time is much longer. This implies a F-N mechanism with substantially lower electric field. The combination of channel hot electrons programming and F-N erase minimizes the problem of Q_{inj} at high fields that is becoming the most severe limitation on F-N programming high density Flash products. This is a reliability and quality differentiator between the two programming mechanisms.

Tab. 3.13 summarizes the above discussion. It compares channel hot electron to F-N tunneling as the programming mechanisms for multilevel applications. The parameters of comparison are electric field, disturb, programming current, cell size and parallel programming. For multilevel applications the first two points are priority keys. The parallel programming is the only advantage for the F-N tunneling, as it is for the single bit case.

Table 3.13 Comparison of F-N tunneling and CHE programming mechanisms for multilevel applications. CHE seems to be more suitable.

FEATURE	CHE	FN
Electric Fields	Lower	Higher
Disturbs	Less	More
Programming Current	High	Low
Cell Size	Same	Same
Parallel Programming	Poor	Good

Knowing that the multilevel concept is extremely difficult to implement in any Flash concept, the advantage of less disturb and lower electric fields are more significant than the F-N parallel programming advantage. Our conclusion is that channel hot electron programming is a much more suitable programming mechanism for multilevel applications.

In the following sections the various Flash concepts already described in Section 3.2 are analyzed from the point of view of their multilevel compatibility. The basic assumption is that channel hot electron programming is an advantage; the focus is on other array related effects and their impact on multilevel applications.

3.3.3 Architectures for Multilevel Flash Memories

Common Ground Array and DINOR

Multilevel is more sensitive to two array issues, namely source resistance and BL to BL coupling. In Fig. 3.42, a schematic diagram of the common ground array architecture is shown.

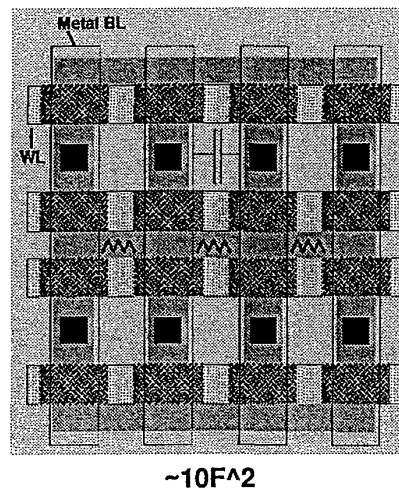
Common Ground

Figure 3.42 The source resistance and BL-BL coupling effects illustrated for a common ground array [1].

The series resistance of the source has an impact on the broadening of the V_t distribution of each of the V_t states. The programmed information along a diffusion source line affects the total current along it, hence the potential drop. This drop along a source line is equivalent to a source bias of the cell, which reduces the applied V_g s and increases the V_{bs} . This is equivalent to a V_t shift seen by the sensing circuit. The problem arises from the fact that the programmed information along any diffusion source line is changed when cells are being programmed. This change occurs after the V_t of a given cell was verified to be within a given window! The delta due to the worst case difference directly adds on to the width of the V_t window. This delta can be reduced by reducing the series resistance of the source line. For a diffusion line with minimum dimensions, there is very little that can be done. In some cases more metal strapping of the source line helps significantly without the disadvantage of added processing complexity.

The second array problem is the BL to BL coupling. As in the source case, due to the change of information during programming, the coupling to the neighbor bit line may change. The difference between byte by byte verify and final verify (after programming) creates another source for V_t window widening. This coupling can be reduced by reducing the bit line to bit line coupling; however, in future generations this will even increase, due to the scaling of BL to BL distances.

For DINOR, both problems are relevant [17, 19]. The local polycide BL with one global metal BL should improve the coupling problem. A wider metal space is the origin of the advantage. In case of source resistance, there is no inherent improvement in the DINOR relative to the common ground case.

Virtual ground - AND, AMG and Triple Poly Array

AND. This architecture is a combination of a virtual ground architecture and a common source architecture [26]. As a common source architecture, the V_t distribution enhancement due to the source resistance is the same as the common ground architecture. As a virtual ground architecture, there are no new issues. The shared local source line only sees the read current of the sensed bit. All other WLs are off and the neighbors are not drawing any current, whether their programmed state is high, low, or in between. Neighbor effects due to bits that share a common WL are solved in this architecture, by the local field isolation. Every local SL and BL are fully isolated from their neighbors.

The second issue with the AND architecture is the bit line coupling. In this case, it is identical to the Common Ground architecture and worse than DINOR.

AMG. The common source series resistance issue exists but is different than in the Common Ground. The AMG source resistance effect is the result of one neighbor along the same WL. The neighbor programming state may affect the total current through the source line. Once the source line current is neighbor programming dependent, the source resistance effect widens the V_t distribution. In case the array precharge state is to ground, this effect does not occur.

A similar effect to the source line widening can happen on the drain side in the AMG architecture. In this case, the programmed state of the neighbor may affect the net current through the drain. It is equivalent to the source side series resistance effect on the V_t distribution widening. The BL to BL coupling problem is less severe in the AMG architecture. The metal bit lines are further apart, hence their coupling capacitor is smaller.

Triple Poly Architecture. This architecture has the same effects as the AMG. The source neighbor effect and the drain neighbor effect are probably less severe in this array due to the very low read currents. The BL to BL coupling problem is the same as the Common Ground array and more severe relative to the AMG array.

NAND Array

The NAND array programming mechanism is F-N channel injection. Much higher voltages are involved in this programming mechanism [15]. The programming inhibit is achieved by channel potential coupling to the WL. Indeed the disturbs in this channel F-N programming are more severe than in the other F-N cases. It just enhances all the problems that make F-N programming less suitable to multilevel applications.

The read through concept, that is described in section on NAND cell and array, further enhances the difficulty in multilevel implementation. This is represented in Fig. 3.43 with a comparison of the V_t window for the NOR and the NAND cases. In the NAND case, a read margin above the highest state has to be created. This implies that the WL level is higher during read than in the NOR case. It adds one more state with narrow V_t distribution.

In the NAND architecture, the read-through concept induces, for each logic state, a non-negligible widening of the V_t distributions. The change in state of the other 15 neighbors increases the V_t distribution in the same way the source resistance affects the NOR. To partially compensate for it, there is a pre-determined sequence of programming in every segment. This makes the effect slightly more predictable. The very low read current and the slow access times help in minimizing all of the above NAND problems. The NAND architecture is the less attractive for multilevel implementation.

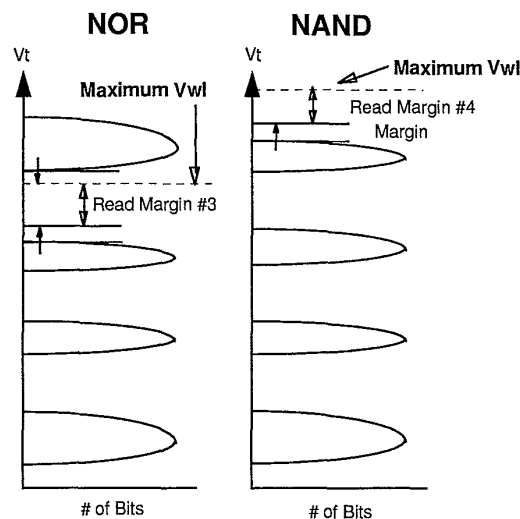


Figure 3.43 The illustration of the larger V_t window for the NAND array due to the read through requirement [1].

Summary of Multilevel Array Concepts

In Tab. 3.14 we summarize our discussion on multilevel architectures for the different Flash concepts. Programming performance, as well as disturb immunity, are optimized in the case of the channel hot electron mechanism. The V_t window is a reflection of the read mechanism and the array impact on the V_t distributions. The requirement for high voltage is linked to the programming mechanism. After all, the multilevel concept is just another way of obtaining a cell size scaling concept. Only the combined effects of small geometry with multilevel result in a small cell per bit. The one clear conclusion is that the NAND array architecture is absolutely unattractive for multilevel applications. Architectures with channel hot electrons programming are having better overall ratings, based on our discussion.

Table 3.14 Summary of various Flash concepts with regard to multilevel implementation. The lowest the grade, the better the performances.

FEATURE	Common Ground	NAND	DINOR	AND	Split Gate	AMG
Programming	1	3	2	2	1	1
Disturbs	1	3	3	3	1	1
VT Window	1	3	1	1	2	2
Array Impact	2	3	2	2	2	2
High Voltage	1	3	2	2	3	1
Total Multilevel	6	15	10	10	9	7
Cell Size	3	1	1	2	1	1
TOTAL	9	16	11	12	10	8

3.3.4 Scaling and Trade-Offs for Multilevel

Analyzing the benefits of multilevel as a scaling concept, one has to make sure that its implementation does not result in a scaling compromise. The significant difficulty in multilevel implementation clearly makes its scaling much more difficult than the scaling of the single bit architecture. The debate whether single or multilevel is the best approach is unresolved.

The arguments of the debate are application specific. In voice recording products there are already cells with 256 different levels. For the digital world, even two bits per cell is a major issue, and adding more levels becomes exponentially more difficult.

A different approach to multilevel is multibit cells, shown in Fig. 3.44 [59]. In this case, two floating gates per cell are used to store the two bits. It may have some advantages over multilevel, but this type of cell requires new process array architecture and design.

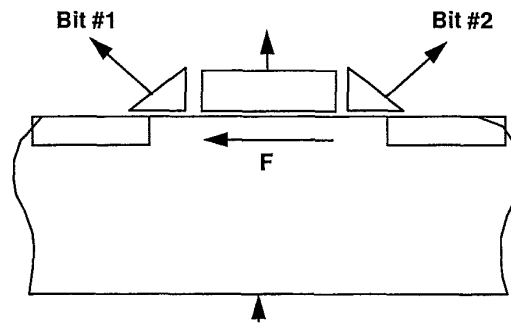


Figure 3.44 A simplified device cross-section for the multibit cell concept. Two different floating gates are merged in the same transistor [66].

In conclusion, multilevel is a very attractive scaling concept for Flash cells. It adds a lot of complication to the array architecture and its scaling. The only possibility that it will become dominant Flash implementation is in the applications or methodologies that will allow implementation of the “non-perfect Flash” concept. We already saw that the scaling of Flash is going to face major device and process obstacles. The attractiveness of multilevel is central to the continuous drive to reducing the cost per bit for Flash memories.

The future is going to show how quickly the combination of the system support and Flash scaling can go hand in hand and help in the introduction of multilevel concept as a main line approach.

References

- [1] Eitan B., Kazerounian R. and Roy A. (1996) “Multilevel Flash cells and their trade-offs”. *IEDM Technical Digest*, p. 169.
- [2] Eitan B. (1995) “Cell concepts and array architectures”. Flash memory tutorial, *NVSM Workshop '95*.
- [3] (1990) “IEDM short course: non-volatile memory”. *IEDM '90*.
- [4] Pavan P. *et al.* (1997) “Flash memory cells - An overview”. *Proc. IEEE*, **85**, p. 1248.
- [5] Kawahara T. *et al.* (1995) “High reliability electron-injection method for high density Flash memories”. *IEEE JSSC*, **30**, p. 1554.
- [6] Crisenza G. *et al.* (1996) “NVM issues challenges and trends for the 2000's scenario”. *Proc. ESSDERC '96*, p. 121.

- [7] Kuo C. *et al.* (1994) "A 512 Kb Flash EEPROM for a 32 Bit microcontroller". *VLSI Symp.*, paper 9-3, p. 87.
- [8] Mukherjee S., Chang T., Chang R., Knecht M. and Hu D. (1985) "A single transistor EEPROM cell and its implementation in a 512 KCMOS EEPROM". *IEDM Technical Digest*, p. 616.
- [9] Baker A. *et al.* (1994) "A 3.3 V 16 Mb Flash memory with advanced write automation". *ISSCC*, p. 146.
- [10] Venkatesh B., Chung M., Govindachar S., Santurkar V., Bill C., Gutala R., Zhou D., Yu J., Van Buskirk M., Kawamura S., Kurihara K., Kawashima H. and Watanabe H. (1996) "A 55 ns 0.35 μ m 5 V-only 16 Mb Flash memory with deep-power-down". *ISSCC Technical Digest*, TP 2.7, p. 44.
- [11] Miahara M. *et al.* (1994) "Row-redundancy scheme for high-density Flash memory". *ISSCC*, p. 150.
- [12] Masuoka F., Momodomi M., Iwata Y. and Shirota R. (1987) "New ultra high density EPROM and Flash EEPROM with NAND structure cell". *IEDM Technical Digest*, 25.6, p. 552.
- [13] Iwata Y. *et al.* (1995) "A 35 nsec cycle-time 3.3 V only 32 Mb NAND Flash EEPROM". *IEEE JSSC*, **30**, p. 1157.
- [14] Suh K.D. *et al.* (1995) "A 3.3 V 32 Mb NAND memory with incremental step pulse programming scheme". *IEEE JSSC*, **30**, p. 1149.
- [15] Jung T.S., Choi J.Y.J., Suh K.D., Suh B.H., Kim J.K., Lim Y.H., Koh Y.N., Park J.W., Lee K.J., Park J.H., Park K.T., Kim J.R., Lee J.H. and Lim H.K. (1996) "A 3.3 V 128 Mb multi-level NAND Flash memory for mass storage applications". *IEEE JSSC*, **31**, p. 1575.
- [16] Tanzawa T. *et al.* (1996) "A compact on-chip ECC for low cost Flash memories". *VLSI Circuits Symp.*, paper 7.4, p. 74.
- [17] Kobayashi S., Mihara M., Miyawaki Y., Ishii M., Futatsuya T., Hosogane A., Ohba A., Terada Y., Ajika N., Kunori Y., Yuzuriha K., Hatanaka M., Miyoshi H., Yoshihara T., Uji Y., Matsuo A., Taniguchi Y. and Kiguchi Y. (1995) "A 3.3 V-only 16 Mb DINOR Flash memory". *ISSCC Technical Digest*, TA 7.2, p. 122.
- [18] Onoda H., Kunori Y., Kobayashi S., Ohi M., Fukumoto A., Ajika N. and Miyoshi H. (1992) "A novel cell structure suitable for a 3 volt operation, sector erase Flash memory". *IEDM Technical Digest*, paper 24.3, p. 599.

- [19] Kobayashi S. *et al.* (1994) "Memory array architecture and decoding scheme for 3 V only sector erasable DINOR Flash memory". *IEEE JSSC*, **29**, p. 454.
- [20] Ohnakado T., Takada H., Hiyashi K., Sugahara K., Satoh S. and Abe H. (1996) "Novel self-limiting program scheme utilizing N-channel select transistors in P-channel DINOR Flash memory". *IEDM Technical Digest*, paper 7.4, p. 181.
- [21] Tsuji N., Ajika N., Yuzuriha N., Kinori Y., Hatanaka M. and Miyoshi H. (1994) "New erase scheme for DINOR Flash memory enhancing erase/write cycling endurance characteristics". *IEDM Technical Digest*, paper 3.4, p. 53.
- [22] Kato M. *et al.* (1994) "A $0.4\mu\text{m}^2$ self-aligned contactless memory cell technology suitable for 256-Mbit Flash memories". *IEDM Technical Digest*, paper 3.8, p. 921.
- [23] Hisamune Y.S., Kanamori K., Kubota T., Suzuki Y., Tsukiji M., Hasegawa E., Ishitani A. and Okazawa T. (1993) "A high capacitive-coupling ratio (HiCR) cell for 3V-only 64Mbit and future Flash memories". *IEDM Technical Digest*, paper 2.3, p. 19.
- [24] Miwa H., Tanaka T., Oshima K., Nakamura Y., Ishil T., Ohba A., Kouro Y., Furukawa T., Ikeda Y., Tsuchiya O., Hori R. and Miyazawa K. (1996) "A 140mm^2 64Mb AND Flash memory with a $0.4\mu\text{m}$ technology". *ISSCC Technical Digest*, TP 2.2, p. 34.
- [25] Kim K.S., Kim J.Y., Yoo J.W., Choi Y.B., Kim M.K., Nam B.Y., Park K.T., Ahn S.T. and Kwon O.H. (1995) "A novel dual string NOR (DuSNOR) memory cell technology scalable to the 256 Mbit and 1 Gbit Flash memories". *IEDM Technical Digest*, paper 11.1, p. 263.
- [26] Ohkawa M., Sugawara H., Sudi N., Tsukiji M., Nakagawa K., Kawata M., Oyama K., Takeshima T. and Ohya S. (1996) "A 98mm^2 die size 3.3 V 64 Mb Flash memory with FN-NOR type four-level cell". *IEEE JSSC*, **31**, p. 1584.
- [27] Ali S.B., Nguyen D., Sani D., Hu C., Ma Y., Kazerounian R., Shubat A. and Eitan B. (1989) "A new staggered virtual ground array architecture implemented in a 4 Mb CMOS EPROM". *VLSI Circuit Symp. Technical Digest*.
- [28] Cernea R., Lee D.J., Molidi M., Chang E.Y., Chien W.Y., Goh L., Fong Y., Yuan J.H., Samachisa G., Guterman D.C., Mehrotra S., Sato K., Onishi H.,

- Ueda K., Noro F., Miyamoto K., Morita M., Umeda K. and Kubo K. (1995) "A 34Mb 3.3 V serial Flash EEPROM for solid-state disk applications". *ISSCC Technical Digest*, TA 7.4, p. 126.
- [29] Harari E., "Highly compact EPROM and Flash EEPROM devices". US Patent 5,554,553.
- [30] Lee D.J., Cernea R.A., Mofidi M., Mehrotra S., Chang E.Y., Chien W.Y., Goh L., Yuan J.H., Mihnea A., Samachisa G., Fong Y., Guterman D.C. and Norman R.D. (1994) "An 18 Mb serial Flash EEPROM for solid-state disk applications". *VLSI Circuits Symp. Technical Digest*, paper 6.1, p. 59.
- [31] Harari E., "Flash EEPROM system cell array with more than two storage states per memory cell". US Patent 5,434,825.
- [32] Kazerounian R., Eitan B., Ma Y. and Ali S. (1988) "A 5 volt high density poly-poly erase Flash EPROM cell". *IEDM Technical Digest*, p. 436.
- [33] Eitan B. *et al.* (1991) "Alternate metal virtual ground (AMG) - A new scaling concept for very high-density EPROM's". *IEEE EDL*, **12**, 8, p. 450.
- [34] Kazerounian R., Bergemont A., Roy A., Wolstenholme G., Irani R., Shamay M., Gaffur H., Rezvani G.A., Anderson L., Haggag H., Shacham E., Kauk P., Nielson P., Kablanian A., Chhor K., Perry J., Sethi R. and Eitan B. (1991) "Alternate metal virtual ground EPROM array implemented in a 0.8 μm process for very high density applications". *IEDM Technical Digest*, p. 311.
- [35] Roy A., Kazerounian R., Irani R., Prabhakar V., Nguyen S., Slezak Y., Trinh C., Kauk P., Agarwal M., Eitan B., Annunziata R., Camerlenghi E., Campisi A., Cappelletti P., Colabella E., Crisenza E., Dallabora M., Fontana G., Pividori L. and Tosi M. (1997) "A new Flash architecture with a 5.8 F² scalable AMG Flash cell". *VLSI Tech Symp. Technical Digest*, p. 67.
- [36] Roy A. and Kazerounian R., "Flash EEPROM and EPROM arrays with select transistors within the bit line pitch". US Patent 5,557,124.
- [37] Bergemont A. *et al.* (1996) "Low voltage NVG: a new high performance 3 V/5 V Flash technology for portable computing and telecommunications applications". *IEEE Trans. on Elect. Dev.*, **43**, p. 1510.
- [38] Pasternak J., Fasoli L., Matarresse S., Advani M., Agarwal M., Golabi M., Prabhakar V., Gopinath P., Roy A., Kazerounian R., Lista V., Rizzari G.,

- Odierna P., Annunziata R., Geraci A., Paulino C., Campisi A. and Dal-labora M. (1998) "4Mb alternate metal, virtual ground Flash memory". *NVSM Workshop*, Monterey, California (USA).
- [39] Kianian S. *et al.* (1994) "A novel 3 volts-only, small sector erase, high density Flash EEPROM". *VLSI Tech. Symp.*, 6A.4, p. 71.
- [40] Van Houdt J.F. *et al.* (1995) "Investigation of the soft-write mechanism in source-side injection Flash EEPROM devices". *IEEE Trans. on Elect. Dev. Lett.*, **16**, p. 181.
- [41] Wu A.T., Chan T.Y., Ko P.K. and Hu C. (1986) "A novel high speed, 5 V programming EPROM structure with source-side injection". *IEDM*, 26.2, p. 584.
- [42] Ma Y., Pang C.S., Chang K.T., Tsao S.C., Frayer J.E., Kim T., Jo K., Kim J., Choi I. and Park H. (1994) "A dual-bit split-gate EEPROM (DSG) cell in contactless array for single-V_{cc} high density Flash memories". *IEDM Technical Digest*, paper 3.5, p. 57.
- [43] Ma Y.Y. and Chang K.T., "Self-aligned dual-bit split gate (DSG) Flash EEPROM cell". US Patent 5,278,439.
- [44] Kolodny A. *et al.* (1986) "Analysis and modeling of floating gate EEPROM cells". *IEEE Trans. on Elect. Dev.*, **33**, p. 835.
- [45] Van Tran H. *et al.* (1996) "A 2.5 V 256 level non-volatile analog storage device using EEPROM technology". *ISSCC Technical Digest*, FP 16.6, p. 270.
- [46] Makimoto T. (1996) "Market and technology trends in the nomadic age". *VLSI Symp. Technical Digest*, p. 6.
- [47] Bude J.D. (1995) "EEPROM/Flash sub 3.0 V drain-source bias hot carrier writing". *IEDM Technical Digest*, paper 3.7, p. 989.
- [48] Fischer B. *et al.* (1997) "Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages". *IEEE Trans. on Elect. Dev.*, **44**, p. 288.
- [49] Hirano H. *et al.* (1997) "2-V/100-ns 1T/1C non-volatile ferroelectric memory architecture with bitline-driven read scheme and nonrelaxing reference cell". *IEEE JSSC*, **32**, p. 649.

- [50] Shoji K. *et al.* (1996) "A $7.03 \mu\text{m}^2$ $V_{cc}/2$ -plate nonvolatile DRAM cell with a Pt/PZT/Pt/TiN capacitor patterned by one-mask dry etching". *VLSI Technology Symp. Technical Digest*, paper 3.4, p. 28.
- [51] Evans J.T. *et al.* (1988) "An experimental 512 bit non-volatile memory with ferroelectric storage cell". *IEEE JSSC*, **23**, p. 1171.
- [52] Guo L., Leobandung E. and Chou S.Y. (1996) "Si single electron MOS memory with nanoscale floating-gate and narrow channel". *IEDM Technical Digest*, paper 21.10, p. 955.
- [53] Welser J.J. *et al.* (1997) "Room temperature operation of a quantum-dot Flash memory". *IEEE Trans. on Elect. Dev. Lett.*, **18**, p. 278.
- [54] Roy A. (1989) "Ph.D. dissertation". Lehigh University, Bethlehem, PA.
- [55] French M.L. and White M.H. (1995) "Scaling of multielectric nonvolatile SONOS memory structures". *Solid State Elect.*, **37**, p. 1913.
- [56] French M.L. *et al.* (1994) "Design and scaling of a SONOS multielectric device for nonvolatile memory applications". *IEEE Trans. on Comp. Pack.*, **17**, p. 390.
- [57] Bauer M., Alexis R., Atwood G., Baltar B., Fazio A., Frary K., Hensel M., Ishac M., Javanifard J., Landgraf M., Leak D., Loe K., Mills D., Ruby P., Rozman R., Sweha S., Talreja S. and Wojciechowski K. (1995) "A multilevel-cell 32Mb Flash memory". *ISSCC Technical Digest*, p. 132.
- [58] Choi Y.J. *et al.* (1996) "A high speed programming scheme for multi-level NAND Flash memory". *VLSI Symp.*, 16.2, p. 170.
- [59] Kazerounian R. and Eitan B., Patent Pending.

4 PHYSICAL ASPECTS OF CELL OPERATION AND RELIABILITY

Luca Selmi¹, Claudio Fiegna²

¹ Dipartimento di Ingegneria Elettrica, Università di Udine
Via delle Scienze 208, 33100 Udine, Italy
uca@picolit.diegm.uniud.it

² Dipartimento di Ingegneria, Università di Ferrara
Via Saragat 1, 44100 Ferrara, Italy
cfiegna@ing.unife.it

Abstract: This chapter overviews the basic physical effects involved in programming and erasing of Flash memory cells, to provide the background for a deeper understanding of their operation and reliability. In particular, tunneling and high field transport are treated and the associated phenomena in MOS-FETs and Flash cells are described by means of measurements and simulations. Device degradation induced by charge injection into thin silicon dioxide layers is also briefly discussed.

4.1 INTRODUCTION

Flash memories are emerging in the non-volatile memory market as excellent candidates for embedded and mass storage applications. Similarly to EPROM and E²PROM cells, Flash ones are based on MOS transistors with a floating gate electrode between the accessible control gate and the channel. Thanks to the potential energy barrier arising at Si-SiO₂ interfaces the floating gate behaves as a potential well capable to retain charge for long times. Negative

charge stored on the floating gate increases the threshold voltage and prevents the formation of a conductive channel between source and drain. Therefore, presence or absence of charge on the floating gate can be associated with boolean values, but efficient and reliable charge injection and removal mechanisms must be found in order to change the stored data.

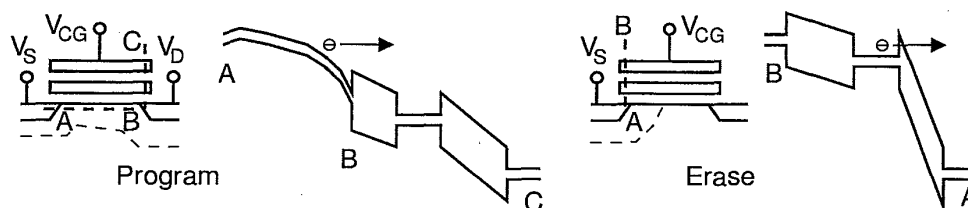


Figure 4.1 Schematic representation of programming and erasing mechanisms for the industry standard Flash cell.

In Flash memories these mechanisms are hot carrier emission and tunneling, as sketched in Fig. 4.1. In the former, electrons gain from the lateral field enough energy to surmount the barrier. In the latter, the thickness of the energy barrier between Si and SiO₂ is reduced by electrical means to allow the electrons to cross it. Hot carriers and tunneling are also responsible for spurious currents and device degradation, with direct consequences on the electrical characteristics and long term reliability of the cell. All these effects feature non-trivial dependencies on geometry, doping, bias. Understanding and correctly modeling them is of great importance for cell design and optimization.

Unfortunately, simplified models, often used to understand major trends and to steer device optimization, are becoming less accurate, and are sometimes stretched beyond their limits of applicability, while aspects that were of minor importance in the past, are sometimes becoming critical issues for cell operation and optimization.

In order to help developing a physically-based but intuitive understanding of these phenomena without sacrificing physical correctness, we approach the description of tunneling and hot-carrier transport starting from a brief overview of the fundamental Schrödinger and Boltzmann equations. Far from aiming to rigorous mathematical treatments, this starting point allows us to highlight the main assumptions underlying the models most often employed. Then, tunneling, hot carrier and degradation effects in MOSFETs and Flash cells are illustrated extensively by means of measurements and simulations.

According to this plan, the chapter has been organized in eight major Sections. Readers mostly interested in phenomenological aspects are referred to

Sections 4.4, 4.6, 4.7, and 4.8. The remaining Sections report definitions and more fundamental physical and modeling topics.

4.2 ELECTRONIC PROPERTIES OF CARRIERS AND MOS STRUCTURES

4.2.1 Electrons in Crystals

From the point of view of quantum mechanics electrons in crystals can be described by wave functions $\Psi_0(\vec{r}, t)$ obeying the single electron Schrödinger equation:

$$j\hbar \frac{\partial \Psi_0}{\partial t} = -\frac{\hbar^2}{2m_0} \nabla^2 \Psi_0 + [U(\vec{r}) + U_C(\vec{r}) + U_S(\vec{r}, t)] \Psi_0(\vec{r}, t) . \quad (4.1)$$

where $m_0 = 9.16 \times 10^{-31} \text{ kg}$ is the free electron mass, $\hbar = 1.05 \times 10^{-34} \text{ J}\cdot\text{s}$ is the reduced Planck constant. The potential term is the sum of three components: 1) $U_C(\vec{r})$ is the crystal potential generated by the ions and the other electrons; 2) $U_S(\vec{r}, t)$ is the scattering potential, representing intense but localized and short perturbations caused by impurities, lattice vibrations, etc.; 3) $U(\vec{r})$ is the potential due to externally applied voltages and macroscopic space charge distributions, which is assumed to be a slowly varying function of $\vec{r}$ compared to U_C and U_S . The probability of finding the electron in $\vec{r}$ at time t is given by the probability density:

$$\rho_{\Psi} = |\Psi_0(\vec{r}, t)|^2 = \Psi_0^*(\vec{r}, t) \Psi_0(\vec{r}, t) , \quad (4.2)$$

where Ψ_0^* is the complex conjugate of Ψ_0 . To evaluate (4.2), Ψ_0 must be calculated solving (4.1) with appropriate boundary conditions.

In a perfect infinite crystal $U_C(\vec{r})$ has the periodicity of the lattice. In the absence of scattering, and of applied and built-in potential gradients (that is, $U_S = 0$ and constant U), the solution of (4.1) with the periodic boundary conditions imposed by $U_C(\vec{r})$ is a Bloch wave:

$$\Psi_0(\vec{r}, t) = u_{\vec{k}}(\vec{r}) \exp(j\vec{k} \cdot \vec{r}) \exp(-jEt/\hbar) , \quad (4.3)$$

where $\vec{k}$ is the wave vector, $u_{\vec{k}}(\vec{r})$ is a rapidly oscillating function of position with the periodicity of the lattice, and $E = E(\vec{k})$ is the total electron energy, that is, the sum of its potential (U) and kinetic (E_K) energy, which is a function known as the band structure of the material.

