FLASH MEMORIES

edited by
Paulo Cappelletti
Carla Golla
Piero Olivo
Enrico Zanoni

KLUWER ACADEMIC PUBLISHERS

APPLE INC.
EXHIBIT 1008 - PAGE 0001

FLASH MEMORIES

APPLE INC.
EXHIBIT 1008 - PAGE 0002

FLASH MEMORIES

By

Paulo Cappelletti

Carla Golla

Piero Olivo

Enrico Zanoni

KLUWER ACADEMIC PUBLISHERS
Boston/Dordrecht/London

APPLE INC.
EXHIBIT 1008 - PAGE 0003

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, photo-copying, 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

APPLE INC.
EXHIBIT 1008 - PAGE 0004

Contents

1
FLASH MEMORIES: AN OVERVIEW 1
P. Olivo, E. Zanoni

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
P. Pavan, R. Bez

APPLE INC.
EXHIBIT 1008 - PAGE 0005

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

APPLE INC.
EXHIBIT 1008 - PAGE 0006

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

APPLE INC.
EXHIBIT 1008 - PAGE 0007

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
M. Branchetti, G. Campardo, S. Commodaro, S. Ghezzi, A. Ghilardelli, C. Golla, M. Maccarrone, I. Martines, R. Micheloni, J. Mulatti, M. Zammattio, S. Zanardi

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" Bost 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

APPLE INC.
EXHIBIT 1008 - PAGE 0008

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
G. Torelli, M. Lanzoni. A. Manstretta, B. Riccò
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
P. Cappelletti, A. Modelli
7.1 Introduction 399
7.2 Memory Array V_t Distributions and Tunnel Oxide "Defects" 401
7.3 Main Yield and Reliability Issues 409

APPLE INC.
EXHIBIT 1008 - PAGE 0009

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
G. Casagrande
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

APPLE INC.
EXHIBIT 1008 - PAGE 0010

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
B. Beverina, P. Bergé, with contributions by C. Kunkel, G. Moy, A. Damiano, R. Ferrara, A. Re
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

APPLE INC.
EXHIBIT 1008 - PAGE 0011

1 FLASH MEMORIES: AN OVERVIEW

Piero Olivo[1], Enrico Zanoni[2]

[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

APPLE INC.
EXHIBIT 1008 - PAGE 0012

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-

APPLE INC.
EXHIBIT 1008 - PAGE 0013

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0014

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

$$V_T = K - Q/Cox \tag{1.1}$$

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 Cox 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/Cox. 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:

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

APPLE INC.
EXHIBIT 1008 - PAGE 0015

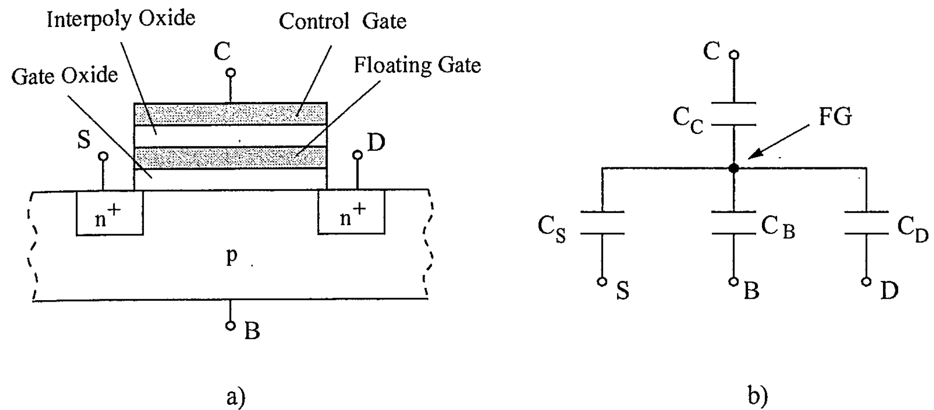


Figure 1.1 a) Schematic cross section of a generic floating gate device; b) electrical model of a floating gate device (junction capacitances are neglected).

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

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

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

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

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:

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

APPLE INC.
EXHIBIT 1008 - PAGE 0016

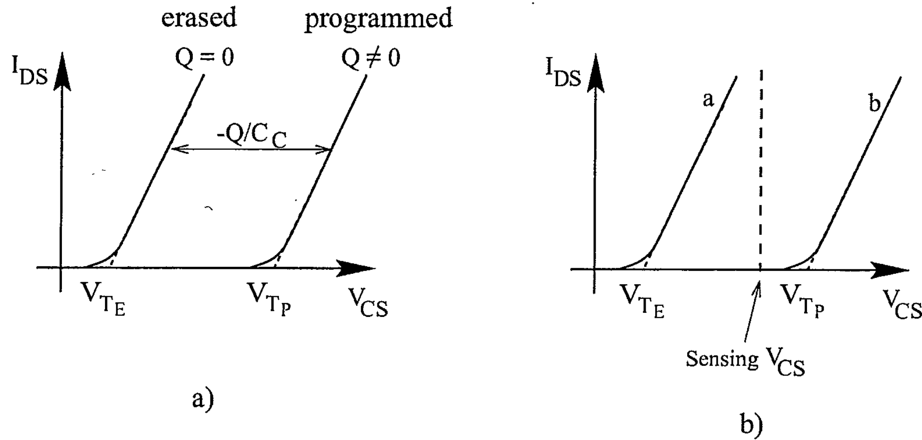


Figure 1.2 a) I-V trans-characteristics of a FG device for two different values of charge stored within the FG ($Q = 0$, and $Q < 0$), denoting two different states, respectively: erased and programmed; b) reading operation of a FG device: a suitable control gate voltage ($V_{T_E} < V_{CS} < V_{T_P}$) 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

$$V_{T_{CS}} = \frac{1}{\alpha_C} V_{T_{FS}} = V_{T_E} \; , \qquad (1.6)$$

$$V_{T_{CS}} = \frac{1}{\alpha_C} V_{T_{FS}} - \frac{Q}{C_C} = V_{T_P} \; , \qquad (1.7)$$

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 $V_{CS} > V_{T_E}$ and the device is ON; when the device has been previously programmed (curve b), $V_{CS} < V_{T_P}$ and the device is OFF.

APPLE INC.
EXHIBIT 1008 - PAGE 0017

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

APPLE INC.
EXHIBIT 1008 - PAGE 0018

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

APPLE INC.
EXHIBIT 1008 - PAGE 0019

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 μm^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

APPLE INC.
EXHIBIT 1008 - PAGE 0020

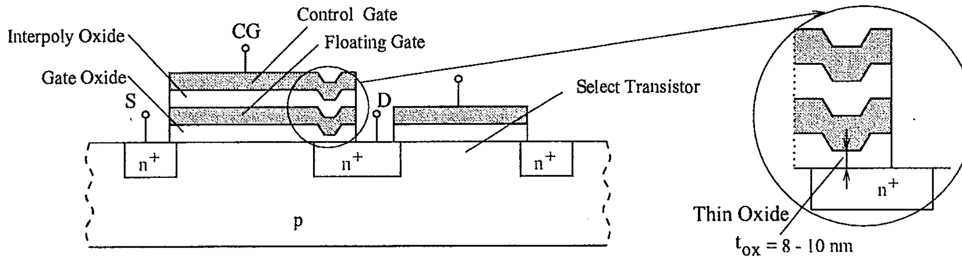


Figure 1.3 Schematic section of a FLOTOX cell including the select transistor.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0021

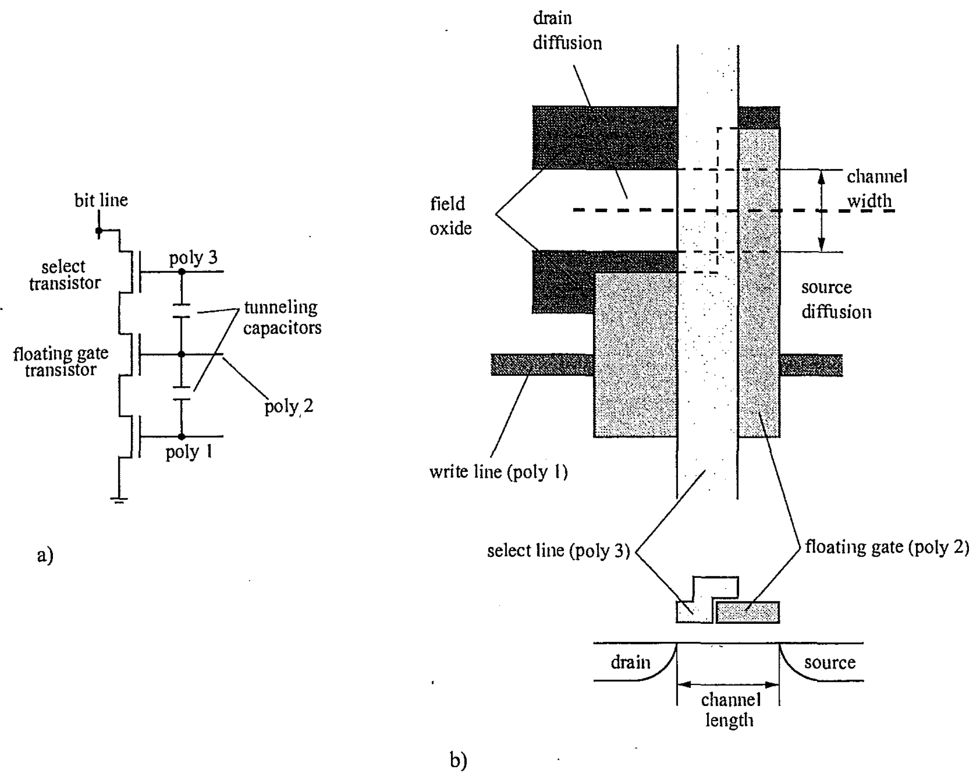


Figure 1.4 Equivalent circuit (a), layout and schematic cross-section (b) of a textured poly EEPROM cell

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

APPLE INC.
EXHIBIT 1008 - PAGE 0022

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.

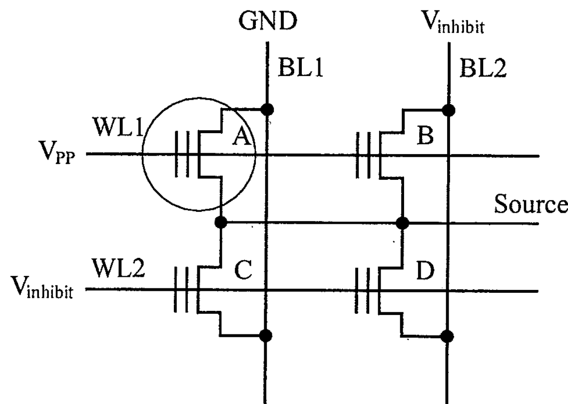


Figure 1.5 Simple 2 × 2 NOR-EEPROM architecture showing a program disturb on cell C when cell A is programmed.

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0023

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, $Pb(Zr,Ti)O_3$ (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.

APPLE INC.
EXHIBIT 1008 - PAGE 0024

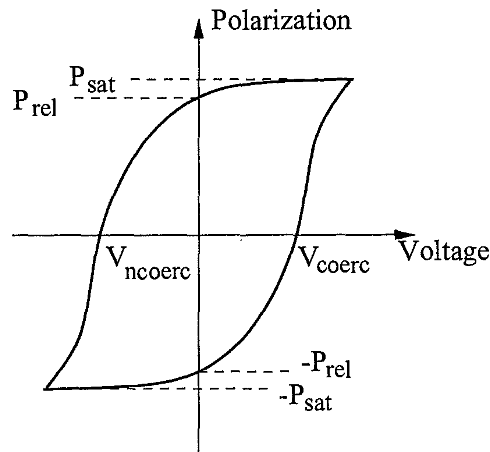


Figure 1.6 Typical hysteresis curve of a ferroelectric capacitor, identifying polarization states.

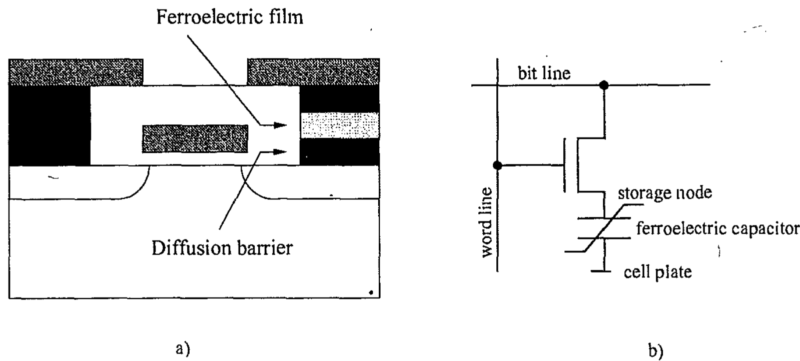


Figure 1.7 a) Schematic cross-section of a ferroelectric nonvolatile DRAM; b) equivalent circuit.

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0025

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 (sacrifying 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 $SrBi_2Ta_2O_9$ (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 $ETOX^{TM}$ (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

APPLE INC.
EXHIBIT 1008 - PAGE 0026

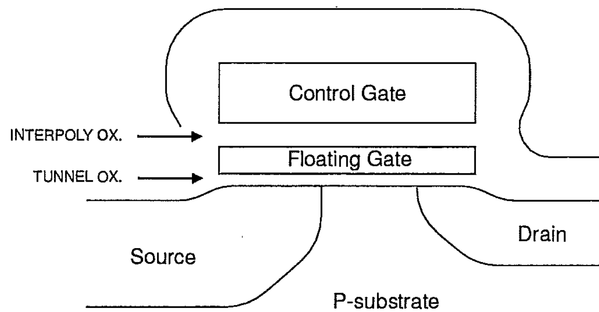


Figure 1.8 Schematic cross-section of an industrial Flash cell.

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

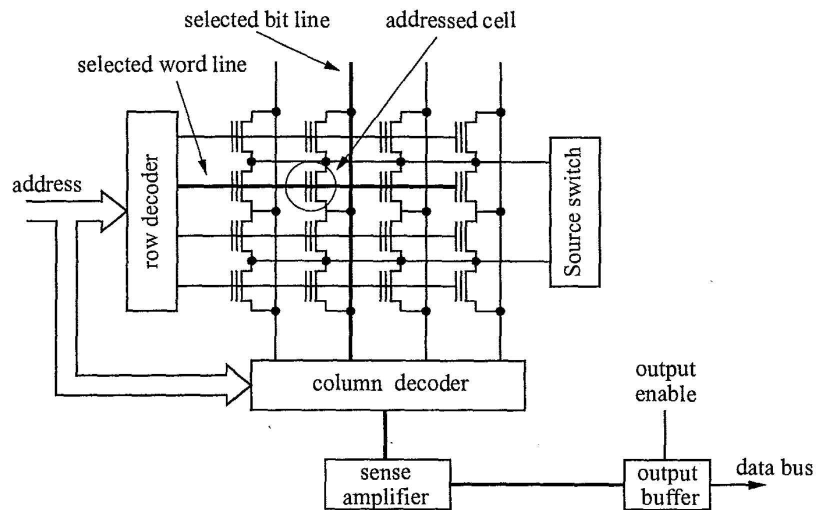


Figure 1.9 Schematic structure of the read path in a NOR organization. Only one bit at a time is here considered as addressable.

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 constrains.

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,

APPLE INC.
EXHIBIT 1008 - PAGE 0028

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 unadressed), 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 ÷ 7 V) and gate (at -8 ÷ -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 ($V_T > V_{Te_M}$), 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

APPLE INC.
EXHIBIT 1008 - PAGE 0029

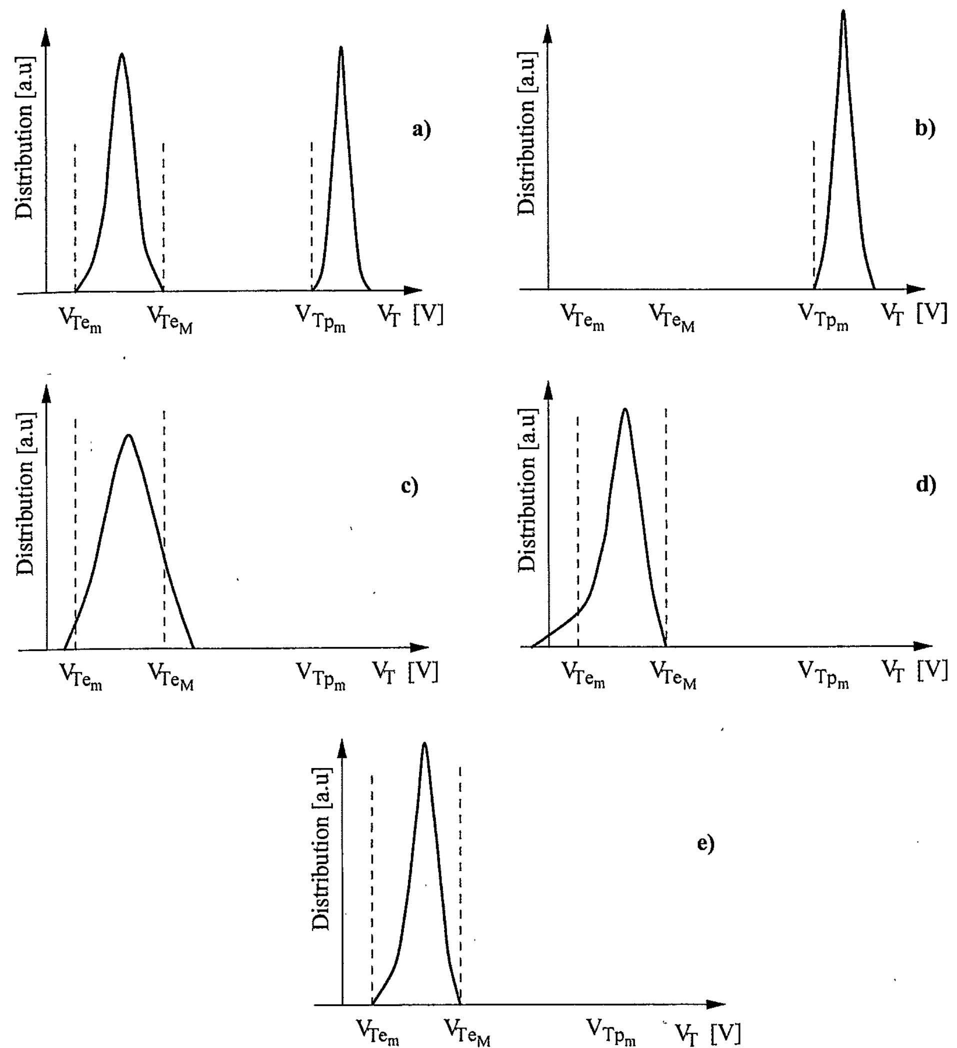


Figure 1.10 Schematic distribution of the threshold voltages during an erase operation: a) before erase; b) after a "program all-0" operation; c) after a single erase pulse; d) after the erase verify procedure has been successfully performed; e) after soft-programming.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0030

(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_{Te_m} and V_{Te_M} (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.

APPLE INC.
EXHIBIT 1008 - PAGE 0031

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.

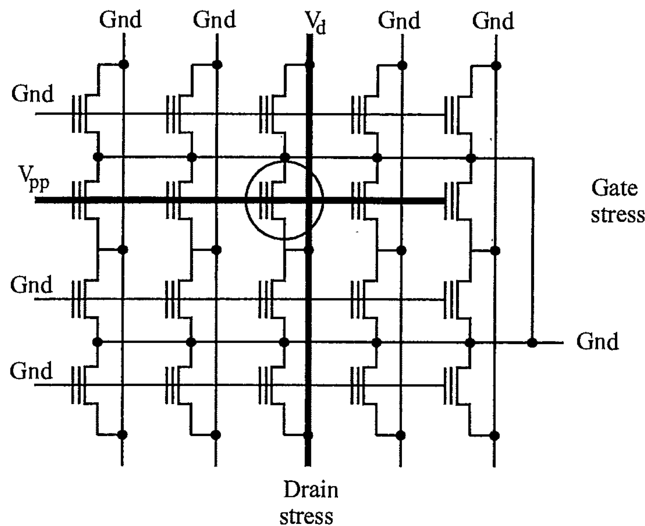


Figure 1.11 Program disturbs.

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:

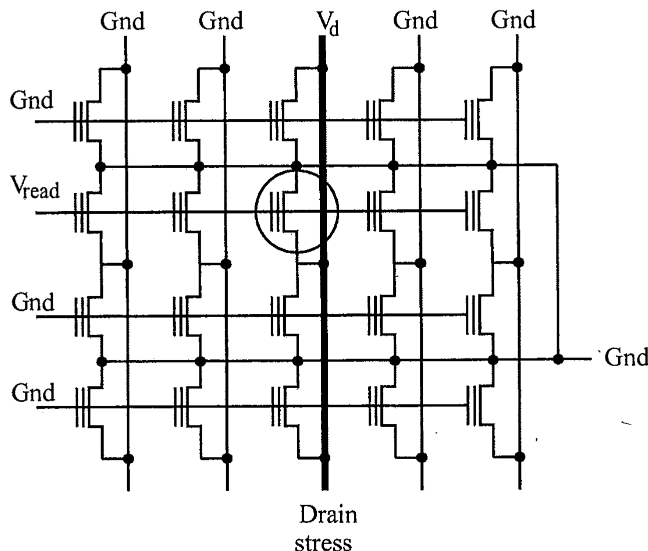


Figure 1.12 Read disturbs.

APPLE INC.
EXHIBIT 1008 - PAGE 0033

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 perfoming 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

APPLE INC.
EXHIBIT 1008 - PAGE 0034

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.

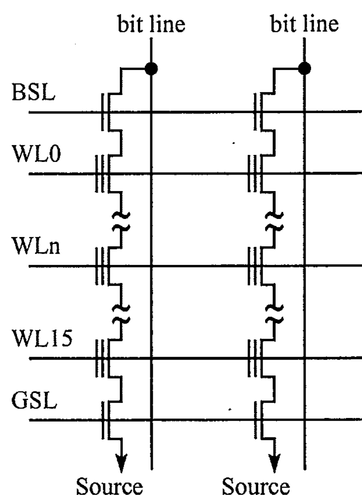


Figure 1.13 Basic structure of a NAND architecture.

APPLE INC.
EXHIBIT 1008 - PAGE 0035

Reading a current through a series of several cells and select transistors is a slow operation (the random access time is $\approx 10\mu s$): therefore, NAND architectures are limited to mass storage only.

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}.

To erase a sector, electrons are injected from the floating gate of all cells to the channel by means of Fowler-Nordheim tunneling. This task is again possible since high fields are applied to the thin gate oxides by grounding all the word lines and forcing a high voltage (up to 20 V) to the array p-well, to be kept fully separated from that of peripheral circuits. To prevent from erasing cells belonging to unselected sectors, the same high voltage is applied to the word lines controlling their gates. With respect to NOR architectures, erase operation is much faster, since the slow "program all 0" step is not to be performed.

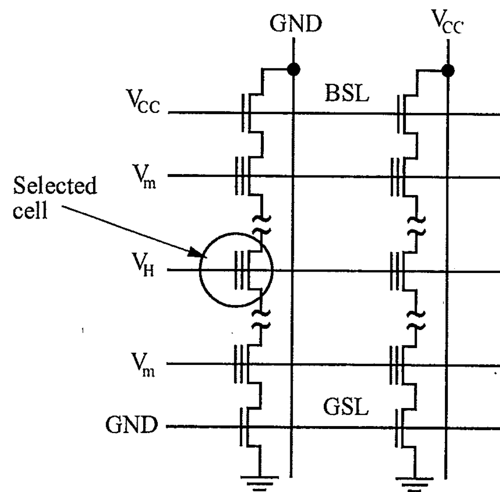


Figure 1.14 Program voltages in a NAND architecture.

APPLE INC.
EXHIBIT 1008 - PAGE 0036

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

APPLE INC.
EXHIBIT 1008 - PAGE 0037

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

APPLE INC.
EXHIBIT 1008 - PAGE 0038

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/SiO_2 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

APPLE INC.
EXHIBIT 1008 - PAGE 0039

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0040

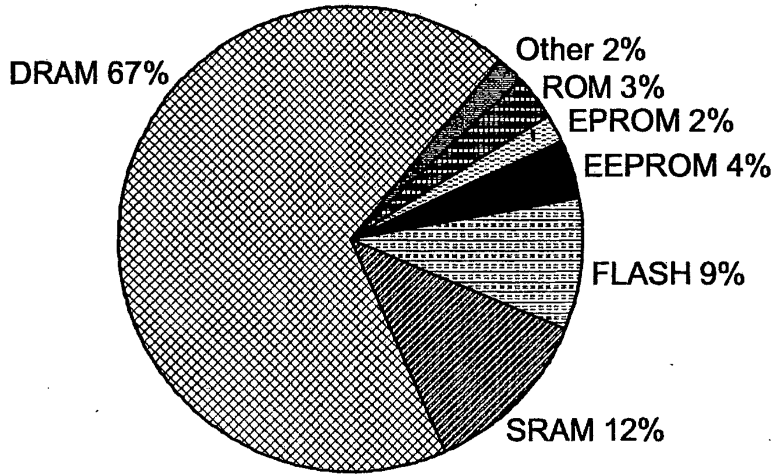


Figure 1.15 1997 worldwide MOS memory market. The total market is 33 B$

	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$).

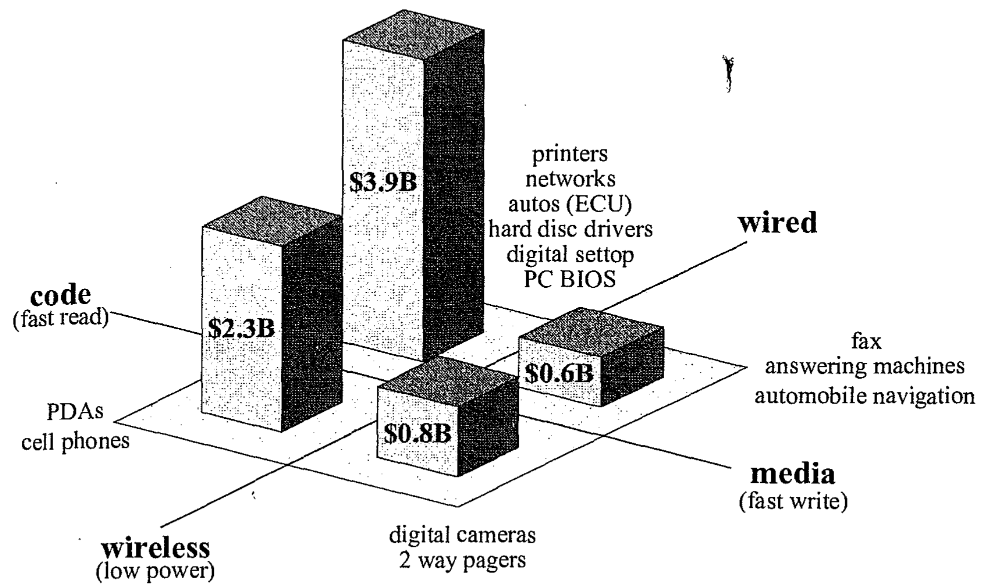


Figure 1.17 Flash market shares divided by application, 2002 forecast [24].

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,

APPLE INC.
EXHIBIT 1008 - PAGE 0042

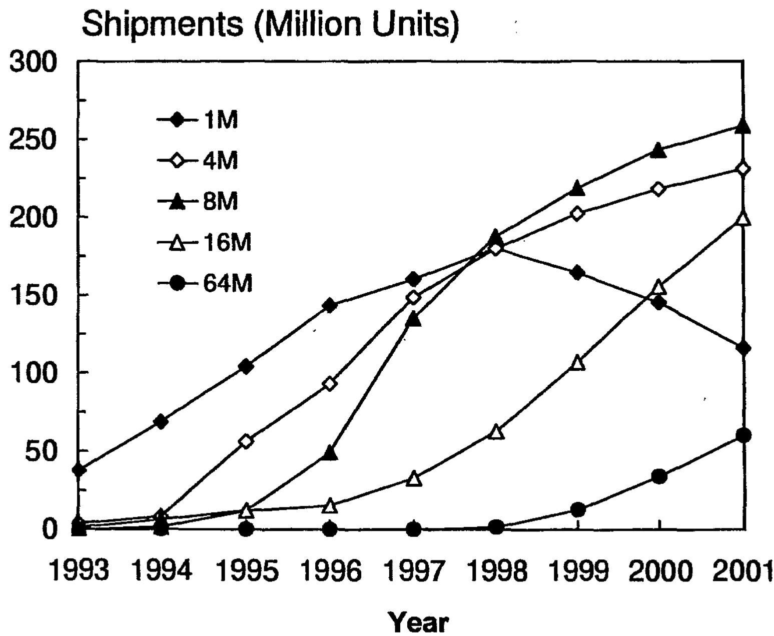


Figure 1.18 Growth of Flash memory shipments, 1993-2001 forecast

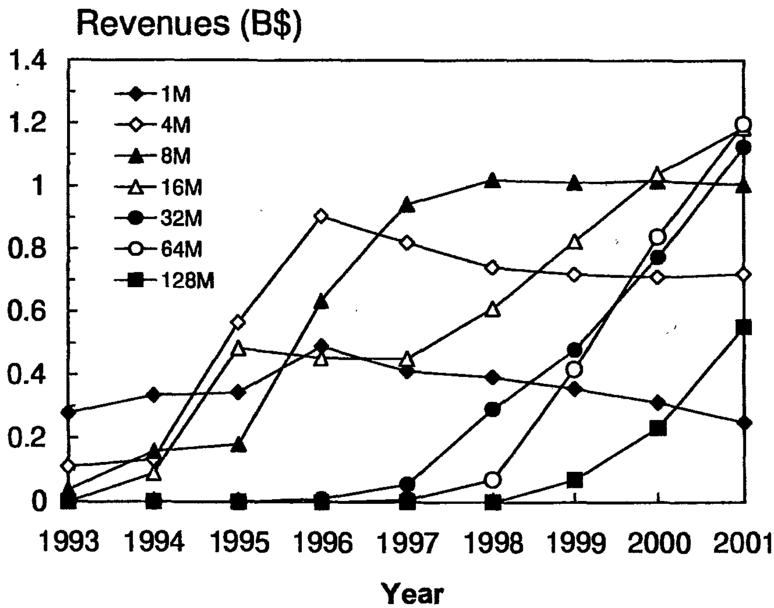


Figure 1.19 Flash market revenues divided by density, 1993-2001 forecast

APPLE INC.
EXHIBIT 1008 - PAGE 0043

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0044

[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 nonvolatile 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 E^2PROM 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0045

[17] Olivo P., Nguyen T. and Riccò B. (1988) "High-field-induced degradation in ultra thin SiO_2 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 μm 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 μm 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0046

2 THE INDUSTRY STANDARD FLASH MEMORY CELL

Paolo Pavan[1], Roberto Bez[2]

[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.

APPLE INC.
EXHIBIT 1008 - PAGE 0047

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 folating 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, staked 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]):

APPLE INC.
EXHIBIT 1008 - PAGE 0048

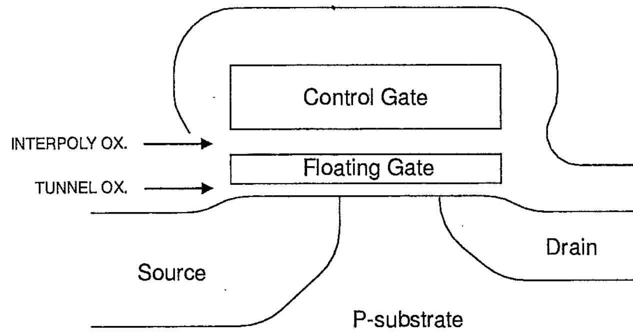


Figure 2.1 Schematic cross section of a Flash cell.

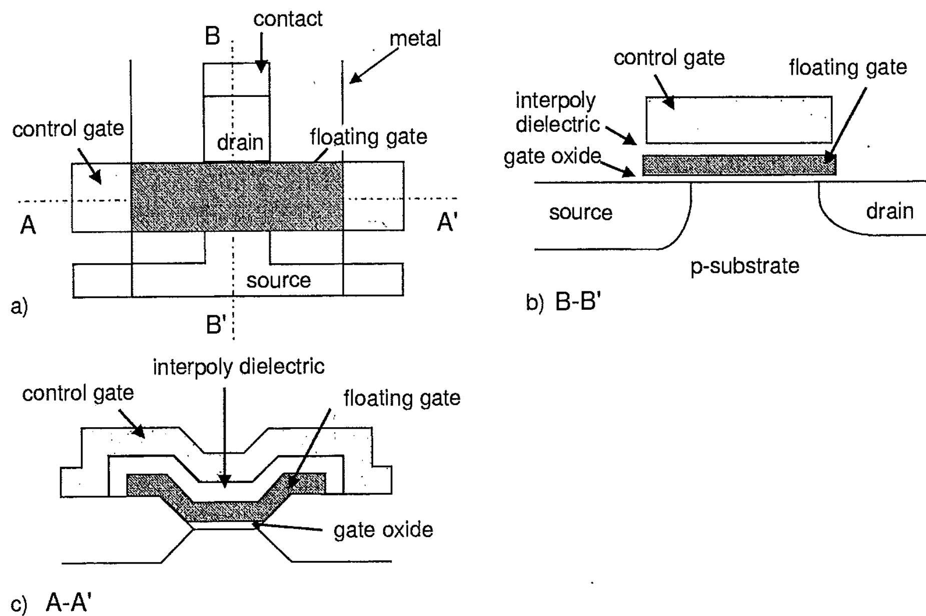


Figure 2.2 a) Layout of a T-shaped double polysilicon stacked gate Flash cell. The schematic cross sections along perpendicular directions are also shown in b) and c).

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

APPLE INC.
EXHIBIT 1008 - PAGE 0049

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 overlaying 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0050

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 μm, 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

APPLE INC.
EXHIBIT 1008 - PAGE 0051

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. $\overline{Q}=0$:

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

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 $C_T=C_{\mathrm{FC}}+C_D+C_S+C_B$ 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:

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

As an example, in a 0.4 μm 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:

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

where:

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

APPLE INC.
EXHIBIT 1008 - PAGE 0052

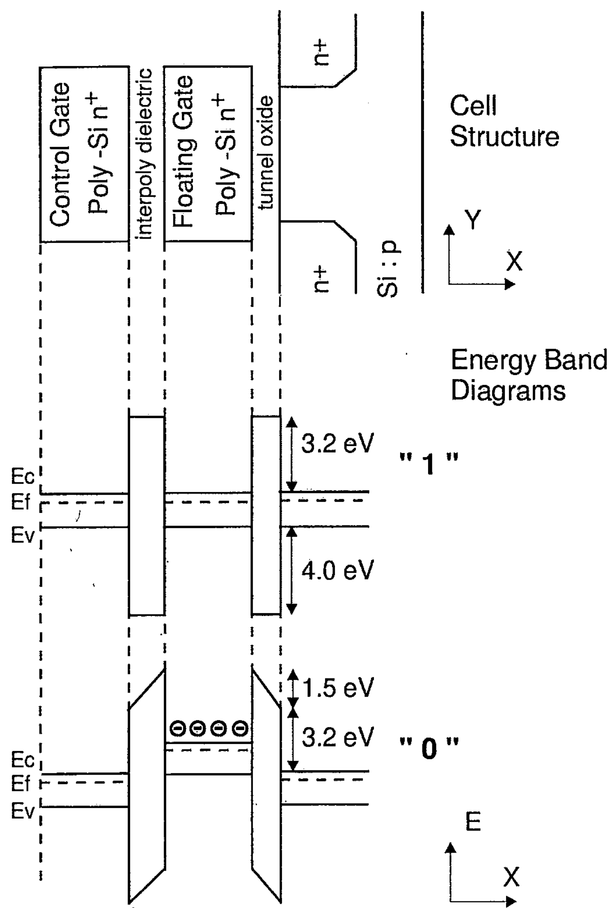


Figure 2.3 Cell structure and energy band diagrams of a Floating Gate transistor. The floating gate stores charge in state "0".

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 $V_{DS} = 0$:

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

APPLE INC.
EXHIBIT 1008 - PAGE 0053

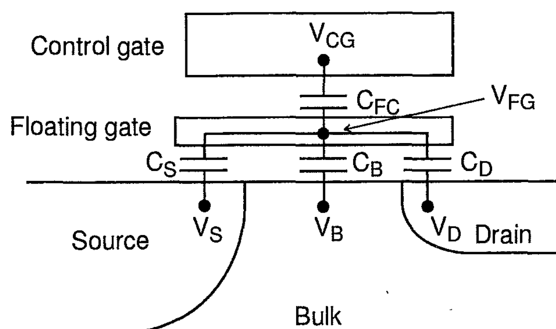


Figure 2.4 Schematic cross section of a Floating Gate transistor. The model using the capacitances between the Floating Gate and the other electrodes is described.

and

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

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 $|V_{\mathrm{DS}}| < |V_{\mathrm{GS}} - V_T|$

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

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

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

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

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

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

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

APPLE INC.
EXHIBIT 1008 - PAGE 0054

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 $\beta(\text{control}-\text{gate}) = \beta^{CG}$ and $V_T(\text{control}-\text{gate}) = V_T^{CG}$.

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).