In crystals the allowed electron energies are grouped in bands separated by gaps. The first band featuring empty states at 0K is the conduction band. For

E close to the bottom of the conduction band, $E(\vec{k})$ is commonly approximated as a spherical parabolic band:

$$E(\vec{k}) = E_{C0} + \frac{\hbar^2}{2m^*} |\vec{k} - \vec{k}_0|^2, \quad (4.4)$$

where E_{C0} denotes the band bottom energy, $\vec{k}_0$ is the corresponding wave vector, m^* is the electron effective mass and $\hbar^2|\vec{k} - \vec{k}_0|^2/2m^*$ is the kinetic energy E_K . With a suitable choice of coordinates $\vec{k}_0$ will be taken equal to zero in the following. Therefore, within the approximation (4.4) $E(\vec{k})$ becomes independent of $\vec{k}$ direction.

Once $E(\vec{k})$ is known, the density of states (that is the number of available electron states per unit volume and energy) can be computed as:

$$g(E) = \frac{1}{4\pi^3} \int \delta(E - E(\vec{k})) d\vec{k}, \quad (4.5)$$

where $\delta(x)$ denotes the Dirac distribution and $1/4\pi^3$ is the number of electron states per unit volume in the six-dimensional $\vec{r}$ and $\vec{k}$ space. $g(E)$ is a very important property of the material which comes into place whenever macroscopic quantities are computed, e.g. carrier concentration or current density.

In practical situations the applied and built in potential is never constant. Compositional material changes and external stimuli result in spatially varying $U(\vec{r})$ and $E_{C0}(\vec{r})$ superimposed to U_C , that make the exact solution of (4.1) very difficult. Exploiting the knowledge of $E(\vec{k})$ in each material, a simpler effective mass equation can be written, in which the periodic U_C does not explicitly appear. For spherical parabolic bands (Eq. (4.4)) the effective mass equation reads [1]:

$$j\hbar \frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m^*} \nabla^2 \Psi + [E_{C0}(\vec{r}) + U(\vec{r}) + U_S(\vec{r}, t)] \Psi(\vec{r}, t), \quad (4.6)$$

where m^* is the electron effective mass, $E_{C0} + U$ is the spatially varying conduction band edge and Ψ is the envelope function that relates to Ψ_0 as $\Psi_0(\vec{r}, t) = u_{\vec{k}}(\vec{r}) \Psi(\vec{r}, t)$. Since Ψ is sensitive only to the smoothly changing built in and applied potentials, it oscillates with a much larger period than $u_{\vec{k}}(\vec{r})$. Thus, Ψ is a smooth function describing the envelope of Ψ_0 . Neglecting the short range oscillations of Ψ_0 , the probability density is expressed as $\rho_\Psi = |\Psi(\vec{r}, t)|^2$.

4.2.2 Electrons as Classical Particles

In quantum mechanics, a classical particle can be described by a wave packet, that is by the superposition of solutions of (4.6) featuring different wave vectors.

The wave packet probability density is significantly non-zero only in a restricted space region which approximately identifies the particle position. This latter is well represented by the centroid ($\vec{r}_{wp}$) of the wave packet probability density ρ_Ψ .

If the potential energy gradient is approximately constant in the space region where the wave packet extends, the average wave vector of the packet satisfies Newton's law of motion:

$$\hbar \frac{d\vec{k}_{wp}}{dt} = -q\vec{F}, \quad (4.7)$$

where $\vec{F}$ is the field associated to built-in and applied potentials. The effect of the crystal, that is the force between electrons and the atom's nuclei, is assumed to be completely described by the band structure of the material $E(\vec{k})$. The effects of scattering potentials are usually approximated by instantaneous collisions that suddenly modify the particle momentum.

Since the particle is actually described by a wave packet, its velocity is the group velocity of the wave packet evaluated for the average wave vector of the wave packet:

$$\vec{v}_g = \frac{d\vec{r}_{wp}}{dt} = \frac{1}{\hbar} \nabla_{\vec{k}} E(\vec{k})_{\vec{k}=\vec{k}_{wp}}, \quad (4.8)$$

and not the phase velocity of the wave packet component at $\vec{k} = \vec{k}_{wp}$. $\vec{v}_g$ can be derived directly from the band structure $E(\vec{k})$ and can reach values in excess of 10^8 cm/s in silicon. Notice that for parabolic bands (Eq. (4.4)) $\vec{v}_g = \vec{p}/m^*$, which is the well known result for a classical particle with mass equal to m^* .

4.2.3 Silicon

Crystalline silicon (Si) is made of spatially ordered atoms with face centered cubic symmetry. The corresponding $E(\vec{k})$ is a complex three dimensional function of k_x, k_y, k_z [4] and the energy density of states $g(E)$, reported in Fig. 4.2, features an energy gap between conduction and valence bands $E_G(\text{Si}) \approx 1.12$ eV at room temperature. The gap is indirect, i.e. the bottom of the conduction band is at a different $\vec{k}$ than the top valence band. Within the first few hundreds meV from its bottom, the conduction band is often approximated with a simple spherical parabolic band (Eq. (4.4)) with an effective mass $m^* \simeq 0.32m_0$.

4.2.4 Silicon Dioxide

Crystalline silicon dioxide (SiO_2) exists in a few allotropic forms. The most important for electronics, due to its resemblance with that of gate oxides, is that of α -quartz. α -quartz features a perfectly ordered arrangement of Si atoms located at the center of tetrahedra, and oxygen atoms at the vertexes.

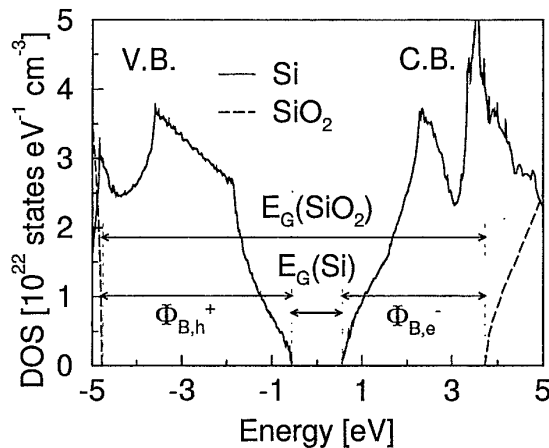


Figure 4.2 Density of states (DOS) of perfect Si and SiO₂ crystals (α -quartz). In this plot $E_G(\text{Si}) = 1.12\text{eV}$ and $E_G(\text{SiO}_2) = 8.5\text{eV}$. $E = 0$ at the Si midgap. Data from [2, 3].

Each oxygen atom occupies a bridging location and forms two chemical bonds with Si atoms belonging to adjacent tetrahedra. The energy density of states of α -quartz is shown in Fig. 4.2 and features an indirect energy gap $E_G(\text{SiO}_2)$ whose reported values range between 8.0eV [5] and 9.2eV [2]. The conduction band can be approximated up to a few eV by a single spherical parabolic band with effective mass $m_{\text{ox}}^* \approx 0.5m_0$ [2].

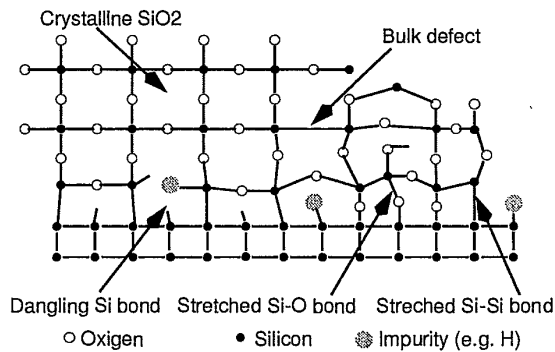


Figure 4.3 Schematic representation of the Si-SiO₂ interface and of defects leading to the formation of extra bulk and interface states.

SiO₂ films used as gate dielectric feature an amorphous (vitreous) form, meaning that the atomic structure is ordered only over short distances. Oxygen

atoms are often placed in non-bridging locations, i.e. they do not share bonds with the silicon atoms of two adjacent tetrahedra, as in a perfect SiO_2 crystal. In other words, a certain fraction of chemical bonds within the bulk of the oxide is stretched or broken (dangling), as illustrated in Fig. 4.3. The corresponding material structure is looser and has a smaller density than that of α -quartz ($\approx 2.2 \text{ g/cm}^3$ instead of $\approx 2.65 \text{ g/cm}^3$), making it possible for a variety of small impurity atoms (hydrogen, sodium, etc.) to enter the oxide and diffuse through it, with important consequences on the electrical properties and reliability of the Si-SiO₂ interface, as will be further discussed in Sections 4.2.6 and 4.8. The partly random distribution of atoms and atomic bonds in gate dielectrics produces a random perturbation of the otherwise periodic crystal potential, which affects mobility, and other oxide electrical properties.

4.2.5 Silicon - Silicon Dioxide Interface

The interruption of the Si crystal continuity at the Si-SiO₂ interface generates extra allowed energy levels confined within a few angstroms of the surface and associated to surface atoms [6, 7]. These interface states are also referred to as fast surface states because they can rapidly exchange charge with the inversion layer. They arise primarily from dangling silicon and oxide bonds at the surface (Fig. 4.3).

In an ideal Si-SiO₂ interface all Si bonds are saturated by oxygen atoms. However, the compositional material change generates an abrupt built-in potential difference $\Phi_{B,e^-} \simeq 3.15 \text{ eV}$ between the Si and SiO₂ conduction band edges, and of $\Phi_{B,h^+} \approx 4.3 \text{ eV}$ between the valence band maxima. These barriers oppose carrier motion from Si to SiO₂ and for most purposes can be modeled as a step like E_{C0} , as shown in Fig. 4.4. Therefore, from now on they will be incorporated in the applied potential term U (see Eq. (4.6)).

4.2.6 Oxide and Interface Traps

Stretched and dangling bonds in the bulk of the oxide and at the Si-SiO₂ gate and channel interfaces are believed to produce sufficiently strong localized potential perturbations to generate bound states for electrons or holes with energy levels in the forbidden gap (Fig. 4.4). These levels act as localized electron or hole traps, i.e. they can capture or release carriers.

Traps can be categorized in two types: acceptor traps are neutral when empty and negatively charged when occupied; donor traps are neutral when occupied and positively charged when empty. In MOS structures, interface traps are believed to be mostly of acceptor type in the upper half of the band gap and of donor type in the lower half [8, 9].

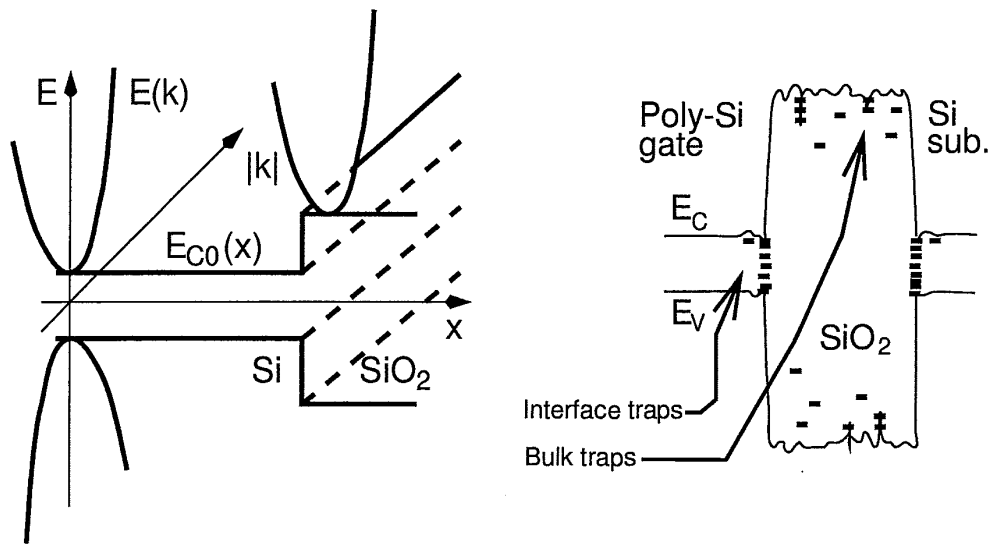


Figure 4.4 Left: Schematic representation of band structure transition at the Si-SiO₂ interface assuming $U(\vec{r}) = 0$. Right: Schematic representation of the perturbation potentials and of the interface and bulk states in a Si-SiO₂-Si stack.

Carrier trapping and detrapping can occur through a large number of processes: 1) thermal capture and emission; 2) photon assisted transitions; 3) particle assisted transitions (impact release of carriers and Auger recombinations); 4) elastic tunneling in and out of the traps [10].

The steady state occupancy of the traps, hence their charge state, changes with bias according to the difference between their energy and the quasi-Fermi level. Under dynamic conditions, instead, trap occupancy and trapping time constants are determined by the efficiency of the filling/emptying processes previously mentioned, and can change by several orders of magnitude depending on bias, oxide quality and growth conditions and temperature.

The charge statically or dynamically accumulated in the traps perturbs the built in potential and can have a strong impact on the electrical characteristics of the interface and of the SiO₂ insulating film. This point will be further discussed in Section 4.7.

In high quality interfaces, most bonds are actually satisfied by oxygen, silicon or impurity atoms as roughly sketched in Fig. 4.3, thus reducing significantly the density of available traps (D_{it}) compared to the number of surface bonds ($N_{it} \approx 6.8 \times 10^{14} \text{cm}^{-2}$ for $\langle 100 \rangle$ surfaces). Impurities, and mostly hydrogen migrating through the loose oxide structure, or intentionally incorporated during

processing, are very effective in passivating the interface bonds, thus further reducing the trap concentration.

However, stretched bonds and bonds formed by contaminants and hydrogen with the underlying crystalline silicon, can be much weaker than well oriented Si-O bonds. Therefore, they are easily broken in the interaction with radiation and hot carriers, generating interface traps that degrade the electronic properties of the surface and impact the electrical performance of the devices (threshold voltage, transconductance, current, etc. [11]).

As an example, it is easy to verify that a fraction of charged interface traps D_{it} as small as 10^{-3} of N_{it} distributed within a distance comparable to that of one monolayer from the channel/oxide or gate/oxide interfaces ($\Delta x \approx 2\text{\AA}$) can induce non-negligible threshold voltage shifts ($\Delta V_T = qD_{it}(1 - \Delta x/t_{ox})/C_{ox} \approx 0.3\text{V}$ and $\Delta V_T = qD_{it}\Delta x/(t_{ox}C_{ox}) \approx 6\text{mV}$, respectively, in a 10nm thick gate oxide MOSFET). Appropriate interface passivation is thus crucial for stable device operation. These problems will be further discussed in Sections 4.7 and 4.8.

4.3 FUNDAMENTALS OF TUNNELING PHENOMENA

4.3.1 Basic Concepts and the WKB Approximation

Tunneling through energy barriers is of great importance in Flash memories, as it is exploited to inject or extract charge in and out of the floating gate (see Chapter 2). In essence tunneling consists in the possibility for electrons to cross classically prohibited regions in which the electron energy (E) is lower than the effective potential energy (U , that includes band edge transitions at the interfaces).

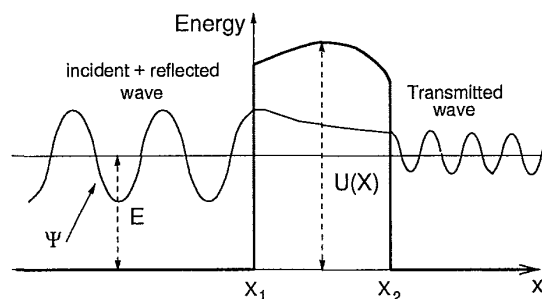


Figure 4.5 Transmission of an electron represented by the envelope function Ψ through a potential energy barrier. x_1 and x_2 are the so called classical turning points.

This process is exemplified in Fig. 4.5 reporting a qualitative sketch of the envelope function of an electron in the presence of an energy barrier. As can be seen, Ψ penetrates the region where $E < U$, decays exponentially therein, and emerges on the other side of the barrier. Since the amplitude of the envelope function (hence, ρ_Ψ) is non zero for $x > x_2$, a finite probability exists that the electron penetrates the barrier.

Tunneling can be explained only by attributing wavelike properties to electrons. Therefore, even a first order analysis of the phenomenon requires the solution of (4.6). To this purpose, neglecting scattering ($U_S = 0$, that is, considering only elastic tunneling processes), and separating the variables $\vec{r}$ and t , it is immediately found that the envelope function obeys the time-independent Schrödinger equation in the effective mass approximation [1]:

$$-(\hbar^2/2m^*)\nabla^2\Psi(\vec{r}) + U(\vec{r})\Psi(\vec{r}) = E\Psi(\vec{r}) \quad , \quad (4.9)$$

where $U(\vec{r})$ incorporates $E_{C0}(\vec{r})$. For a one dimensional barrier as that in the direction perpendicular to the Si-SiO₂ interface of a MOS capacitor, separation of variables can be applied to (4.9) leading to a one-dimensional problem. Nevertheless, solving (4.9) is difficult and expensive, so that several approximate methods have been developed to this purpose. Among these, the WKB (Wentzel-Kramers-Brillouin) method is that most frequently used for first order analysis of tunneling in the MOS systems of interest here.

Assuming slowly varying potential energy profiles, the WKB method approximates the incident wave propagating in the positive x direction as [12]:

$$\Psi(x) \simeq A \frac{1}{\sqrt{k_\perp(x)}} \exp\left(+j \int_{-\infty}^x k_\perp(x) dx\right) \quad , \quad (4.10)$$

where $\hbar k_\perp(x) = p_x = \sqrt{2m^*E_\perp(x)}$ is the momentum component in the direction perpendicular to the interface, and $E_\perp(x) = p_x^2/2m^* = E - U(x) - (p_y^2 + p_z^2)/2m^*$ is the associated component of the kinetic energy. Since zero field is assumed in the direction parallel to the interface the parallel momentum and the corresponding energy $E_\parallel = (p_y^2 + p_z^2)/2m^*$ are conserved along the tunneling path. Therefore, $E_\parallel$ will be treated as a constant in the following.

4.3.2 Transmission Coefficient

For application purposes it is necessary to evaluate the transmission coefficient, which represents the probability that an electron described by the envelope function Ψ crosses the region between x_1 and x_2 (see Fig. 4.5). This probability is given by:

$$T = \frac{J_\Psi(x_2)}{J_\Psi(x_1)} \quad , \quad (4.11)$$

where $J_\Psi(x)$ is the probability current density defined as:

$$J_\Psi(x) = \frac{-j\hbar}{2m^*} \left[\Psi^* \frac{\partial \Psi}{\partial x} - \Psi \frac{\partial \Psi^*}{\partial x} \right] . \quad (4.12)$$

The probability current density of the envelope function (4.10) is:

$$J_\Psi(x) = \frac{\hbar k_\perp(x)}{m^*} |\Psi(x)|^2 = v_\perp \rho_\Psi , \quad (4.13)$$

which can be intuitively interpreted as the product of the electron velocity in the tunneling direction ($v_\perp$) by the probability density (ρ_Ψ).

Substituting (4.13) into (4.11) we get:

$$T = \frac{v_\perp(x_2)}{v_\perp(x_1)} \frac{|\Psi(x_2)|^2}{|\Psi(x_1)|^2} \approx \frac{|\Psi(x_2)|^2}{|\Psi(x_1)|^2} = \exp \left(+2j \int_{x_1}^{x_2} k_\perp(x') dx' \right) , \quad (4.14)$$

where the velocity ratio is often assumed equal to unity. Notice that $E_\perp < 0$ in the classically forbidden region between x_1 and x_2 , thus $k_\perp(x)$ is imaginary and an exponential attenuation of the envelope function occurs, leading to $T \leq 1$.

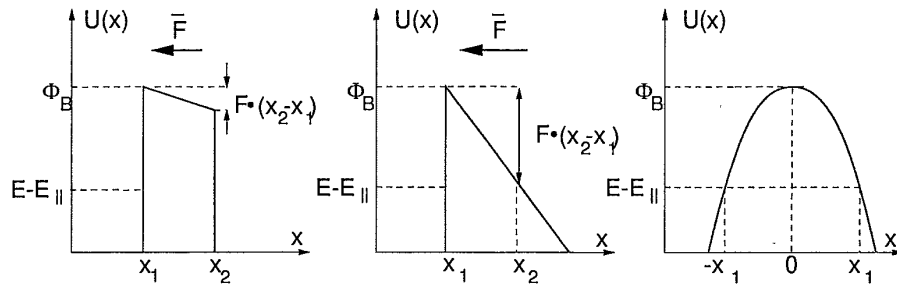


Figure 4.6 Trapezoidal, triangular and parabolic potential energy profiles.

Fig. 4.6 shows simple trapezoidal, triangular and parabolic energy barriers for which the tunneling probability can be analytically derived with the WKB method. The first two profiles schematically represent the energy barrier perpendicular to the Si-SiO₂ interface under the bias conditions typical of program (trapezoidal) and erase (triangular) operations in an industry standard Flash memory cell (see Chapter 2). The parabolic barrier, as will be seen in Section 4.4.3, represents the effective barrier for electrons tunneling from the valence into the conduction band in band-to-band transitions.

Eq. (4.14) leads to the following expressions of the tunneling coefficient (in standard units):

$$T_{\text{trap.}}(\xi) \simeq \exp \left(-\frac{4(2m^*)^{1/2}}{3\hbar q} \cdot \frac{\xi^{3/2} - [-qF(x_2 - x_1) + \xi]^{3/2}}{F} \right) \quad (4.15)$$

$$T_{\text{triang.}}(\xi) \simeq \exp\left(-\frac{4(2m^*)^{1/2}}{3\hbar q} \cdot \frac{\xi^{3/2}}{F}\right), \quad (4.16)$$

$$T_{\text{parab.}}(\xi) \simeq \exp\left(-\frac{\pi x_1(2m^*)^{1/2}}{\hbar} \cdot \xi^{1/2}\right), \quad (4.17)$$

where F denotes the absolute value of electric field for the triangular or trapezoidal barriers, and $\xi = \Phi_B - (E - E_{\parallel}) = \Phi_B - p_x^2/2m^* \geq 0$ is the effective barrier height. Notice that the wave function (4.10) is not defined at the classical turning points x_1 and x_2 where $k_{\perp} = 0$. Nevertheless T is finite. In the WKB approximation $T = 1$ for $\xi < 0$, although quantum-mechanical reflections actually occur also for $\Phi_B < E - E_{\parallel}$. For a fixed total energy E , ξ increases and T decreases exponentially as $E_{\parallel}$ increases, because less energy and momentum are available in the tunneling direction.

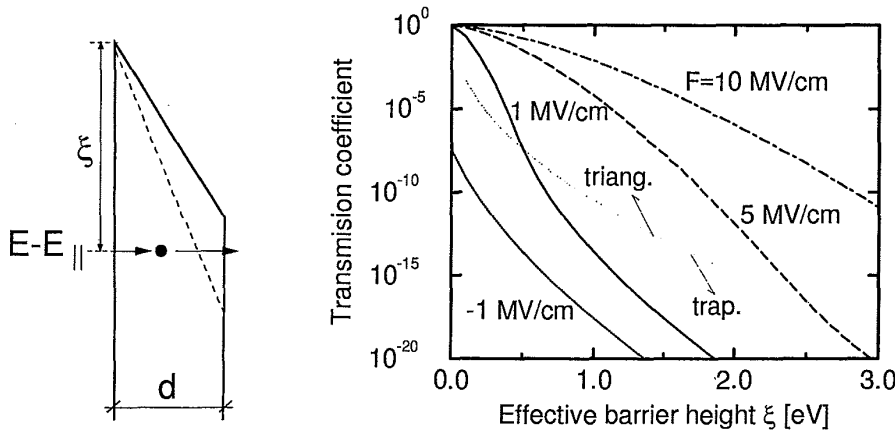


Figure 4.7 Transmission coefficient through an energy barrier as a function of the effective barrier height ξ for different field strengths and an barrier thickness $d = 5\text{nm}$. The transition between triangular and trapezoidal barrier shape is indicated by the dotted line.

Fig. 4.7 reports $T(\xi)$ computed according to (4.15) and (4.16) for different values of the electric field F . $T(\xi)$ decreases rapidly for increasing ξ , and it is a weaker function of energy for trapezoidal barriers than for triangular ones, as evidenced by the curve for $F = 1\text{MV/cm}$. The transition between triangular and trapezoidal barrier takes place for $\xi = F(x_2 - x_1)$, a condition for which (4.15) and (4.16) give the same T value.

The transmission coefficients change with temperature only because of the weak temperature sensitivity of the effective mass. As we will see, this is one of the reasons why most tunneling phenomena in devices are weakly dependent on temperature.

4.3.3 Tunneling Current

The tunneling current from region 1 ($x < x_1$ in Fig. 4.5) to region 2 ($x > x_2$) can be computed summing the contributions of all filled states in region 1 weighted by the corresponding tunneling probability T . The elemental contribution of electrons with x-component of momentum p_x and energy E can be expressed as [12]:

$$dJ_{1 \rightarrow 2} = q \frac{p_x}{m^*} T g_1^*(E, p_x) f_1(E) (1 - f_2(E)) dp_x dE, \quad (4.18)$$

where p_x/m^* is the x-component of the velocity of particles hitting the interface, $g_1^*(E, p_x) dp_x$ is the number of available electron states with momentum p_x inside a bin dp_x (thus, $\int_{-\infty}^{\infty} g^*(E, p_x) dp_x$ is the energy density of states, $g(E)$), $f_1(E)$ is the occupation probability of the initial state in region 1 (that is, the distribution function introduced in more detail in Section 4.5.1) and $(1 - f_2(E))$ is the probability of finding an empty state in region 2 at the same energy E (as only elastic tunneling is considered). In quasi equilibrium conditions, f_1 and f_2 are given by Fermi-Dirac statistics, while in strong off equilibrium conditions they should be calculated with the methods discussed in Section 4.5. Eq. (4.18) does not include the density of states in region 2, thus implicitly assuming that for each electron tunneling from region 1, a final state exists in region 2 ensuring conservation of energy and transverse momentum.

In general, we must also take into account the reverse current from region 2 to region 1 so that the net current flowing in the positive x direction is given by:

$$J = \int dJ_{1 \rightarrow 2}(E, p_x) - \int dJ_{2 \rightarrow 1}(E, p_x). \quad (4.19)$$

In many practical cases, however, one of the two terms in (4.19) is negligible with respect to the other one, because f_1 (f_2) is much smaller than f_2 (f_1) for a given energy and $T_{1 \rightarrow 2}$ is much smaller than $T_{2 \rightarrow 1}$ for a given occupation probability f , or vice versa. Notice also that since T is almost independent of lattice temperature, J is a function of T_L mainly through the f and g terms.

4.4 TUNNELING PHENOMENA IN MOSFETS

Tunneling is of utmost relevance for MOS transistors and Flash cells. Under suitable bias conditions, carriers can tunnel through the tunnel or inter-poly oxides, giving rise to gate currents that charge/discharge the storage node of

floating gate devices. Tunneling electrons emerge at the anode with energy high above the bottom of the conduction band and can generate typical hot carrier effects such as impact ionization (see Section 4.6.4).

In addition, at low gate and high source or drain voltages band-to-band (Zener) tunneling between valence and conduction bands can occur in the gate-source (gate-drain) overlap regions, providing a major contribution to sub-threshold leakage currents.

Finally, at the high fields involved in tunneling problems, charge flow in the oxide is accompanied by carrier trapping and generation of new traps, with important consequences on device reliability.

In the following, some of these phenomena that are particularly relevant for Flash cells will be analyzed in more detail.

4.4.1 Fowler-Nordheim and Direct Tunneling Through Gate Oxides

Fig. 4.8 represents a schematic view of the energy bands of a MOS system upon application of a positive gate voltage. As V_G increases from V_{G1} to V_{G2} ,

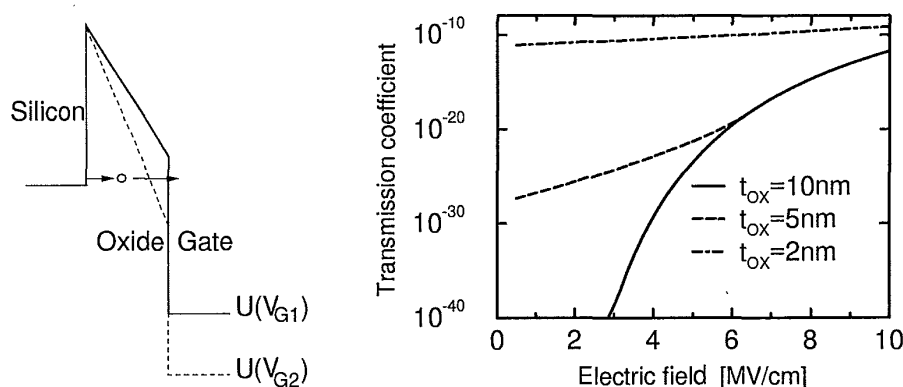


Figure 4.8 Left: Schematic representation of direct and Fowler-Nordheim tunneling conditions. Right: Transmission coefficient as a function of electric field for $\xi = 3.15\text{eV}$.

the oxide field increases and the potential barrier seen by thermal electrons near the band bottom changes from trapezoidal to triangular. Eventually, a detectable electron current starts to flow through the oxide. Currents through trapezoidal barriers are referred to as direct tunneling currents, because electrons are injected directly into the gate electrode. The triangular barrier case, instead, is referred to as Fowler-Nordheim tunneling because of its analogy with

the emission of electrons from cold cathodes theoretically studied by these two authors [13]. In thick oxides the bias dependence of the tunneling current is dominated by that of the transmission coefficient on the oxide field, and can be controlled by electrical means acting on V_G . In very thin oxides ($< 3\text{nm}$) the bias dependence of f and g in (4.18) also play a significant role.

Fig. 4.8 also reports the transmission coefficient as a function of the oxide field for different oxide thicknesses and $\xi = 3.15\text{eV}$ as in a MOS system. As can be seen, the tunneling probability does not depend on oxide thickness in the Fowler-Nordheim regime ($F_{\text{OX}}t_{\text{OX}} \geq \xi$). In addition, thin oxides exhibit enhanced transmission coefficients (hence leakage currents) at low field due to direct tunneling. In order to achieve long retention times in the memory cell, T must be very small at low fields; therefore, the oxide thickness must be large enough to keep direct tunneling negligible (see Chapter 2). For this reason the oxide thickness of Flash cells can not be scaled below $\approx 7\text{--}8\text{nm}$ [14]. For $t_{\text{ox}} \geq 7\text{nm}$, detectable currents can be observed only for $F \geq 6\text{MV/cm}$ which falls well into the Fowler-Nordheim regime. Tunneling currents suitable for fast erase of Flash cells typically require a tunneling probability larger than 10^{-10} ; hence, $F \approx 10\text{MV/cm}$ (Fig. 4.8). This is the typical range of oxide fields used in applications.

4.4.2 Modeling the Tunnel Current of MOS Structures

In order to calculate the current-voltage characteristics of MOS structures the procedure outlined in Section 4.3.3 should be followed. However, several complications arise, that are schematically illustrated in Fig. 4.9 and discussed below.

- The potential energy drops in the semiconductor (Φ_{Si}) and polysilicon gate (Φ_P), the actual oxide thickness (t_{ox}) and the barrier height (Φ_B) must be accurately known in order to derive precisely the oxide field ($F_{\text{OX}} = (V_G - V_{\text{FB}} - \Phi_{\text{Si}} - \Phi_P)/t_{\text{ox}}$), hence the correct transmission coefficient T .
- The integral in (4.14) requires the knowledge of the dispersion relationship $E(\vec{k})$ within the forbidden gap, a quantity of questionable physical meaning for an amorphous material such as SiO_2 , which is often approximated extrapolating into the gap ($E < U$) the parabolic dispersion relationship of the oxide conduction band, i.e. $E - U = \hbar^2 k^2 / 2m_{\text{ox}}^*$ [15]. More accurate dispersion relationships that match the oxide conduction and valence band effective masses ($m_C^* \approx 0.5m_0$, $m_V^* \approx 5m_0$ [2]) have been derived in [15, 16].

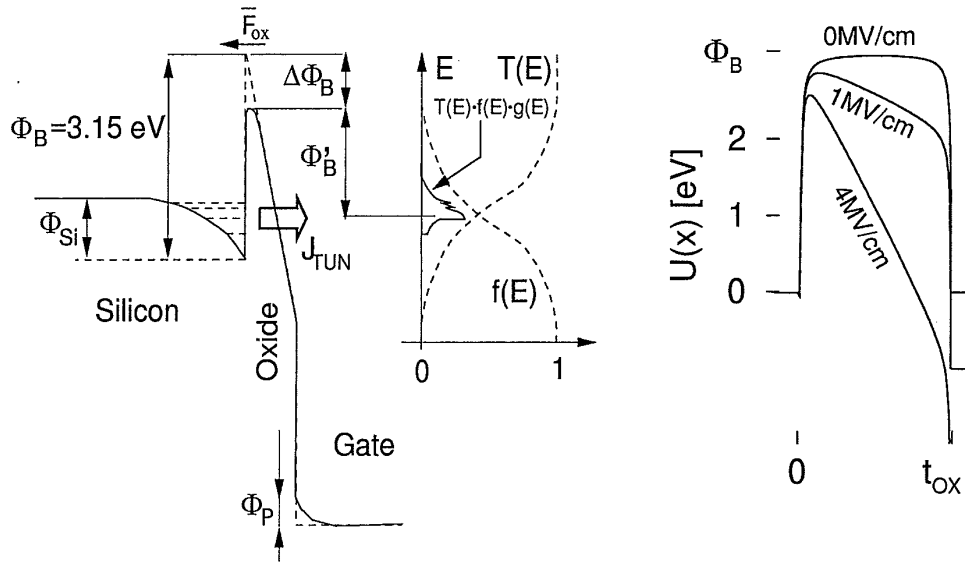


Figure 4.9 Left: Band diagram of a MOS structure in the FN regime. Dashed lines represent the quantized energy levels in the injecting electrode and the ideal barrier profile. Center: Schematic representation of the distribution function f , the transmission probability T and the product $f \cdot g \cdot T$ at the substrate-oxide interface. Right: Field induced barrier lowering according to (4.20); $\Phi_B = 3.15 \text{ eV}$, $t_{ox} = 10 \text{ nm}$.

- Electrons crossing the oxide locally perturb the potential. These perturbations can be treated at first order as image force corrections [17, 18, 19]. Sketching the highly doped polysilicon gate and the inversion layer as perfectly conductive surfaces ($\epsilon \rightarrow \infty$), the electron as a point charge and considering only the first order images with respect to the gate and channel interfaces, the potential energy in the oxide ($0 < x < t_{ox}$) is given by [18]:

$$U(x) \simeq \Phi_B - q|F_{OX}|x - \frac{q^2}{16\pi\epsilon_\infty x} - \frac{q^2}{16\pi\epsilon_\infty(t_{ox} - x)}, \quad (4.20)$$

where ϵ_∞ is generally taken as the high frequency permittivity of the oxide ($\simeq 1.9 \times 10^{-13} \text{ F/cm}$, [20]). The fourth term in the right hand side of (4.20) is typically much smaller than the third one for $t_{ox} > 5 \text{ nm}$, as is the case in Flash technologies. By neglecting it, we can estimate the image force induced reduction of the effective barrier height as $\Delta\Phi_B = \sqrt{q^3 F_{OX} / 4\pi\epsilon_\infty}$. As illustrated in Fig. 4.9, the correction term amounts to several hundreds meV for F_{OX} of a few MV/cm. Since T depends on the integral of $k_\perp$ in the forbidden region (Eq. (4.14)), image force

corrections to the shape of the top of the barrier become increasingly important as the electron energy increases.

- Since T is an increasing function of energy while $f \cdot g^*$ is a decreasing one, the product $T \cdot f \cdot g^*$ in (4.18), measuring the contribution of each energy level to the total tunneling current, peaks at an energy higher than the bottom of the conduction band (Fig. 4.9). Consequently, even if Φ_B is accurately known, the effective barrier height Φ'_B (i.e. the difference between the top of the barrier and the energy level providing the maximum contribution to the tunneling current) is lower than Φ_B .
- Quantization modifies the density of states with respect to its bulk values. Since quantized energy levels change with bias, Φ'_B is a bias dependent quantity [21]. Numerical calculations for accumulated layers indicate that, neglecting image force corrections, this effect can be modeled as a field dependent barrier lowering $\Phi'_B = \Phi_B - (d_0 + d_1|F_{OX}|)$ where $d_0 = 0.112\text{eV}$ and $d_1 = 1.69 \times 10^{-8}\text{eV cm/V}$ in the temperature range 77–300K [22].
- Trapped charges modify the barrier shape, leading to lower or higher currents depending on their sign. Furthermore, trap assisted tunneling can enhance the current considerably in very thin or heavily stressed oxides.

Obviously, the effects above can be accounted for only by numerical means. In particular, in the presence of quantum confinement at the interface and of image force effects, T can be efficiently calculated through a piecewise linear approximation of the potential profile [23], possibly coupled with a self consistent solution of the Poisson and Schrödinger equations [24, 25, 26].

A simplified analytical expression for the MOS tunneling current density has been given in [27] for triangular barriers. Starting from the Fowler-Nordheim theory of electron tunneling from a metal electrode [13], assuming $f_2(E) = 0$ (i.e. final states at the classical turning point in the oxide are always available), a parabolic energy-wave vector dispersion relationship in the oxide gap, a lattice temperature $T_L = 0\text{K}$ (so that f_1 is a simple step function), and neglecting the reverse current from the oxide to the silicon, we get [27]:

$$J_{\text{TUN}} = \frac{q^3 F_{\text{OX}}^2 m_0}{16\pi^2 \hbar m_{\text{ox}}^* \Phi'_B} \exp \left(\frac{-4(2m_{\text{ox}}^*)^{1/2}}{3\hbar q} \frac{\Phi_B'^{3/2}}{F_{\text{OX}}} \right), \quad (4.21)$$

where Φ'_B is an effective energy barrier accounting for the actual barrier shape and the distribution of injecting levels from the bottom of the conduction band.

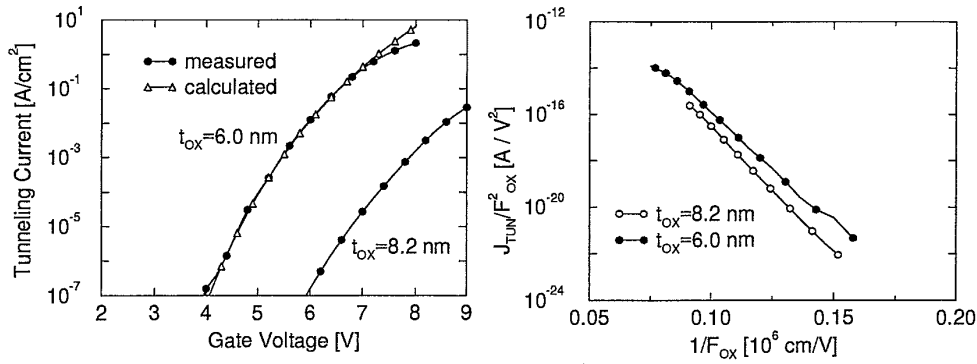


Figure 4.10 Left: Experimental tunneling current through thin gate oxides as a function of gate voltage. The current calculated with (4.21) and $\Phi'_B = 2.85\text{eV}$ is also shown for comparison. Right: Fowler-Nordheim plot of the J-V characteristic shown on the left.

Notice that, due to the 0K assumption, the exponential term in (4.21) is essentially the tunneling probability (4.16), while the pre-exponential term stems from the integration (4.19) of (4.18) under the assumptions outlined above.

Fig. 4.10 shows typical tunneling curves as a function of gate bias and compares them with (4.21). Φ'_B has been adjusted to optimize agreement with experiments. By taking Φ'_B equal to the commonly accepted value ($\Phi_B \simeq 3.15\text{eV}$, [28]) the calculated current would be much smaller than experimental values, because of the correction effects mentioned earlier. At high current, the calculations exceed the measured values because of the voltage drop on the parasitic resistance in series with the MOS capacitor and of charge trapping in the oxide (see also Section 4.7.1).

Fig. 4.10 also reports the tunneling current in the so called Fowler-Nordheim plot, that is as $J_{\text{TUN}}/F_{\text{OX}}^2$ versus $1/F_{\text{OX}}$. According to (4.21) the data should follow a straight line, and Φ'_B could be extracted from the slope of the curves ($S_{\text{FN}} = d(\log J_{\text{TUN}}/F_{\text{OX}}^2)/d(1/F_{\text{OX}}) = -4\sqrt{2m_{\text{ox}}^*}\Phi_B'^{3/2} \log_{10} e$) once $m_{\text{ox}}^* \simeq 0.5m_0$ is known. It has been argued, however, that in presence of channel quantization the Fowler-Nordheim plot essentially remains a straight line, but S_{FN} does not have the physical meaning suggested by (4.21). Therefore, only the effective Φ'_B value needed in (4.21) can be extracted from such plots [29].

4.4.3 Band-to-band and Trap-to-band Tunneling

If a very high electric field is forced in silicon ($F_{\text{Si}} \approx 1\text{MV/cm}$) generation of electron-hole pairs by electron tunneling from the valence to the conduction

band occurs. This phenomenon, known as band-to-band (BBT) or Zener tunneling [30], has been extensively studied in the past for its relevance on the operation of tunnel diodes [31] and it has gained renewed interest recently for its impact on the operation of scaled MOS devices and Flash cells [32].

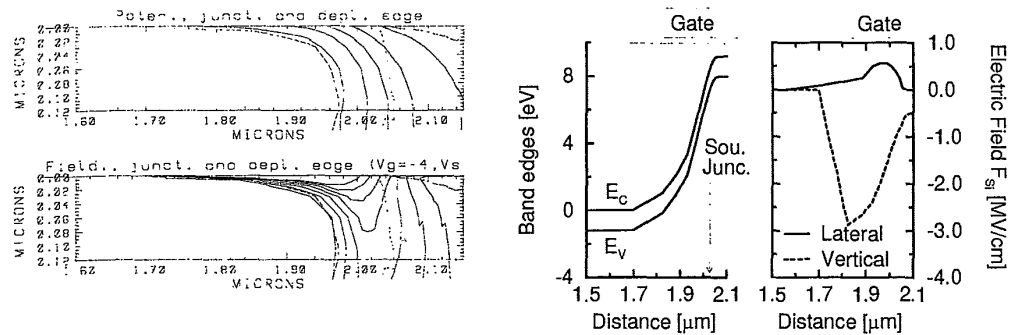


Figure 4.11 Left: simulated potential (upper plot), electric field (lower plot) and depletion region contours (dashed line) at the source side of a Flash memory cell for a typical erase bias ($V_G = -4V$, $V_S = 8V$). The Si-SiO₂ interface is located at $y = 0$, the gate extends for $x \geq 1.82\mu\text{m}$. Notice the deeply depleted layer in the gate-source overlap region on the left of the junction (represented by the dotted line at $x \simeq 2.05\mu\text{m}$). Right: Band diagram and electric field components along the surface of the same Flash cell of the graph on the left.

In MOS structures, BBT typically occurs at high source or drain voltage and low (floating) gate voltage. In Flash devices, these conditions take place in cells under erase operations, or in unselected cells sharing the same bit line with a cell under programming (see Chapter 2). BBT contributes to the so called Gate Induced Drain Leakage current, GIDL [33, 34], which can be a significant fraction of the subthreshold drain leakage current and can compromise proper functioning of the substrate bias generators.

Fig. 4.11 illustrates the basic mechanism of BBT during erase operation. A negative V_G brings the channel into accumulation and a strong vertical field depletes the gate-source (gate-drain) overlap region. The formation of a hole inversion layer, however, is prevented by the lateral field that sweeps minority carriers towards the accumulated channel, thus keeping their concentration below the equilibrium value. The deep depletion region thus formed hosts a large vertical field. The tunneling distance between conduction and valence bands is reduced and the transmission coefficient becomes large enough to originate interband transitions. The point of maximum hole generation rate is at the interface, while that of maximum electron generation typically lies inside the

source (or drain), several tens of angstroms below the interface. Electrons flow towards the source (drain), while holes are pushed towards the accumulated channel and then towards the substrate, giving rise to a parasitic leakage current.

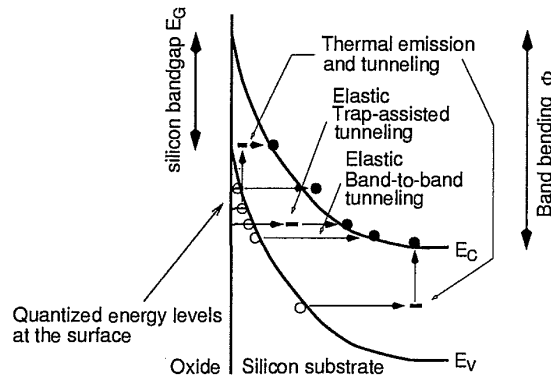


Figure 4.12 Band diagram of a MOS structure along the vertical direction through the depletion region illustrating band-to-band and trap-to-band tunneling mechanisms.

Notice that this current can not be explained by junction breakdown, because it is observed also for small or zero reverse junction bias [34].

Fig. 4.12 illustrates a few mechanisms responsible of interband transitions. Elastic band-to-band and trap-to-band (TBT) tunneling occur only if the band bending (Φ) is larger than the energy gap (E_G), whereas trap assisted inelastic processes can be effective also for $\Phi < E_G$ [33, 34]. Since the indirect gap in silicon ($\simeq 1.12\text{eV}$ at room temperature) is much smaller than the direct one ($\simeq 3.5\text{eV}$) BBT is expected to be an indirect process assisted by phonons, impurities or traps. Consistently with this picture, measured threshold band bendings for BBT-TBT $\Phi \approx 1\text{--}1.5\text{V}$ have been reported [34].

Localized traps within the gap often provide empty states and shorter tunneling paths for a variety of mixed tunneling and thermal excitation mechanisms, also shown in Fig. 4.12. These indirect processes enhance, and sometimes give the dominant contribution to the subthreshold drain leakage current. Because of their sensitivity to trap concentration, they have been exploited to monitor the interface trap density in the overlap region [35, 36].

Since most of V_{GS} (V_{GD}) is dropped across the oxide, a band bending of 1–1.5V in the deep depletion layer arises only if the doping in the overlap region is smaller than $\approx 10^{19}\text{cm}^{-3}$. On the other hand, only for dopant concentrations larger than $\approx 10^{18}\text{cm}^{-3}$ the tunneling distance is sufficiently small to allow for

significant generation. For these reasons, BBT-TBT is maximum for doping concentrations in the $\approx 10^{18}$ – 10^{19} cm $^{-3}$ range [34, 37].

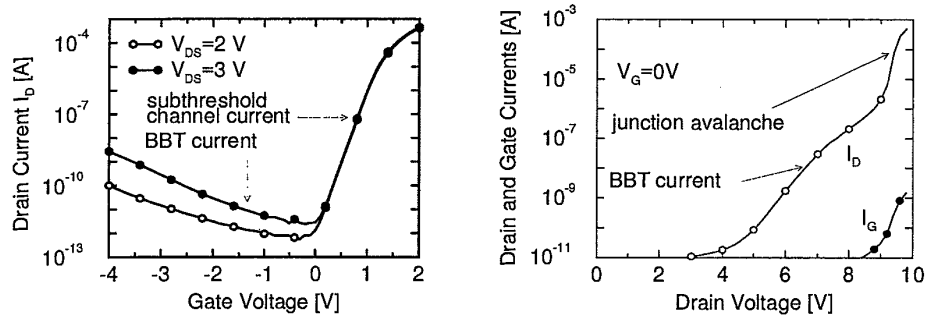


Figure 4.13 MOSFET drain and gate currents in the BBT-TBT regime. $L = 0.5\mu\text{m}$, $t_{\text{ox}} = 10\text{nm}$.

BBT-TBT is weakly dependent on channel length, channel doping and temperature [33] (as expected for tunneling phenomena, Section 4.3.1), whereas it is very sensitive to the gradient of the doping profile in the overlap region [33, 38]. In particular, the steep drain junction of Flash cell exhibits a much larger tunneling current than the graded source junction.

The drain current of a MOSFET in the BBT-TBT regime, qualitatively representative of that of Flash cells too, is shown in Fig. 4.13. For fixed V_{DS} (left plot) the current decreases exponentially in the subthreshold region. The BBT-TBT contribution stands out as V_{G} is made negative and the high field deep depletion region is formed. In this regime the current is a strong function of the gate-drain (gate-source) potential drop and of the oxide thickness but it depends weakly on the drain-bulk (source-bulk) voltage (for fixed V_{GD} or V_{GS}). This is because the field is roughly vertical in the deep depletion region, and its component normal to the interface is directly determined by the corresponding one in the oxide. In formulae: $F_{\text{SI}} \simeq F_{\text{OX}} \epsilon_{\text{ox}} / \epsilon_{\text{si}} \simeq V_{\text{GD}} \epsilon_{\text{ox}} / \epsilon_{\text{si}} t_{\text{ox}}$. An exception to this rule of thumb is given by very abrupt junctions featuring a high lateral field, in which the drain bulk voltage plays a non negligible role.

For increasing V_{DS} the large lateral field induced by the source (drain) to substrate potential drop allows generated holes to gain considerable energy and to achieve a high probability to impact ionize, thus leading the junction to a surface avalanche regime [39]. This is evidenced in Fig. 4.13 by the abrupt increase of the drain current at high V_{DS} . BBT- and impact ionization-generated holes can become so energetic to overcome the Si-SiO $_2$ barrier, thus originating

a detectable gate current (line with filled symbols in the right plot of Fig. 4.13) [33, 40]. Although electrons tunneling from the poly-silicon gate can also be responsible of this current [41], hole injection is particularly harmful for device reliability.

Indeed, as will be discussed in more detail in Section 4.8, hole trapping and interface state generation accompany hot hole injection in the oxide [42]. Newly generated interface states can further enhance the BBT-TBT current through the mechanisms illustrated in Fig. 4.12, while trapped holes locally modify the oxide field, affect the tunnel current and eventually cause program or erase problems [43, 44]. Holes reaching the floating gate recombine with available electrons, causing write disturbs such as charge loss and a decrease of the programmed threshold voltage [45] (see also Chapter 7). Since hole heating is mostly due to the lateral field, careful engineering of the source and drain junction profiles is mandatory to control the associated detrimental effects. Drain erase of Flash cells with limited avalanche and hole-induced degradation can be achieved biasing the device at reduced drain voltage and negative gate voltage [46].

4.4.4 Modeling the Band-to-band and Trap-to-band Tunneling Current

From the modeling point of view, BBT and TBT pose serious challenges and problems not completely solved so far [47, 48]. A simplified equation for the BBT current based on a one-dimensional analysis will be developed below following the procedure described in Section 4.3.1.

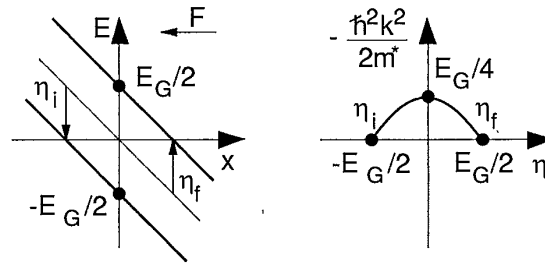


Figure 4.14 Energy band diagram and equivalent barrier height for BBT tunneling.

To this goal, let us consider only elastic transitions in a direct gap semiconductor (that is, no phonons involved in the tunneling process). The net BBT current can be expressed as the difference between the current of electrons tunneling from the valence towards the conduction band (generation) and that in the opposite direction (recombination) (see Eq. (4.19), with regions 1 and 2 corresponding to the valence and conduction band, respectively). Accord-

ing to (4.18) the former is proportional to $Tg_v g_c f_v (1 - f_c)$ and the latter to $Tg_c g_v f_c (1 - f_v)$, where T is the transmission coefficient, g_c, g_v are the conduction and valence band density of states, f_c, f_v are the corresponding occupation probabilities, respectively. Therefore, subtracting the two contributions:

$$dJ_{\text{BBT}} \propto Tg_v g_c (f_v - f_c) . \quad (4.22)$$

To compute the transmission coefficient in (4.22) by means of (4.14) $k_{\perp}(x)$ must be known along the tunneling path. In order to fulfill the requirement $k = 0$ at the band edges, a parabolic dispersion relationship across the gap is often assumed:

$$\hbar^2 k^2 / 2m^* = \hbar^2 k_{\perp}^2 / 2m^* + \hbar^2 k_{\parallel}^2 / 2m^* = - \left((E_G/2)^2 - \eta^2 \right) / E_G , \quad (4.23)$$

where $m^* = 2(m_c^{-1} + m_v^{-1})^{-1}$ is a suitable effective mass [49] ($m_c = 0.32m_0$ and $m_v = 0.16m_0$ or $0.49m_0$ for light and heavy holes, respectively), and η is the electron energy measured from the center of the energy gap (Fig. 4.14) and ranges between $-E_G/2$ and $E_G/2$ along the tunneling path. Notice that (4.23) has the correct behavior at the band edges but it does not match the different effective masses of the silicon valence and conduction bands. Assuming a constant field of absolute value F and taking the origin at $\eta = 0$ we have $\eta = qFx$. Therefore, (4.23) transforms in a parabolic barrier as that of Fig. 4.6 and Eq. (4.17) with $x_1 = \sqrt{(E_G/2)^2 + E_G E_{\parallel}} / qF$ and $\xi = \Phi_B - (E - E_{\parallel}) = E_G/4$. Hence, the transmission coefficient is [50]:

$$T_{\text{BBT}} \simeq \exp \left(- \frac{\pi (2m^*)^{1/2} E_G^{1/2}}{4\hbar q} \cdot \frac{E_G + 4E_{\parallel}}{F} \right) . \quad (4.24)$$

Similarly to the tunnel probability of triangular and trapezoidal barriers (Eqs. (4.15) and (4.16)) T_{BBT} has an exponential dependence on $1/F$ for elastic tunneling.

Phonon assisted BBT transitions, although probably much more likely than direct ones, can not be dealt with the WKB approximation. However, the transmission coefficient of phonon assisted processes has a similar functional dependence on $\exp(1/F)$ as (4.24) [47, 51]. Therefore, combining (4.24) with (4.22) and assuming $f_v = 1, f_c = 0$, the BBT current can be expressed as:

$$J_{\text{BBT}} \approx A_{\text{BB}} F^{\nu} \exp(-B_{\text{BB}}/F) , \quad (4.25)$$

where A_{BB} and B_{BB} are constants and ν ranges between 1 and $5/2$ depending on the considered transitions (direct or phonon assisted) and on the assumptions made in integrating (4.22).

It must be pointed out that some of the approximations behind (4.25) have been proven inadequate to describe the phenomenon quantitatively [48], so that this expression should be regarded as a semi-empirical formula with parameter values which lack of a sound physical basis. The reported spread in A_{BB} and B_{BB} values indirectly confirms this statement. For example, B_{BB} ($\approx 23\text{MV/cm}$ according to (4.24)) ranges between 18MV/cm [33] and 27MV/cm [39] in experiments. Much larger B_{BB} values ($\approx 36\text{MV/cm}$) have been attributed to indirect BBT in the absence of traps [52].

The approximations made to derive J_{BBT} clearly suggest that numerical analysis is the only viable solution for accurate evaluation of BBT in devices. For example, numerical analysis can account for the bi-dimensionality of the potential distribution in the generation region. However, a fully two-dimensional calculation of tunneling rates is quite difficult. Since the barrier width required for significant BBT is often less than $10\text{--}15\text{nm}$, whereas the curvature radius of the potential contour lines in the deeply depleted region typically exceeds 100nm owing to doping profile, carrier concentration and bias, simulators most often employ a one dimensional generation rate computed in rectangular or radial coordinate systems, along the electric field flowlines [39, 53].

An additional problem stems from the assumption of constant electric field made in deriving BBT rates. Since the depth of the depletion region is not much larger than the tunneling distance, the electric field changes considerably along the tunneling path. Detailed analysis demonstrated that the hypothesis of vertical tunneling paths with constant field leads to order of magnitude errors on the estimated BBT generation rate. Much more accurate values are obtained approximating the variable field with a suitable effective field, such as the average field along the path [53].

Finally, the high electric field at the silicon interface quantizes the available energy levels (Fig. 4.12). Therefore, the threshold band bending increases in a bias dependent way, which raises the problem of computing the electric field self-consistently with the quantized levels. This latter problem has not been tackled so far in ways appropriate for device analysis.

4.5 FUNDAMENTALS OF CARRIER TRANSPORT

4.5.1 The Distribution Function

For the purpose of state of the art device analysis, transport can be tackled by means of a semiclassical approach in which electrons (and holes) are described as an ensemble of particles weakly interacting with each other, and characterized in a statistical way by a distribution function $f(\vec{r}, \vec{k}, t)$. For the sake of simplicity, in the following we will make reference to electrons, but most

concepts and definitions are applicable to holes as well, provided minor and straightforward changes are made to the basic equations.

$f(\vec{r}, \vec{k}, t)$ represents the fraction of available electron states of wave vector $\vec{k}$ that is actually occupied in $\vec{r}$ at time t , i.e. the occupation probability of electron states in the phase space $(\vec{r}, \vec{k})$. Therefore, $f(\vec{r}, \vec{k}, t)$ is a number between 0 and 1. The number of electrons per unit volume centered in $\vec{r}$, at time t , with wave vector between $\vec{k}$ and $\vec{k} + d\vec{k}$, is given by:

$$\mathcal{F}(\vec{r}, \vec{k}, t)d\vec{k} = \frac{1}{4\pi^3} f(\vec{r}, \vec{k}, t)d\vec{k} , \quad (4.26)$$

where $g = 1/4\pi^3$ is the density of states, that is the number of available states per unit volume in $\vec{k}$ space.

In the following we will often make use of the energy distribution function $f(\vec{r}, E, t)$, that is the occupation probability of electron states at energy E . $f(\vec{r}, E, t)$ provides a more intuitive representation of the electron population than $f(\vec{r}, \vec{k}, t)$, and it is obtained averaging the latter over equi-energy surfaces of $\vec{k}$ space. Accordingly, the number of electrons per unit volume with energy between E and $E + dE$ is given by:

$$\mathcal{F}(\vec{r}, E, t)dE = f(\vec{r}, E, t)g(E)dE , \quad (4.27)$$

where $g(E)$ denotes the energy density of states (Eq. (4.5)).

$f(\cdot)$ and $\mathcal{F}(\cdot)$ are often indicated as distributions. However, as explained above, they have different physical meanings and different units. In the following we will refer to f as the distribution function and to $\mathcal{F}$ as the carrier distribution. In addition, for the sake of simplicity, we will consider only steady state conditions, so that distributions will not depend on time explicitly.

According to the definitions above, the electron concentration and the electron current density are obtained summing the contributions of all occupied electron states:

$$n(\vec{r}) = \frac{1}{4\pi^3} \int f(\vec{r}, \vec{k})d\vec{k} , \quad (4.28)$$

$$\vec{J}(\vec{r}) = -\frac{q}{4\pi^3} \int \vec{v}_g(\vec{k})f(\vec{r}, \vec{k})d\vec{k} . \quad (4.29)$$

Neglecting anisotropy effects in $\vec{k}$ space (i.e. assuming that $f(\vec{r}, \vec{k}, t)$ depends only on $|\vec{k}|$ and not on $\vec{k}$'s direction), n and $\vec{J}$ can be expressed using functions of the carrier energy:

$$n(\vec{r}) = \int f(\vec{r}, E)g(E)dE , \quad (4.30)$$

$$\vec{J}(\vec{r}) = -q \int \vec{v}_g(E) f(\vec{r}, E) g(E) dE, \quad (4.31)$$

where $\vec{v}_g$ can be computed averaging (4.8) over equi-energy surfaces. (Rigorously speaking, $\vec{v}_g(E)$ should be calculated weighting the average with the distribution function. In most practical cases however, the suggested average provides a result accurate enough).

Once f is known, not only integral quantities such as n and $\vec{J}$ can be calculated, but also complex physical phenomena related to small subsets of the carrier population can be analyzed, such as gate and substrate currents in MOSFETs (Sections 4.6.5 and 4.6.7).

4.5.2 The Boltzmann Transport Equation

In order to determine the distribution function, a relation between f and the built-in, applied and scattering potentials is necessary. This relation is given by the Boltzmann transport equation (BTE), whose validity limits are discussed, for example, in [54, 55]. In steady state conditions the BTE reads:

$$\frac{d\vec{r}}{dt} \cdot \nabla_{\vec{r}} f + \frac{d\vec{k}}{dt} \cdot \nabla_{\vec{k}} f = \left(\frac{\partial f}{\partial t} \right)_{\text{coll}}, \quad (4.32)$$

where $d\vec{r}/dt = \vec{v}_g$ is the electron group velocity, and $d\vec{k}/dt$ is related to the electric field $\vec{F}$ by the classical Newton equation of motion (4.7).