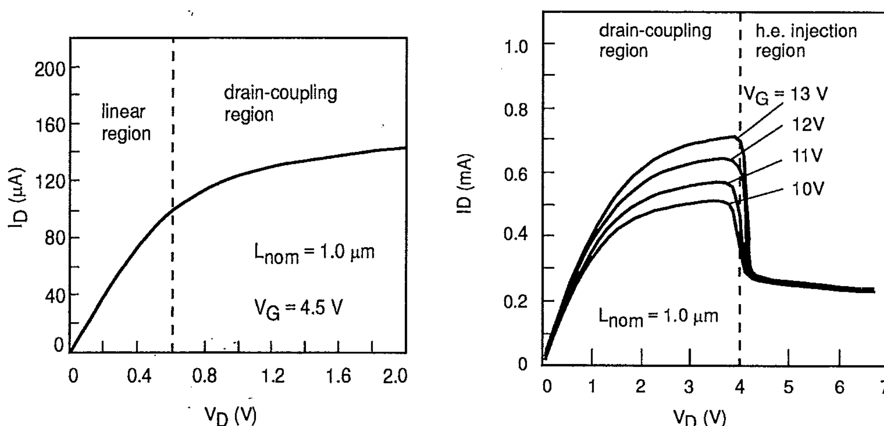


Figure 2.5 I-V characteristics of a floating gate transistor. In the right figure the effect of programming is also shown.

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

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

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

4) The transconductance in SR is given by:

$$g_m = \frac{\partial I_{\mathrm{DS}}}{\partial V_{\mathrm{GS}}}_{(V_{\mathrm{DS}}=\mathrm{constant})} = \alpha_G \beta(V_{\mathrm{GS}} + fV_{\mathrm{DS}} - V_T) \quad (2.12)$$

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 ($f = \alpha_D/\alpha_G = C_D/C_{\mathrm{FC}}$) and its value can be verified by:

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

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. $\overline{Q} \neq 0$. All the hypotheses made in Section 2.2 hold true, and the following modifications need to be included.

APPLE INC.
EXHIBIT 1008 - PAGE 0056

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

$$V_{\mathrm{FG}} = \alpha_G V_{\mathrm{GS}} + \alpha_D V_{\mathrm{DS}} + \frac{\overline{Q}}{C_T} \tag{2.14 }$$

$$V_T^{\mathrm{CG}} = \frac{1}{\alpha_G} V_T^{\mathrm{FG}} - \frac{\overline{Q}}{C_T \alpha_G} = \frac{1}{\alpha_G} V_T^{\mathrm{FG}} - \frac{\overline{Q}}{C_{\mathrm{FC}}} \tag{2.15 }$$

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

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

$$\Delta V_T = V_T - V_{\mathrm{T0}} = -\frac{\overline{Q}}{C_{\mathrm{FC}}} \tag{2.17 }$$

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

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 $V_{\mathrm{GS}} \simeq 5\,\mathrm{V}$, $V_{\mathrm{DS}} \simeq 1\,\mathrm{V}$), 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 $100\,\mu\mathrm{A}$ 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:

$$\Delta V_T(t) = V_T(t) - V_{\mathrm{T0}} = -\frac{\overline{Q(t)}}{C_{\mathrm{FC}}} \tag{2.18 }$$

APPLE INC.
EXHIBIT 1008 - PAGE 0057

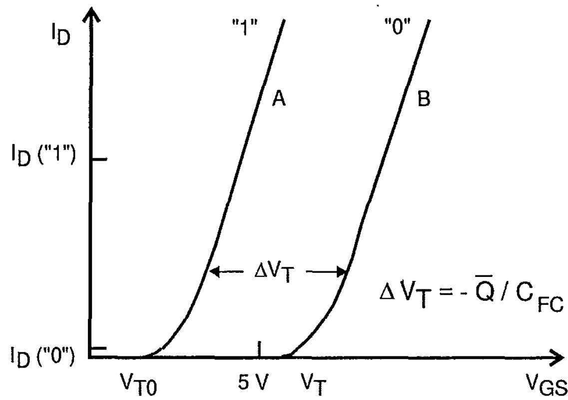


Figure 2.6 I-V curves of a floating gate device when there is no charge stored in the FG ("1" - curve A) and when a negative charge $\overline{Q}$ is stored in the FG ("0" - curve B) [25].

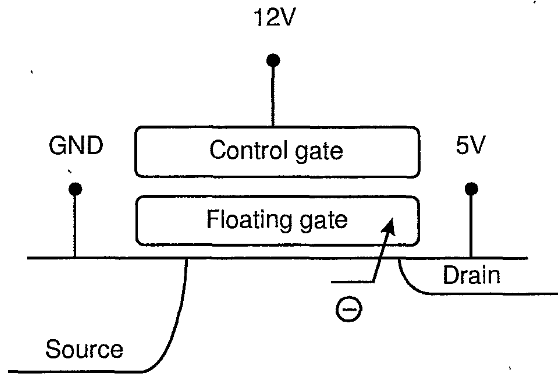


Figure 2.7 Schematic of the cell bias during program.

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:

$$\overline{Q(t)} = \int_0^t J_G(\tau)\, d\tau \tag{2.19}$$

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

APPLE INC.
EXHIBIT 1008 - PAGE 0058

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

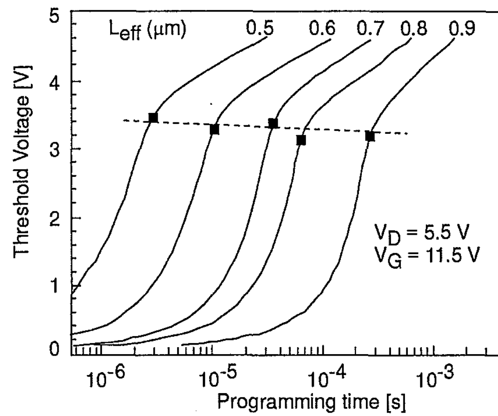


Figure 2.8 Programming curves of Flash cells with different channel lengths; t_{ox} is 12 nm. Black squares indicate ΔV_T^* [26].

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 ($L_{\text{eff}} = 0.7\,\mu$m, $V_G = 11.5$ 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:

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

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0059

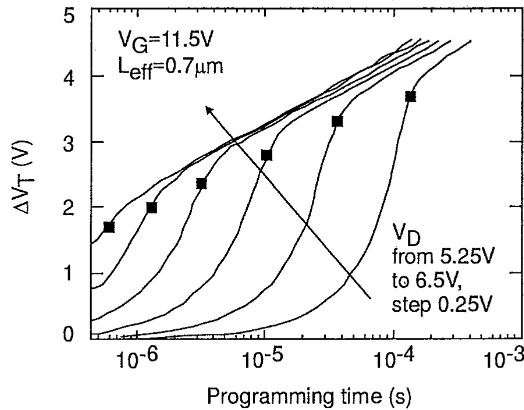


Figure 2.9 Programming curves of Flash cells for different drain voltages; t_{ox} is 12 nm. Black squares indicate ΔV_T^* [26].

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0060

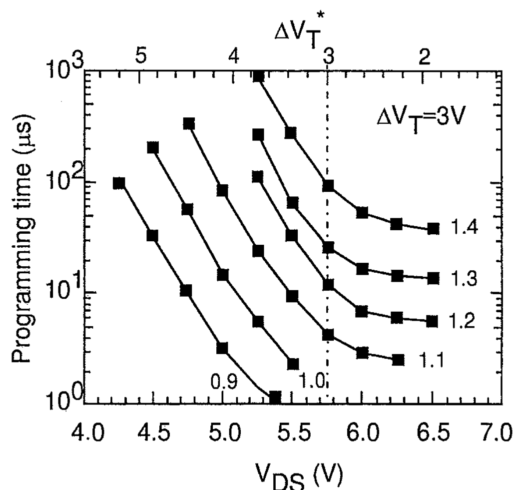


Figure 2.10 The programming time to obtain $\Delta V_T = 3\,\mathrm{V}$ is plotted vs. V_{DS} and ΔV_T^* for $V_{\mathrm{CG}} = 11.5\,\mathrm{V}$. The number near each curve marks the nominal channel length (μm) [26].

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}:

$$\gamma_{\mathrm{pr}} = \frac{I_G}{I_D} \tag{2.21}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0061

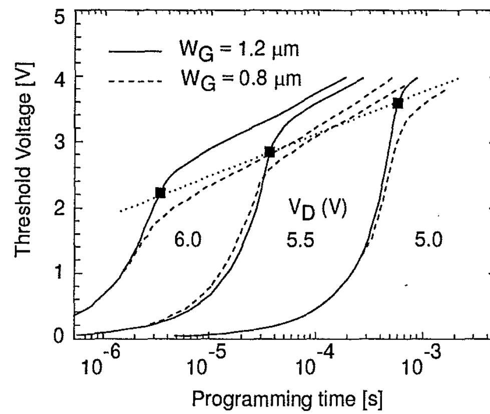


Figure 2.11 Programming curves of Flash cells with different coupling ratios and at different V_D; $L_{\text{eff}} = 0.7\,\mu$m, $t_{\text{ox}} = 12\,$nm [26].

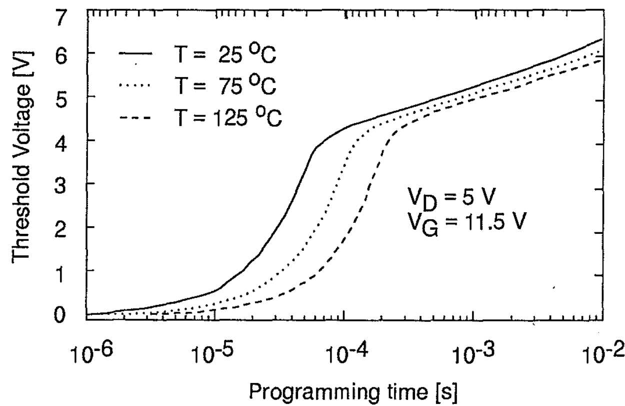


Figure 2.12 Programming curve of a Flash cell at different temperatures; $L_{\text{eff}} = 0.7\,\mu$m, $t_{\text{ox}} = 12\,$nm.

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, $I_D \approx 1\,$mA, is required to have the gate current, $I_G \approx 1\,$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:

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

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 μs–1 ms, and $C_{\rm FG}$ 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.

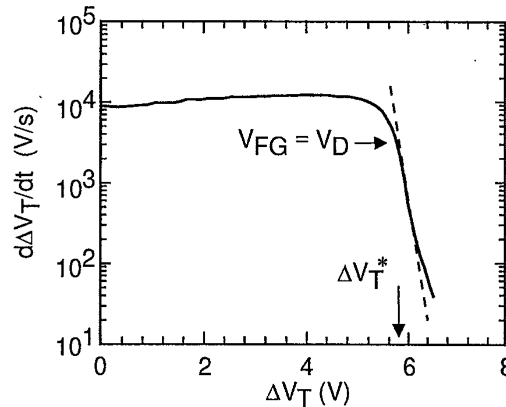


Figure 2.13 Time derivative of the programming curve, $d\Delta V_T/dt$ vs. ΔV_T. The point ΔV_T^* corresponds to the physical condition $V_{\rm FG} = V_D$ [22].

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:

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

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

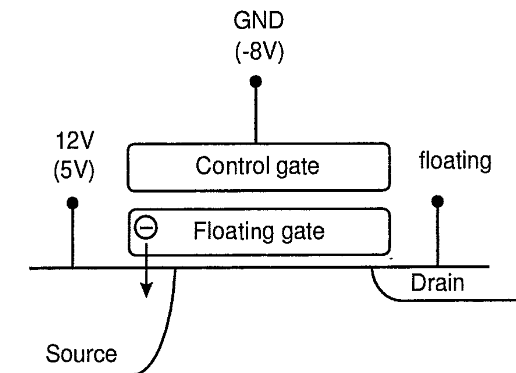


Figure 2.14 Schematic cross section of a Flash cell during erase. Voltages on electrodes refer to Dual Voltage operation mode, and in brackets to Single Voltage.

Continuity of the displacement vector at the surface gives:

$$\epsilon_{ox} E_{ox} = \epsilon_{Si} E_{Si} \tag{2.24}$$

where $\epsilon_{ox}, \epsilon_{Si}$ 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:

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

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/SiO$_2$ 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

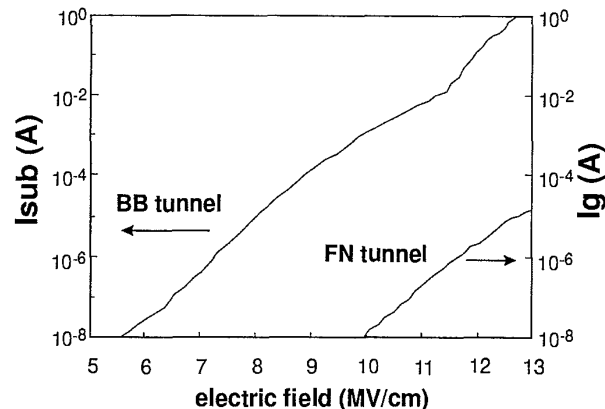


Figure 2.15 Relation between substrate and gate currents as a function of electric field measured on a large perimeter gated diode, with the p-n junction having the same profile of the source junction of a Flash cell [27].

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/SiO_2 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

APPLE INC.
EXHIBIT 1008 - PAGE 0065

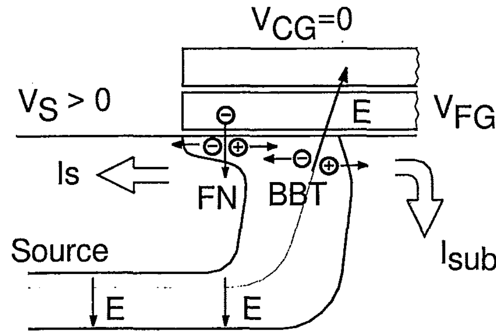


Figure 2.16 Pictorial description of Band-to-Band Tunneling (BBT) and Fowler-Nordheim (FN) currents ($V_S > 0$, and $V_{CG} = 0\,V$) occurring in the n^+ source junction of a Flash cell [30].

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:

$$E_{ox} \approx \frac{|V_{FG} - V_S|}{T_{ox}} \tag{2.26}$$

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:

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

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

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

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

APPLE INC.
EXHIBIT 1008 - PAGE 0066

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

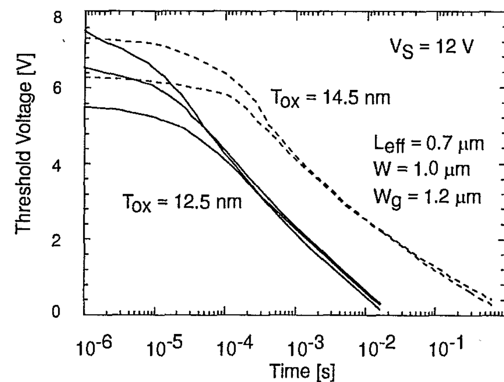


Figure 2.17 Erase curves of two Flash cells having different oxide thicknesses, and same L_{eff}.

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 100 ms–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}):

$$\gamma_{\text{er}} = \frac{I_G}{I_{\text{SUB}}} \tag{2.29 }$$

APPLE INC.
EXHIBIT 1008 - PAGE 0067

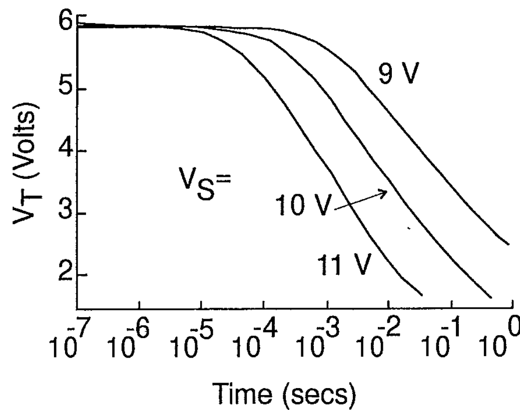


Figure 2.18 Erase curves of a cell with $T_{ox} = 12\,\mathrm{nm}$, when different source voltages are applied [30]. .

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 $1.2\,\mu\mathrm{m}$ technology, while today production is peaked on the $0.35\,\mu\mathrm{m}$ technology (Fig. 2.20), comparing with the $0.25\,\mu\mathrm{m}$ 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 $0.25\,\mu\mathrm{m}$ and $0.18\,\mu\mathrm{m}$ technologies.

APPLE INC.
EXHIBIT 1008 - PAGE 0068

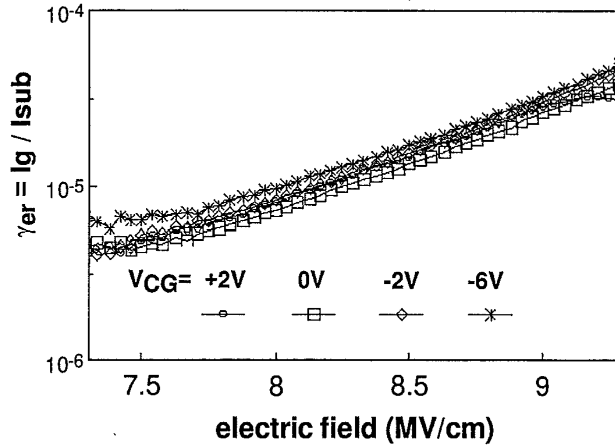


Figure 2.19 Erase efficiency as a function of electric field, when different gate voltages are applied.

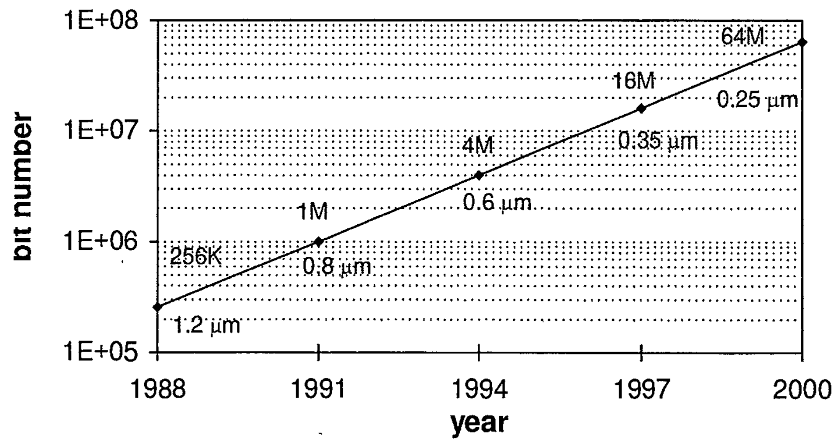


Figure 2.20 Evolution of Flash products in production.

APPLE INC.
EXHIBIT 1008 - PAGE 0069

The basic process flow for actual Flash device (16 Mb in 0.35 μm) 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

APPLE INC.
EXHIBIT 1008 - PAGE 0070

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

To this aims different LOCOS (L) schemes, proposed for 0.5 μm to 0.25 μm 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.