Fig. 4.15 provides a pictorial representation of the BTE in a two-dimensional real and wave vector space, and exemplifies how this apparently complex mathematical expression is essentially a continuity equation in the multi dimensional phase space $(\vec{r}, \vec{k})$. Particles move in real space at the group velocity $\vec{v}_g(\vec{k})$ and in wave vector (that is momentum) space, because they are accelerated (decelerated) by the electric field. This is expressed by the two terms on the left hand side of (4.32), that represent the net particle flux out of the elemental volume $d\vec{r}d\vec{k}$ in phase space.

In addition, inter-particle and lattice-particle collisions suddenly change the momentum, whereas generation recombination processes increase and decrease the carrier population in $d\vec{r}d\vec{k}$. Collisions are taken into account by the right hand side of (4.32), that represents the perturbation of the distribution function caused by scattering potentials and generation-recombination processes. The collision term is the difference between the carrier flux entering state $(\vec{r}, \vec{k})$ from any other state $(\vec{r}, \vec{k}')$ $((\partial f / \partial t)_{\text{in}})$, and the carrier flux outgoing from state $(\vec{r}, \vec{k})$ towards any other state $((\partial f / \partial t)_{\text{out}})$. This balance can be expressed as:

$$\left(\frac{\partial f}{\partial t} \right)_{\text{coll}} = \left(\frac{\partial f}{\partial t} \right)_{\text{in}} - \left(\frac{\partial f}{\partial t} \right)_{\text{out}} =$$

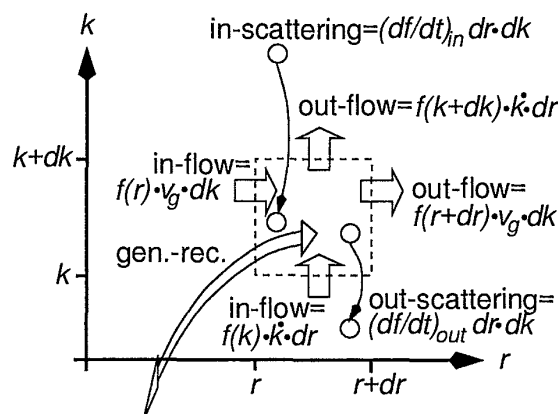


Figure 4.15 Schematic representation of particle flow in one-dimensional real and one-dimensional wave vector space according to the Boltzmann transport equation [55]. $\dot{r} = dr/dt = \vec{v}_g$ and $\dot{k} = dk/dt = -qF/\hbar$. Fluxes in real space are proportional to the surface in momentum space (dk) and vice versa. The number of particles scattered in and out of the elemental volume are proportional to the volume itself ($drdk$).

$$\begin{aligned}
 &= \sum_{\vec{k}'} S(\vec{r}, \vec{k}', \vec{k}) f(\vec{r}, \vec{k}') (1 - f(\vec{r}, \vec{k})) - \\
 &- \sum_{\vec{k}} S(\vec{r}, \vec{k}, \vec{k}') f(\vec{r}, \vec{k}) (1 - f(\vec{r}, \vec{k}')) , \quad (4.33)
 \end{aligned}$$

where $S(\vec{r}, \vec{k}, \vec{k}')$ is the probability of a scattering from state $(\vec{r}, \vec{k})$ to state $(\vec{r}, \vec{k}')$ per unit time and has units of s^{-1} , while the products $f(\vec{r}, \vec{k})(1 - f(\vec{r}, \vec{k}'))$ and $f(\vec{r}, \vec{k}')(1 - f(\vec{r}, \vec{k}))$ indicate the probability to have a carrier in the initial state $\vec{k}$ (or $\vec{k}'$) and an available empty final state $\vec{k}'$ (or $\vec{k}$), respectively, and the two sums are performed over all possible final states.

In the so called relaxation time approximation, the collision term is assumed proportional to the deviation of the distribution function (f) from its equilibrium value (f_0) through a phenomenological relaxation time τ_f :

$$\left(\frac{\partial f}{\partial t} \right)_{\text{coll}} = \frac{f - f_0}{\tau_f} . \quad (4.34)$$

In summary, (4.32) is easily derived equating to zero the algebraic sum of the flux components depicted in Fig. 4.15.

4.5.3 Scattering

Scatterings have a very relevant effect on the distribution function as they modify the orientation of electron's momentum and dissipate (or, with a lower probability, increase) electron's kinetic energy. The most important scattering mechanisms in silicon devices are acoustic and optical phonon absorption and emission (representing interactions between the carriers and the lattice vibrational modes), ionized impurity scattering (due to the presence of ionized dopant atoms), impact ionization (that is the generation of electron hole pairs), Coulomb interactions between particles. These can be classified in long range (carrier-plasmon) and short range (carrier-carrier) interactions.

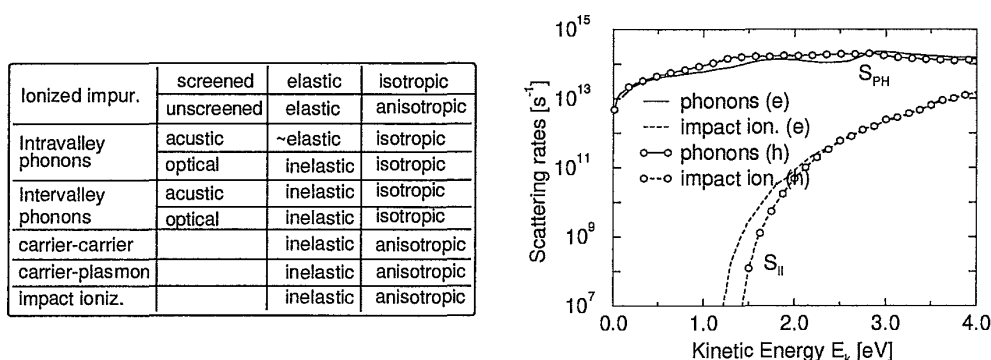


Figure 4.16 Left: Basic properties of the main electron scattering mechanisms in bulk silicon assuming a spherical parabolic band structure. Right: Phonon (S_{PH}) and impact ionization (S_{II}) scattering rates in silicon as a function of carrier energy before scattering. The non-monotonicity of the phonon rate reflects deviations of the band structure from the simple parabolic approximation. Data from [56, 57].

Fig. 4.16 lists the main characteristics of each scattering type specifying its effect on electron's energy (elastic or inelastic collision) and on the redirection induced on carrier's momentum (isotropic redirection or anisotropic with a non-uniform distribution of scattering angle). It reports also the total phonon and impact ionization scattering rates as a function of carrier energy, that is, the average number of collisions per second suffered by electrons with given energy. These rates are obtained integrating $S(\vec{k}, \vec{k}')$ over the $\vec{k}$ states with energy E and, for each initial state, over all final states fulfilling energy and momentum conservation. For scattering type i , S_i is a property of the semiconductor material strictly related to the band structure, and possibly dependent on temperature, impurity or carrier concentration.

Phonon and ionized impurity scatterings largely dominate over impact ionization, but they change the carrier energy by a very small amount in each collision (typically a few tens of meV). On the other hand, impact ionization and Coulomb interactions can exchange large amounts of energy. Therefore, they also play a relevant role in determining the particle distribution and can act as efficient dissipative mechanisms for high energy electrons.

Once the S_i 's are known, it is useful to introduce a few parameters which summarize the effect of scattering on the carrier status.

The first one is the energy relaxation time, which is a measure of the rate at which scattering dissipates the carrier energy. The energy relaxation time is obtained weighting each collision by the fractional change in energy. In formulae:

$$\tau_E(E) = \left[\sum_i \sum_{\vec{k}(E)} \sum_{\vec{k}'} S_i(\vec{k}, \vec{k}') \left(1 - \frac{E(\vec{k}')}{E(\vec{k})} \right) \right]^{-1}. \quad (4.35)$$

The second parameter is the momentum relaxation time, which is a measure of the average time that scattering takes in randomizing the carrier momentum with respect to a reference direction x , and it is given by:

$$\tau_p(E) = \left[\sum_i \sum_{\vec{k}(E)} \sum_{\vec{k}'} S_i(\vec{k}, \vec{k}') \left(1 - \frac{p_x(\vec{k}')}{p_x(\vec{k})} \right) \right]^{-1}. \quad (4.36)$$

Energy and momentum relaxation times are complex functions of the carrier wave vector (hence of carrier energy). In silicon they have average values $\tau_E \simeq 0.3\text{--}0.4\text{ps}$, $\tau_p \simeq 0.01\text{--}0.1\text{ps}$, for electrons [58], and $\tau_E \simeq 0.2\text{--}0.4\text{ps}$ [59, 60], $\tau_p \simeq 0.03\text{--}0.1\text{ps}$ for holes [59], respectively.

4.5.4 The Carrier Distribution in Thermal Equilibrium

As exemplified by Fig. 4.15, the state of the ensemble of electrons (that is, f) is the result of the balancing actions of the applied electric field $\vec{F}$ and of the scattering events that move carriers in and out of each elemental volume of the phase space $d\vec{r}d\vec{k}$.

In thermal equilibrium f does not depend explicitly on the carrier momentum, but only on the total carrier energy $E = U(\vec{r}) + E_K(\vec{k})$. For $E \gg E_F$, as in non-degenerate silicon, f is well approximated by the Maxwell-Boltzmann distribution:

$$f(E) \simeq \exp \left(- (E - E_F) / k_B T_L \right), \quad (4.37)$$

where E_F denotes the Fermi energy. The distribution is spherical in momentum space and the average velocity of the ensemble, the so called drift velocity, defined as

$$\vec{v}_d(\vec{r}) = -\frac{\vec{J}}{qn} = \frac{\int \vec{v}_g(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k}}{\int f(\vec{r}, \vec{k}) d\vec{k}}, \quad (4.38)$$

is zero. Hence, the current density $\vec{J} = -qn\vec{v}_d$ (Eqs. (4.28) and (4.38)) is also zero, as expected at equilibrium. The Maxwell-Boltzmann distribution predicts a decrease of the carrier concentration by approximately 17 orders of magnitude per eV of energy at 300K. Given this rapid decrease of the number of carriers for increasing energy, the average kinetic energy of the particle ensemble

$$w(\vec{r}) = \frac{\int E_K(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k}}{\int f(\vec{r}, \vec{k}) d\vec{k}}, \quad (4.39)$$

is very small ($w = 3k_B T_L / 2 \approx 40 \text{ meV}$ at room temperature).

In the general non-equilibrium case f should be computed solving the BTE with a suitable set of boundary conditions. Unfortunately, in most practical cases the BTE (4.32) can be solved only by means of numerical techniques, such as those based on the spherical harmonics expansion of the distribution function [61], the cellular automata [62], the scattering matrix approach [63], or the powerful Monte Carlo method [64]. In the following, distribution functions will be analyzed in a few cases relevant to understand hot carrier effects in MOSFETs and Flash cells. Unless otherwise specified, the reported simulation results have been computed by means of the Monte Carlo method.

4.5.5 Carrier Distributions in Homogeneous Electric Fields

In a constant (homogeneous) field electrons gain momentum in the direction of the field according to the classical law of motion (4.7) $\hbar d\vec{k}/dt = d\vec{p}/dt = -q\vec{v}_E$ and their kinetic energy increases as described by the band structure of the material ($E(\vec{k})$). However, inelastic scattering events such as optical phonon emission and impact ionization, change (most often reduce) carrier energy, and contribute with all other scatterings in randomizing $\vec{k}$. Since the scattering rate is essentially an increasing function of energy and new scattering mechanisms set on at high energy (e.g. impact ionization, see Fig. 4.16) a new steady state is attained in which the average energy loss due to more frequent scatterings is balanced by the energy gained from the electric field. Consequently, f deviates from its equilibrium Maxwellian shape.

Fig. 4.17 reports calculated electron energy distributions for a few electric fields and temperatures. Since the Monte Carlo method used to compute these distributions is a statistical technique to solve the BTE, the distributions

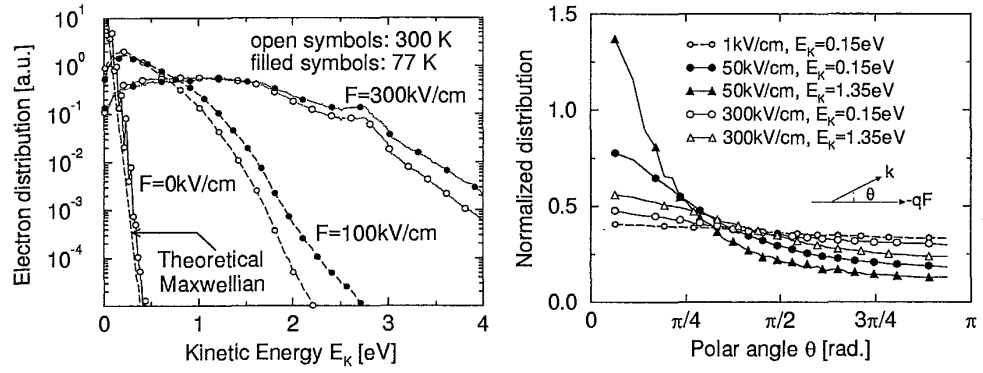


Figure 4.17 Left: Electron distributions normalized to unit concentration in homogeneous silicon slabs doped $1 \times 10^{17} \text{cm}^{-3}$ at various electric fields. A theoretical Maxwell-Boltzmann distribution at 300K is also shown for reference. Right: Distributions of carriers at given energy normalized to unit area as a function of the polar angle (θ) between the direction of momentum and that of the driving force $-q\vec{F}$.

are inherently affected by statistical noise, particularly evident in the zero field case. As can be seen, the number of high energy carriers is orders of magnitude larger than in equilibrium. Furthermore, since optical phonon emission (the most frequent inelastic scattering mechanism in silicon) decreases at low temperature, enhanced distribution tails are observed at 77K. In momentum space f is not spherical any more, but becomes elongated in the direction opposite to the field (because of the negative electron charge). Therefore both the drift velocity $\vec{v}_d$ and the current density $\vec{J}$ are non zero.

The elongation of the distribution increases for increasing F up to $\approx 5 \times 10^4 \text{V/cm}$, and so does the drift velocity v_d . This is clearly seen in the right plot of Fig. 4.17, showing the angular distribution of electrons for different fields. For $F \gg 5 \times 10^4 \text{V/cm}$ scattering becomes so effective in randomizing momentum that it limits the elongation of f and any further growth of v_d . This is evidenced in Fig. 4.17 by the relatively flat polar distributions at $F = 300 \text{kV/cm}$ compared to those at 50kV/cm . Thus, v_d saturates to an almost constant value $v_s \approx 10^7 \text{cm/s}$, called saturation velocity, that eventually limits the maximum electron current density through the homogeneous slab $\vec{J} = -qn\vec{v}_d$. The random velocity component $|\vec{v} - \vec{v}_d|$ has a spherical distribution of directions and it can reach values as large as $\approx 10^8 \text{cm/s}$, that is much larger than the saturated velocity $v_s \approx 10^7 \text{cm/s}$. Notice also that the high energy population of the distribution (curves with triangles in the right plot of

Fig. 4.17) is always more collimated in the driving force direction than the low energy one (curves with circles).

4.5.6 The Effective Temperature Model

In homogeneous conditions, a widely adopted approximation describes the distribution function f as a displaced Maxwellian distribution [55]:

$$f(\vec{r}, \vec{k}) \approx \exp \left((F_n(\vec{r}) - U(\vec{r})) / k_B T_e \right) \exp \left(- |\hbar \vec{k} - m^* \vec{v}_d|^2 / 2m^* k_B T_e \right), \quad (4.40)$$

where F_n denotes the quasi Fermi level. In simple terms, we can think of (4.40) as a generalized form of the equilibrium Maxwell-Boltzmann distribution (4.37): F_n replaces E_F , E_K is expressed according to a simple spherical parabolic band, and a carrier temperature $T_e \geq T_L$ and an average carrier velocity $\vec{v}_d \neq 0$ are introduced to account for the increased number of energetic electrons and for a non-zero current density. The displaced Maxwellian distribution retains the spherical shape of the equilibrium distribution around an average momentum $\vec{p}_d = m^* \vec{v}_d$, and represents a reasonable approximation only if frequent collisions almost completely randomize the momentum gained in the field direction.

Assuming the validity of (4.40), the average kinetic energy w (Eq. (4.39)) turns out to be the sum of a drift (or convective) and a thermal energy:

$$w = \frac{1}{2} m^* v_d^2 + \frac{3}{2} k_B T_e. \quad (4.41)$$

Since randomizing scatterings limit the maximum v_d , the second term in (4.41) dominates at high fields, where w is well approximated by the thermal energy alone. For example, at $F = 10^5 \text{ V/cm}$, $w \approx 0.5 \text{ eV}$ whereas $m^* v_d^2 / 2 \approx 30 \text{ meV}$ [65].

In these conditions, carriers behave as a high temperature gas of particles ($T_e \gg T_L$) interacting through scattering with the much cooler hosting lattice. The term hot carriers expresses this concept, and refers to the presence in the device of a significant number of particles with kinetic energy largely exceeding the average equilibrium value $3k_B T_L / 2$.

At high homogeneous fields $|\vec{p}| \gg |m^* \vec{v}_d|$. Hence (4.40) can be further approximated as:

$$f(\vec{r}, E) \approx \exp \left((F_n(\vec{r}) - U(\vec{r})) / k_B T_e \right) \exp \left(- E_K / k_B T_e \right), \quad (4.42)$$

where $E_K = |\vec{p}|^2 / 2m^* = \hbar^2 |\vec{k}|^2 / 2m^*$ is the carrier kinetic energy in the assumed spherical parabolic band structure. Eq. (4.42) is often referred to as

the effective temperature model and, thanks to its simple analytic form allows for a first order evaluation of major trends regarding hot carrier effects in devices. According to (4.42) $\ln(f)$ should decrease linearly with E_K with slope proportional to $1/T_e$. As can be inferred from Fig. 4.17, the shape of actual distributions does not allow the carrier temperature to be unambiguously defined, because the absolute value of the logarithmic slope increases for increasing energy. However, the electron temperature concept is still useful for heuristic investigation of hot carrier problems.

4.6 HOT CARRIER EFFECTS IN MOSFETS

Carrier heating is of great importance for the operation of Flash devices because on one hand, it is exploited for programming and, on the other hand, it is responsible for reliability concerns.

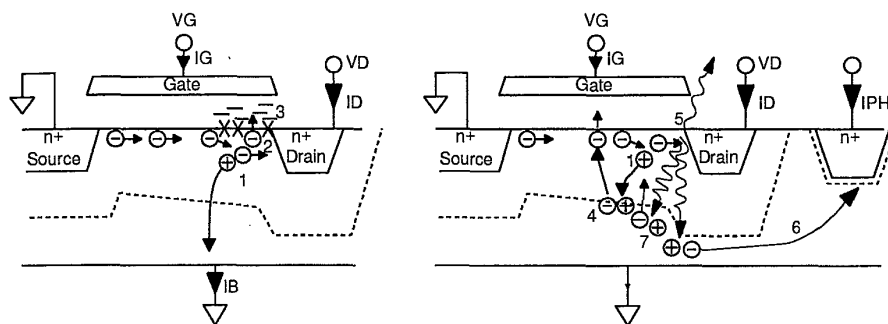


Figure 4.18 Qualitative sketch of the main hot carrier effects in an n-MOSFET.

Fig. 4.18 schematically illustrates most hot carrier related phenomena typically encountered in MOSFETs. Carriers heated by the channel electric field create electron-hole pairs by impact ionization (process 1). The secondary generated holes that flow towards the substrate contact cause a voltage drop that tends to forward bias the source-substrate junction, leading to parasitic bipolar effects. In addition, hot electrons are injected into the gate oxide (process 2), damage the interface (surface states generation, indicated by $\times$ symbols) and the insulator properties (charge trapping, denoted by $-$ symbols) (process 3). Holes heated in the vertical field and photons emitted by channel hot electrons relaxing to low energy states (process 5) generate electron-hole pairs in the depletion region and in the substrate (process 4). The generated electrons are responsible of long range interactions between junctions within the same chip (process 6) or can be reinjected towards the gate (process 7).

In the following, a few of these processes that are most relevant to the operation of Flash cells will be described in more detail. To this end, carrier heating and hot carrier distributions in MOSFETs need to be discussed first.

4.6.1 Carrier Heating in MOSFETs and Flash Cells

Fig. 4.19 reports the lateral electric field along the interface of two saturated MOSFETs. Clearly, the field is highly inhomogeneous. Carriers emitted from

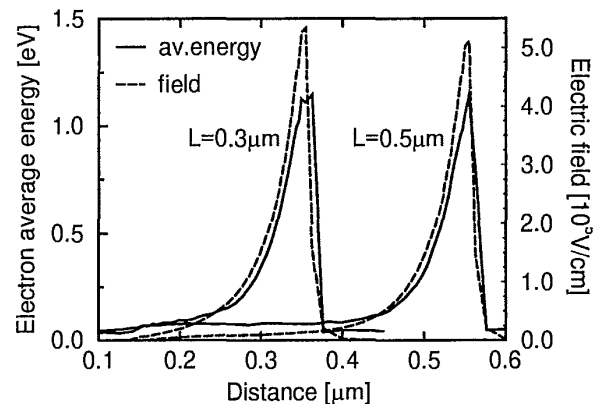


Figure 4.19 Electric field and average energy along the channel of sub-micrometer n-MOSFETs biased at $V_{GS} = V_{DS}/2 = 1.5V$, and featuring the same doping profiles and oxide thickness ($t_{ox} = 10nm$).

the source junction barrier initially experience a region of small field followed by an abrupt spike near the drain junction. The average carrier energy (w) increases rapidly in this spike. Since carriers need a finite time and distance to gain energy, the heating process is spatially delayed and w attains its maximum value beyond the point of maximum field. The displacement is slightly larger in short devices than in long ones. Then, upon entering the low-field drain region, hot carriers merge with the residing population of thermal electrons. Their excess kinetic energy is dissipated by phonon, impact ionization and carrier-carrier interactions. Correspondingly, the average energy decays to its equilibrium value.

Fig. 4.19 clearly points out that, due to the delay in the heating process, and differently from the homogeneous field case treated previously, no unique relationship exists between electric field and average energy. This means that in non homogeneous conditions, at every point in the channel the carrier energy distribution $f(\vec{r}, \vec{k})$ is not simply a function of local quantities (e.g. field or

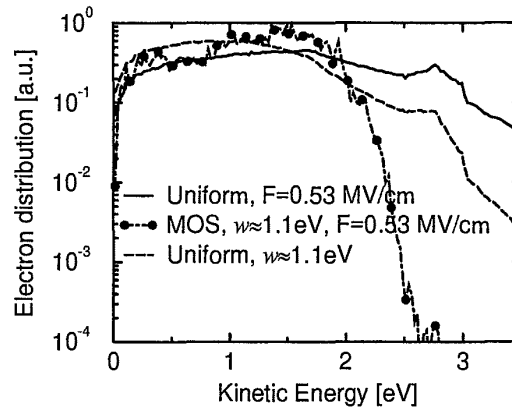


Figure 4.20 Electron distribution in a uniform field $F = 0.53 \text{ MV/cm}$ (solid line); in a uniform field featuring $w \approx 1.1 \text{ eV}$ (dashed line); at a point along the channel of a $0.3 \mu\text{m}$ MOSFET biased at $V_{\text{GS}} = V_{\text{DS}}/2 = 1.5 \text{ V}$, and featuring $w \approx 1.1 \text{ eV}$ and $F = 0.53 \text{ MV/cm}$ (dash-dotted line). Distributions are normalized to the carrier concentration, that is to unit area.

concentration gradient) but it becomes a complex *non-local* property (that is, a functional) of the whole field profile [66]. This point is further demonstrated in Fig. 4.20, which compares the electron distribution calculated at a point in the channel of a MOSFET to two distributions evaluated in homogeneous silicon slabs and featuring either the same field or the same mean energy as the MOSFET distribution. As can be seen, carrier distributions are quite dissimilar from each other, especially in the high energy tails, that give a minor contribution to w . Evidently, local and average quantities such as the electric field or the electron mean energy are insufficient to identify unambiguously the high-energy carrier population, hence to predict hot carrier effects accurately.

The assumption that the distribution stays essentially spherical because of frequent scattering, inherent to the displaced Maxwellian approximation (4.40), is often violated in the region of increasing field, where carriers can travel ballistically over a non negligible distance compared to device dimensions and very few carriers are back scattered against the field direction. As a result, collimated distributions of particles flowing in a preferential direction can be found in spatially varying conditions. Because of the skewed shape of the distribution, the drift component of the average energy ($w_d = m^* v_d^2/2$) can be comparable to the thermal one $w_t = w - w_d$, a situation never observed in homogeneous slabs [65]. In addition, the drift velocity (v_d) can exceed the

saturation velocity ($v_s \approx 10^7$ cm/s), giving rise to the phenomenon known as velocity overshoot [65].

For all these reasons, analytical approximations of $f(\vec{r}, \vec{k})$ such as (4.40) represent an oversimplified picture in the case of real devices. The inherent non-local nature of the carrier heating process is such that even assuming *a priori* the validity of (4.40), T_e and $\vec{p}_d$ can not be expressed as well defined functions of the local electric field or concentration gradient. Only in the presence of gradually changing fields the state of the carrier ensemble (i.e. $f(\vec{r}, \vec{k})$) can be predicted with reasonable accuracy by local or average quantities.

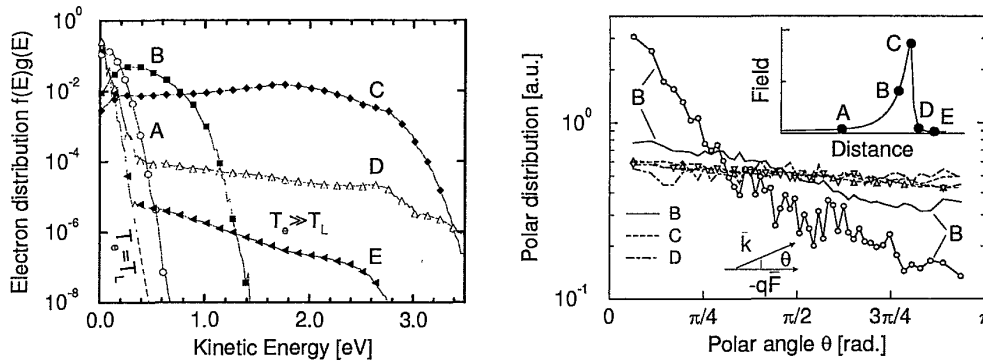


Figure 4.21 Left: Electron distributions at different positions along the channel of a MOSFET. Right: Normalized polar distributions at $E_K = 0.1$ eV (lines) and $E_K = 1.5$ eV (lines with symbols).

The evolution of the carrier distribution along the channel of a MOSFET is quite complex. If the low field region just ahead of the source is longer than several mean free paths ($\lambda \approx 10$ nm for electrons [20], and $\lambda \approx 4$ nm for holes [67, 68]), the quasi-equilibrium behavior described in Section 4.5.5 is approximately verified in the first section of the device.

Distributions along the remaining high field and drain regions are illustrated in Fig. 4.21. As can be seen, the distribution is gradually heated while electrons enter the high field region (points A, B and C). Eventually, the population of channel electrons coming from the source merges with the cooler population of electrons inside the drain, giving rise to a peculiar distribution with a “cold” branch ($T_e \simeq T_L$) and a hot part featuring $T_e \gg T_L$ (points D and E).

Angular distributions for two different energies are shown in the right plot of Fig. 4.21. The polar distribution of high energy carriers is highly collimated in the channel region where the electric field rapidly increases (point B).

Then, since the time constant for momentum relaxation (τ_p , Eq. (4.36)) is much smaller than that for energy relaxation (τ_E , Eq. (4.35)), momentum is randomized well before carriers lose significant fractions of their energy. Therefore, the distribution recovers an approximately spherical shape well before being cooled off by the dissipative scattering mechanisms prevailing inside the drain (points C, D and E). The polar distributions at the drain junction (where the electric field is maximum, point C) are already almost flat, while the energy distributions feature very enhanced tails.

4.6.2 MOSFET Design and Carrier Heating

An important issue to be addressed is the relationship between carrier heating and MOSFET technological parameters. At this regard, Fig. 4.19 shows that the maximum carrier mean energy is not very sensitive to a reduction of the gate length when other relevant device dimensions (gate oxide thickness and junction depth) are kept constant. This is a general result [69] that holds true unless very short gate lengths are considered, leading to a loss of the gate control over the channel potential, and to an extension of the high field drain region towards the source (Drain Induced Barrier Lowering, DIBL).

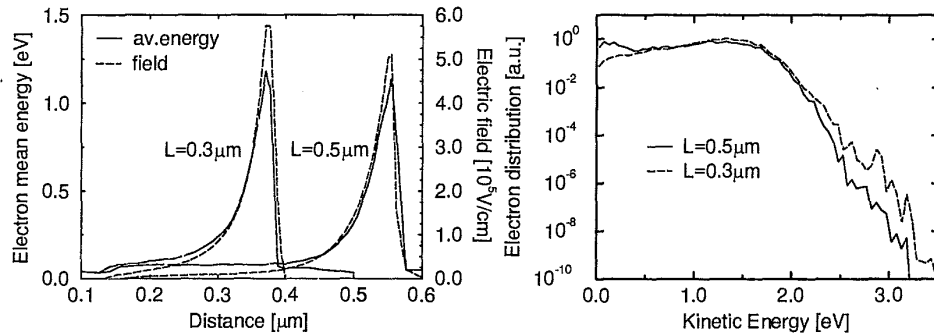


Figure 4.22 Left: Electric field and average energy along the channel of n-MOSFETs biased at $V_{GS} = V_{DS}/2 = 1.5\text{V}$. Gate length $L = 0.5\mu\text{m}$, $t_{ox} = 10\text{nm}$, junction depth $x_J = 100\text{nm}$, and $L = 0.3\mu\text{m}$, $t_{ox} = 7\text{nm}$, $x_J = 70\text{nm}$, respectively. Right: Corresponding electron energy distributions at the points of maximum average energy.

A substantial increase of the peak electric field is obtained, instead, if the gate oxide thickness and junction depth are scaled together with the gate length (left plot in Fig. 4.22) as required by the most common scaling rules. In spite of the larger peak field value, the electron mean energy is not appreciably enhanced in the shorter device with respect to the long one. On the other

hand, carrier heating is actually stronger in the short device, as demonstrated by the electron energy distributions at the points of maximum mean energy shown in the right plot of Fig. 4.22. In particular, an enhanced high energy tail is obtained in the short device. Indeed, conventional MOSFET scaling also requires the reduction of power supply voltage in order to keep hot-carrier effects under control.

4.6.3 Simplified Models of Carrier Heating

As pointed out in Section 4.6.1, the average energy does not give sufficient information on the actual amount of high energy carriers. Distributions with the same average energy and significantly different tails can be found at different locations since the latter are sensitive to the details of the field profile. In spite of this well known fundamental issue, the cost and difficulty of solving the BTE promoted the derivation of many simplified models to predict $f(\vec{r}, E)$ from average quantities such as w .

Most of these models rely on a two-step procedure. The first step aims at calculating $w(\vec{r})$; the second one at reconstructing the whole distribution by imposing a fixed relationship between w and the shape of $f(\vec{r}, E)$. These models partly or totally neglect the non-local nature of carrier heating described by the BTE and are intrinsically doomed to loose accuracy as device scaling evolves towards steeper profiles.

4.6.3.1 Average Energy. Most simplified approaches start from the BTE (4.32) in the relaxation time approximation (4.34) and employ the moment's method [55] to derive, to different degrees of approximation, equations for n , $\vec{v}_d$ and w [70, 71]. These equations are solved either self-consistently (as in hydrodynamic and energy balance models) or as a post-processing step of conventional drift-diffusion analysis.

In one-dimension, neglecting generation-recombinations and heat fluxes, the simplest of these energy balance equations reads [71, 72]:

$$\frac{dw}{dx} = \frac{3}{5} qF(x) - \frac{w(x) - w_0}{\lambda_w}, \quad (4.43)$$

where $w_0 = 3k_B T_L/2$, $\lambda_w = 5v_d\tau_E/3$, v_d being the drift velocity and τ_E the energy relaxation time (Eqs. (4.38) and (4.35), respectively). In principle the energy relaxation length λ_w is a function of energy. In practice, it is often taken as a constant and adjusted to optimize agreement with experiments. Typical values are $\lambda_w \approx 65\text{--}80\text{nm}$ for electrons [73, 74, 75] and $\lambda_w \approx 55\text{--}100\text{nm}$ for holes [59, 75].

Integrating (4.43) along a current path computed (non self-consistently) with a conventional drift diffusion simulator, $\Delta w(x) = w(x) - w_0$ is obtained

as:

$$\Delta w(x) = \frac{3}{5} \int_0^x qF(\xi) \exp\left(-\int_\xi^x \frac{dz}{\lambda_w(z)}\right) d\xi, \quad (4.44)$$

where $w(0) = w_0 = 3k_B T_L/2$. This is an efficient procedure to calculate w , that accounts in a simple way for the first order non-local effects. More complex and accurate balance equations have been derived in [70].

4.6.3.2 Carrier Distribution. The second step in building a simplified model of carrier heating is to establish a relationship between the average energy and the shape of the distribution function. The simplest approach is to hold valid also in non-homogeneous conditions the displaced Maxwellian approximation (4.40) or the even simpler effective temperature model (4.42) and the approximate relation $w = 3k_B T_e/2$. For slowly varying fields (small changes within several mean free paths, that is $dw/dx \simeq 0$), and assuming λ_w constant, (4.43) reduces to:

$$T_e = T_L + \frac{2}{5} \frac{qF\lambda_w}{k_B}, \quad (4.45)$$

which relates the effective temperature of the distribution to the local electric field value, and therefore totally neglects the non-local nature of hot carrier transport. An approximate version of (4.45), $T_e = qF\lambda/k_B$, can also be derived comparing the effective temperature model with the Lucky Electron Model [76, 77, 78].

A difficulty of models based on (4.40) is the assumption of a parabolic band structure. Non parabolicity results in carrier distributions with depressed high energy tails even in homogeneous conditions [79]. To overcome these limitations, improved expressions for carrier distributions have been proposed [80, 81], that have the general form:

$$f(E) = A \exp\left(-\frac{E^\xi}{\eta(k_B T_e)^\nu}\right), \quad (4.46)$$

where T_e is usually derived from energy balance solutions under the approximation $w \simeq 3k_B T_e/2$.

The models resulting from combining the two steps described above rely on a few parameters ($\xi, \eta, \nu, \lambda_w$) to be adjusted by comparison with experiments (gate or substrate currents). These models have been applied to the simulation of MOSFETs, EPROM and Flash cells [80, 81, 82, 83, 84] with gate length down to a few tenths of a micron and high drain voltage during the programming phase ($V_{DS} > 5V$). For shorter devices and smaller drain biases, relevant non local effects make these models less reliable.

4.6.4 Impact Ionization

Impact ionization is the process by which energetic carriers cause the transition of valence band electrons into the conduction band, thus generating electron-hole pairs. Impact ionization is a complex phenomenon, whose probability depends on the $\vec{k}$ -state of the primary electron. To compute the ionization probability per unit time, that is the scattering rate (S_{II}), it is necessary to evaluate all the possible transitions from the initial state leading to the creation of a secondary electron and a secondary hole, fulfilling energy and momentum conservation laws.

Energy conservation requires a primary electron energy larger than $E_G(\text{Si})$. If momentum conservation is also enforced and only direct transitions are considered (no phonons involved) the minimum energy becomes $3E_G/2$ [50], which is often reported as the threshold energy for impact ionization. Phonon assisted impact ionization events relax this constrain, so that the actual impact ionization scattering rate increases smoothly from $E_K = E_G$ with a soft threshold (see Fig. 4.16).

The impact ionization generation rate, that is the number of electron-hole pairs generated per unit time and volume at point $\vec{r}$ by one carrier type is given by:

$$G_{II}(\vec{r}) = \frac{1}{4\pi^3} \int S_{II}(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k} , \quad (4.47)$$

where S_{II} represents the rate of impact ionization per unit time for an electron with initial wave vector $\vec{k}$. If the energy of the primary carrier is low, S_{II} is highly anisotropic, and changes as much as 2 orders of magnitude depending on $\vec{k}$ direction [85]. At high energy, instead, the spread over $\vec{k}$ directions is significantly reduced [86].

In order to gain a more intuitive understanding of the ionization process it is useful to average $S_{II}(\vec{k})$ over equi-energy surfaces in $\vec{k}$ -space and to express G_{II} in terms of an energy dependent scattering rate $S_{II}(E)$ as:

$$G_{II}(\vec{r}) = \int_{E_G}^{\infty} S_{II}(E) f(\vec{r}, E) g(E) dE . \quad (4.48)$$

Adding the contributions of both carriers, G_{II} is often written as:

$$G_{II}(\vec{r}) = \alpha_n \frac{|\vec{J}_n(\vec{r})|}{q} + \alpha_p \frac{|\vec{J}_p(\vec{r})|}{q} , \quad (4.49)$$

where the ionization coefficients α_n and α_p are defined as the number of electron-hole pairs generated by a single electron (hole) per unit traveled distance. In essence, the product $\alpha \cdot dx$ represents the probability that a carrier

suffers impact ionization while covering the distance dx . Equating (4.47) to either one of the terms at the right hand side of (4.49) and accounting for (4.29), α_n and α_p can be expressed as:

$$\alpha_{n,p}(\vec{r}) = \frac{\int S_{II}(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k}}{\int f(\vec{r}, \vec{k}) v_g(\vec{k}) d\vec{k}}, \quad (4.50)$$

where S_{II} , f and v_g refer to the considered carrier. Eq. (4.50) clearly points out that $\alpha_n(\vec{r})$ and $\alpha_p(\vec{r})$ are functions of the local electron and hole distributions respectively, that is, non-local functions of the electric field. Since S_{II} is lower for holes than for electrons (especially at low energy [56], Fig. 4.16), and hole distributions are cooler than electron ones because of the higher phonon scattering rate (Fig. 4.16 and [87]), α_p is smaller than α_n .

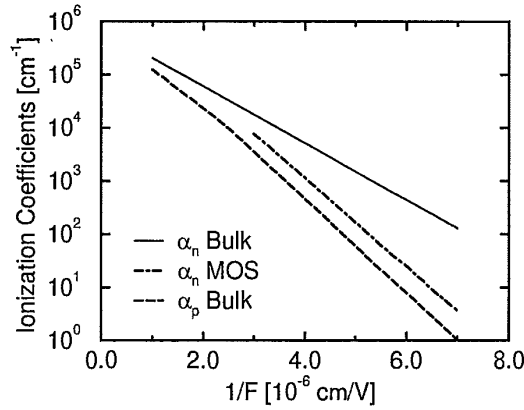


Figure 4.23 Electron and hole impact ionization coefficients at room temperature. Data from [88] (solid line), [89] (dot-dashed line), [88] (dashed line).

In homogeneous conditions (constant field), there is a unique relation between the distribution and the field, so that the alphas become unique functions of the electric field, well approximated by Chynoweth equation [90]:

$$\alpha_{n,p} = \alpha_{n,p}^{\infty} \exp(-F_{n,p}^{\text{crit}}/F), \quad (4.51)$$

with $\alpha_n^{\infty} = 7.03 \times 10^5 \text{ cm}^{-1}$, $F_n^{\text{crit}} = 1.231 \text{ MV/cm}$, $\alpha_p^{\infty} = 1.582 \times 10^6 \text{ cm}^{-1}$, $F_p^{\text{crit}} = 2.036 \text{ MV/cm}$, for $F < 4 \times 10^5 \text{ V/cm}$ and $\alpha_p^{\infty} = 6.71 \times 10^5 \text{ cm}^{-1}$, $F_p^{\text{crit}} = 1.693 \text{ MV/cm}$ for $F > 4 \times 10^5 \text{ V/cm}$ [88, 91]. Fig. 4.23 reports room temperature impact ionization coefficients in bulk silicon as a function of electric field, measured in conditions for which F is slowly varying with position.

In a highly non-uniform electric field, f is not a well defined function of the local field intensity because carrier heating lags behind the local electric field (Fig. 4.19). Therefore, impact ionization coefficients can not be modeled by simple relationships as (4.51). In the framework of the simplified models discussed in Section 4.6.3, approximate expressions for α_n and α_p in terms of average electron energy can be derived replacing F with an effective field $F_{\text{EFF}} = 5(w - w_0)/3q\lambda_w$, which is consistent with (4.43) under the assumption of slowly varying field ($dw/dx \simeq 0$). The more general expressions (4.47) and (4.48) should be used in higher order models.

Since scattering rates are tightly related to the band structure of the material ($E(\vec{k})$), and the latter is modified by the presence of an interface (Section 4.2.5), all scattering rates at the surface of a MOS system may be expected to be different than in bulk Si. Experimental evidence supporting this conclusion with regard to S_{II} was first given in [89], and it is reported in Fig. 4.23. The reduced ionization coefficient of MOSFETs described in [89] has been tentatively explained invoking channel quantization [57]. Recently, however, it has been argued that the discrepancy between surface (MOS) and bulk ionization coefficients can be attributed to the local models used to derive α from experiments [92], so that no clear evidence of significant differences between bulk and surface ionization coefficients exists, at least in the most common bias range for state of the art technologies. The great importance of non local effects in determining impact ionization in state of the art devices is now generally agreed upon [74].

4.6.5 Substrate Current

Impact ionization is responsible of a few detrimental effects that eventually limit the maximum drain voltage applicable to MOSFETs and Flash cells.

In fact, as V_{DS} is increased the ionization probability in the reverse-biased drain junction approaches unity and leads to an uncontrolled increase of the current (junction breakdown) that may cause the destruction of the device due to excessive power dissipation. Well before junction breakdown, holes generated by impact ionization and swept towards the substrate contact by the vertical electric field (process 1 in Fig. 4.18) give rise to a measurable substrate current (I_B) that can adversely affect the operation of substrate bias generators and the subthreshold characteristics of neighboring devices [93]. In addition, I_B gives rise to ohmic drops that can forward-bias the source-substrate junction and lead to parasitic bipolar phenomena such as snap-back and latch-up.

Assuming that all the generated holes contribute to I_B (i.e. neglecting carriers' recombination and hole injection into the source), the substrate current can

be expressed integrating the generation rate of (4.48) over the device volume:

$$I_B = q \int_V G_{II}(\vec{r}) d\vec{r} = q \int_{E_G}^{\infty} S_{II}(E) \left(\int_V f(\vec{r}, E) g(E) d\vec{r} \right) dE . \quad (4.52)$$

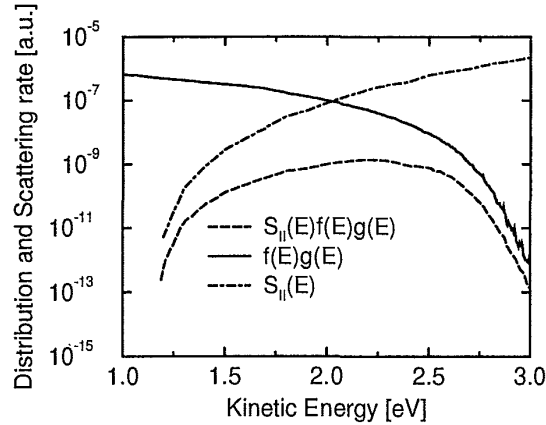


Figure 4.24 Impact ionization rate (dot-dashed line), electron distribution integrated over the device (solid line) and corresponding product (dashed line) in a $L = 0.5\mu\text{m}$ MOSFET biased at $V_G = V_D/2 = 1.5\text{V}$. $t_{\text{ox}} = 10\text{nm}$. $E = 0$ at the bottom of the conduction band.

Fig. 4.24 reports $f(E)g(E)$ and $S_{II}(E)f(E)g(E)$ integrated over the volume of a $0.5\mu\text{m}$ MOSFET. As the result of the trade-off between the decreasing $f(E)g(E)$ and the increasing ionization rate $S_{II}(E)$, the largest contribution to impact ionization comes from electrons between 2 and 2.5 eV, but no obvious threshold energy can be identified. In general terms, the energy of primary electrons giving the maximum contribution to impact ionization depends on how steeply the distribution function decays in energy. Progressively cooler distributions result in an average energy of ionizing electrons that approaches the bandgap energy.

Fig. 4.25 reports the results of an accurate numerical analysis of impact ionization in a MOSFET biased at $V_{GS} = V_{DS}/2$, that is, close to the condition of maximum substrate current. Contour lines represent the loci of equi-impact ionization generation rate. Due to the spatially retarded electron heating, G_{II} is maximum inside the drain, and not at the drain junction where the field is maximum, as it might be expected based on local considerations. In addition, since for high drain voltage and low gate voltage in proximity of the drain the current flows far from the interface due to carrier diffusion and to the repulsive

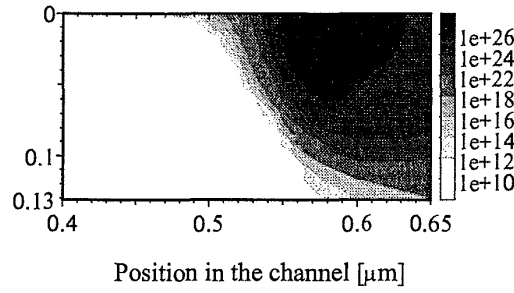


Figure 4.25 Impact ionization generation rate contours for electrons at the drain end of a $L = 0.5\mu\text{m}$ MOSFET biased at $V_G = V_D/2 = 1.5\text{V}$. $t_{\text{ox}} = 10\text{nm}$. The solid line represents the drain junction.

vertical field, G_{II} is maximum somewhat inside the bulk of the material and not at the interface.

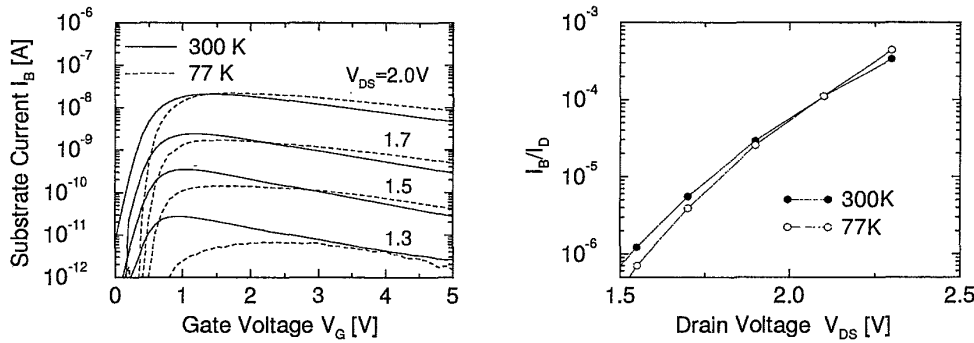


Figure 4.26 Left: Substrate current as a function of gate and drain voltages and temperature. Data from [94]. Right: Normalized substrate current as a function of drain voltage at 300 and 77K. Data from [95].

Fig. 4.26 reports typical substrate current curves versus gate and drain voltages. I_B rapidly increases for increasing V_{DS} , as an expected consequence of the increased field in the channel, resulting in enhanced high energy tails of the distribution function (as suggested by (4.42) and (4.45)). As V_{GS} is raised above the threshold voltage (V_T) a progressively larger number of carriers flows in the channel. Therefore, $f(\vec{r}, E)g(E)$ in (4.52) (whose integral over all energies is the carrier concentration $n(\vec{r})$) increases, and I_B increases too. At the same time, however, the maximum electric field at the drain junction decreases,

leading to reduced carrier heating, hence, to a smaller T_e in the effective temperature model. These two effects typically balance each other for $V_{GS} \approx V_{DS}/2$ where a maximum of I_B is obtained (see again the left plot in Fig. 4.26). For $V_{GS} > V_{DS}/2$ the substrate current decreases due to the dominant effect of reduced carrier heating. Impact ionization increases as the gate length and oxide thickness are scaled, due to the increase of the electric field in the channel (Fig. 4.22).

The temperature dependence of I_B is mainly determined by three factors:

- At fixed bias, a low temperature raises the threshold voltage (V_T), hence reduces the carrier flux and slightly enhances the peak electric field.
- Optical phonon scattering, typically the dominant inelastic scattering mechanism, is reduced at low T_L . Hence, the distribution function is hotter than at room temperature for a given field (see, for example, Fig. 4.17).
- The impact-ionization scattering rate decreases at low temperature because the energy gap, which is the minimum energy for impact ionization, increases at low temperature.

The combined action of these factors results in the temperature dependence shown in Fig. 4.26. The first effect reduces impact ionization around threshold regardless of V_{DS} (left plot). The second one dominates far from threshold at high V_{DS} . Together with the increase of the peak field, it is responsible for the increased maximum I_B , typically observed at low T_L (left plot, and right plot for $V_{DS} > 2.25V$). For drain voltages approaching the bandgap voltage E_G/q , instead, the third effect dominates, and makes the substrate current at low T_L lower than at room temperature [95, 96, 97], as observed in the right plot of Fig. 4.26 for $V_{DS} < 2V$.

4.6.6 Hot Carrier Injection into SiO_2

In quite general terms, the electron current density from Si to SiO_2 at a point x along the channel of a MOSFET can be analytically expressed as:

$$J_{inj,n}(x) = q \int_0^\infty v_\perp(x, E) f_\perp(x, E) g(E) P(x, E) dE \quad , \quad (4.53)$$

where $f_\perp(x, E)$ is the hemi-distribution of electrons that hit the interface between x and $x + dx$, $P(x, E)$ is the injection probability, and $v_\perp(x, E)$ is the the electron velocity component perpendicular to the interface and directed towards it. The integrand in (4.53) is conceptually similar to (4.18), and points out that the injected current is the combined result of a few factors: 1) the number ($f_\perp \cdot g$) and velocity ($v_\perp$) of electrons directed towards the interface;

2) the electron energy and momentum distributions (because only carriers with velocity directed towards the interface contribute to $J_{inj,n}$); 3) the probability of injection from Si to SiO₂ ($P(x, E)$). Differently from (4.18), in the two dimensional case considered here most quantities are explicit functions of the position along the interface, here denoted with x . Let us discuss some of them in more detail.

4.6.6.1 Distribution Function. If the distribution is a quasi equilibrium Maxwell-Boltzmann (or Fermi-Dirac) function, the number of available carriers decreases so rapidly for increasing energy that significant injection occurs only from the bottom of the conduction band. In essence, this is the “pure tunneling” case analyzed in Section 4.4.1. For $V_{DS} > 0$ this situation can occur at the source side of a MOS, if V_{GS} is large or if t_{ox} is very thin, as in extremely advanced devices [98].

On the other hand, if the tail of the energy distribution is largely populated, injection is mostly due to carriers in proximity or above the top of the barrier, because low energy ones have a small injection probability. This is the most common situation at the drain end of MOSFETs in the “hot carrier injection” conditions (high V_{GS} and V_{DS}) exploited to program the industry standard Flash devices.

4.6.6.2 Injection Probability. As for the injection probability, it should be noticed that several physical phenomena are lumped into this term.

If an abrupt effective potential barrier is assumed at the interface (for example because image force corrections are neglected), $P(x, E)$ essentially becomes the tunneling probability through the barrier ($P(x, E) = T(x, E)$, where T is the transmission coefficient). Since the potential profile at the drain end of the channel is highly two-dimensional, this probability is harder to evaluate than in the one-dimensional case treated in Section 4.3.2. Often, a gradual channel approximation is used in which the point by point applicability of the one-dimensional tunneling rates is assumed (see, e.g. [81, 95]).

Moreover, it has to be remembered that transmission coefficients (including those of (4.16) and (4.15)), are calculated assuming a simple parabolic band approximation whose validity is questionable for high energy electrons near the top of the barrier.

In addition, if image force barrier lowering is accounted for, P actually models multiple processes in series, namely: emission from Si into SiO₂, transport within the SiO₂ conduction band, and, possibly, tunneling through the SiO₂ barrier. A satisfactory microscopic description of most of these processes has not been reached yet (see, for example, [15, 57]), also because of the essentially unknown properties of the band structure at the interface.

4.6.7 Gate Current

From the macroscopic point of view, carrier injection from silicon into silicon dioxide in MOSFETs is revealed by the existence of measurable gate currents (I_G) and by charge collection in the floating gate of memory cells. I_G is related to the injected current density in a non trivial way, essentially because of two reasons.

- Injected carriers can be reflected back towards the silicon. Indeed, carriers that enter the oxide conduction band therein suffer collisions that cause energy loss and directional changes of momentum. Although in principle electrons can access all oxide regions that are energetically admissible, even if this implies moving uphill against a repulsive oxide electric field, the combined effects of energy losses and opposing field can result in significant re-emission towards the silicon. Since in many cases I_G is limited by injection in the oxide conduction band and not by oxide transport, modest effort has been devoted so far to properly include this effect in two dimensional device simulators. However, they are certainly important for $V_{GS} < V_{DS}$ [99].
- Both electrons and holes contribute to I_G . Indeed, at most of the bias points of interest for gate current analysis, significant impact ionization takes place in the substrate. Depending on bias conditions, generated holes can reach the interface with enough energy to be injected into the oxide and possibly reach the gate.

Therefore, in quite general terms we can write:

$$I_G = W \int_0^L [\eta_n(x) J_{inj,n}(x) - \eta_p(x) J_{inj,p}(x)] dx, \quad (4.54)$$

where W and L are the gate dimensions, while η_n and η_p denote the efficiency (≤ 1) by which injected carriers actually reach the gate.

In MOSFETs and Flash cells J_{inj} and η change with position and bias, giving rise to a complex physical picture, illustrated by the typical gate current curves of Fig. 4.27. In the following we will interpret these curves with the aid of the simulated potential profiles and carrier distributions of Fig. 4.28.

4.6.7.1 Channel Hot Electron Injection. At $V_{GS} \simeq V_{DS}$ (case (a) in Fig. 4.28), the vertical electric field attracts electrons towards the interface along most of the channel and, in particular, around the point of maximum carrier heating, typically located slightly beyond the drain junction. The oxide field (F_{OX}) at the drain end of the channel is zero or weakly positive. Hence,

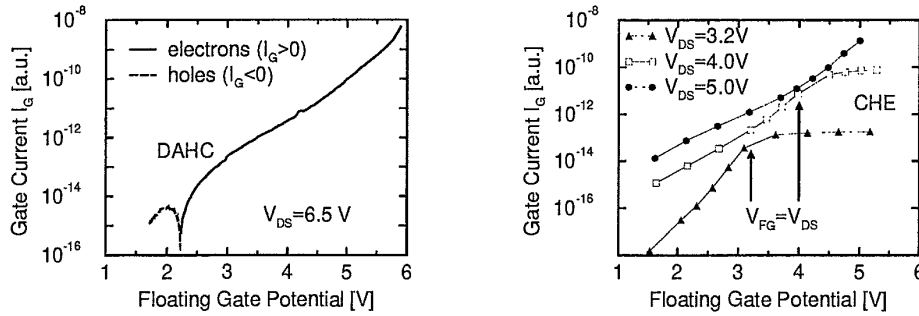


Figure 4.27 Left: Gate current of a Flash cell at high V_{DS} demonstrating electron and hole injection. $L_{EFF} = 0.4\mu\text{m}$. Data from [100]. Right: Gate current of a Flash cell as a function of floating gate and drain potentials. $L_{EFF} = 0.3\mu\text{m}$.

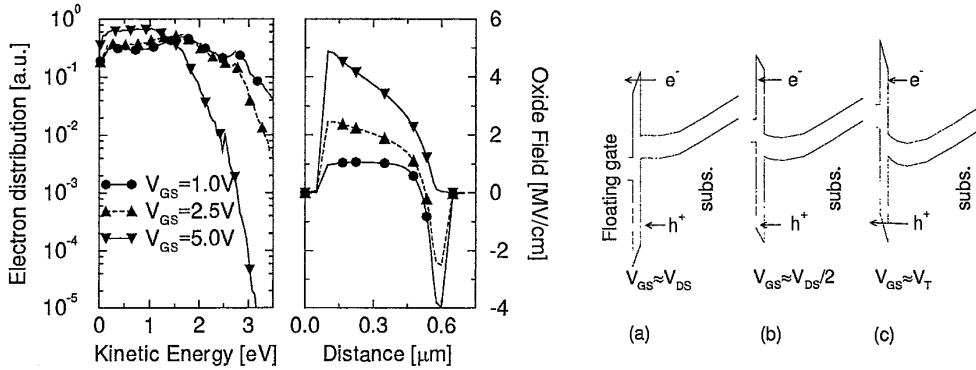


Figure 4.28 Left: Electron distributions at the point of maximum carrier heating in a $0.5\mu\text{m}$, $t_{ox} = 10\text{nm}$ MOSFET biased at: a) $V_{GS} = 5\text{V}$; b) $V_{GS} = 2.5\text{V}$; c) $V_{GS} = 1\text{V} \approx V_T$. $V_{DS} = 5\text{V}$, $V_{SB} = 0\text{V}$. Center: Oxide field along the interface. Right: Vertical band diagram at the drain junction. Arrows indicate carriers with equal kinetic energy at the interface.

$\eta_n \simeq 1$. Generated holes, instead, are pushed away from the interface by the repulsive vertical electric field.

Moving towards the source F_{OX} increases and enhances $P(x, E)$ (Fig. 4.28) but the distribution function becomes much cooler (low T_e). Since this latter effect dominates, injection is rapidly extinguished, and $J_{inj}(x)$ is sharply peaked at the drain end of the channel, as documented in Fig. 4.29. This injection regime is generally referred to as Channel Hot-Electron injection (CHE), and

it is widely adopted for Flash programming because for a given V_{DS} it nearly corresponds to the largest achievable electron gate current (I_G , see the right plot in Fig. 4.27) and injection efficiency (I_G/I_D).

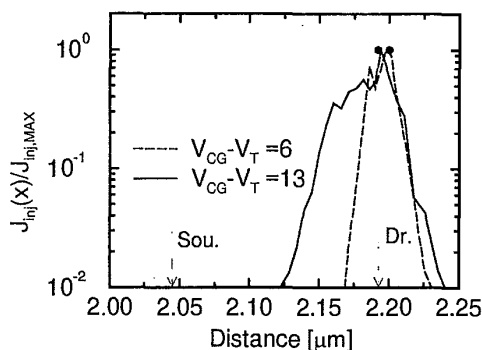


Figure 4.29 Simulated gate current density along the channel of a Flash cell normalized to its maximum value. Filled points indicate the abscissae of maximum injection.

Increasing V_{GS} above V_{DS} increases F_{OX} , hence $P(x, E)$, over the whole channel. However, the peak field decreases, thus depressing the high energy tail of f . Correspondingly, the injection area widens towards the source (Fig. 4.29). I_G can decrease, increase or stay essentially constant (as in the right plot of Fig. 4.28) depending on which of the two effects dominates [101].

Upon application of a source substrate voltage, the lateral field in the channel slightly increases in the pinch-off region, thus enhancing the high energy population of the distribution. Therefore, the gate current slightly increases for increasing V_{SB} in the CHE regime.

4.6.7.2 Drain Avalanche Hot Carrier Injection. The injection regime from $V_{GS} < V_{DS}$ down to $V_{GS} \approx V_T$ is often referred to as Drain Avalanche Hot Carrier injection (DAHC, [102]), as it partly involves holes generated by impact ionization. This terminology originates from studies on gated diodes [103, 104] and, although not strictly applicable to the MOSFET case [105], is still widely used.

As illustrated in Fig. 4.28 (case (b)), if V_{GS} is lowered below V_{DS} , the oxide field at the drain end of the channel becomes repulsive for electrons and attractive for holes, but reverses to attractive for electrons and repulsive for holes as we move towards the source. The carrier distribution is hotter than at a larger V_{GS} , essentially because the peak field is higher. However, the opposing

oxide field reduces the carrier concentration, increases the effective barrier for electron injection (that is, decreases P exponentially) and reflects the electrons injected through the interface back to the channel ($\eta_n \ll 1$). Since only a much smaller fraction of hot electrons than in the CHE regime has enough energy to surmount the barrier or to find its way to the gate, I_G decreases rapidly if V_{GS} is reduced below V_{DS} , as evidenced in the right plot of Fig. 4.27.