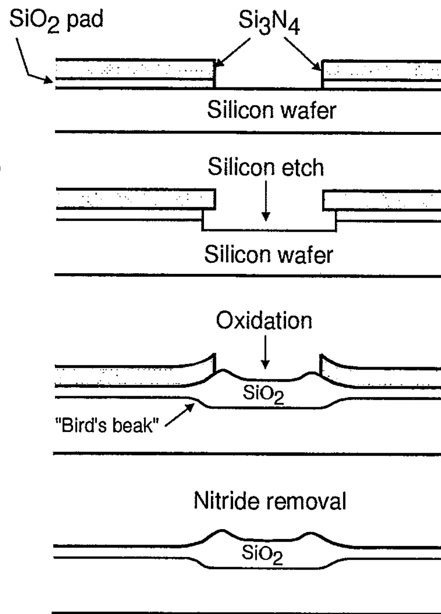


Figure 2.21 Process steps to obtain recessed LOCOS isolation.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0071

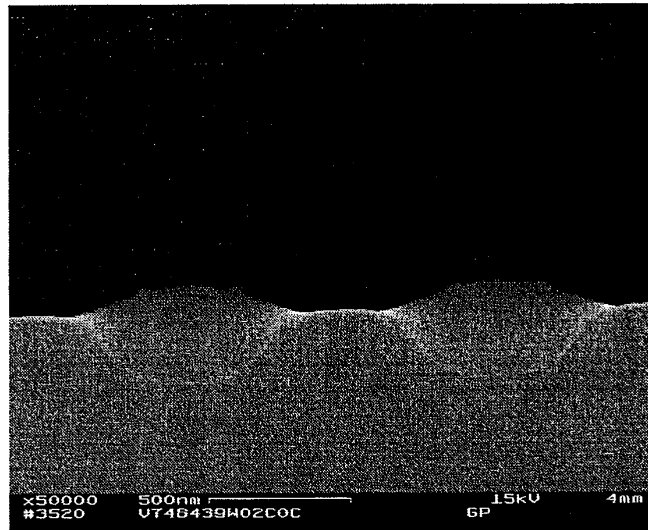


Figure 2.22 SEM micrograph of recessed LOCOS isolated devices.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0072

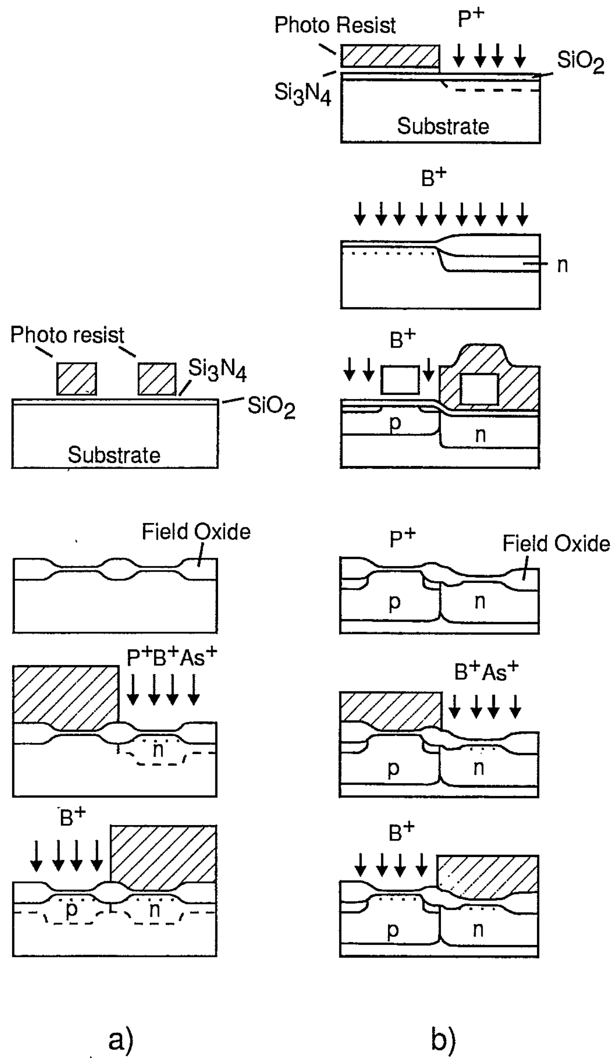


Figure 2.23 Isolation techniques: a) retrograde well, and b) conventional process flows.

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].

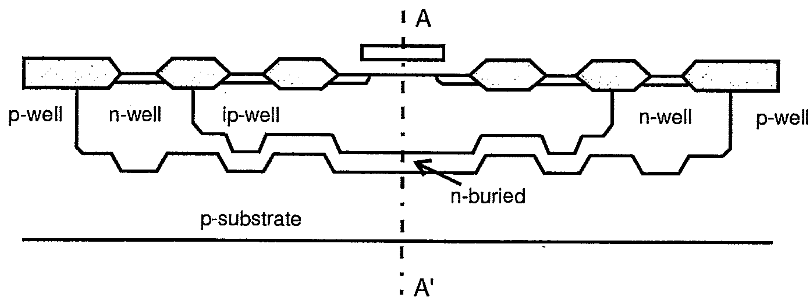


Figure 2.24 Triple well structure [41].

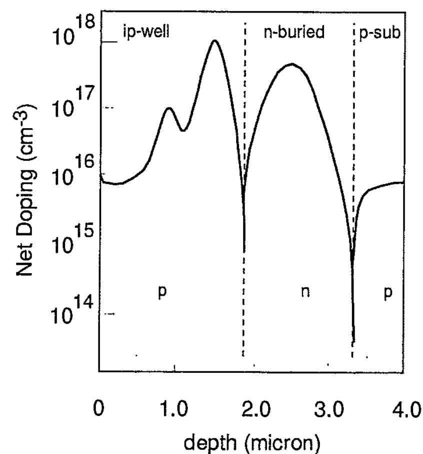


Figure 2.25 Doping profile of a triple well structure along A-A' section of the previous figure [40].

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

APPLE INC.
EXHIBIT 1008 - PAGE 0074

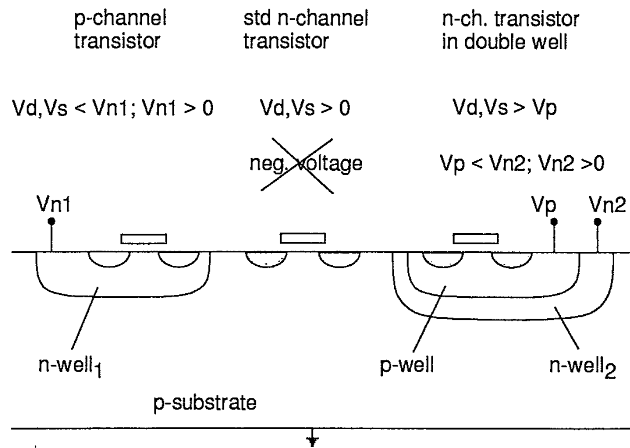


Figure 2.26 Schematic cross section of transistors in a triple well process [40].

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

APPLE INC.
EXHIBIT 1008 - PAGE 0075

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 ad 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 ($> 1100°C$), 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0076

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

APPLE INC.
EXHIBIT 1008 - PAGE 0077

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-

APPLE INC.
EXHIBIT 1008 - PAGE 0078

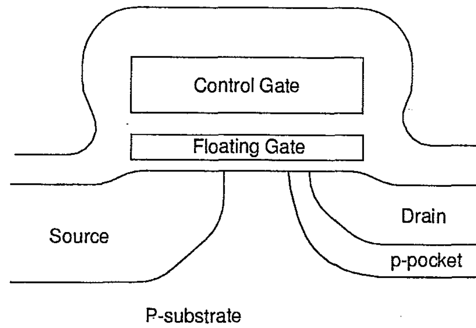


Figure 2.28 Cross section of a generic Flash device with p-pocket.

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].

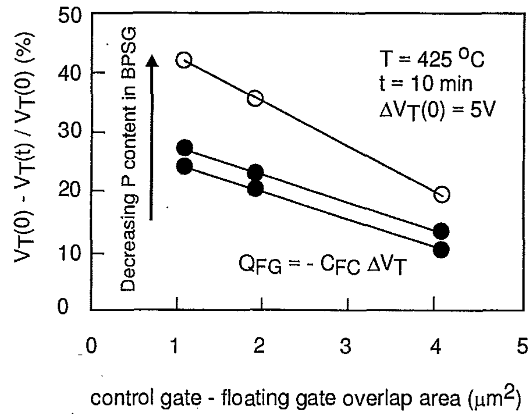


Figure 2.29 Charge loss as a function of control gate/floating gate overlapping area in cells with the same gate oxide thickness but different P concentration in BPSG [55].

APPLE INC.
EXHIBIT 1008 - PAGE 0079

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-

APPLE INC.
EXHIBIT 1008 - PAGE 0080

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-

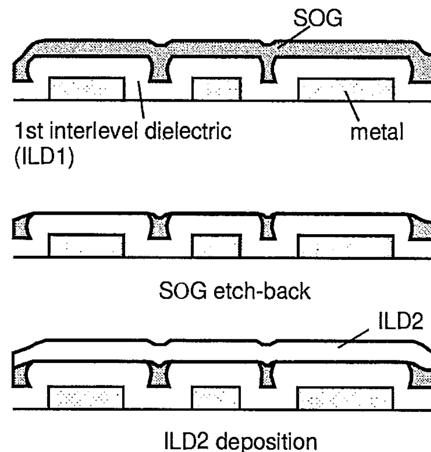


Figure 2.30 Planarization techniques: Spin on Glass (SOG).

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0081

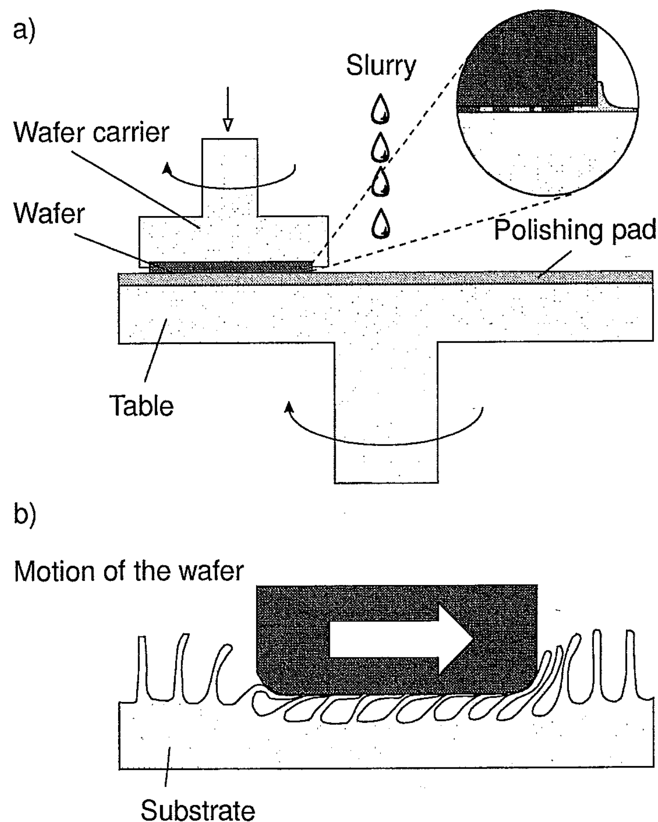


Figure 2.31 Planarization techniques: Chemical Mechanical Polishing (CMP). a) Schematic of the machinery; b) detail of the wafer surface during the procedure [58].

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 $1.7\,\mu\Omega$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 clorine 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].

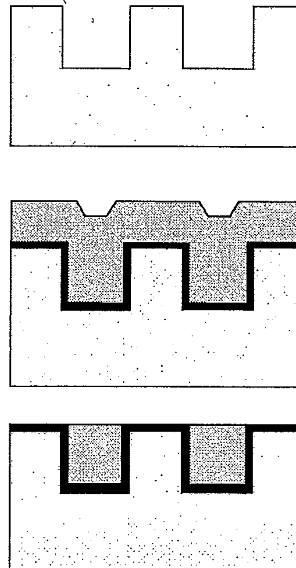


Figure 2.32 Sequence of damascene technique; top: trench definition and etch; middle: copper deposition; bottom: interconnections after chemical-mechanical-polishing [60].

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0083

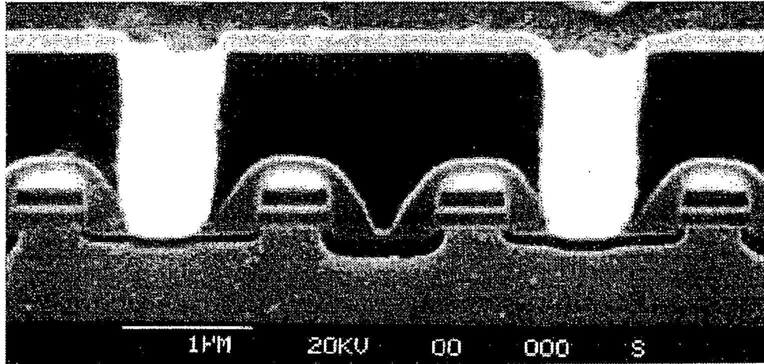


Figure 2.33 SEM micrograph of a few cells in an array.

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0084

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 $0.4\,\mu$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).

APPLE INC.
EXHIBIT 1008 - PAGE 0085

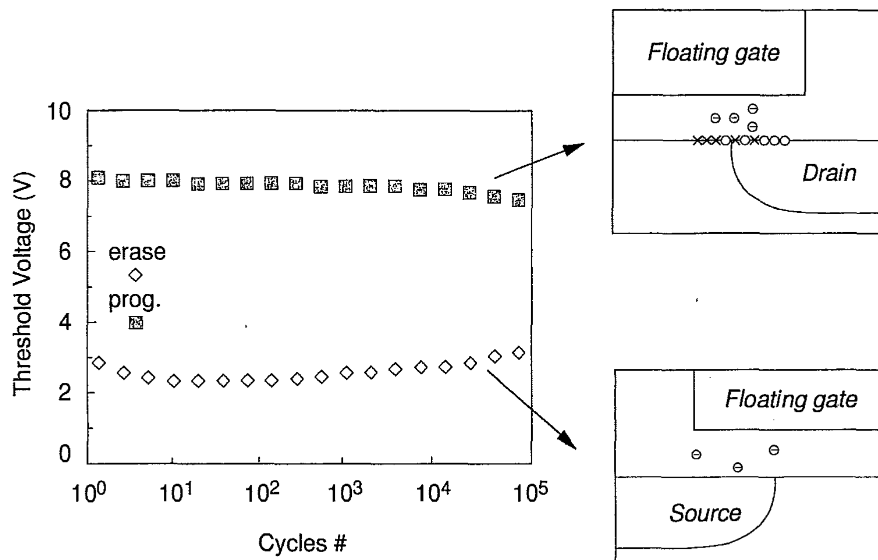


Figure 2.34 Threshold voltage window closure as a function of program/erase cycles on a single cell [61].

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0086

2.5.4 Programming Disturbs

Considering an array as in Fig. 2.35, if we want to program the highlighted transistor, a high voltage ($V_{pp} = 12\,V$) is applied to the WL, and a sufficiently high voltage ($V_{dd} = 5 - 7\,V$) 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.

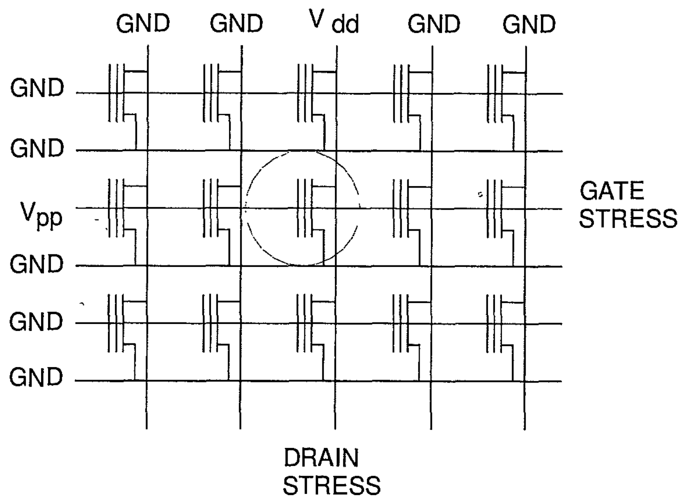


Figure 2.35 Programming disturbs.

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0087

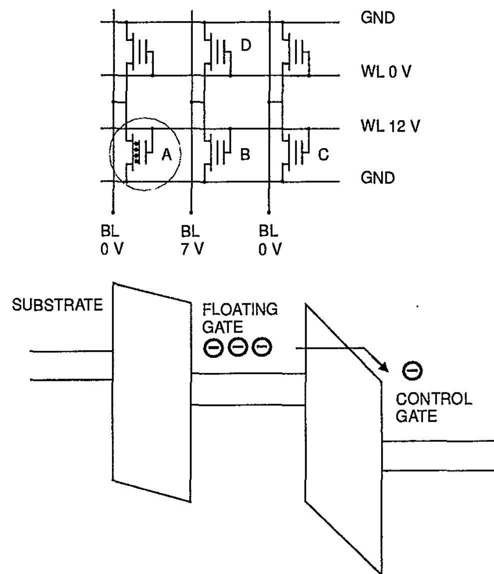


Figure 2.36 Programming disturbs - DC erasing of a programmed cell. Cell A is programmed, cell C is non-programmed.

A relatively high voltage ($V_{dd} = 5 - 7\,\mathrm{V}$) 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 1 Mbit array, a thousand repetitions. Disturb influence becomes more and more important on increasing the number of reading-programming cycles, or programming-erasing cycles.

APPLE INC.
EXHIBIT 1008 - PAGE 0088

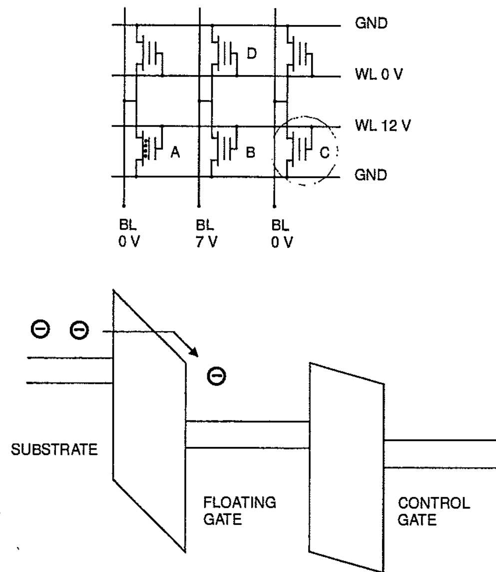


Figure 2.37 Programming disturbs - DC programming of a non-programmed cell. Cell A is programmed, cell C is non-programmed.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0089

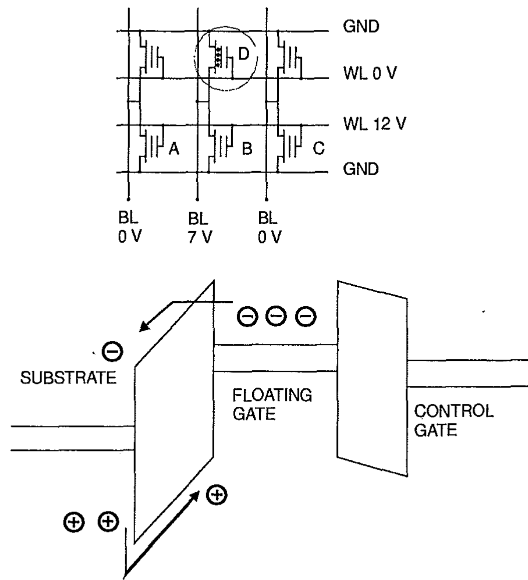


Figure 2.38 Programming disturbs - Drain disturb. Cell D is programmed.

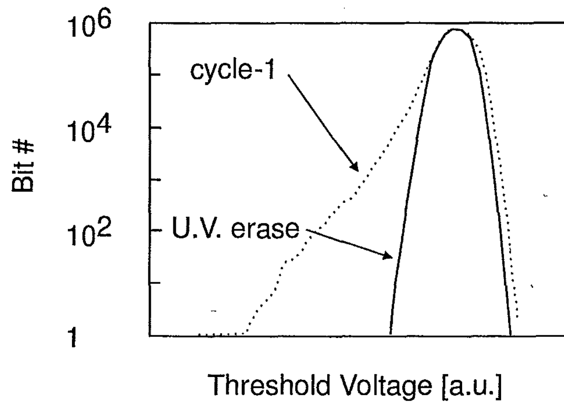


Figure 2.39 Threshold voltage distribution after different erase procedures: UV erase (solid curve), after the first cycle (dotted line) [61].

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-

APPLE INC.
EXHIBIT 1008 - PAGE 0090

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

APPLE INC.
EXHIBIT 1008 - PAGE 0091

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

APPLE INC.
EXHIBIT 1008 - PAGE 0092

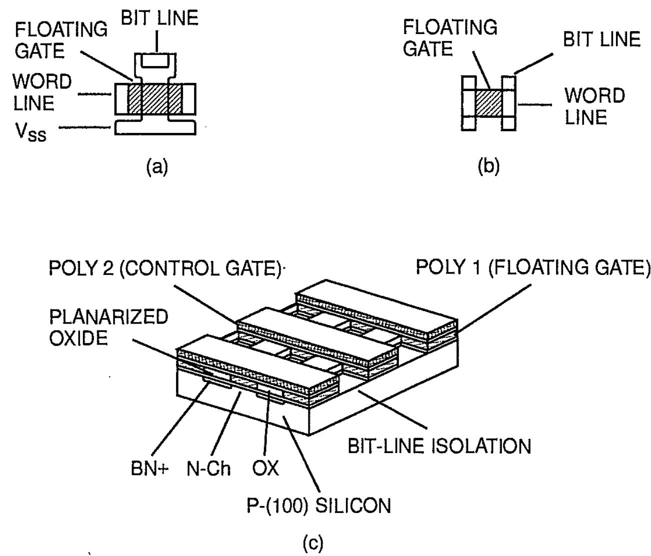


Figure 2.40 Layout of a T-shaped staked gate Flash cell (a). A contactless (Virtual Ground) cell is also shown in (b) and (c) [73].

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,*

APPLE INC.
EXHIBIT 1008 - PAGE 0093

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 E^2PROM 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 μm 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 μm 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0094

[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 μm 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 subthreshold 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0095

[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.

APPLE INC.
EXHIBIT 1008 - PAGE 0096

[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 256 Mbit DRAM cells". *IEDM Tech. Dig.*, pp. 279–282.

[38] Pfiester J.R., Kenkare P.U., Subrahmanyan 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 4 Mb 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0097

[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 SiO_2 films in N_2O 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 N_2O 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 SiO_2 by oxidizing Si substrate in N_2O". *Applied Physics Letters*, **57**, p. 26.

[48] Bellafiore N., Pio F. and Riva C. (1993) "Thin oxide nitridation in N_2O 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 E^2PROM 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0098

[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. Kranemburg (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.

APPLE INC.
EXHIBIT 1008 - PAGE 0099

[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.

APPLE INC.
EXHIBIT 1008 - PAGE 0100

3 BINARY AND MULTILEVEL FLASH CELLS

Boaz Eitan[1], Anirban Roy[2]

[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

APPLE INC.
EXHIBIT 1008 - PAGE 0101

requested by the customer. Product access times are in the range of 25–100 nS, programming time of 10 μs/byte and erase time of a few hundred milliseconds to few seconds. The supply voltage in operation is typically 5 V±10%, with many emerging requirements for 3 V±10% and even 2 V±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.

APPLE INC.
EXHIBIT 1008 - PAGE 0102

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

APPLE INC.
EXHIBIT 1008 - PAGE 0103

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 Vt is a fixed parameter based on cell coupling ratios and floating gate threshold voltage. The electrical erased Vt 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0104

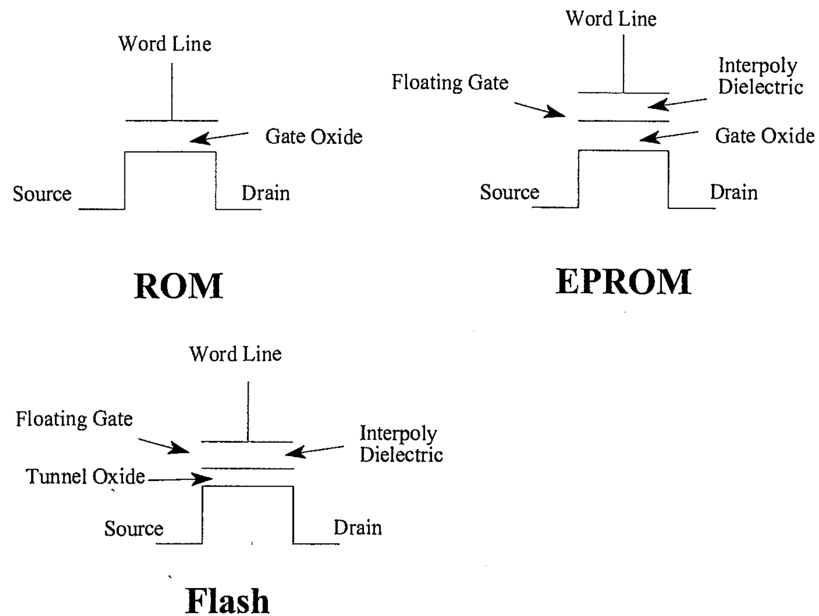


Figure 3.1 Device schematics for ROM, EPROM and Flash cells.

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 Vt 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0105

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 standalone 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-EPROM, 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

APPLE INC.
EXHIBIT 1008 - PAGE 0106

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 Vt 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 tayloring 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0107

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 Vt 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 Vt or a low Vt 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 Vt) 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 Vt (same as IEEE standard), while the DINOR and AND arrays have "programmed" state which is low Vt (so the IEEE standard calls this state "erased").

APPLE INC.
EXHIBIT 1008 - PAGE 0108

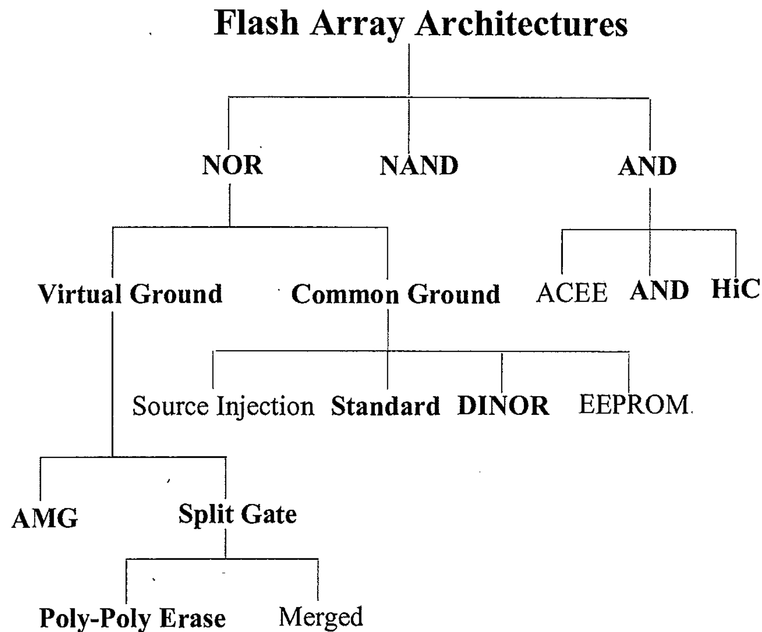


Figure 3.2 The family tree of Flash array architectures. The NOR architecture variations are the majority of the concepts. The Common Ground is the predominant concept in production today. The ACEE (Array Contactless EEPROM) and HiC (High Coupling ratio) are among the many variants of the Flash cells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

APPLE INC.
EXHIBIT 1008 - PAGE 0109

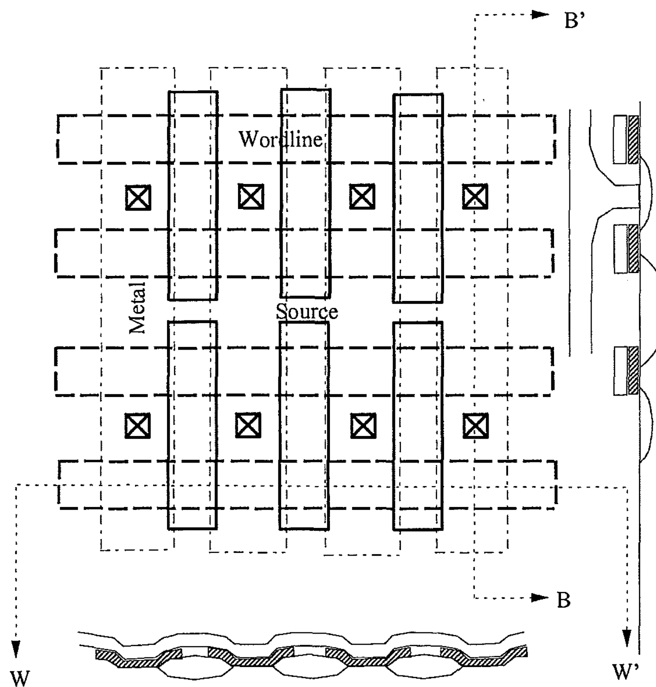


Figure 3.3 The schematic layout of the industry standard common ground NOR array [3]. The polycide wordlines traverse left to right (dashed borders in the figure), while the metal bitlines (dash-and-dot borders) traverse top to bottom. Notice the sharing of bitline contact between two cells and the definition 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

APPLE INC.
EXHIBIT 1008 - PAGE 0110

size. The 0.6 μm cell size is 3.6 μm^2, the 0.5 μm generation is $\approx$ 2.8 μm^2 and the 0.35 μm generation is 1.2 μm^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

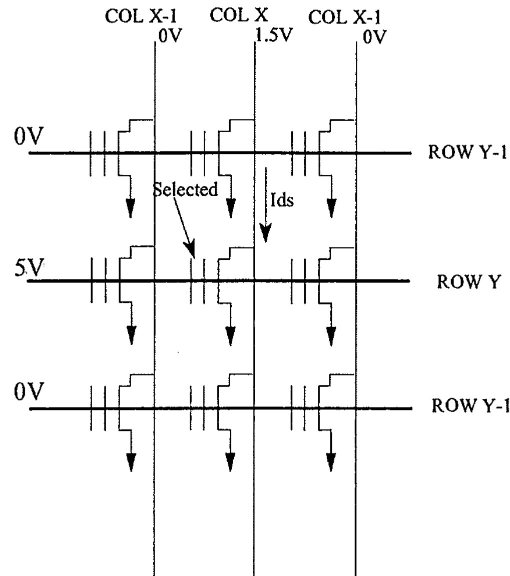


Figure 3.4 The device schematic of the layout in Fig. 3.3 [3]. The cell during read/programming is selected with a combination of wordline and bitline, while the erase is done for the whole block of the array.

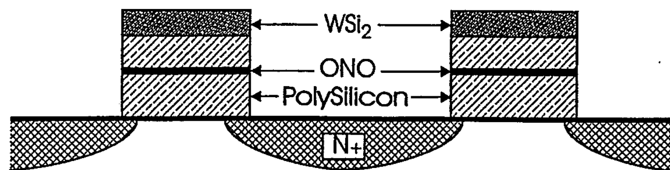


Figure 3.5 The basic process steps involved in fabricating the standard common ground NOR array [3]. The figure illustrates the misalignment critical masking required for the graded source junction.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0112

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 Vt distribution control is one of the most significant problems in implementing this Flash concept. In order to enlarge the tolerance to the Vt distribution, the maximum Vt 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**

APPLE INC.
EXHIBIT 1008 - PAGE 0113

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 $10\,F^2$ to $6\,F^2$. 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.

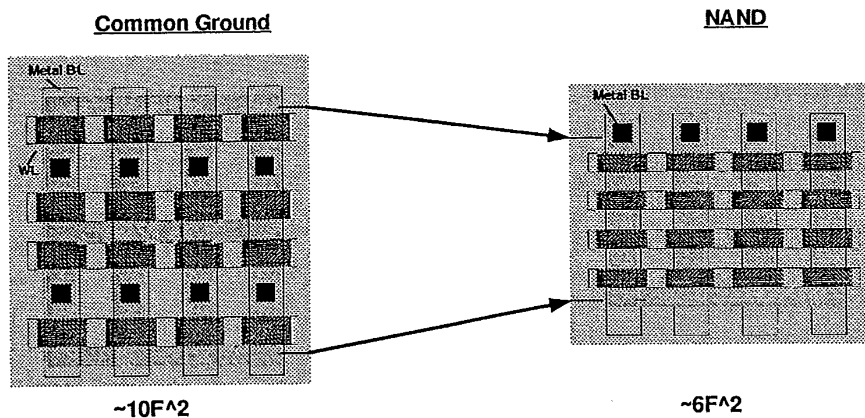


Figure 3.6 Illustration of the wordline pitch scaling in the NAND array. The cell size scales down from a $10\,F^2$ for the standard common ground to $6\,F^2$ for the NAND [1]. Note the reduced number of contacts.

APPLE INC.
EXHIBIT 1008 - PAGE 0114

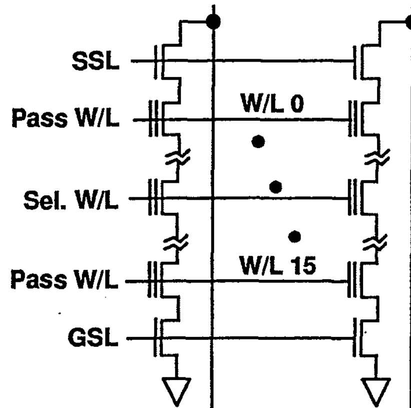


Figure 3.7 The device schematic diagram of the NAND array. The group of 16 cells are connected to the global bitline with select transistors and the selected cell access requires the neighboring cells to provide a read through [14].

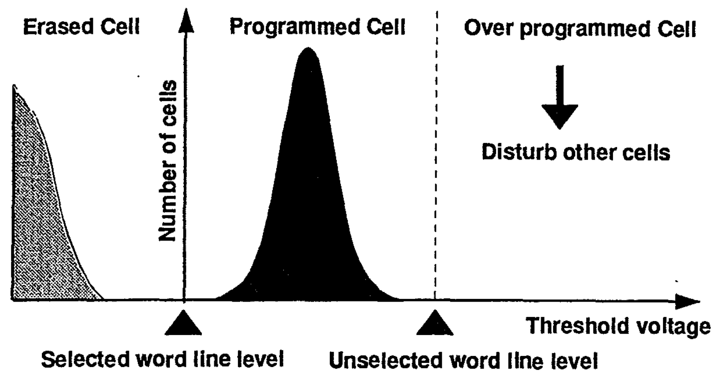


Figure 3.8 The programmed and erased array Vt distribution for the NAND array. Over-programmed cells can potentially affect the read access of the other cells [13].

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

APPLE INC.
EXHIBIT 1008 - PAGE 0115

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 μs. 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 Vt. Erasing requires biasing the p-well and the n-substrate with a high positive voltage ($\approx$ 20 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 Vcc, 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0116

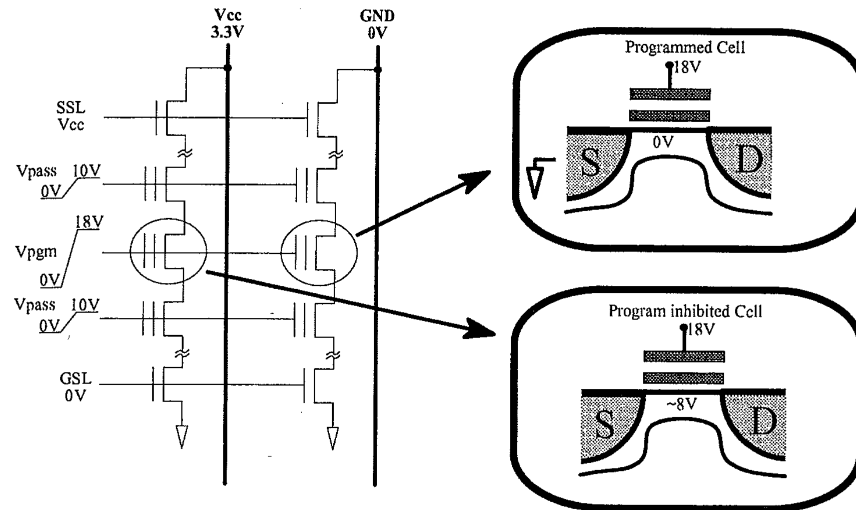


Figure 3.9 Dynamic program inhibit concept for the NAND array. Deep-depletion under the floating gate of the cells belonging to the wordline being programmed reduces the voltage stress for the gate disturb [14].

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

APPLE INC.
EXHIBIT 1008 - PAGE 0117

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 16 Mbit 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 (DIvided 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

APPLE INC.
EXHIBIT 1008 - PAGE 0118

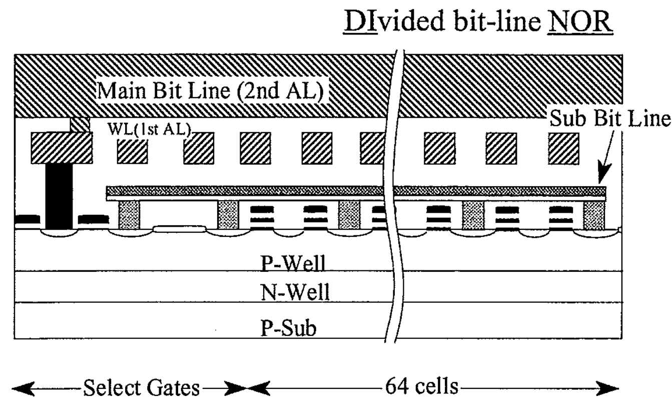


Figure 3.10 DINOR array device cross-section showing the triple well, triple poly and double metal process required to realize this array [17]. There are salicided local bitlines and the first metal is used for wordline strapping.