In addition, secondary holes generated by impact ionization near the drain experience a two dimensional field that heats and pushes them towards the interface [106]. The hole distribution at the interface features a tail of energetic holes that have non-negligible probability of injection because of the attractive F_{OX} (hence, $\eta_p \simeq 1$). Therefore, electron and hole injection coexist and partially compensate each other [99].

The centroids of injected electrons and holes are slightly offset with respect to each other, but it is believed that a certain fraction of the interface experiences both electron and hole injection currents, although the intensity of the two currents can be very different. The simultaneous injection of both electrons and holes is especially important for the hot carrier degradation mechanisms that will be discussed in Section 4.8.

As V_{GS} is further reduced, $\eta_p J_{inj,p}$ increases and $\eta_n J_{inj,n}$ decreases until the total gate current goes to zero. This happens almost at the same V_{GS} for which the substrate current reaches the maximum value provided the drain voltage is large enough to allow for substantial hole heating as it is the case in the left plot of Fig. 4.27. Notice that the condition $I_G = 0$ actually corresponds to large values of $\eta_n J_{inj,n}$ and $\eta_p J_{inj,p}$ that balance each other.

Finally, for V_{GS} approaching V_T (case (c) in Fig. 4.28) carrier heating increases, but electron injection is suppressed by the opposing oxide electric field and only the negative hole current is measured at the gate electrode. This current is observed in Fig. 4.27 for $V_{FG} < 2.2V$, and it is orders of magnitude smaller than the CHE current because of the larger barrier height for holes than for electrons ($\Phi_{B,h+} \approx 4-5eV$ and $\Phi_{B,e-} \simeq 3.15eV$, respectively), of the larger scattering rate (smaller mean free path) that limits hole heating, and of the sudden drop of the hole density of states at energies comparable to the hole barrier height [60].

Experimentally, it is observed that V_{SB} enhances the electron gate current also in the DAHC regime, through mechanisms that are presently under investigation [105].

4.6.7.3 Secondary Generated Hot Electron Injection. SGHE denotes a few mechanisms by which electrons created inside or just outside the depletion region below the gate are accelerated towards the interface and possibly injected

in the oxide by the vertical field, thus providing additional and sometimes dominant contributions to the gate current [102].

Among the mechanisms responsible of SGHE we mention:

- photons emitted by channel hot carriers while relaxing to low energy states (mechanism 5 in Fig. 4.18). These photons penetrate in the bulk, where they can be reabsorbed with creation of carrier pairs [107, 108, 109]. The generated electrons that escape recombination diffuse to the edge of the depletion region and then are pushed to the interface (mechanism 7 in Fig. 4.18). Alternatively, they diffuse towards adjacent n^+ regions (mechanisms 6);
- holes generated by impact ionization in proximity of the interface can re-ionize while drifting towards the substrate, thus creating tertiary electrons deep inside the depletion region (mechanism 4 in Fig. 4.18) [102, 110];
- band-to-band tunneling in heavily doped substrates triggered by the application of a substrate potential [40, 111, 112];
- thermal generation inside the depletion region [112].

In general, SGHE injection is very sensitive to the applied source-substrate voltage (V_{SB}), since the latter controls directly the maximum potential drop and the intensity of the electric field experienced by the carriers. The importance of SGHE related effects is expected to increase in the future, because the triggering mechanisms listed above are enhanced by the high fields, doping and doping gradients, typical of scaled technologies [113, 114].

Since in proximity of the drain the vertical potential drop from the substrate to the interface is larger than the lateral one from source to drain (because of the added built in potential of the drain junction and of the applied source-substrate voltage), the maximum potential energy drop available to substrate generated carriers is larger than that available to channel electrons. Therefore, SGHE effects are best observed at low bias, and become progressively more important as supply voltages are reduced and the available potential energy drop becomes the limiting factor for hot carrier effects. This point will be further discussed in Section 4.6.8.

4.6.7.4 Substrate Hot Electron Injection. Often referred to as homogeneous injection, SHE was proposed in [20, 115, 116] as a simple mean to study various hot carrier effects.

As sketched in Fig. 4.30a, an n-MOS structure is biased above threshold ($V_{GS} > V_T$), at zero drain-source voltage (V_{DS}), and with a large source-substrate voltage (V_{SB}) that gives rise to a wide depletion region below the

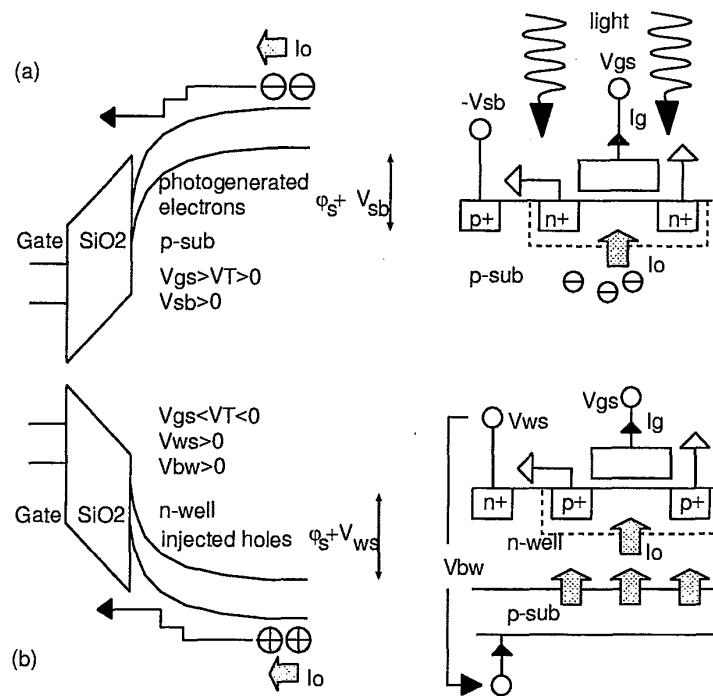


Figure 4.30 Schematic representation of the substrate hot electron (a) and hole (b) injection experiment.

gate. In this condition an essentially one-dimensional electric field perpendicular to the interface is achieved below the gate. Since the surface potential (ϕ_s) is almost constant, the oxide field ($F_{OX} = (V_G - V_{FB} - \phi_s)/t_{ox}$, where V_{FB} is the flat band voltage) is controlled only by V_{GS} . V_{SB} , instead, separately controls the potential drop and the field in the semiconductor depletion region.

Cold electrons are either optically generated exposing the sample to a photon flux [20] or injected by forward biasing a buried junction [68, 115, 117], or generated by band-to-band tunneling [40]. Upon entering the depletion region, they are accelerated towards the interface by the substrate electric field. Those electrons that gain enough energy are injected into the gate oxide, while the remaining ones are collected by the source and drain terminals (I_S, I_D). Only electrons reach the interface, and their fluence can be easily monitored acting on the light intensity or on the buried junction forward bias. Hole injection experiments can be performed similarly on p-MOSFETs (see Fig. 4.30b).

Fig. 4.31 reports normalized electron and hole gate currents measured in homogeneous injection experiments on n and p-MOSFETs featuring comparable

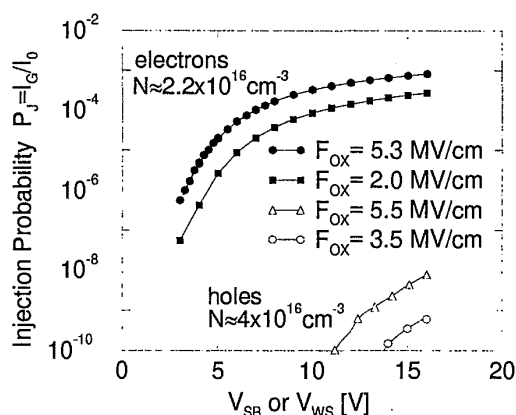


Figure 4.31 Measured homogeneous electron and hole gate currents normalized to the injected current I_G/I_0 as a function of accelerating substrate voltage and oxide field in n- and p-MOS devices featuring comparable substrate doping (N). Data from [68].

substrate doping concentration. I_G/I_0 increases for increasing V_{SB} because the high energy tail of the distribution is enhanced by the increasing field. I_G/I_0 is also a function of F_{OX} because $T(E)$ increases (Fig. 4.8). For a given V_{SB} , and F_{OX} , hole currents are smaller than electron ones because of the shorter mean free path and the higher Si-SiO₂ barrier with respect to electrons.

4.6.7.5 Implications for Device Operation. From the practical point of view, all the injection regimes described above are of the utmost importance for the operation and the reliability of Flash cells.

In general terms, we notice that in hot carrier generation conditions, most electrons do not gain enough energy for injection and are eventually collected by the drain terminal. Therefore, the injection efficiency (I_G/I_D) is typically small, and becomes a concern for single voltage Flash devices powered at reduced supply. On the contrary, no channel current is needed for tunneling injection, which in principle results in a much larger efficiency, mainly limited by band-to-band and tunnel induced generation and multiplication at the junction (Chapter 2).

More specifically, CHE injection is routinely used for programming the industry standard Flash device. A memory cell based on SGHE injection that can be programmed in less than 10 μ s with drain voltages below 3V has been proposed in [113]. The same phenomenon is exploited in [118] to achieve self convergence of over-erased cells. A few other self convergence schemes based on the balance between electron and hole currents in the DAHC regime have

been proposed in [119, 120]. SHE injection at zero V_{DS} has been indicated as a mean to achieve efficient programming at low voltages [111, 121].

In addition, column read and program disturbs are caused by low level hot carrier injection into the floating gate. Substrate electron injection has also been identified as a dangerous source of disturbs [112, 122].

4.6.8 Hot Carrier Effects at Low Voltages

If supply voltages are lower than those corresponding to the thresholds for injection over the Si-SiO₂ energy barrier and for impact ionization ($\Phi_{B,e^-}/q = 3.15V$ and $\Phi_{B,h^+}/q \approx 4V$, $\Phi_{II}/q \simeq E_G/q \simeq 1.1V$, respectively), the energy acquired by carriers drifting in the applied potential becomes insufficient to justify thermionic emission or impact ionization, respectively. On the other hand, as

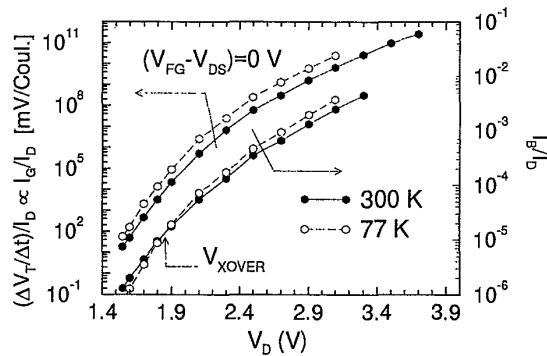


Figure 4.32 $\Delta V_T/\Delta t I_D$, proportional to I_G/I_D , and I_B/I_D of a Flash memory cell at low V_{DS} . Data from [123].

shown in Fig. 4.32, no dramatic decrease of hot-carrier effects has been observed when V_{DS} decreases below these thresholds. [96, 123, 124, 125]. Therefore, additional and efficient energy-gain mechanisms capable to populate the tail of the distribution for $E > qV_{DS}$ must be invoked to explain the observed gate and substrate currents.

Several mechanisms have been suggested to explain these evidences: 1) net phonon absorption [126, 127]; 2) carrier-carrier interaction [95, 97, 128, 129]; 3) impact ionization feedback [110]; 4) photon assisted processes; 5) Auger recombinations [124]. Understanding these mechanisms is becoming of increasing importance to preserve the injection efficiency of single voltage non volatile memories in spite of reduced supply voltages.

Low voltage hot carrier effects impact directly also the reliability of Flash cells. In particular, spurious hot carrier injection at low V_{DS} ($\approx 1V$, that is

$\ll \Phi_{B,e-}/q$) occurs during read operations and it is responsible of a read disturb known as soft programming which can lead to a threshold voltage drift in the reference cells of the sense amplifiers, and to data loss in erased bits subject to repeated read operations (see also Chapter 7).

Soft programming can not be attributed to Fowler-Nordheim tunneling of cold carriers, since it becomes undetectably small as V_{DS} tends to zero [94]. Recent experiments and theoretical calculations indicate that electron-electron interaction and tunneling injection of mildly hot carriers might be the dominant mechanisms responsible of this disturb [95, 130].

4.7 OXIDE DEGRADATION DUE TO HIGH FIELD STRESS

The continuous application of a high electric field across the gate dielectric of a MOS structure, as it is necessary during tunneling erase or program of Flash cells, progressively degrades the insulating properties of the oxide (wear-out phase). Eventually, a conductive path is locally generated and the electrostatic energy accumulated in the MOS capacitor is abruptly discharged by a strong current (breakdown). In most cases the associated self-heating vaporizes the oxide and part of the gate material on top of it, thus leaving a stable conductive path across the oxide that prevents proper device operation.

The simplest characterization of oxide robustness to stress at high electric fields employs the so called time zero breakdown experiments (TZB), in which a fast voltage ramp ($\approx 1\text{V/s}$) is forced on the MOS capacitor and the breakdown voltage (V_{BD}), or field $F_{BD} \approx V_{BD}/t_{ox}$, is recorded. Based on this technique, breakdown events can be classified as extrinsic or intrinsic. The former are related to major defects of the native oxide (contaminants, oxygen precipitates, etc.), that cause almost instantaneous ruptures at small electric fields ($< 5\text{MV/cm}$). The latter occur at much larger fields (typically $> 10\text{MV/cm}$), and are related to the intimate disordered nature of the oxide structure, that inherently incorporates weak spots (see also Section 4.2.6). Gross extrinsic defects are rare in state of the art ultraclean processes and in small area devices ($\approx 10^{-4}\text{cm}^2$ or less). Thus, studies have soon focused upon the detailed analysis of intrinsic breakdown and on the wear-out phase that precedes it.

4.7.1 Oxide Wear-out and SILC

In order to characterize oxide wear-out, time dependent experiments are performed in which the current and voltage across the MOS capacitor are recorded as a function of time.

The evolution of the gate current density, J_G (gate voltage, V_G) during constant voltage (current) stress experiments is illustrated in Fig. 4.33. A transient of positive charge formation thins out the barrier causing the initial current in-

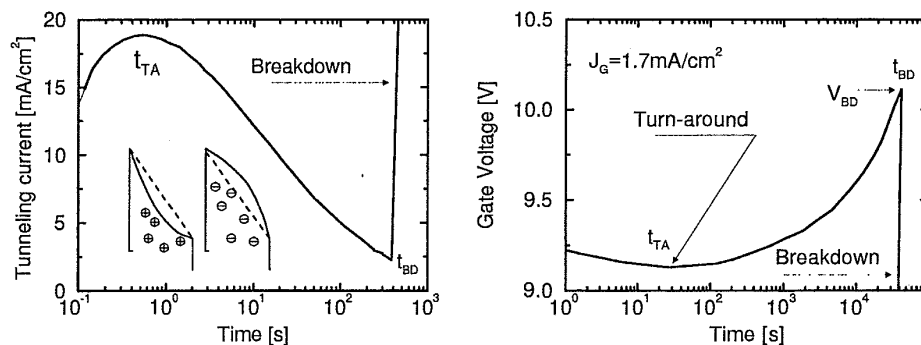


Figure 4.33 Typical gate current (left) and voltage (right) during constant voltage (current) stress experiments on MOS capacitors. The inset shows the deformation of the oxide barrier shape due to positive and negative net charges.

crease (voltage decrease), for $t < t_{TA}$. Eventually, negative charges accumulate in the oxide, thickening the barrier and reducing the current (increasing the voltage), as observed for $t > t_{TA}$. Depending on oxide thickness, electric field and current density, either one or both regimes are observed.

The origin of the positive and negative charges causing these transients has been much debated and ascribed by different authors mainly to three type of effects.

First, it has been observed that if n-MOS transistors are chosen for the stress experiments instead of the MOS capacitors, a large hole substrate current (J_P) is measured in addition to the gate electron tunneling and the source/drain currents [132]. This current is a clear indication that a large number of holes is generated during the stress phase. Fig. 4.34 illustrates the oxide field dependence of this current, and sketches some of the hole generation mechanisms postulated to explain it. In particular, these are: a) impact ionization of energetic electrons entering the anode and subsequent hole tunneling towards the cathode [133]; b) impact ionization in the oxide [134]; c) valence band electron tunneling [132]. Mechanism (b) is expected to be ineffective in thin oxides ($t_{ox} < 10\text{nm}$) where the voltage drop is insufficient to provide electrons of the required energy ($E \approx E_G(\text{SiO}_2) > 8\text{eV}$) [132]. Mechanism (c) is inadequate to account quantitatively for the hole current and for its dependence on oxide thickness [135, 136] at least down to $t_{ox} \approx 4\text{nm}$. Therefore, not without controversial opinions [137], process (a) is commonly reported as the most likely responsible for hole generation [138]. Due to their large effective mass m^* , low mobility [139] and high trapping probability, holes can be easily captured in

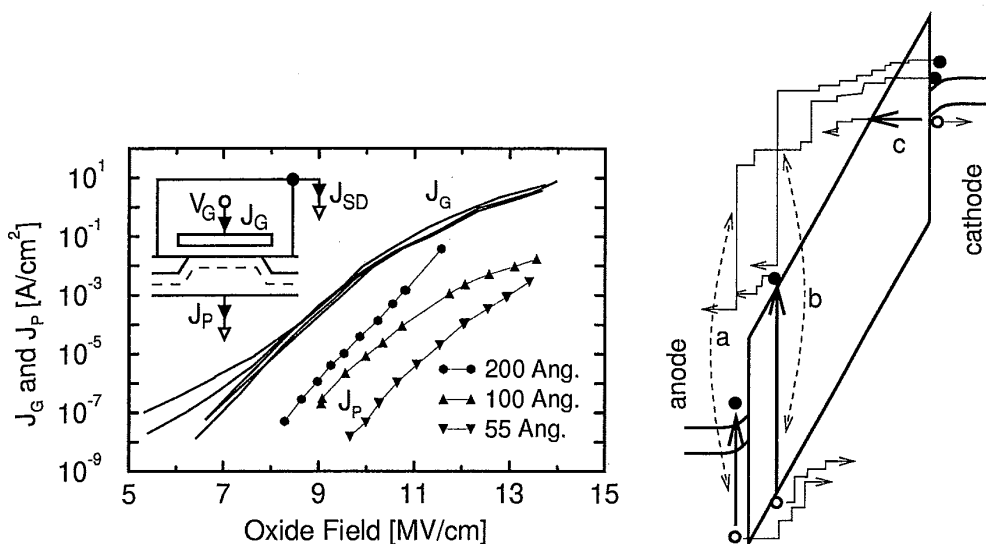


Figure 4.34 Left: density of the electron gate current (J_G) and of the hole substrate current (J_P) during tunneling experiments. Data from [131]. Right: Schematic representation of a few processes leading to the generation of holes during tunnel injection.

the oxide. Thus, hole generation and trapping provides a first mechanism of positive charge accumulation in the oxide.

Second, two electric field dependent processes compete in determining the net space charge in the oxide and the tunneling current, namely carrier trapping and removal in and out of existing oxide traps [140]. The balance between these mechanisms is such that only a fraction of the traps is occupied at a given field. This fraction decreases rapidly for increasing oxide field [118, 141]. Depending on the trap type and charge state, positive or negative charges can appear in the oxide. In addition, sudden F_{OX} changes (as those occurring while passing from stress to characterization phases and vice versa) trigger charge redistribution transients in and out of the traps, with a wide variety of time constants (depending on trap location and energy, oxide thickness, injected current density, etc. [140, 141, 142]). These transients interfere with the accurate evaluation of trapped charges and trap density, causing widespread opinions on their origin and relative importance.

Third, a continuous generation of defects, acting as electron or hole traps, takes place during stress. Charges captured in the newly generated traps influence the oxide field and the tunnel current. The generation rate increases rapidly for increasing oxide field. Among other mechanisms, defect/trap generation has been attributed to: 1) impact ionization inside the oxide near the

anode [143]; 2) recombination of cathode injected electrons with anode injected holes, producing neutral electron traps [144]; 3) high field emission of electrons from the bulk of the oxide and from the Si-SiO₂ interfaces, generating positively charged traps that subsequently change their charge state [145]. Since the generated traps are stable, i.e. they are not annihilated by successive charging/discharging as would happen if they were holes or electrons, atom displacement and impurities (N, Ar, C, Cl, and, in particular, hydrogen) are likely involved in their formation [9, 145]. Trapping of the tunneling electrons in the newly generated trap sites is believed to be responsible for the non-saturating current decrease for $t > t_{TA}$.

Defect/trap generation and charge build-up processes continue throughout the whole wear-out phase and are accompanied by a detectable increase of the noise component of the tunneling current [146].

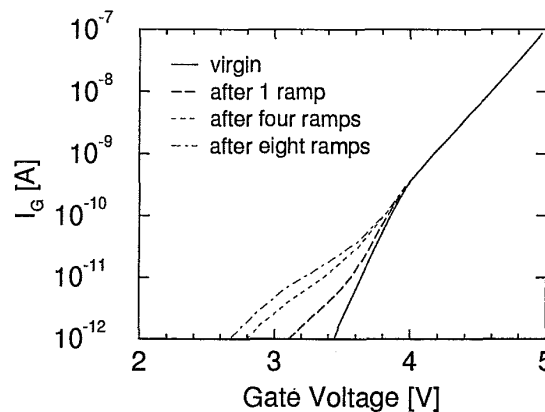


Figure 4.35 I-V characteristics of thin oxide capacitors showing low field leakage current increase after one, four and eight stress voltage ramps. Data from [147].

4.7.1.1 Stress Induced Leakage Currents (SILC). If the oxide thickness is reduced below $\approx 10\text{nm}$ (as in today's technologies) high-field stress manifests through a degradation of the insulating properties of the oxide well before the occurrence of breakdown. Namely, as illustrated Fig. 4.35, a large leakage current appears at low electric fields as a result of stress at high field [148, 149]. The latter can be induced by repeated measurements of the IV characteristic up to a field slightly lower than the breakdown field, as in Fig. 4.35, or by constant voltage (or current) stress. Stress-induced leakage currents (SILC) are responsible for serious data retention problems in non-volatile memories

(see Chapter 7) [150] and represent the main concern in the process of scaling the tunnel oxide thickness to lower values.

SILC occurs independently of the stress polarity and, therefore, does not depend on the injecting interface [147]. If SILC is induced by repeated stress ramps with a constant (moderately high) stop voltage, the leakage current increases after each ramp but this increase tends to saturate gradually (Fig. 4.35). A much larger increase of SILC occurs after successive stress ramps with progressively larger stop voltages until, eventually, the oxide breaks destructively [147]. If SILC is induced by a constant voltage stress, instead, the leakage current first grows and then saturates at a value which is an increasing function of the stress voltage, until breakdown occurs [147]. Regardless of the stress procedure, SILC becomes undetectable when the oxide thickness is raised to 10nm and above.

Several SILC models have been proposed in the literature. One model attributes SILC to tunneling through weak spots related to defects that reduce locally the tunneling barrier. In order to explain the time evolution of the observed degradation, the average spot area is assumed to increase as stress proceeds [147]. An alternative model claims that SILC is due to a reduction of the barrier height and a corresponding enhancement of the tunneling probability caused by positive charges generated inside the oxide during stress [148]. Trap assisted tunneling, qualitatively represented in Fig. 4.36, is now generally accepted as the dominant SILC mechanism [151, 152, 153, 154]. Recently, it has been pointed out that the tunneling process might be strongly inelastic [155].

The trap-assisted tunneling mechanism is supported by recent studies, evidencing that SILC includes a DC and a transient component [156, 157, 158]. The DC component dominates in very thin oxides (5nm and below), and it is explained by electron tunneling through the whole oxide assisted by bulk oxide traps generated during the stress phase. The transient component, instead, is the dominant one in relatively thick oxides ($\simeq 10\text{nm}$), and it is caused by tunnel charging/discharging of the traps located close to the silicon and gate electrode interfaces. These transients are qualitatively analogous to those occurring during high field stress/characterization cycles of thicker oxides [140, 141].

4.7.2 Oxide Breakdown

Breakdown is the sudden catastrophic event that ends the life of a stressed MOS structure. Sudden breakdowns are observed only in relatively thick oxides, in which a large electrostatic energy is stored because of large oxide voltage drops. In thin oxides soft-breakdowns manifest as a progressive degradation of MOS characteristics [159].

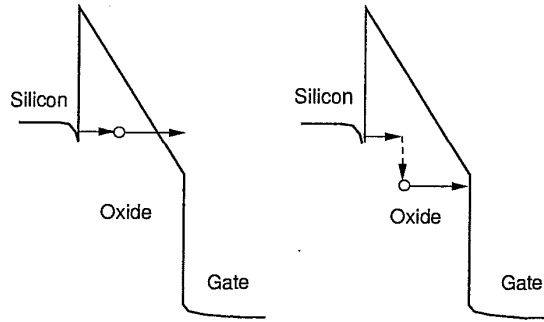


Figure 4.36 Possible SILC mechanisms in thin oxides. Left: elastic trap assisted tunneling; Right: inelastic trap assisted tunneling.

No generally accepted model exists of oxide breakdown. A widely cited theory argues that the positive charge generation mechanisms mentioned in the previous Section 4.7.1 result in the build-up of positive oxide charge that increases the cathode field and the injected current, thereby triggering a positive feedback. The charge build-up (hence, the current density) should be enhanced at localized weak spots or extrinsic defects associated to impurities, oxygen precipitates, oxide thinning or structural defects [138, 160]. According to this model, the breakdown event should happen when a critical hole fluence ($Q_{P,BD}$) is reached. Therefore:

$$Q_{P,BD} = \int_0^{t_{BD}} J_P(t) dt , \quad (4.55)$$

where $J_P(t)$ denotes the hole current density measured at the substrate contact and t_{BD} is the time to breakdown.

In agreement with this model, several experiments indicate that for a given oxide thickness $Q_{P,BD}$ is almost constant, regardless of oxide fields and injected current density (Fig. 4.37, [131, 161]). However, the hole induced breakdown model was proven inadequate to explain several experimental evidences in [162]. Furthermore, $Q_{P,BD}$ decreases significantly for thinner oxides [160], and exhibits a puzzling temperature dependence which is difficult to reconcile with the essential features of the model [161].

An alternative theory assumes a gradual degradation and weakening of the oxide structure caused by the continuous charge flow, until a critical defect/trap density is reached, that creates a conductive path between anode and cathode [163, 164, 165, 166, 167]. According to this model, breakdown should correlate to the total electron charge flown through the insulator [163]:

$$Q_{BD} = \int_0^{t_{BD}} J_G(t) dt , \quad (4.56)$$

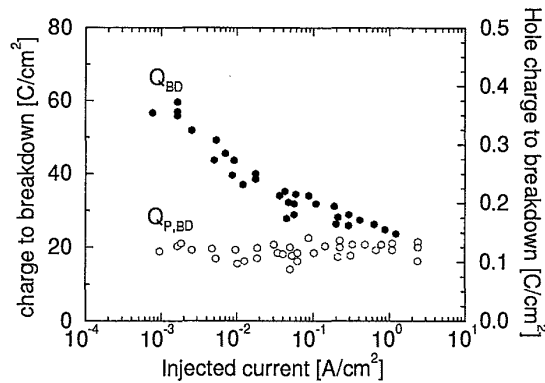


Figure 4.37 Charge to breakdown and hole charge to breakdown as a function of stress current density and stress field. $t_{\text{ox}} = 10.7\text{nm}$. Data from [131].

where J_G denotes the electron tunneling current.

Q_{BD} is larger for injection from the substrate (positive V_G) compared to injection from the gate (negative V_G), a fact that has been correlated to cathode smoothness [168]. Indeed, smoother cathodes produce a higher Q_{BD} and fewer breakdown events at low field, probably due to less asperities and to a more uniform distribution of the generated stress [145]. Q_{BD} decreases for increasing J_G above $\approx 10^{-3}\text{A/cm}^2$ (Fig. 4.37), and it also appears to increase for decreasing oxide thickness at constant J_G , indicating that thinner oxides may be more breakdown resistant than their thicker counterparts [160, 166].

Recent experimental results provided indications to reconcile these two models. First, it has been shown that a unique relationship exists between the generated oxide trap density and the hole fluence Q_P , independent of oxide field and thickness [167]. Therefore, the critical hole fluence required for thin oxide breakdown according to the first model also corresponds to a critical trap density as suggested by the second model. In addition, it was reported that electron injection in oxides containing trapped holes results in the generation of electron traps [9, 136, 144], suggesting that a two step process combining hole trapping and electron recombination with the trapped hole might be at the origin of trap generation and oxide wear-out.

4.7.3 Lifetime Evaluation Models

Oxide degradation at high electric fields is a continuous process that eventually terminates with a catastrophic breakdown event. However, due to wear-out and SILC, the insulating properties of the oxide can become unbearably poor well before breakdown. Therefore, the lifetime of a device can be limited by charge

trapping, SILC or breakdown, depending on its specific function within the circuit (see Chapter 7).

Although thin oxide SILC and pre-breakdown wear-out seem to share similar microscopic degradation processes, they manifest in very different ways. Present understanding of these microscopic mechanisms is still unsatisfactory, and no unified model is currently available of oxide wear-out and breakdown including SILC, also because historically they have been separately addressed. As a consequence, different and largely empirical models are presently used to estimate SILC- and breakdown-limited device lifetime.

4.7.3.1 SILC Lifetime Evaluation Model. An analytical expression for the prediction of SILC-related lifetime for Flash memory cells stressed by repeated write/erase cycles, was presented in [169]. Lifetime was defined as the time needed for a certain amount of charge loss, that is of threshold voltage shift, induced by the low field SILC currents flowing during read operations. The method relies on the observation that a one-to-one correlation exists between the DC component of SILC (J_{SS}) and the density of the oxide traps (D_{OT}) created by cell erase operations at high electric field. According to experimental results this correlation reads:

$$J_{SS} \propto D_{OT} J_{TUN}(\Phi_{eff}) , \quad (4.57)$$

where $J_{TUN}(\Phi_{eff})$ represents the Fowler-Nordheim tunneling current (Eq. (4.21)) through a reduced barrier whose effective height Φ_{eff} has been estimated to range between 0.9eV and 1.1eV depending on technology. The trap density is empirically related to the cumulative stress by:

$$D_{OT}(t) \propto J_{stress}^{\beta} Q_{inj}^{\alpha}(t) \quad (4.58)$$

where $\alpha \simeq 0.5$, $\beta \simeq 0.2$ (possibly dependent on oxide characteristics), and J_{stress} and Q_{inj} are the current during stress (i.e. during erase operation) and the total injected charge, respectively. Substituting (4.58) into (4.57) it is possible to estimate the built-up of traps during write/erase cycles, and the related increase of SILC. Integrating the SILC current in time and setting a threshold for the maximum acceptable floating gate charge loss, that is, the maximum acceptable threshold voltage shift, the cell lifetime can be estimated [169].