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 Vt, is usually called programming by DINOR developers. Erasing is the opposite operation,

APPLE INC.
EXHIBIT 1008 - PAGE 0119

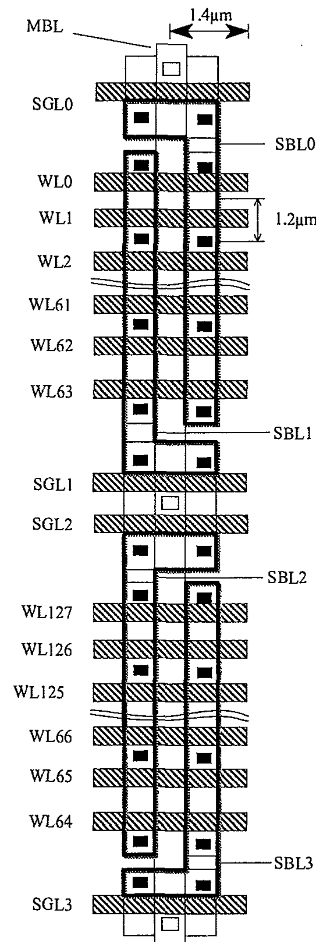


Figure 3.11 Schematic layout of the DINOR array [17]. The select transistors connect the global second metal lines to the salicided local bitlines. There is no double-diffused junction, and this helps in scaling the array pitch.

carried out via FN tunneling from the channel to the floating gate; it results in high Vt 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

APPLE INC.
EXHIBIT 1008 - PAGE 0120

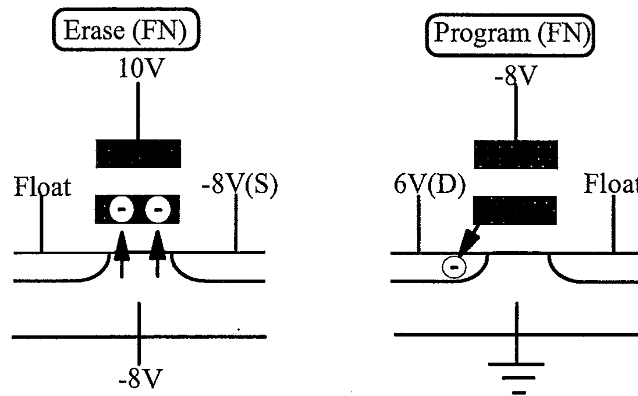


Figure 3.12 The erase and programming tunneling operations in the DINOR array [17]. Notice the programmed Vt is the low Vt state, a major departure from the standard common ground.

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 $1.68\,\mu m^2$ in a $0.5\,\mu m$ 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 $\approx 20\,\mu s$ (see Fig. 3.13a). Since programming to a low Vt 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 Vt 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

APPLE INC.
EXHIBIT 1008 - PAGE 0121

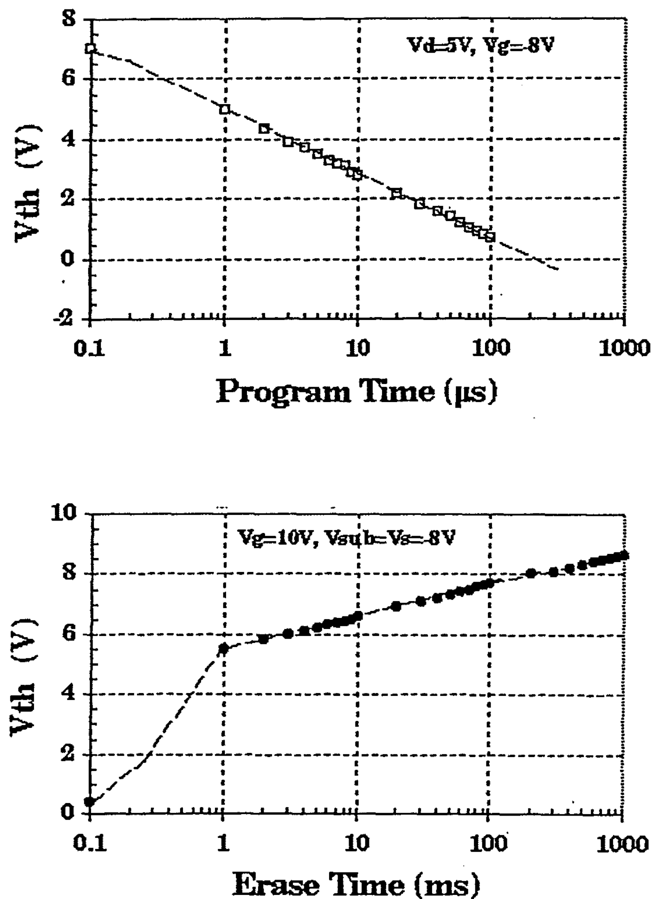


Figure 3.13 a) The programming time, and b) erase time characteristics for the DINOR array. To reduce the programming time as many as 256 bytes are programmed in parallel [18].

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 Vt; 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0122

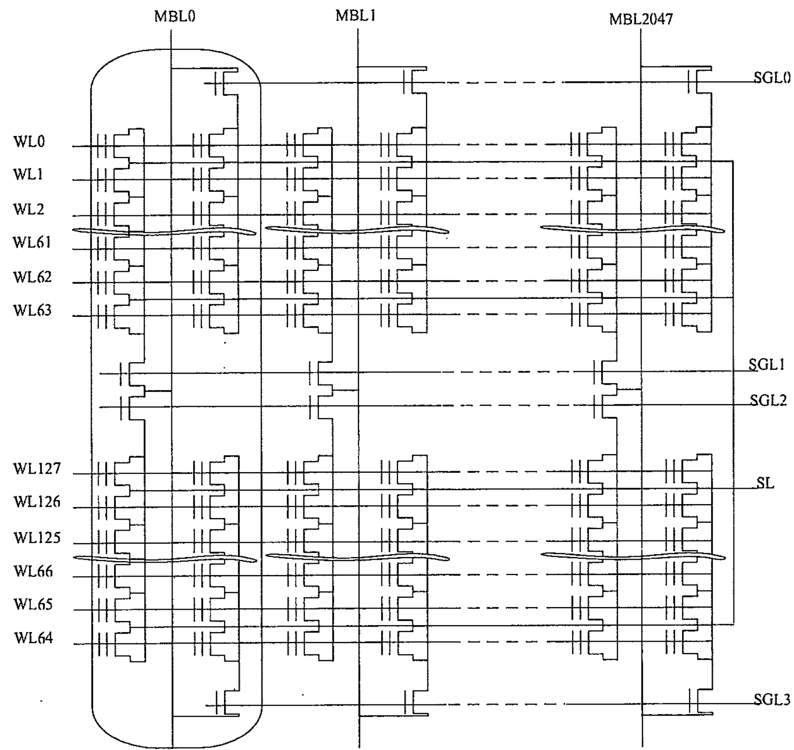


Figure 3.14 The DINOR array device schematic diagram showing blocks of 64 cells, accessible via select transistors [17].

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 Vt. 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-

APPLE INC.
EXHIBIT 1008 - PAGE 0123

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

APPLE INC.
EXHIBIT 1008 - PAGE 0124

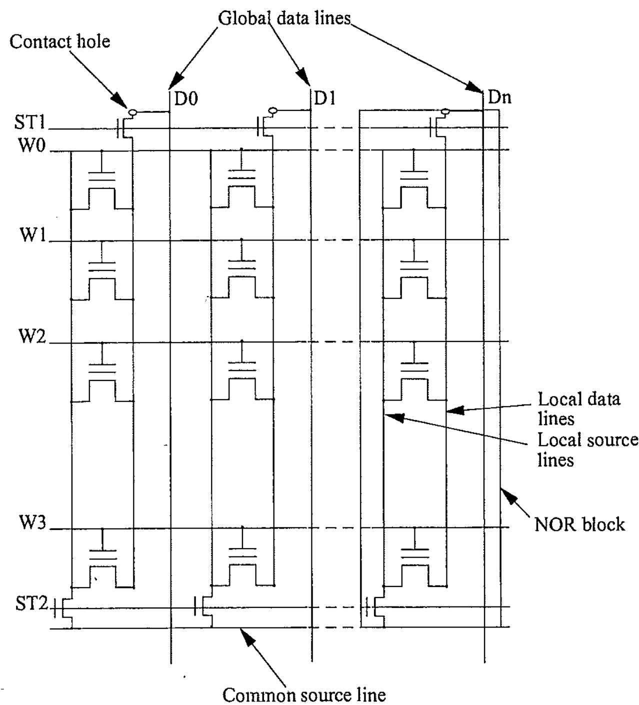


Figure 3.15 AND array device schematic showing 32 cell blocks accessible with select transistors. There are local diffusion bit lines and source lines [22]. The sector overhead for the 64 cell sector is 30%.

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 Vt) 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 Vt. This definitions are the same as DINOR, and the opposite of the industry standard NOR array. Typically, programming is accomplished

APPLE INC.
EXHIBIT 1008 - PAGE 0125

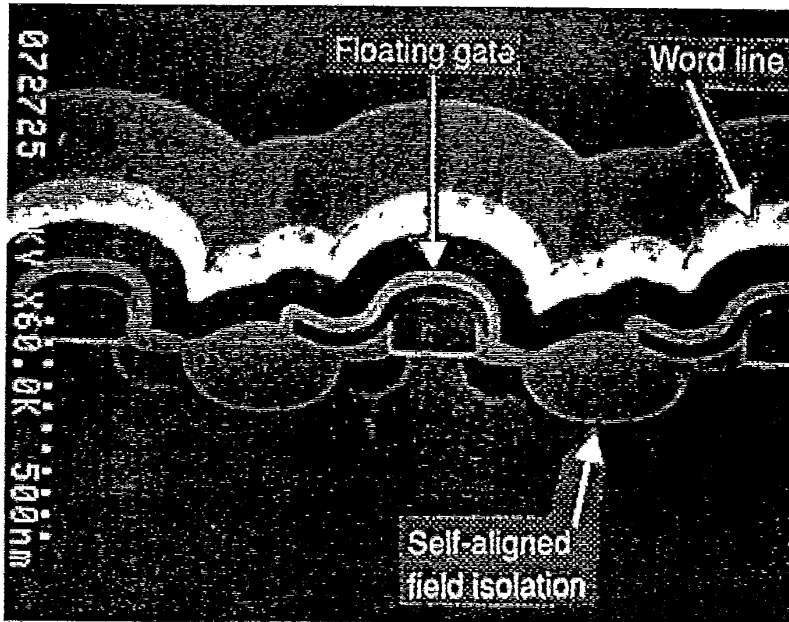


Figure 3.16 AND array device cross-section showing the self aligned field isolation, and the increased gate coupling with triple poly technology [22]. The wordline RC of the array is about 4 times than that of the common ground array.

with −9 V on the word line and +3 V on the local bit line (source floating), while erasing is carried out with bitline and source line grounded, and +13 V applied to the wordline. All voltages applied to the wordline are internally generated starting from Vdd = 3 V. Using program and verify, the low Vt 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

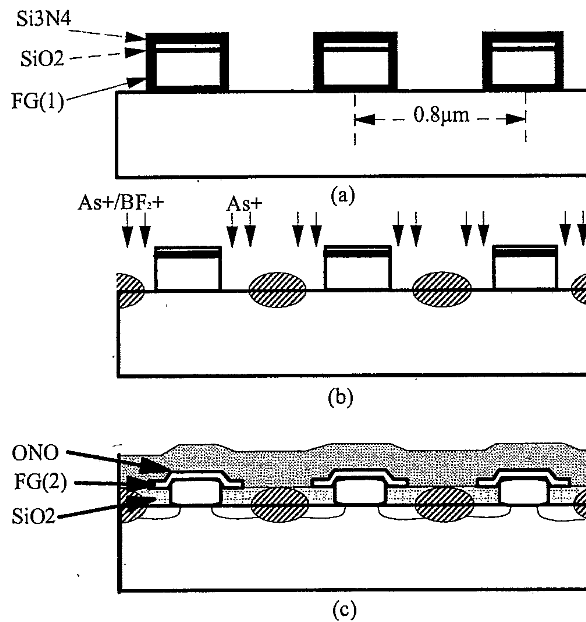


Figure 3.17 Few of the process steps in the formation of the AND array [22]. The process steps worthy of note are: a) double poly floating gate, b) tunnel oxide formation before the array field oxide, c) very thin poly-poly dielectric.

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 $8\,\text{F}^2$, and a $0.4\,\mu\text{m}^2$ cell for $0.25\,\mu\text{m}$ technology at 6.4F^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.

APPLE INC.
EXHIBIT 1008 - PAGE 0127

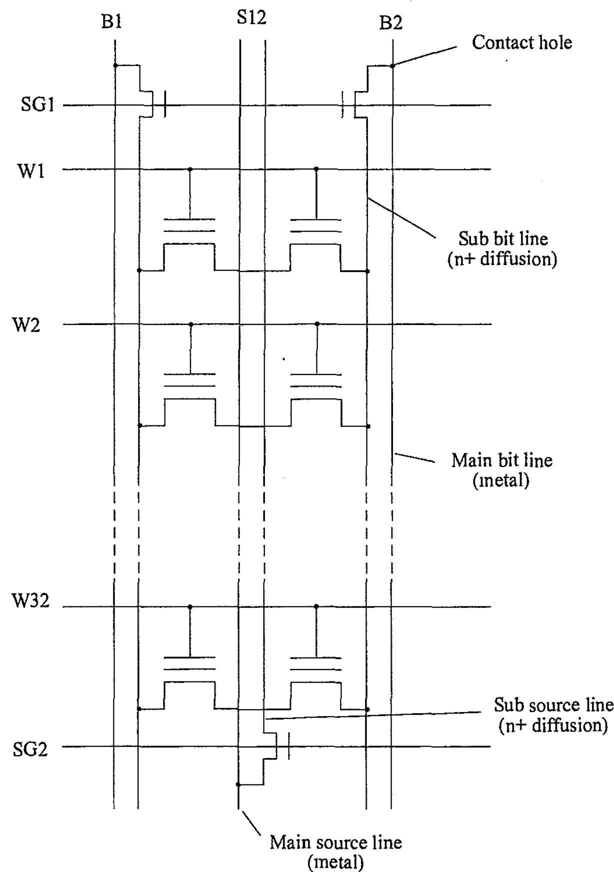


Figure 3.18 Device schematic for a modified AND array [23]. There is a dedicated source and contacts to bit lines and source lines are staggered to provide better scaling.

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0129

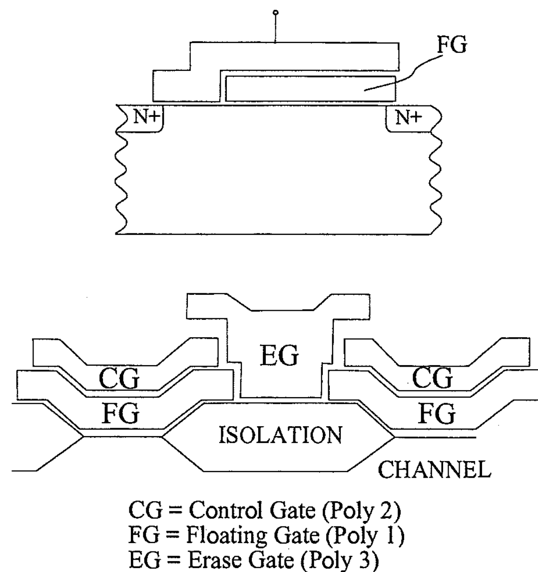


Figure 3.19 Triple poly split gate Flash cell device cross-sections: a) along the channel length, and b) along the channel width directions [28]. The first poly layer forms the floating gate, the second poly layer forms the wordlines and the third poly layer forms the erase line.

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 $0.55\,\mu$m technology the cell size is $2.1\,\mu\text{m}^2$ or $6.9\,\text{F}^2$.

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 interpolysilicon 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 Vt of the cell tends asymptotically to the Vt value of the transfer gate. While the floating gate Vt can be negative (overerased), the device as a whole is immune from this over erased state.

APPLE INC.
EXHIBIT 1008 - PAGE 0130

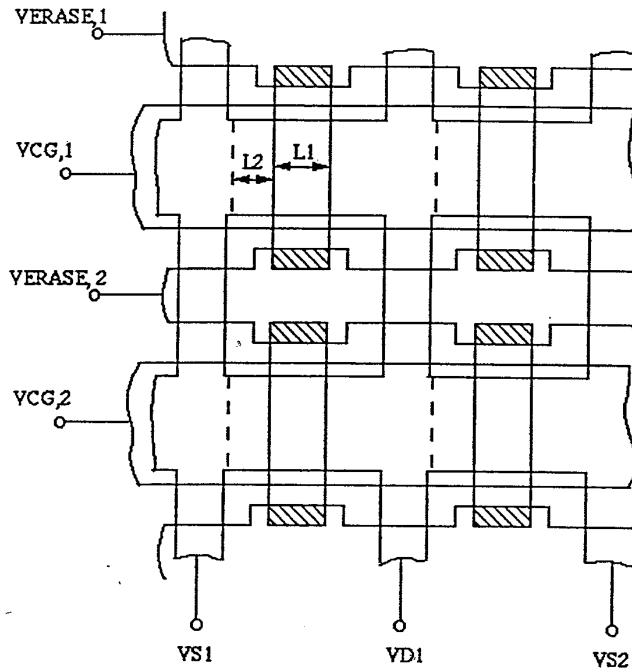


Figure 3.20 Cell layout of the triple poly split gate cell [29]. The erase line overlap on the floating gate is very misalignment sensitive.

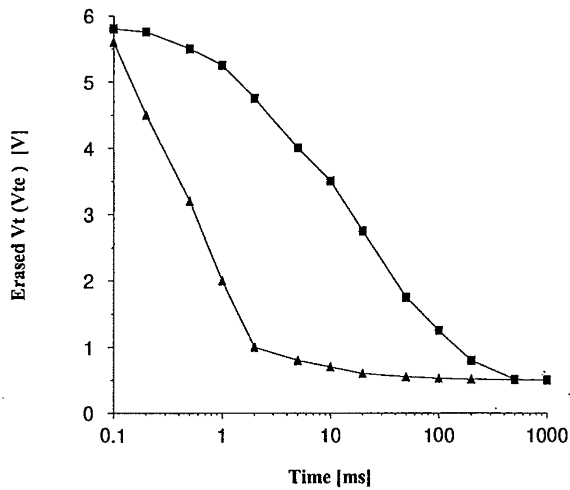


Figure 3.21 The saturated erase characteristics of the split gate cell [32]. The over-erase complications during read are avoided.