Fig. 4.38 shows a comparison between experimental and simulated disturb characteristics for two different Flash cells: a standard one (test cell) and an optimized one, that were preliminarily stressed with 10^3 and 10^4 write/erase cycles, respectively [169]. The time needed for obtaining a threshold voltage shift of 500mV due to SILC currents during read is reported for the two cells

as a function of $1/V_J$, where V_J is the erase junction voltage used as a stress acceleration factor.

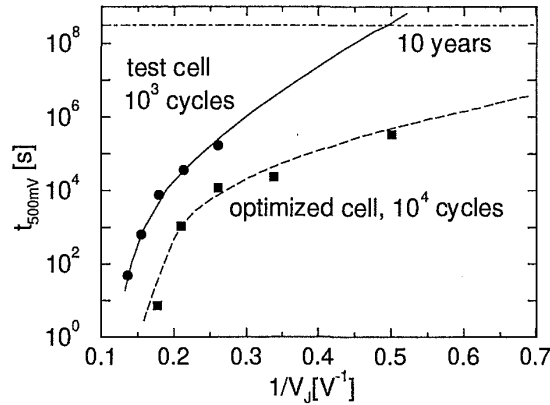


Figure 4.38 Experimental (symbols) and simulated (lines) SILC induced disturb characteristics for two flash cells. Data from [169].

4.7.3.2 Breakdown Lifetime Evaluation Model. The evaluation of breakdown lifetime is generally based on Q_{BD} or t_{BD} measurements. Q_{BD} or t_{BD} are measured at high electric fields and the lifetime at the operating F_{OX} is estimated by linear extrapolation on a linear-logarithmic scale (Fig. 4.39). Although extrapolation as a function of F_{OX} is still used [170, 171], plots as a function of $1/F_{OX}$ are preferred because they predict a less conservative operating field and, as shown in Fig. 4.39, they exhibit a more linear behavior than those versus F_{OX} . A semi-empirical justification of this procedure was given in the framework of the debated hole injection breakdown model, and will be summarized in the following [138, 160].

In essence, according to this model breakdown occurs when $Q_{P,BD}$ is reached. Assuming that both J_G and J_P stay essentially constant during the whole stress phase, we have

$$t_{BD} = \frac{Q_{BD}}{J_G} = \frac{Q_{P,BD}}{J_P} \quad (4.59)$$

On the other hand:

$$J_P = \alpha_P \theta_P J_G \quad (4.60)$$

where $\alpha_P \theta_P$ is the probability for a hole to be generated and to tunnel through the barrier (mechanism (a) in Fig. 4.34).

The gate tunnel current density, J_G , is roughly an exponential function of $1/F_{OX}$ (Eq. (4.21)). The same holds for the hole tunneling probability (θ_P)

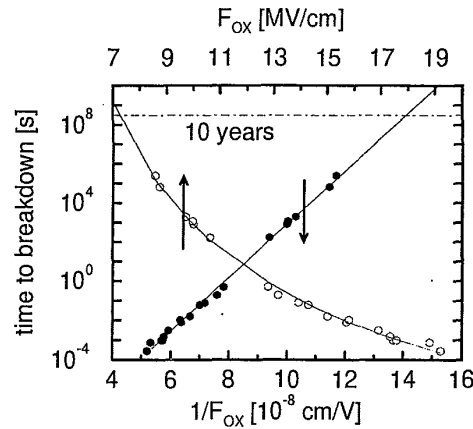


Figure 4.39 Time to breakdown data plotted as a function of stress oxide field and inverse stress oxide field. $t_{ox} = 7.9\text{nm}$. Data from [138].

that, if modeled as the electron one, is proportional to $\exp(-B\Phi'_{B,h+}/F_{OX})$, where B is a constant, and $\Phi'_{B,h+}$ is weakly dependent on the oxide field for both direct and Fowler-Nordheim tunneling (Eqs. (4.15) and (4.16)). Furthermore, $Q_{P,BD}$ is approximately constant at room temperature (Fig. 4.37), while $\alpha_P \approx 0.1$ independently of oxide thickness [160].

Therefore, both $Q_{BD} = Q_{P,BD}/\alpha_P\theta_P$ and the time to breakdown $t_{BD} = Q_{BD}/J_G$ turn out to be roughly exponential functions of $1/F_{OX}$. This dependence has been confirmed by several authors in the range of electric fields typically used for accelerated stress experiments. Accordingly, linear extrapolations of $\log(t_{BD})$ versus $1/F_{OX}$ are used to extract oxide lifetime.

Since nominally identical oxides are never perfectly the same, t_{BD} , Q_{BD} and $Q_{P,BD}$ are actually statistical variables described by a breakdown distribution, that is, by the percentage of devices that failed within given t_{BD} , Q_{BD} and $Q_{P,BD}$ values under specified stress conditions. This distribution is believed to be of the Weibull type.

A model capable to explain Q_{BD} distributions in thin oxides has been recently proposed [167, 172]. According to this model, breakdown occurs when a conductive path forms between adjacent traps, that is when their distance is sufficiently small to allow carrier hopping. Therefore, the statistical spread of Q_{BD} and t_{BD} inherently increases for small oxide thicknesses, because on average a smaller number of traps is required to generate a conductive path in thin oxides than in thick ones. In addition, since shorter paths have a higher probability of being well oriented to cause breakdown, the critical electron trap

density to cause breakdown (hence the critical $Q_{P,BD}$) decreases for thin oxides, a fact that has been experimentally confirmed in [167].

More sophisticated models have also been proposed based on the idea that breakdown is triggered when a critical trap density is reached, but they are not suitable for lifetime extrapolation purposes [164, 173].

4.8 OXIDE AND INTERFACE DEGRADATION DUE TO HOT CARRIER INJECTION

As discussed in Section 4.6.1, upon application of a high drain voltage a population of hot carriers is generated in proximity of the drain junction of MOSFETs and Flash cells, and carrier injection into the oxide takes place. These hot carriers (electrons and holes) progressively degrade the interface and oxide quality, reducing the performance of the device and undermining its reliability through different microscopic mechanisms. The highly non-uniform field profile typically encountered at the drain end of saturated MOSFETs is obviously unfavorable to study the details of the degradation phenomena because of the lack of control on carrier fluence, carrier heating and oxide electric field. Separating the effects of hot electrons from those of hot holes is also difficult because they are simultaneously present (Section 4.6.6).

4.8.1 Homogeneous Hot Carrier Degradation

Due to the difficulties mentioned above, in depth studies of the basic physical mechanisms of hot carrier degradation are more easily carried out by means of the substrate hot carrier injection experiments described in Section 4.6.7, since they provide: 1) injection of only one carrier type; 2) one-dimensional electric field profiles; 3) decoupling of the oxide and silicon fields; 4) easy and accurate calculation of the fields; 5) tight control on the carrier fluence.

4.8.1.1 n-channel Devices. Electron injection experiments in n-MOSFETs revealed essentially three types of damage: 1) the generation of acceptor type interface traps; 2) electron capture in existing bulk oxide traps; 3) the generation of new bulk oxide traps.

Both electron trapping and interface trap generation are relatively low probability process, with approximately one trapped electron and one new interface state every $\approx 10^5$ – 10^7 injected electrons [174, 175, 176]. All these degradation modes are strongly activated by the oxide field, with thresholds of approximately 1.5MV/cm for interface trap generation [174], and between 1.5MV/cm [175] and 4MV/cm [174] for bulk trap generation. To observe the oxide field activation of the damage, care must be paid in the characterization phase.

Since trap occupancy decreases at high fields [141], inconsistent results can be obtained if the bias dependence of the trap occupation probability is not accounted for in the evaluation of damage from the apparent shift of the electrical device parameters [176].

As for the dependence of carrier trapping on source-substrate bias (that is, on the energy of carriers impinging on the interface), controversial results have been reported in the literature [175, 176, 177]. Recent studies indicate that below 2×10^{19} injected charges per square centimeter, the observed V_{SB} dependence of degradation should be attributed to changes in the injected current rather than to the carrier energy distribution at the interface [176].

4.8.1.2 p-channel Devices. Homogeneous injection of holes in p-MOSFETs has also been studied [68, 178]. The associated damage has been characterized as a function of the silicon and oxide electric fields [42, 117, 179]. Differently from the case of electron injection, the hole-induced damage is weakly activated by the oxide field, and it is practically independent of source-substrate voltage (that is hole energy) and current density [117]. The effective number of trapped holes (measured by the threshold voltage shift) is always a large fraction of the number of injected ones (10 to 20%) so that hole injection must be regarded as particularly harmful for the oxide quality. Upon hole injection, new interface traps are also generated with an efficiency of approximately 1 interface trap per 40 trapped holes [42]. Interface trap generation is observed also for long times after hole injection, with a fraction of the trapped holes being detrapped or transformed into interface traps [117].

4.8.2 Non-homogeneous Hot Carrier Degradation

In the non-homogeneous conditions typically encountered in MOSFETs, electron and hole injections always take place simultaneously, although at slightly different locations along the interface and with different relative intensities. In general, it is difficult to separate the effects of trapped charges from those of interface traps and those of electrons from those of holes. These experiments, however, reproduce the actual operating conditions and, differently from substrate injection techniques, are suited to evidence the combined effects of both carrier types.

Fig. 4.40 reports typical i-V curves of a MOSFET before and after hot carrier stress at $V_{GS} \simeq V_{DS}/2$. The damaged region affects the device characteristics causing a long term drift of all the important electrical parameters (transconductance g_m , current I_D , subthreshold slope S , threshold voltage V_T , multiplication factor $M = I_B/I_D$, interface states density D_{it}). Most of these quantities, however, are sensitive to both trapped charges and interface trap

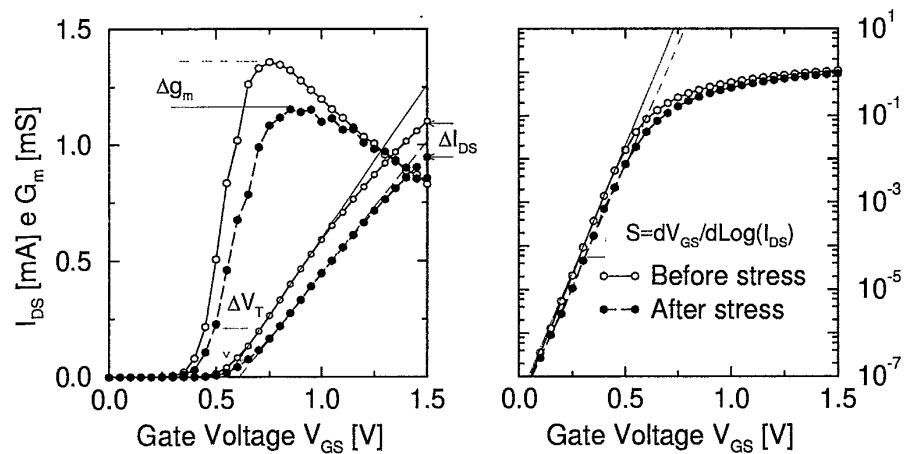


Figure 4.40 Degradation of linear MOSFET characteristics (in linear and logarithmic scales) after hot-carrier stress at $V_{GS} = V_{DS}/2$.

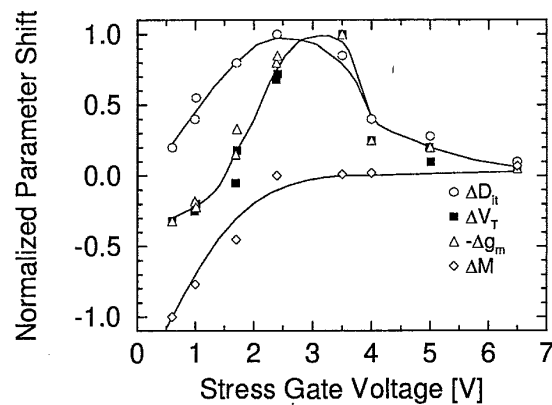


Figure 4.41 Interface state density (measured with the charge pumping technique), threshold voltage, transconductance and multiplication factor shift during stress experiments at constant drain voltage as a function of the stress gate voltage. Data from [180, 181].

generation and it is difficult to separate their effects, unless suitable characterization techniques are employed, such as charge pumping experiments [182]. In addition, due to the two-dimensional nature of the electric field distribution at the drain junction simple one dimensional models can mislead the interpretation of degradation results even at low V_{DS} [183].

For a fixed stress time and drain voltage, degradation of the electrical parameters depends largely on the gate voltage during stress. As shown in Fig. 4.41 three main regimes can be distinguished.

- Stress at low V_{GS} leads to negative threshold voltage shifts and positive transconductance shifts ($\Delta V_T < 0$, $\Delta g_m > 0$). In these conditions, significant hole injection takes place (Section 4.6.7). As discussed in the previous section, holes are trapped with high efficiency. The positive charge thus formed raises the electron concentration in the channel section below the holes trapped at the drain side. The drain is thus effectively extended towards the source, leading to a reduction of the effective channel length, hence to a higher g_m .
- For stress at a bias $V_{GS} \approx V_{DS}/2$ we have $\Delta V_T > 0$, $\Delta g_m < 0$, and degradation reaches its maximum. Hence, this is the worst case stress condition for n-MOSFETs. Both ΔV_T and Δg_m are well correlated to the increase of the charge pumping current (proportional to D_{it}), indicating that generation of acceptor type interface traps is the dominant degradation mechanism. In these conditions, the large number of generated traps is progressively charged as V_{GS} increases, thus reducing the subthreshold swing (Fig. 4.40). In addition, the negative charge, accumulated in the occupied acceptor traps or trapped in the oxide, increases locally the threshold voltage, degrades the mobility and increases the series resistance, thus lowering the extrinsic g_m .
- At $V_{GS} \approx V_{DS}$ electron injection dominates. Electron trapping and interface trap generation coexist but are quantitatively smaller than for $V_{GS} \approx V_{DS}/2$.

In essence, two mechanisms have been proposed to explain the worse case degradation phenomenology in n-MOSFETs. According to the first of them, hot electrons break weak silicon-hydrogen bonds in proximity of the interface, generating trivalent silicon atoms that act as electron interface traps [184, 185]. Since the Si-H bond energy is $\Phi_{Si-H} \approx 0.3-0.5\text{eV}$, an interface trap generation threshold energy $\Phi_{ITG} = \Phi_{B,e^-} + \Phi_{Si-H} \approx 3.5-4\text{eV}$ is expected to exist for device degradation. This value is consistent with the results of a few recent model studies [186, 187].

The second degradation model assumes a two-step process. In the first step, a strained silicon-oxygen bond captures a hole and, due to the positive coulomb charge, it is converted into an efficient electron trap. In a second phase, the trap site captures an electron. The energy released in the electron-hole recombination process produces a permanent damage that transforms the site in an acceptor type trap [9, 99].

4.8.3 Lifetime Evaluation Models

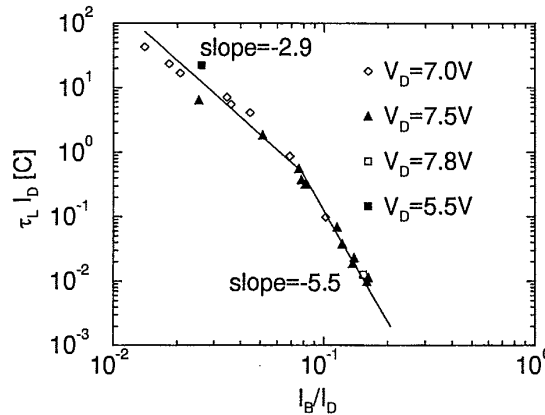


Figure 4.42 Lifetime plot versus I_B/I_D . Lifetime is defined by D_{it} shifts. Data from [181].

Lifetime evaluation models are based on empirical equations that describe the time dependence of degradation during accelerated stress experiments. In the initial phase, hot carrier induced degradation (as monitored by ΔV_T , ΔI_D , Δg_m or ΔD_{it}) can be described by:

$$\Delta D(t) = (D_0 t)^n \quad (4.61)$$

Here, $\Delta D(t)$ represents the absolute parameter shift during aging and D_0 is a function of the stress conditions related to the specific degradation monitor taken into consideration and to the rate of damage generation. The constant n has been empirically related to stress conditions. In particular, $n \simeq 0.5 \div 0.7$ for interface trap generation, and for threshold voltage shift at $V_{GS} \approx V_{DS}/2$, while $n \simeq 0.2 \div 0.3$ for ΔV_T at $V_{GS} \approx V_{DS}$ and $V_{GS} \approx V_T$ [180]. For long stress times ΔD tends to saturate, possibly because of the feedback of the trapped charge on the field at the injection point [188]. However, when degradation saturates ΔD has often exceeded the maximum acceptable limit, so that in most cases (4.61) is sufficient for lifetime evaluation purposes.

To derive an equation for lifetime extraction, we start assuming that both I_B and I_G are proportional to the number of carriers above an effective threshold energy (Φ_{Ieff} and Φ_{Beff} , respectively). Based on the effective temperature model (4.42) and a few approximations, the following relationships can be derived for gate and substrate currents [185]:

$$\frac{I_B}{I_S} = \exp(-\Phi_{Ieff}/k_B T_e) \quad (4.62)$$

$$\frac{I_G}{I_S} = C(F_{OX}) \exp(-\Phi_{Beff}/k_B T_e) , \quad (4.63)$$

where I_S is the channel current and $C(F_{OX})$ is a function of the oxide field that accounts for back scattering in the image force potential well. Thus,

$$\frac{I_G}{I_S} = C(F_{OX}) \left(\frac{I_B}{I_S} \right)^{\left(\frac{\Phi_{Beff}}{\Phi_{Ieff}} \right)} . \quad (4.64)$$

Eq. (4.64) predicts a linear relationship between I_B/I_S and I_G/I_S in a double logarithmic plot. This correlation between I_B and I_G , whose existence was first suggested in [189], is well verified over a broad range of technologies and bias conditions [125, 190].

Assuming that interface traps are generated only by those carriers featuring an energy in the silicon larger than Φ_{ITG} , and that interface trap generation is the only active degradation mechanism (a reasonable approximation for the worse case condition $V_{GS} \simeq V_{DS}/2$), we can exploit again (4.42) and (4.61) to express D_0 and ΔD , respectively as:

$$D_0 \propto I_S \exp(-\Phi_{ITG}/k_B T_e) , \quad (4.65)$$

$$\Delta D = H \left(I_S \exp \left(-\frac{\Phi_{ITG}}{k_B T_e} \right) t \right)^n , \quad (4.66)$$

where H is an empirical parameter.

Therefore, eliminating T_e from (4.66) through (4.62), the device lifetime (that is the time, τ_L , required to reach the maximum admissible degradation ΔD^*) is given by:

$$\tau_L = \frac{C_\tau}{I_S} \left(\frac{I_B}{I_S} \right)^{-\frac{\Phi_{ITG}}{\Phi_{Ieff}}} . \quad (4.67)$$

Eq. (4.67) yields a straight line when $\tau_L I_S$ is plotted as a function of I_B/I_S in a log-log graph. Following this line, the results of accelerated stress experiments at large I_B/I_S can be extrapolated to the real operating conditions. Notice that in most cases plots are presented in terms of $\tau_L I_D$ versus I_B/I_D , thus implicitly neglecting parasitic bipolar action and assuming a low avalanche multiplication condition (that is, $I_D = I_S$).

A typical lifetime plot is reported in Fig. 4.42, where τ_L is defined monitoring the increase of interface trap density, ΔD_{it} . A slope $-(\Phi_{ITG}/\Phi_{Ieff}) \approx -2.9$ is obtained for stress at $V_{GS} \simeq V_{DS}/2$ and above, which yields $\Phi_{ITG} \simeq 3.8\text{eV}$ if $\Phi_{Ieff} \simeq 1.3\text{eV}$ is assumed. This value, in reasonable agreement with the expectations of the Si-H bond breaking degradation model [185], is often attributed to hot-electron generated traps. A slope of approximately -5.5 , instead, is

observed for stress at low V_{GS} and high I_B/I_D . This latter value has been related to the generation of fast interface traps by hole injection [180], although different interpretations have also been proposed [191]. Parasitic bipolar action can also increase the slope of the lifetime plot at high I_B/I_D [192].

A less accurate but simpler extrapolation procedure can be derived noticing that in quasi equilibrium conditions T_e is a linear function of the electric field (Eq. (4.45)) and that the maximum field in the device (the one having the largest influence on carrier heating and degradation according to a local point of view of carrier transport) can be expressed as $F_{\max} = \theta(V_{DS} - V_{DS,SAT})$ where $V_{DS,SAT}$ is the MOSFET saturation voltage [193]. Therefore, from (4.61) a straight line should be obtained when plotting $\tau_L I_S$ as a function of $1/(V_{DS} - V_{DS,SAT})$ in a semi-log graph. This latter extrapolation method is easier to apply because it does not require to measure the substrate current, but relies on the extraction of $V_{DS,SAT}$, a non trivial exercise especially in the case of short MOSFETs [194]. Sometimes, an even more approximate version of the method is applied, and lifetimes are simply plotted as a function of $1/V_{DS}$ [195].

Lifetime prediction at stress bias conditions different from $V_{GS}/V_{DS}/2$ has been addressed in [196], where a practical empirical equation is proposed and verified over a broad range of experimental conditions.

Acknowledgments

We would like to thank Roberto Bez, Antonio Abramo and David Esseni for carefully proofreading the manuscript.

References

- [1] Datta S. (1989) *Quantum Phenomena*. Addison Wesley, New York.
- [2] Chelikowsky J.R. and Schluter M. (1977) "Electron States in α -quartz: A self-consistent pseudopotential calculation". *Physical Review B*, **15**, 8, p. 4020.
- [3] Venturi F. and Ghetti A. (1997) "Assessment of accuracy limitations of full band Monte Carlo device simulation". *Proc. SISPAD*, p. 343.
- [4] Chelikowsky J.R. and Cohen M.L. (1976) "Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blend semiconductors". *Physical Review B*, **14**, p. 556.
- [5] Powell R.J. and Morad M. (1978) "Optical absorption and photoconductivity in thermally grown SiO_2 Films". *Journal of Applied Physics*, **49**, 4, p. 2499.

- [6] Tamm I. and Blochinzev D. (1933) "Über die austrittsarbeit der elektronen aus metallen". *Phys. Z. der Sowjetunion*, **1**, p. 733.
- [7] Shockley W. (1939) "On the surface states associated with a periodic potential". *Physical Review*, **56**, p. 317.
- [8] Mc Whorter P.J. and Winokur P.S. (1986) "Simple technique for separating the effects of interface traps and trapped oxide charge in metal-oxide-semiconductor transistors". *Journal of Applied Physics*, **48**, 2, p. 133.
- [9] Lai S.K. (1983) "Interface trap generation in silicon dioxide when electrons are captured by trapped holes". *Journal of Applied Physics*, **54**, p. 2540.
- [10] Sah C.T. (1990) "Models and experiments on degradation of oxidized silicon". *Solid State Electronics*, **33**, p. 147.
- [11] Grunthaner F.J., Lewis B.F., Zamini N., Maserjan J. and Madhukar A. (1980) "XPS studies of structure-induced radiation effects at the SiO₂ interface". *IEEE Trans. on Nuclear Science*, **27**, p. 1640.
- [12] Fromhold A.T. (1981) *Quantum Mechanics for Applied Physics and Engineering*. Dover Publications Inc., New York.
- [13] Fowler R.H. and Nordheim L. (1928) "Electron emission in intense electric fields". *Proc. Royal Society London Series A*, **119**, p. 173.
- [14] Yoshikawa K., Mori S., Sakagami E., Arai N., Kaneko Y. and Ohshima Y. (1990) "A Flash EEPROM cell scaling including tunnel oxide limitations". *Proc. European Solid State Device Res. Conf.*, p. P/2.
- [15] Weinberg Z.A. (1982) "On tunneling in metal-oxide-silicon structures". *Journal of Applied Physics*, **53**, p. 5052.
- [16] Franz W. (1956) "Dielektrischer durchschlag". *Handbuch der Physik*, S. Flugge (Ed.), Springer-Verlag, Berlin, vol. XVII, p. 155.
- [17] Hartstein A. and Weinberg Z.A. (1979) "Unified theory of internal photoemission and photon-assisted tunneling". *Physical Review B*, **20**, 4, p. 1335.
- [18] Kleefstra M. and Herman G.C. (1980) "Influence of the image force on the band gap in semiconductors and insulators". *Journal of Applied Physics*, **51**, p. 4923.

- [19] Puri A. and Schaich W.L. (1983) "Comparison of image force potential theories". *Physical Review B*, **28**, p. 1781.
- [20] Ning T.H., Osburn C.M. and Yu H.N. (1977) "Emission probability of hot electrons from silicon into silicon dioxide". *Journal of Applied Physics*, **48**, p. 286.
- [21] Weinberg Z.A. (1977) "Tunneling of electrons from Si into thermally grown SiO₂". *Solid State Electronics*, **20**, p. 11.
- [22] Suñe J., Olivo P. and Riccò B. (1992) "Quantum mechanical modeling of accumulation layers in MOS structures". *IEEE Trans. on Electron Devices*, **7**, p. 1732.
- [23] Lui W. and Fukuma M. (1986) "Exact solution of the Schrödinger Equation across an arbitrary one-dimensional piecewise-linear potential barrier". *Journal of Applied Physics*, **60**, 5, p. 1555.
- [24] De Castro E. and Olivo P. (1985) "Quantum effects in accumulation layers of Si-SiO₂ interfaces in the WKB effective mass approximation". *Phys. Status Solidi (b)*, **132**, p. 153.
- [25] Suñe J., Olivo P. and Riccò B. (1991) "Self-consistent solution of Poisson and Schrödinger Equations in accumulated semiconductor-insulator interfaces". *Journal of Applied Physics*, **70**, p. 337.
- [26] Schenk A. (1996) "Modeling tunneling through ultra-thin gate oxides". *Proc. SISPAD*, p. 7.
- [27] Lenzlinger M. and Snow E.H. (1969) "Fowler-Nordheim tunneling into thermally grown SiO₂". *Journal of Applied Physics*, **40**, p. 278.
- [28] Powell R.J. (1970) "Interface barrier energy determination from voltage dependence of photoinjected currents". *Journal of Applied Physics*, **41**, 6, p. 2424.
- [29] Olivo P. Suñe J. and Riccò B. (1991) "On the determination of the Si-SiO₂ barrier height from Fowler-Nordheim plot". *IEEE Electron Device Letters*, **12**, p. 620.
- [30] Zener C. and Wills H.H. (1934) "A theory of the electrical breakdown of solid dielectrics". *Proc. Royal Society*, **A145**, p. 523.
- [31] Esaki L. (1958) "New phenomenon in narrow germanium p-n junctions". *Physical Review*, **109**, p. 603.

- [32] Hu C. (1933) "Future CMOS scaling and reliability". *Proc. of the IEEE*, **81**, p. 682.
- [33] Chan C. and Lien J. (1987) "Corner-field induced drain leakage in thin oxide MOSFETs". *IEDM Technical Digest*, p. 714.
- [34] Chan T.Y., Chen J., Ko P.K. and Hu C. (1987) "The impact of gate-induced leakage current on MOSFET scaling". *IEDM Technical Digest*, p. 718.
- [35] Acovic A., Dutoit M. and Ilegems M. (1990) "Characterization of hot-electron-stressed MOSFETs by low-temperature measurements of the drain tunnel current". *IEEE Trans. on Electron Devices*, **37**, 6, p. 1467.
- [36] Okhonin S., Hessler T. and Dutoit M. (1996) "Comparison of gate-induced drain leakage and charge pumping measurements for determining lateral interface trap profiles in electrically stressed MOSFETs". *IEEE Trans. on Electron Devices*, **43**, 4, p. 605.
- [37] Shirota R., Endoh T., Momodomi M., Nakayama R., Inoue S., Kirisawa R. and Masuoka F. (1988) "An accurate model of sub-breakdown due to band-to-band tunneling and its application". *IEDM Technical Digest*, p. 26.
- [38] Nedev I., Asenov A. and Stefanov E. (1991) "Experimental study and modeling of band-to-band tunneling leakage current in thin-oxide MOSFETs". *Solid State Electronics*, **34**, 12, p. 1401.
- [39] Orlowski M., Sun S.W., Blakey P. and Subrahmanyam R. (1990) "The combined effects of band-to-band tunneling and impact ionization in the off regime and LDD MOSFET". *IEEE Electron Device Letters*, **11**, 12, p. 593.
- [40] Chen I.C., Coleman D.J. and Teng C.W. (1989) "Gate current injection initiated by electron band to band tunneling in MOS devices". *IEEE Electron Device Letters*, **10**, p. 297.
- [41] Chen J., Chan T.Y., Ko P.K. and Hu C. (1989) "Gate current in OFF-state MOSFET". *IEEE Electron Device Letters*, **10**, 5, p. 203.
- [42] Van Den Bosch G., Groeseneken G., Heremans H., Heyns M. and Maes H. (1992) "Hole trapping and hot hole induced interface state trap generation in MOSFETs at different temperatures". *Proc. European Solid State Device Res. Conf.*, p. 477.