APPLE INC.
EXHIBIT 1008 - PAGE 0131

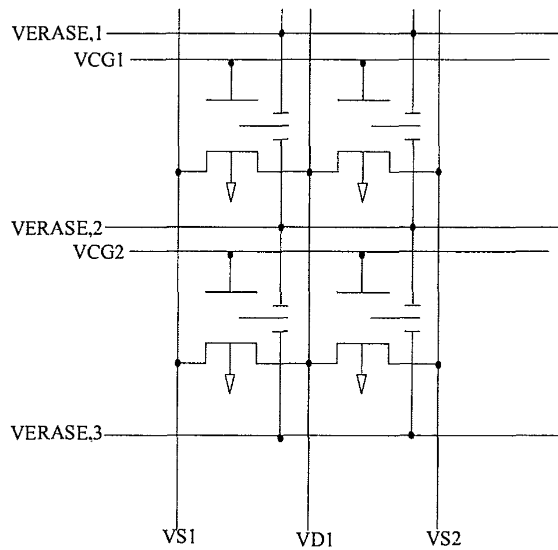


Figure 3.22 The array architecture schematic diagram of the split gate concept [29]. The array only requires positive voltages. The array also offers excellent natural segmentation, a feature useful for disk replacement products.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0132

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 ($4\,F^2$). 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-

APPLE INC.
EXHIBIT 1008 - PAGE 0133

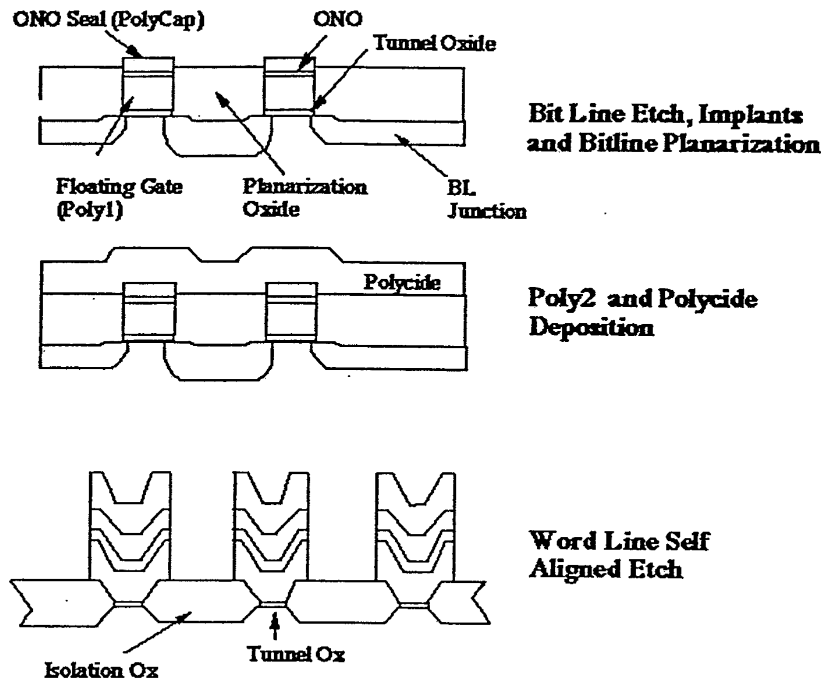


Figure 3.23 AMG Flash cell device cross-sections: a) along the channel length, and b) along the channel width directions [2]. The poly1 pitch is the limiting pitch along the channel length while the poly2 pitch (with gate coupling requirement) determines the pitch along the channel width.

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 $2.1\,\mu m^2$ in $0.6\,\mu m$ 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0134

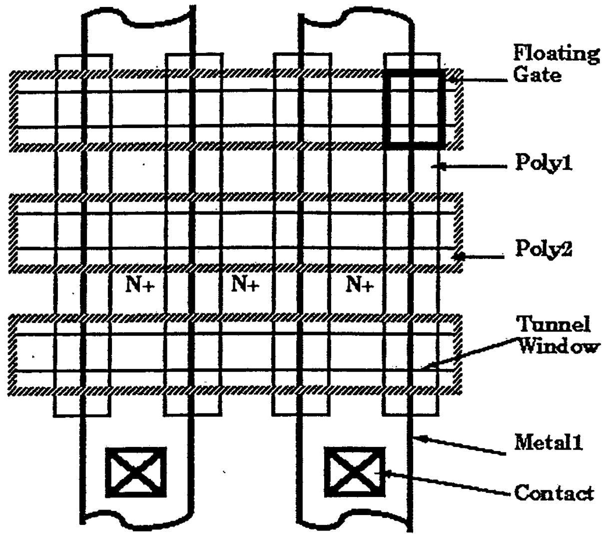


Figure 3.24 AMG Flash array layout [2]. Notice that each metal 1 bitline accommodates two n^+ diffusion bitlines.

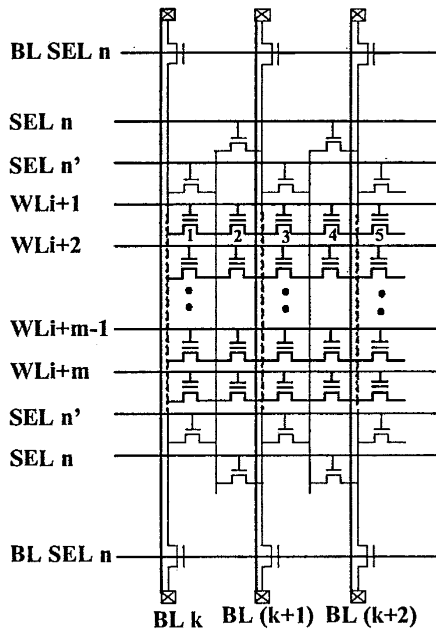


Figure 3.25 Schematic diagram of the AMG array architecture [35]. The natural segmentation of the array is provided with block select transistors BL SEL n, the AMG select transistors SEL n and SEL n′ provide connection to the drain junction.

APPLE INC.
EXHIBIT 1008 - PAGE 0135

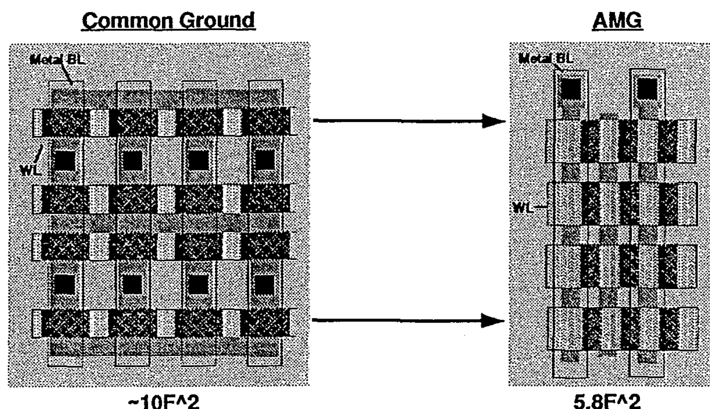


Figure 3.26 Comparison of the common ground and AMG arrays to illustrate the bitline scaling of the latter concept [1]. In the AMG concept, the number of contacts is significantly reduced and only one metal line is required for two bitlines.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0136

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

APPLE INC.
EXHIBIT 1008 - PAGE 0137

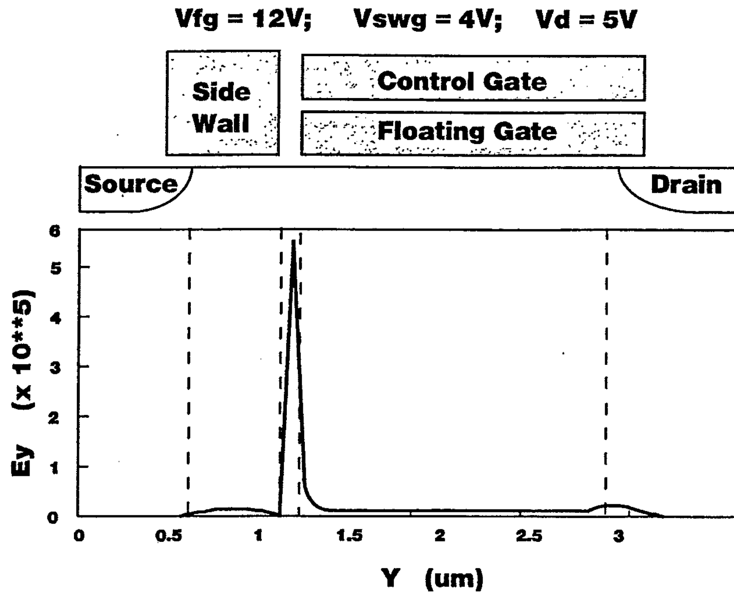


Figure 3.27 Split gate source injection concept is depicted [2, 39]. The electric field concentration on the source side of the floating gate is responsible for increased efficiency channel hot electron injection.

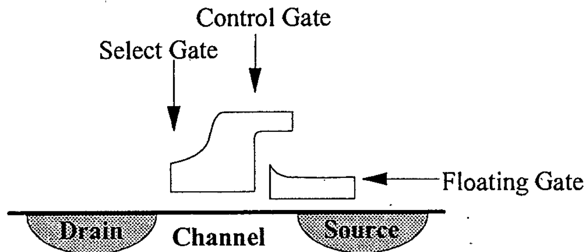


Figure 3.28 Device cross-section of the source injection common ground array. The programming is via source side injection while the erase is via poly-poly FN-tunneling from the sharply defined floating gate feature [39].

create the source lateral field enhancement. The cell in Fig. 3.28 is a 13.6 F^2 in 0.5 μm technology (3.4 μm^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

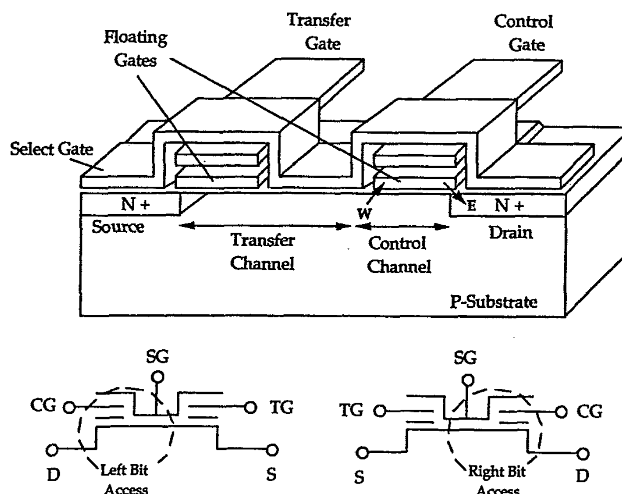


Figure 3.29 Device cross-section of the merged split gate source injection array [42]. Back to back split gates are constructed to share the third poly select gate for better cell size. Only one metal line is required for every two bitlines. The select gate controls the source injection.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0139

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0140

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?

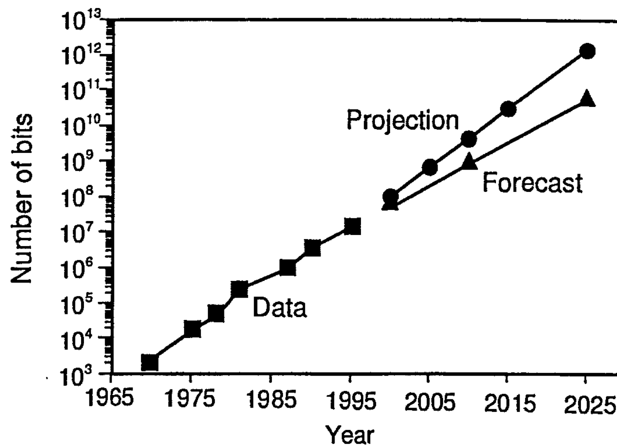


Figure 3.30 Flash density evolution curve. The historical growth rate is at about a doubling every 1.8 years.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0141

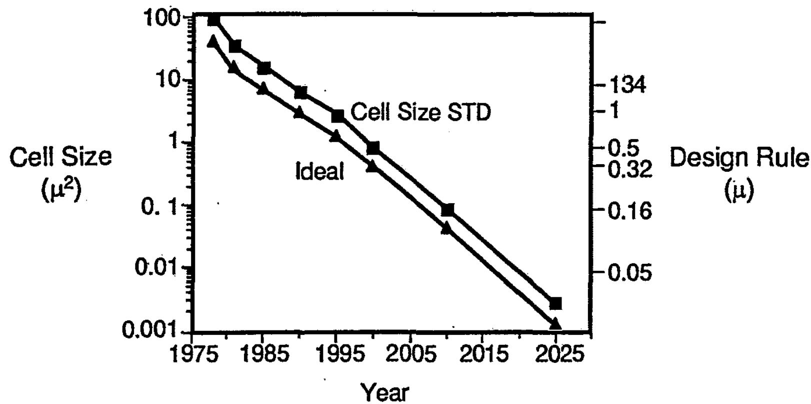


Figure 3.31 The history and future projection of cell size scaling for the common ground and ideal Flash arrays. The year 2025 is chosen to illustrate the demands on the process and device technology, if the present trend continues.

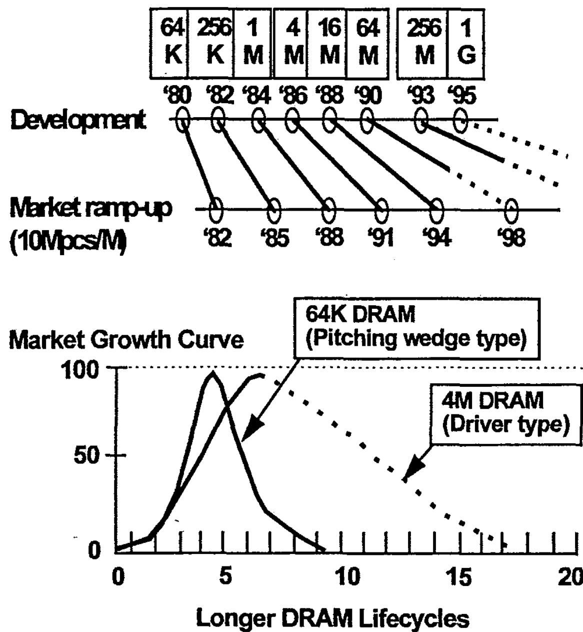


Figure 3.32 DRAM development and market ramp up time schedules and its impact on product lifecycles [46]. The product lifecycles in the future are expected to stretch longer.

the year 2025 results in density of 2×10^{12}, a cell size of 500 Å×500 Å, with a 180 Åfeature 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 Åresults 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 Vd-Vsub and Vg-Vsub 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

APPLE INC.
EXHIBIT 1008 - PAGE 0143

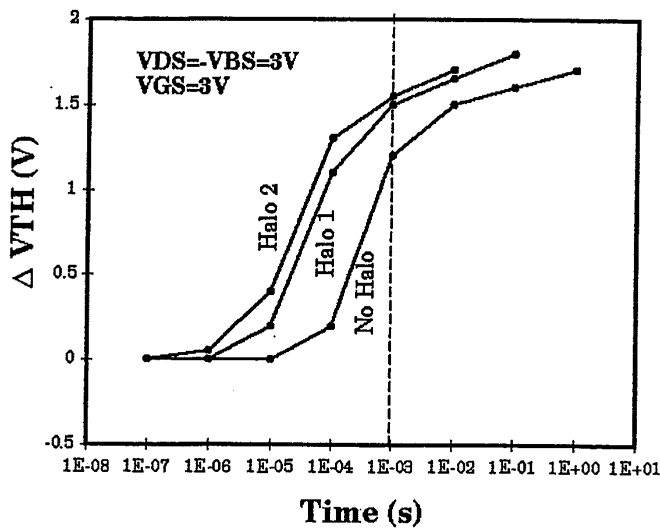


Figure 3.33 Secondary electron injection concept to provide $\pm 3\,\mathrm{V}$ programming in the millisecond time scale [47]. "Halo" refers to different boron halo doses ($5 \cdot 10^{13}$ for halo 1 and 10^{14} for halo 2, respectively) added to increase junction doping.

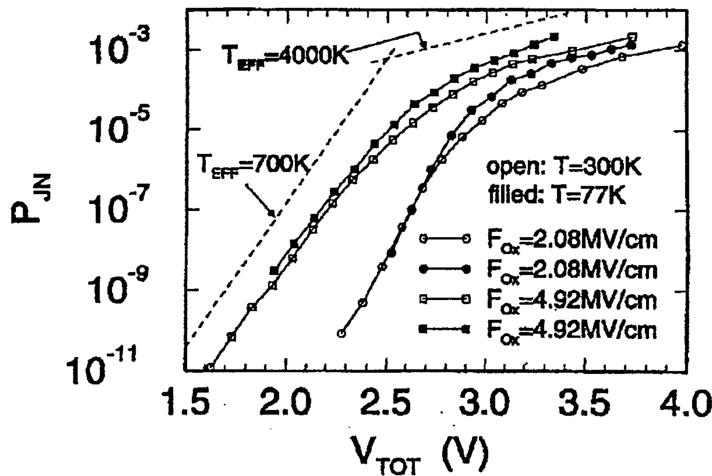


Figure 3.34 Very low accelerating voltages (Vds) Hot Electron programming [48]. P_{JN} represents the measured electron injection probability as a function of the total accelerating voltage. F_{Ox} is the oxide field.

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.

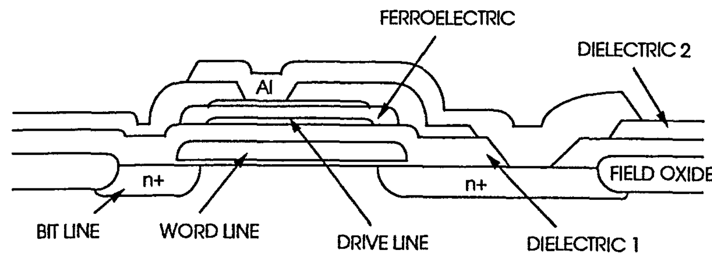


Figure 3.35 The ferroelectric array device cross-section secondary electron injection concept to provide ±3 V programming in the millisecond time scale [3]. The ferroelectric capacitor is integrated above the CMOS array and promises very high endurance. Ferroelectric memory read operation being destructive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

APPLE INC.
EXHIBIT 1008 - PAGE 0145

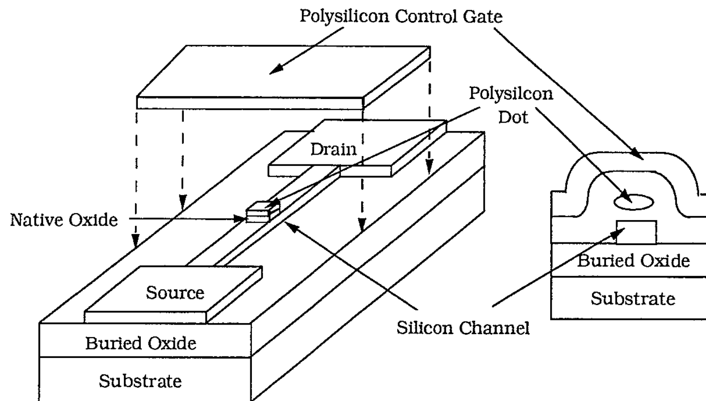


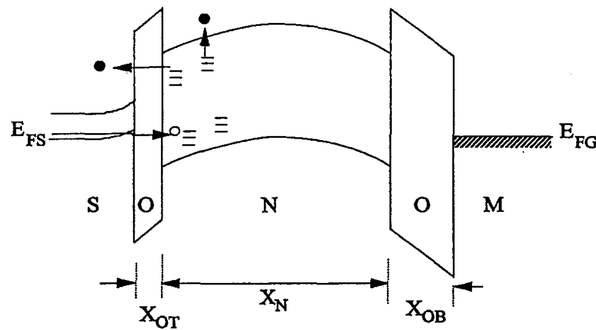
Figure 3.36 Demonstration of the single electron concept with a very narrow polysilicon floating gate structure [52]. The spatial localization of the electron wavefunction leads to quantized energy levels (and subsequently threshold voltage steps).