- [43] Igura Y., Matsuoka H. and Takeda E. (1989) "New device degradation due to *cold* carriers created by band-to-band tunneling". *IEEE Electron Device Letters*, **10**, 5, p. 227.
- [44] Haddad S., Chang C., Swaminathan B. and Lien J. (1989) "Degradations due to hole trapping in Flash memory cells". *IEEE Electron Device Letters*, **10**, 3, p. 117.
- [45] Yoshikawa K., Mori S., Sakagami E., Ohshima Y., Kaneko Y. and Arai N. (1990) "Lucky-hole injection induced by band-to-band tunneling leakage in stacked gate transistors". *IEDM Technical Digest*, p. 577.
- [46] Haddad S., Chang C., Wang A., Bustillo J., Lien J., Montalvo T. and Van Buskirk M. (1990) "An investigation of erase mode dependent hole trapping in Flash EEPROM memory cell". *IEEE Electron Device Letters*, **11**, 11, p. 514.
- [47] Schenk A. (1993) "Rigorous theory and simplified model of the band-to-band tunneling in silicon". *Solid State Electronics*, **36**, 1, p. 19.
- [48] Habas P. (1993) *Analysis of physical effects in small silicon devices*. PhD thesis, Technische Universität Wien, Wien, Austria.
- [49] Kane E.O. (1959) "Zener tunneling in semiconductors". *J. Phys. Chem. Solids*, **12**, p. 181.
- [50] Moll J.L. (1964) *Physics of Semiconductors*. McGraw Hill.
- [51] Kane E.O. (1961) "Theory of Tunneling". *Journal of Applied Physics*, **32**, 1, p. 83.
- [52] Wen D.S., Goodwin-Jonanson S.H. and Osburn C.M. (1935) "Tunneling leakage in germanium pre-amorphized shallow junctions". *IEEE Trans. on Electron Devices*, **35**, 7, p. 1107.
- [53] Habas P., Lugbauer A. and Selberherr S. (1992) "Two-dimensional numerical modeling of interband tunneling accounting for nonuniform electric field". *Proc. NUPAD Conf.*, p. 135.
- [54] Baccarani G. (1982) "Physics of submicron devices". *Large Scale Integrated Circuit Technology*, L. Esaki and G. Soncini (Eds.), Nijhoff, The Hague, p. 647.
- [55] Lundstrom M.S. (1990) *Fundamentals of Carrier Transport*. Addison Wesley, New York.

- [56] Kunikiyo T., Takenaka M., Morifuji M., Taniguchi K. and Hamaguchi C. (1996) "A model for impact ionization due to the primary hole in silicon for a full band Monte Carlo simulation". *Journal of Applied Physics*, **79**, 10, p. 7718.
- [57] Fischetti M.V., Laux S. and Crabbè E. (1995) "Understanding hot electron transport in silicon devices: is there a shortcut?". *Journal of Applied Physics*, **78**, 2, p. 1058.
- [58] Hasnat K., Yeap C.F., Jallepalli S., Shih W.K., Hareland S.A., Agostinelli V.M., Tasch A.F. and Maziar C.M. (1996) "A pseudo lucky electron model for simulation of electron gate current in submicron nMOSFETs". *IEEE Trans. on Electron Devices*, **43**, 8, p. 1264.
- [59] Kuhn T., Reggiani L. and Varani L. (1992) "Coupled Langevin Equations analysis of hot carrier transport in semiconductors". *Physical Review B*, **45**, 4, p. 1903.
- [60] Jallepalli S., Rashed M., Shih W.K., Maziar C.M. and Jr. A. (1997) "A full-band Monte Carlo model for hole transport in silicon". *Journal of Applied Physics*, **81**, 5, p. 2250.
- [61] Ventura D., Gnudi A. and Baccarani G. (1995) "A deterministic approach to the solution of the BTE in semiconductors". *La Rivista del Nuovo Cimento*, **18**, 6, p. 1.
- [62] Kometer K., Zandler G. and Vogl P. (1992) "Lattice-gas cellular-automaton method for semiclassical transport in semiconductors". *Physical Review B*, **46**, p. 1382.
- [63] Alam M.A., Stettler M.A. and Lundstrom M.S. (1993) "Formulation of the Boltzmann Equation in terms of scattering matrices". *Solid State Electronics*, **36**, 2, p. 263.
- [64] Jacoboni C. and Reggiani L. (1983) "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials". *Reviews of Modern Physics*, **55**, p. 645.
- [65] Baccarani G. and Wordeman M.R. (1985) "An investigation of steady-state velocity overshoot in silicon". *Solid State Electronics*, **28**, p. 407.
- [66] Higman J.M., Hess V., Hwang C.G. and Dutton R.W. (1989) "Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFET's". *IEEE Trans. on Electron Devices*, **36**, p. 930.

- [67] Crowell C.R. and Sze S.M. (1966) "Temperature dependence of avalanche multiplication in semiconductors". *Journal of Applied Physics*, **9**, p. 242.
- [68] Selmi L., Sangiorgi E., Bez R. and Riccò B. (1993) "Measurement of the hot hole injection probability from Si into SiO₂ in p-MOSFET's". *IEDM Technical Digest*, p. 333.
- [69] Fiegna C., Iwai I., Wada T., Saito M., Sangiorgi E. and Riccò B. (1994) "Scaling the MOS transistor below 0.1 μ m: methodology, device structures and technology requirements". *IEEE Trans. on Electron Devices*, p. 941.
- [70] Bløtekjær K. (1970) "Transport equations for electrons in two valley semiconductors". *IEEE Trans. on Electron Devices*, p. 38.
- [71] Cook R.K. and Frey J. (1982) "Two-dimensional numerical simulation of energy transport in Si and GaAs MESFETs". *IEEE Trans. on Electron Devices*, **29**, p. 970.
- [72] Goldsman N. and Frey J. (1988) "Efficient and accurate use of the energy transport method in device simulation". *IEEE Trans. on Electron Devices*, **35**, p. 1524.
- [73] Crabbè E.F., Stork J.M.C., Baccarani G., Fischetti M.V. and Laux S.E. (1990) "The impact of non-equilibrium transport on breakdown and transit time in bipolar transistors". *IEDM Technical Digest*, p. 463.
- [74] Slotboom J.W., Streutker G., Woerlee M.P.H., Pruijmboom A. and Gravesteijn D.J. (1991) "Non-local impact ionization in silicon devices". *IEDM Technical Digest*, p. 127.
- [75] Zanoni E., Crabbè E.F., Stork J.M.C., Pavan P., Verzellesi G., Vendrame L. and Canali C. (1992) "Measurements and simulation of avalanche breakdown in advanced Si bipolar transistors". *IEDM Technical Digest*, p. 927.
- [76] Shockley W. (1961) "Problems related to p-n junctions in silicon". *Solid State Electronics*, **2**, 1, p. 35.
- [77] Hu C. (1979) "Lucky-electron modeling of channel hot electron emission". *IEDM Technical Digest*, p. 22.
- [78] Tam S., Hsu F.C., Hu C., Muller R.S. and Ko P.K. (1983) "Hot-electron currents in very short channel MOSFETs". *IEEE Electron Device Letters*, **4**, 7, p. 249.

- [79] Fukuma M. and Iizuka T. (1988) "Advanced physical models for MOS-FETs". *Proc. SISDEP*, p. 57.
- [80] Fiegna C., Venturi F., Melanotte M., Sangiorgi E. and Ricc6 B. (1991) "Simple and efficient modeling of EPROM writing". *IEEE Trans. on Electron Devices*, **38**, p. 603.
- [81] Hasnat K., Yeap C.F., Jallepalli S., Hareland S.A., Shih W.K., Agostinelli V.M., Jr. A. and Maziar C.M. (1997) "Thermionic emission model of electron gate current in submicron nMOSFETs". *IEEE Trans. on Electron Devices*, **44**, 1, p. 129.
- [82] Urai T., Frey J., Peng Z.Z. and Goldsman N. (1990) "Simulation of EPROM programming characteristics". *Electronics Letters*, **26**, p. 716.
- [83] Peng Z.Z., Lin Q., Fang P., Kwan M., Longcor S. and Lien J. (1994) "Accurate simulation of EPROM hot carrier induced degradation using physics based interface and oxide charge generation models". *Proc. Int. Reliability Physics Symp.*, p. 154.
- [84] Concannon A., Piccinini F., Mathewson A. and Lombardi C. (1995) "The numerical simulation of substrate and gate currents in MOS and EPROMs". *IEDM Technical Digest*, p. 289.
- [85] Sano N. and Yoshii A. (1994) "Impact ionization rates near thresholds in Si". *Journal of Applied Physics*, **75**, p. 5102.
- [86] Takagi S. and Toriumi A. (1992) "New experimental findings on hot carrier transport under velocity saturation regime in Si MOSFETs". *IEDM Technical Digest*, p. 711.
- [87] Laux S.E., Fischetti M.V. and Frank D.J. (1990) "Monte Carlo analysis of semiconductor devices: the DAMOCLES program". *IBM Journal of Research and Development*, **34**, p. 466.
- [88] Van Overstraeten R. and De Man H. (1970) "Measurements of the ionization rates in diffused silicon p-n junctions". *Solid State Electronics*, **13**, p. 583.
- [89] Slotboom J.W., Streuker G., Davis G.J.T. and Hartog P.B. "Surface impact ionization in silicon devices". *IEDM Technical Digest*, p. 494.
- [90] Chynoweth A.G. (1958) "Ionization rates for electrons and holes in silicon". *Physical Review*, **109**, p. 1537.

- [91] Selberherr S. (1984) *Analysis and Simulation of Semiconductor Devices*. Springer-Verlag, Wien/New York.
- [92] Jungemann C., Yamaguchi S. and Goto H. (1996) "Is there experimental evidence for a difference between surface and bulk impact ionization in silicon?". *IEDM Technical Digest*, p. 383.
- [93] Sakui K., Wong S.S. and Wooley B.A. (1994) "The effects of impact ionization on the operation of neighboring devices and circuits". *IEEE Trans. on Electron Devices*, **41**, 9, p. 1603.
- [94] Esseni D., Selmi L., Bez R., Sangiorgi E. and Riccò B. (1994) "Bias and temperature dependence of gate and substrate currents in n-MOSFETs at low drain voltage". *IEDM Technical Digest*, p. 307.
- [95] Ghetti A., Selmi L., Bez R. and Sangiorgi E. (1996) "Monte Carlo simulation of low voltage hot-carrier effects in non-volatile memory cells". *IEDM Technical Digest*, p. 379.
- [96] Eitan B., Frohman-Bentchkowsky D. and Shappir J. (1982) "Impact ionization at very low voltages in silicon". *Journal of Applied Physics*, **65**, p. 1244.
- [97] Fischetti M.V. and Laux S. (1995) "Monte Carlo study of sub-bandgap impact ionization in small silicon field effect transistors". *IEDM Technical Digest*, p. 305.
- [98] Momose H.S., Ono M., Yoshitomi T., Ohguro T., Nakamura S., Saito M. and Iwai H. (1994) "Tunneling gate oxide approach to ultra high current drive in small geometry MOSFETs". *IEDM Technical Digest*, p. 593.
- [99] Hofmann K.R., Werner C., Weber W. and Dorda G. (1985) "Hot-electron and hole-emission effects in short n-channel MOSFET's". *IEEE Trans. on Electron Devices*, **32**, p. 691.
- [100] Cappelletti P., Bez R., Cantarelli D. and Fratin L. (1994) "Failure mechanisms of FLASH cell in program/erase cycling". *IEDM Technical Digest*, p. 293.
- [101] Selmi L., Ghetti A., Bez R. and Sangiorgi E. (1997) "Trade-offs between tunneling and hot-carrier injection in short channel floating gate MOS-FET". *Microelectronic Engineering*, **36**, 1-4, p. 293.
- [102] Takeda E. (1984) "Hot carrier effects in submicrometer MOS VLSIs". *Proc. IEE*, **131**, p. 153.

- [103] Bulucea C. (1974) "Avalanche injection into the oxide in silicon gate controlled devices - I: Theory". *Solid State Electronics*, **18**, p. 363.
- [104] Bulucea C. (1974) "Avalanche injection into the oxide in silicon gate controlled devices - II: Experimental results". *Solid State Electronics*, **18**, p. 363.
- [105] Esseni D. and Selmi L. (1999) "A better understanding of substrate enhanced gate current in MOSFETs and Flash cells". *IEEE Trans. on Electron Devices*, n. 2.
- [106] Venturi F., Fiegna C., Abramo A., Sangiorgi E. and Riccò B. (1990) "Hot-holes generation and transport in n-MOSFETs: a Monte-Carlo investigation". *IEDM Technical Digest*, p. 455.
- [107] Childs P.A., Eccleston W. and Stuart R.A. (1981) "Alternative mechanism for substrate minority carrier injection in MOS devices operating in low level avalanche". *Electronics Letters*, **17**, p. 281.
- [108] Tam S., Hsu C., Ko P.K., Hu C. and Muller R.S. (1982) "Hot-electron induced excess carriers in MOSFET's". *IEEE Electron Device Letters*, **3**, p. 376.
- [109] Childs P.A., Stuart R.A. and Eccleston W. (1983) "Evidence of optical generation of minority carriers from saturated MOS transistors". *Solid State Electronics*, **26**, 7, p. 685.
- [110] Bude J.D. (1995) "Gate current by impact ionization feedback in sub-micron MOSFET technologies". *Proc. Symp. on VLSI Technology*, p. 101.
- [111] Chen I.C., Kaya C. and Paterson J. (1989) "Band-to-band tunneling induced substrate hot-electron (BBISHE) injection: A new programming mechanism for non-volatile memory devices". *IEDM Technical Digest*, p. 263.
- [112] Roy A., Kazerounian R., Kablanian A. and Eitan B. (1992) "Substrate injection induced program disturb: A new reliability consideration for Flash-EEPROM arrays". *Proc. Int. Reliability Physics Symp.*, p. 68.
- [113] Bude J.D., Frommer A., Pinto M.R. and Weber G.R. (1995) "EEPROM/Flash sub 3.0V drain-source bias hot carrier writing". *IEDM Technical Digest*, p. 989.
- [114] Esseni D., Selmi L. and Bez R. (1998) "The impact of device design on the substrate enhanced gate current of VLSI MOSFET's". *Proc. European Solid State Device Res. Conf.*

- [115] Verwey J.F. (1973) "Nonavalanche injection of hot carriers into SiO₂". *Journal of Applied Physics*, **44**, 6, p. 2681.
- [116] Ning T.H. and Yu H.N. (1974) "Optically induced injection of hot electrons into silicon dioxide". *Journal of Applied Physics*, **45**, p. 5373.
- [117] Schwerin A.V., Heyns M.M. and Weber W. (1990) "Investigation on the oxide field dependence of hole trapping and interface state generation in SiO₂ layers using homogeneous nonavalanche injection of holes". *Journal of Applied Physics*, **67**, p. 7595.
- [118] Hu C.Y., Kenke L., Banerjee S.K., Richart R., Bandyopadhyay B., Moore B., Ibok E. and Garg S. (1995) "A convergence scheme for over-erased Flash EEPROM's using substrate- bias- enhanced hot electron injection". *IEEE Electron Device Letters*, **16**, p. 500.
- [119] Yamada S., Suzuki T., Obi E., Oshikiri M., Naruke K. and Wada M. (1991) "A self-convergence erasing scheme for a simple stacked gate Flash EEPROM". *IEDM Technical Digest*, p. 307.
- [120] Chi M.H. and Bergemont A. (1997) "A new multi-level erase scheme with self-convergence for Flash memory cell". *Proc. Non Volatile Semic. Memory Workshop*, p. 6.3.
- [121] Tsuji N., Ajika N., Yuzuriha K., Kunori Y., Hatanaka M. and Miyoshi H. (1994) "New erase scheme for DINOR Flash memory enhancing erase/write cycling endurance characteristics". *IEDM Technical Digest*, p. 53.
- [122] Kaya C., Middendorf M., Mehrad F., San K. and Huber B. (1995) "Low-level gate current injections in Flash memories initiated by minority carrier collection of floating terminals". *IEEE Trans. on Electron Devices*, p. 2131.
- [123] Esseni D., Selmi L., Sangiorgi E., Bez R. and Riccò B. (1995) "Temperature dependence of gate and substrate currents in the CHE crossover regime". *IEEE Electron Device Letters*, **16**, p. 506.
- [124] Riccò B., Sangiorgi E. and Cantarelli D. (1984) "Low voltage hot-electron effects in short channel MOSFETs". *IEDM Technical Digest*, p. 92.
- [125] Chung J.E., Jeng N.C., Moon J.E., Ko P.K. and Hu C. (1990) "Low-voltage hot-electron currents and degradation in deep-submicrometer MOSFET's". *IEEE Trans. on Electron Devices*, **37**, p. 1651.

- [126] Sangiorgi E., Venturi F., Fiegna C., Abramo A. and Capasso F. (1992) "Non-local effects on the electron energy distribution in short channel devices under high-field conditions", *Proc. Int. Workshop on Computational Electronics*, p. 221.
- [127] Fischer B., Ghetti A., Selmi L., Bez R. and Sangiorgi E. (1997) "Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages". *IEEE Trans. on Electron Devices*, p. 288.
- [128] Abramo A., Fiegna C. and Venturi F. (1995) "Hot carrier effects in short MOSFETs at low applied voltages". *IEDM Technical Digest*, p. 301.
- [129] Ellis-Monaghan J.J., Hulfachor R.B., Kim K.W. and Littlejohn M.A. (1996) "Ensemble Monte-Carlo study of interface-state generation in low voltage scaled silicon MOS devices". *IEEE Trans. on Electron Devices*, **43**, 7, p. 1123.
- [130] Selmi L., Fischer B., Ghetti A. and Bez R. (1996) "Hot-carriers at low voltages: new experimental evidences and open issues". *IEDM Technical Digest*, p. 375.
- [131] Chen I.C., Holland S. and Hu C. (1986) "Oxide breakdown dependence on thickness and hole current - enhanced reliability of ultra thin oxides". *IEDM Technical Digest*, p. 660.
- [132] Eitan B. and Kolodny A. (1983) "Two-components of tunneling current in metal- oxide- semiconductor structures". *Journal of Applied Physics*, **43**, 1, p. 106.
- [133] Fischetti M.V. (1985) "Model for the generation of positive charge at the Si-SiO₂ interface based on hot-hole injection from the anode". *Physical Review B*, **31**, p. 2099.
- [134] Klein N. and Solomon P. (1976) "Current runaway in insulators affected by impact ionization and recombination". *Journal of Applied Physics*, **47**, p. 4364.
- [135] Weinberg Z.A. and Fischetti M.V. (1985) "Investigation of the SiO₂-induced substrate current in silicon field-effect transistors". *Journal of Applied Physics*, **57**, p. 443.
- [136] Weinberg Z.A., Fischetti M.V. and Nissan-Cohen Y. (1986) "SiO₂ induced substrate current and its relation to positive charge in field effect transistors". *Journal of Applied Physics*, **59**, 3, p. 824.

- [137] DiMaria D.J. (1995) "Hole trapping, substrate currents, and breakdown in thin silicon dioxide films". *IEEE Electron Device Letters*, **16**, p. 184.
- [138] Chen I.C., Holland S.E. and Hu C. (1985) "Electrical breakdown in thin gate and tunneling oxides". *IEEE Trans. on Electron Devices*, **32**, 2, p. 413.
- [139] Hughes R.C. (1978) "High field electronic properties of SiO₂". *Solid State Electronics*, **21**, p. 251.
- [140] Olivo P., Riccò B. and Sangiorgi E. (1983) "Electron trapping-detrapping within thin SiO₂ films in the high field tunneling regime". *Journal of Applied Physics*, **54**, 9, p. 5267.
- [141] Nissan-Cohen Y., Shappir J. and Frohman-Bentchkowsky D. (1986) "Trap generation and occupation dynamics in SiO₂ under charge injection stress". *Journal of Applied Physics*, **60**, 6, p. 2024.
- [142] Scott R.S., Dumin N.A., Hughes T.W., Dumin D.J. and Moore B.T. (1996) "Properties of high-voltage stress generated traps in thin silicon oxide". *IEEE Trans. on Electron Devices*, **43**, p. 1133.
- [143] DiMaria D.J., Arnold D. and Cartier E. (1992) "Degradation and breakdown of silicon dioxide films on silicon". *Applied Physics Letters*, **61**, p. 2329.
- [144] Chen I.C., Holland S.E. and Hu C. (1987) "Electron trap generation by recombination of electrons and holes in SiO₂". *Journal of Applied Physics*, **61**, 9, p. 4544.
- [145] Dumin D.J., Mopuri S.K., Vanchinathan S., Scott R.S., Subramoniam R. and Lewis T.G. (1995) "High field related thin oxide wearout and breakdown". *IEEE Trans. on Electron Devices*, **42**, 4, p. 760.
- [146] Neri B., Olivo P. and Riccò B. (1987) "Low-frequency noise in silicon-gate metal-oxide-silicon capacitors before oxide breakdown". *Journal of Applied Physics*, **51**, 25, p. 2167.
- [147] Olivo P., Nguyen T.N. and Riccò B. (1988) "High-field-induced degradation in ultra-thin SiO₂ films". *IEEE Trans. on Electron Devices*, **35**, p. 2259.
- [148] Maserjan J. and Zamani N. (1982) "Observation of positively charged state generation near the Si/SiO₂ interface during Fowler-Nordheim tunneling". *Journal of Vacuum Science and Technology*, **20**, 3, p. 743.

- [149] Nguyen T.N., Olivo P. and Riccò B. (1987) "A new failure mode of very thin ($< 50\text{\AA}$) Thermal SiO_2 Films". *Proc. Int. Reliability Physics Symp.*, p. 66.
- [150] Naruke N., Taguchi S. and Wada M. (1988) "Stress induced leakage current limiting to scale down EEPROM tunnel oxide". *IEDM Technical Digest*, p. 424.
- [151] Baglee D.A. and Smayling M.C. (1985) "The effects of write/erase cycling on data loss in EPROM's". *IEDM Technical Digest*, p. 624.
- [152] Moazzami R. and Hu C. (1992) "Stress-induced current in thin silicon dioxide films". *IEDM Technical Digest*, p. 139.
- [153] Kimura M. and Koyama H. (1994) "Stress-induced low level leakage mechanism in ultrathin silicon dioxide films cause by neutral oxide trap generation". *Proc. Int. Reliability Physics Symp.*, p. 167.
- [154] DiMaria D.J. (1995) "Stress induced leakage currents in thin oxides". *Microelectronic Engineering*, **28**, p. 63.
- [155] Takagi S., Yasuda N. and Toriumi A. (1996) "Experimental evidence of inelastic tunneling and new I-V model for stress-induced leakage current". *IEDM Technical Digest*, p. 323.
- [156] Sakikabara K., Ajika N., Atanaka M., Miyoshi H. and Yasuoka A. (1997) "Identification of stress-induced leakage current components and the corresponding trap models in SiO_2 films". *IEEE Trans. on Electron Devices*, **44**, 6, p. 986.
- [157] Sakikabara K., Ajika N., Eikyu K., Ishikawa K. and Miyoshi H. (1997) "A quantitative analysis of time-decay reproducible stress-induced leakage current in SiO_2 films". *IEEE Trans. on Electron Devices*, **44**, 6, p. 1002.
- [158] Runnion E.F., Gladstone I.S.M., Scott J.R.S., Dumin D.J., Lie L. and Mitros J.C. (1997) "Thickness dependence of stress-induced leakage currents in silicon oxide". *IEEE Trans. on Electron Devices*, **44**, 6, p. 993.
- [159] Depas M., Nigam T. and Heyns M.M. (1996) "Soft breakdown of ultrathin gate oxide layers". *IEEE Trans. on Electron Devices*, **43**, p. 1499.
- [160] Schuegraf K.F. and Hu C. (1993) "Hole injection oxide breakdown model for very low voltage lifetime extrapolation". *Proc. Int. Reliability Physics Symp.*, p. 7.

- [161] Satake H. and Toriumi A. (1993) "Substrate hole current generation and oxide breakdown in Si MOSFETs under Fowler-Nordheim electron tunnel injection". *IEDM Technical Digest*, p. 337.
- [162] Weinberg Z.A. and Nguyen T.N. (1987) "The relation between positive charge and breakdown in metal-oxide-silicon structures". *Journal of Applied Physics*, **61**, 5, p. 1947.
- [163] Wolters D.R., Van der Schoot J.J. and Poorter T. (1983) "Damage caused by charge injection". *Proc. INFOS*, p. 256.
- [164] Avni E. and Shappir J. (1988) "A model for silicon-oxide breakdown under high field and current stress". *Journal of Applied Physics*, **64**, 2, p. 734.
- [165] Wolters D.R. and Zeegers-van Duijnhoven A.T.A. (1990) "Breakdown of thin dielectrics". *Ext. Abs. Mtg. of Electrochem. Soc.*, p. 272.
- [166] Apte P.P. and Saraswat K.C. (1994) "Modeling ultrathin dielectric breakdown on correlation of charge trap-generation to charge-to-breakdown". *Proc. Int. Reliability Physics Symp.*, p. 136.
- [167] Degraeve R., Groeseneken G., Bellens R., Depas M. and Maes H. (1995) "A consistent model for the thickness dependence of intrinsic breakdown in ultrathin oxides". *IEDM Technical Digest*, p. 863.
- [168] Felsch C. and Rosenbaum E. (1985) "The relation between oxide degradation and oxide breakdown". *Proc. Int. Reliability Physics Symp.*, p. 142.
- [169] Blauwe J., Van Houdt J., Wellekens D., Degraeve R., Roussel Ph., Haspelslagh L., Deferm L., Groeseneken G. and Maes H.E. (1996) "A new quantitative model to predict SILC-related disturb characteristics in Flash E²PROM devices". *IEDM Technical Digest*, p. 343.
- [170] Boyko K.C. and Gerlach D.L. (1989) "Time dependent dielectric breakdown of 210Å oxides". *Proc. Int. Reliability Physics Symp.*, p. 1.
- [171] Lo G.Q., Ito S. and Kwong D.L. (1992) "Charge trapping/detrapping and dielectric breakdown in SiO₂/Si₃N₄/SiO₂ stacked layers on rugged Poly-Si under dynamic stress". *Proc. Int. Reliability Physics Symp.*, p. 42.
- [172] Degraeve R., Roussel P.H., Groeseneken G. and Maes H. (1996) "A new analytical model for the description of the intrinsic oxide breakdown statistics of ultra-thin oxides". *Microelectronic Reliability*, **36**, 11, p. 1639.

- [173] Avni E. and Shappir J. (1988) "Modeling of charge injection effects in metal-oxide-semiconductor structures". *Journal of Applied Physics*, **64**, 2, p. 743.
- [174] Heyns M.M., Rao D.K. and Keersmaecker R. (1989) "Oxide field dependence of the Si-SiO₂ interface state generation and charge trapping during electron injection". *Applied Surface Science*, **39**, p. 327.
- [175] DiMaria D.J. and Stasiak J.W. (1989) "Trap creation in silicon dioxide produced by hot electrons". *Journal of Applied Physics*, **65**, p. 2342.
- [176] Von Schwerin A. and Heyns M.M. (1991) "Oxide field dependence of bulk and interface trap generation in SiO₂ due to electron injection". *Proc. INFOS*, p. 263.
- [177] Zhao S.P., Taylor S., Eccleston W. and Barlow K.J. (1992) "P-well bias dependence of electron trapping in gate oxide n-MOSFETs during substrate hot-electron injection". *Electronics Letters*, **28**, p. 2080.
- [178] Van Den Bosch G., Groeseneken G. and Maes H. (1994) "Critical analysis of the substrate hot-hole injection technique". *Solid State Electronics*, **37**, 3, p. 393.
- [179] Ning T.H. (1976) "Capture cross section and trap concentration of holes in silicon dioxide". *Journal of Applied Physics*, **47**, p. 1079.
- [180] Heremans P., Bellens R., Groeseneken G. and Maes H. (1988) "Consistent model for the hot-carrier degradation in n-channel and p-channel MOSFETs". *IEEE Trans. on Electron Devices*, **12**, p. 2194.
- [181] Wang C.T. (1992) *Hot-Carrier Design Considerations for MOS Devices and Circuits*. Van Nostrand Reinhold, New York.
- [182] Heremans P., Witters J., Groeseneken G. and Maes H. (1989) "Analysis of the charge pumping technique and its application for the evaluation of MOSFET degradation". *IEEE Trans. on Electron Devices*, **36**, p. 1318.
- [183] Haddara H. and Cristoloveanu S. (1987) "Two-dimensional modeling of locally damaged short-channel MOSFETs". *IEEE Trans. on Electron Devices*, **34**, p. 378.
- [184] Nicollian E.H. and Brews J.R. (1983) *MOS Physics and Technology*. Wiley, New York.

- [185] Hu C., Tam S.C., Hsu F.C., Ko P.K., Chan T.Y. and Terrill K.W. (1985) "Hot-electron-induced MOSFET degradation: model, monitor, and improvement". *IEEE Trans. on Electron Devices*, **32**, p. 375.
- [186] Selmi L., Fiegna C., Sangiorgi E., Bez R. and Riccò B. (1993) "A study of injection conditions in the substrate hot electron induced degradation of n-MOSFETs". *Proc. Int. Workshop on VLSI Process and Device Modelling*, p. 156.
- [187] Bude J.D., Iizuka T. and Kamakura Y. (1996) "Determination of threshold energy for hot electron interface state generation". *IEDM Technical Digest*, p. 865.
- [188] Liang C., Gaw H. and Cheng P. (1992) "An analytical model for the self-limiting behavior of hot carrier degradation in $0.25\mu\text{m}$ n-MOSFETs". *IEEE Electron Device Letters*, **11**, p. 569.
- [189] Eitan B. and Frohman-Bentchkowsky D. (1981) "Hot-electron injection into the oxide in n-channel MOS devices". *IEEE Trans. on Electron Devices*, **28**, p. 328.
- [190] Hu H., Jacobs J., Chung J.E. and Antoniadis D. (1994) "The correlation between gate current and substrate current in $0.1\mu\text{m}$ NMOSFETs". *IEEE Electron Device Letters*, **15**, p. 418.
- [191] Choi J.Y., Ko P.K. and Hu C. (1987) "Effect of oxide field on hot-carrier induced degradation of metal-oxide-semiconductor field-effect transistors". *Journal of Applied Physics*, **51**, 17, p. 1188.
- [192] Krieger G., Cuevas P.P. and Misheloff M.N. (1988) "The effect of impact ionization induced bipolar action on n-channel hot-electron degradation". *IEEE Electron Device Letters*, **9**, 1, p. 26.
- [193] Tsividis Y. (1987) *Operation and Modeling of the MOS Transistor*. McGraw Hill.
- [194] Saha S. (1994) "Extraction of substrate current model parameters from device simulation". *Solid State Electronics*, **37**, 10, p. 1786.
- [195] Chen M.L., Leung C.W., Cochran W.T., Jüngling W., Dziuba C. and Yang T. (1988) "Suppression of hot carrier effects in submicrometer CMOS technology". *IEEE Trans. on Electron Devices*, **35**, p. 2210.
- [196] Woltjer R. and Paulzen G.M. (1994) "Improved prediction of interface trap generation in n-MOST". *IEEE Electron Device Letters*, **15**, p. 4.