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 Vt. 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0146



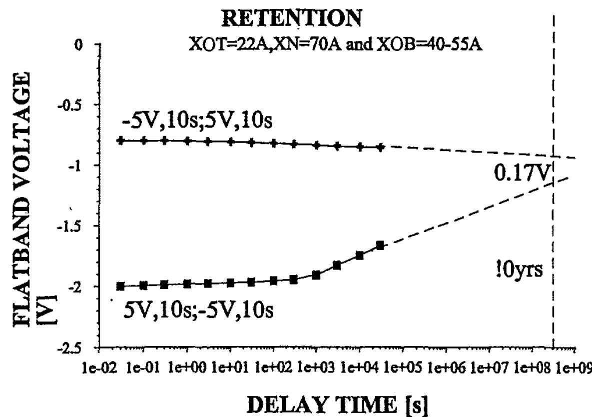


Figure 3.37 Scaled MNOS concept to achieve true 5 V only Flash concept [54]. The charges are trapped in the nitride layer to provide threshold voltage shifts. The bottom figure provides the results of a retention measurement extrapolated to ten years. X_{OT} stands for tunnel oxide, X_N is the nitride 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 Vt and a programmed Vt. For the NOR array, the read Vt window is defined as the difference of the maximum erased Vt and minimum programmed Vt. The read Vt 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 Vt states is shown. To incorporate four states, the total window is larger, while

APPLE INC.
EXHIBIT 1008 - PAGE 0147

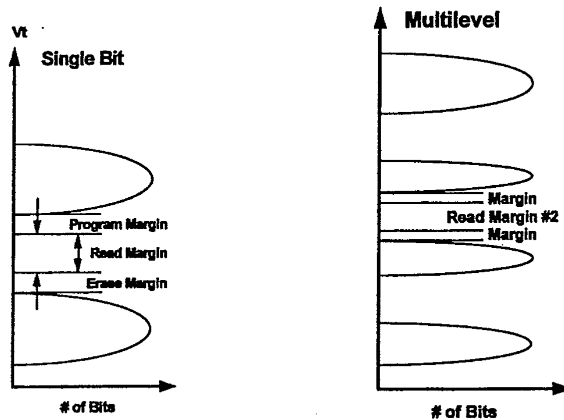


Figure 3.38 The program read and erase margins for a) single bit Flash cell, and b) multilevel Flash cell with 4 states [1].

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 Vt distribution per state. This is due to the need to minimize the total magnitude of the window. The requirement for a smaller Vt 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 Vt 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

APPLE INC.
EXHIBIT 1008 - PAGE 0148

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

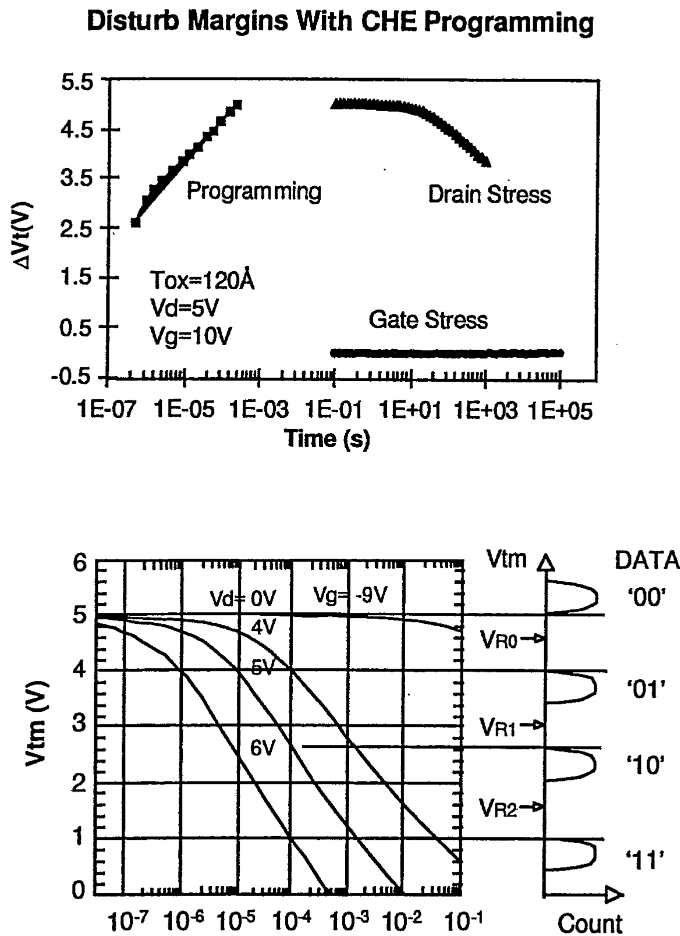


Figure 3.39 a) Drain and gate disturb margins for CHE programming mechanism [1], and b) the drain voltage enhanced gate disturb for F-N tunnel programming mechanism [26].

in the range of 10 μs or less. The programming time is defined as the time to obtain a delta Vt 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 Vt. 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 Vt distribution of each of the Vt 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 Vt 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0150

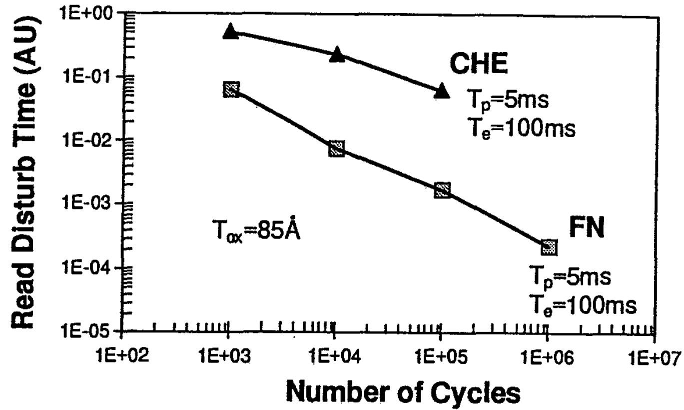


Figure 3.40 A comparison between the disturb due to channel hot electron and Fowler-Nordheim tunneling programming, as a function of program erase cycles [1]. This experiment is performed on the same product for both mechanisms.

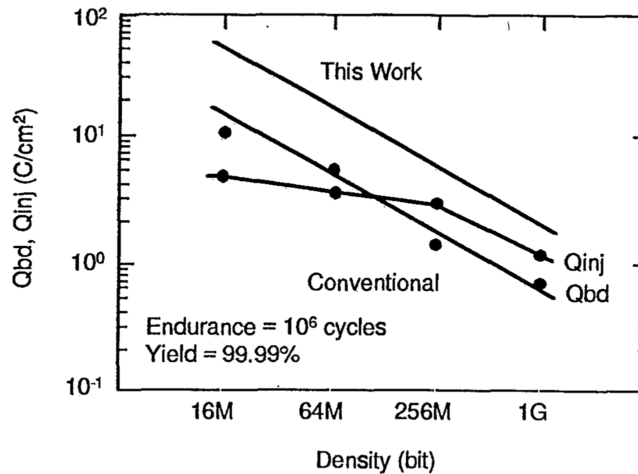


Figure 3.41 Qbd and Qinj as a function of memory density [5]. For very high density, the Qbd limit is reached.

The higher field during programming for the F-N programming enhances the charge to breakdown (Qbd) limitation for endurance. In Fig. 3.41 the Qbd and injected charge (Qinj) 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 Qinj 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0152

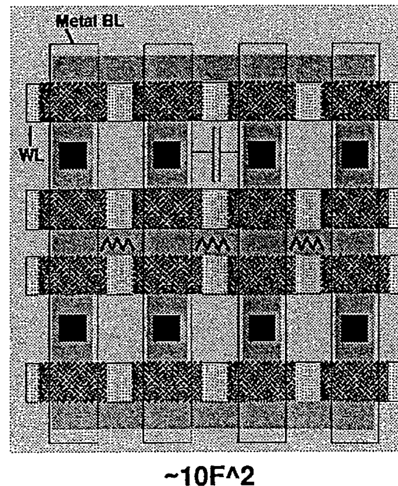


Figure 3.42 The source resistance and BL-BL coupling effects illustrated for a common ground array [1].

The series resistance of the source has an impact on the broadening of the Vt distribution of each of the Vt 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 Vgs and increases the Vbs. This is equivalent to a Vt 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 Vt 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 Vt 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 Vt 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0153

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 Vt 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 Vt 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 Vt 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 Vt 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 Vt distribution.

In the NAND architecture, the read-through concept induces, for each logic state, a non-negligible widening of the Vt distributions. The change in state of the other 15 neighbors increases the Vt distribution in the same way the source resistance affects the NOR. To partially compensate for it, there is a predetermined 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.

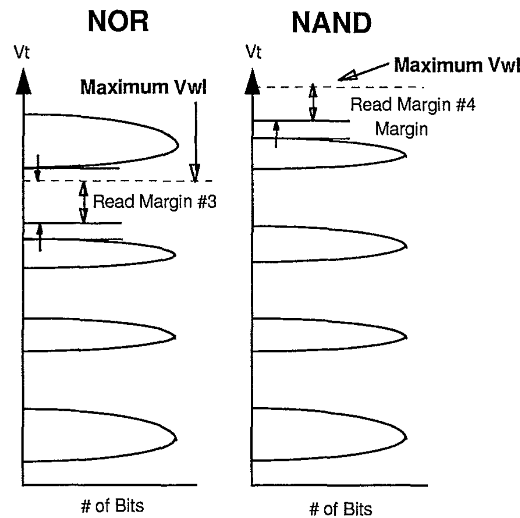


Figure 3.43 The illustration of the larger Vt window for the NAND array due to the read through requirement [1].

APPLE INC.
EXHIBIT 1008 - PAGE 0155

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 Vt window is a reflection of the read mechanism and the array impact on the Vt 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0156

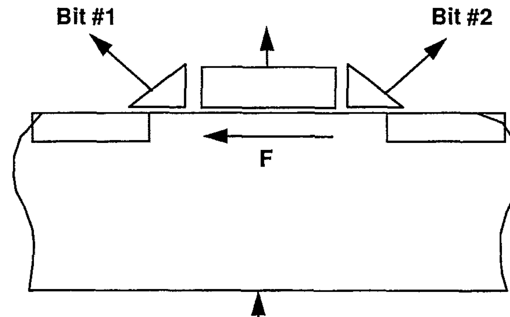


Figure 3.44 A simplified device cross-section for the multibit cell concept. Two different floating gates are merged in the same transistor [66].

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0157

[7] Kuo C. *et al.* (1994) "A 512Kb 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 μm 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0158

[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 μm^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 3 V-only 64 Mbit 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 140 mm^2 64 Mb AND Flash memory with a 0.4 μm 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 98 mm^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.,

APPLE INC.
EXHIBIT 1008 - PAGE 0159

Ueda K., Noro F., Miyamoto K., Morita M., Umeda K. and Kubo K. (1995) "A 34 Mb 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 F^2 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.,

APPLE INC.
EXHIBIT 1008 - PAGE 0160

Odierna P., Annunziata R., Geraci A., Paulino C., Campisi A. and Dallabora M. (1998) "4 Mb 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-Vcc 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0161

[50] Shoji K. *et al.* (1996) "A 7.03 μm^2 Vcc/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 multidielectric nonvolatile SONOS memory structures". *Solid State Elect.*, **37**, p. 1913.

[56] French M.L. *et al.* (1994) "Design and scaling of a SONOS multidielectric 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 32 Mb 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0162

4 PHYSICAL ASPECTS OF CELL OPERATION AND RELIABILITY

Luca Selmi[1], Claudio Fiegna[2]

[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 MOSFETs 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 E^2PROM 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-SiO_2 interfaces the floating gate behaves as a potential well capable to retain charge for long times. Negative

APPLE INC.
EXHIBIT 1008 - PAGE 0163

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.

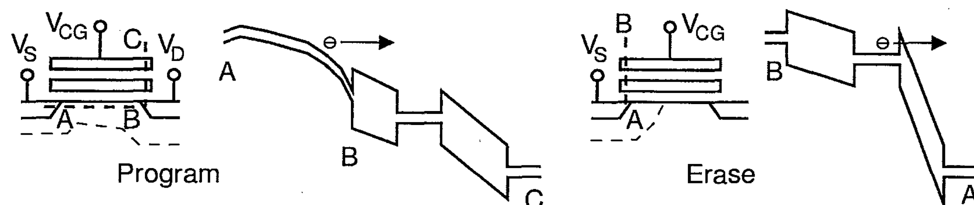


Figure 4.1 Schematic representation of programming and erasing mechanisms for the industry standard Flash cell.

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 SiO_2 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

APPLE INC.
EXHIBIT 1008 - PAGE 0164

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:

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

where $m_0 = 9.16 \times 10^{-31}$kg is the free electron mass, $\hbar = 1.05 \times 10^{-34}$ J·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:

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

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 Block wave:

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

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:

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

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:

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

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]:

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

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.

APPLE INC.
EXHIBIT 1008 - PAGE 0166

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 ($\vec{r}_{\rm wp}$) 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:

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

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:

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

and not the phase velocity of the wave packet component at $\vec{k} = \vec{k_{\rm wp}}$. $\vec{v}_g$ can be derived directly from the band structure $E(\vec{k})$ and can reach values in excess of 10^8cm/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 $E_G(\mathrm{Si}) \approx 1.12$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 $m^* \simeq 0.32 m_0$.

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 $\alpha-$quartz. $\alpha-$quartz features a perfectly ordered arrangement of Si atoms located at the center of tetrahedra, and oxygen atoms at the vertexes.

APPLE INC.
EXHIBIT 1008 - PAGE 0167

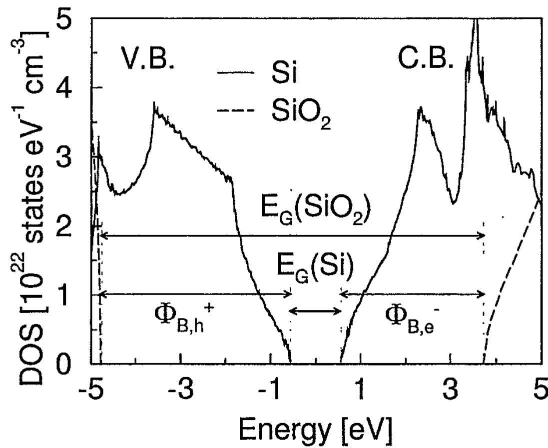


Figure 4.2 Density of states (DOS) of perfect Si and SiO_2 crystals (α-quartz). In this plot $E_G(\mathrm{Si}) = 1.12\mathrm{eV}$ and $E_G(\mathrm{SiO_2}) = 8.5\mathrm{eV}$. $E = 0$ at the Si midgap. Data from [2, 3].

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 $E_G(\mathrm{SiO_2})$ 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 $m^*_{\mathrm{ox}} \approx 0.5m_0$ [2].

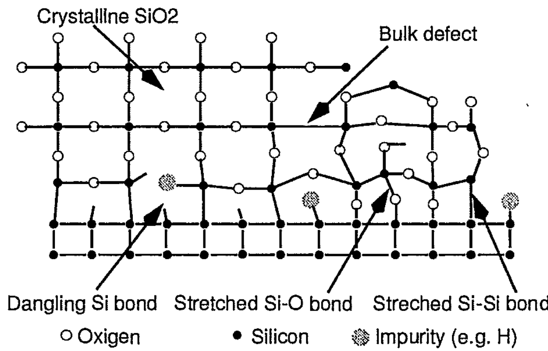


Figure 4.3 Schematic representation of the Si-SiO_2 interface and of defects leading to the formation of extra bulk and interface states.

SiO_2 films used as gate dielectric feature an amorphous (vitreous) form, meaning that the atomic structure is ordered only over short distances. Oxigen

APPLE INC.
EXHIBIT 1008 - PAGE 0168

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 (≈ 2.2 g/cm^3 instead of ≈ 2.65 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-SiO_2 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-SiO_2 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-SiO_2 interface all Si bonds are saturated by oxygen atoms. However, the compositional material change generates an abrupt built-in potential difference $\Phi_{B,e^-} \simeq 3.15$eV between the Si and SiO_2 conduction band edges, and of $\Phi_{B,h^+} \approx 4.3$eV between the valence band maxima. These barriers oppose carrier motion from Si to SiO_2 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-SiO_2 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].

APPLE INC.
EXHIBIT 1008 - PAGE 0169

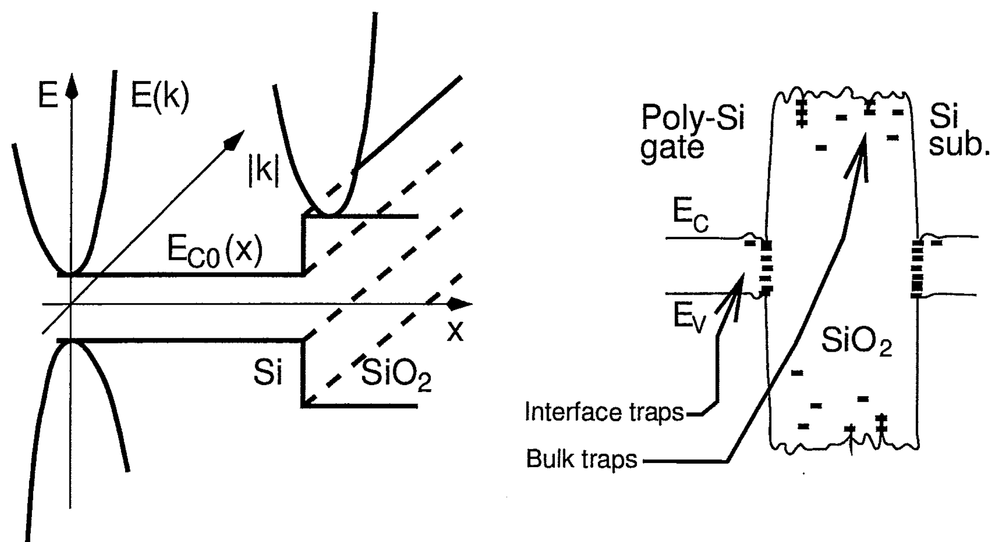


Figure 4.4 Left: Schematic representation of band structure transition at the Si-SiO_2 interface assuming $U(\vec{r}) = 0$. Right: Schematic representation of the perturbation potentials and of the interface and bulk states in a Si-SiO_2-Si stack.

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 SiO_2 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 ($\Delta x \approx 2\text{Å}$) can induce non-negligible threshold voltage shifts ($\Delta V_T = qD_{\mathrm{it}}(1 - \Delta x/t_{\mathrm{ox}})/C_{\mathrm{ox}} \approx 0.3\mathrm{V}$ and $\Delta V_T = qD_{\mathrm{it}}\Delta x/(t_{\mathrm{ox}}C_{\mathrm{ox}}) \approx 6\mathrm{mV}$, 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).

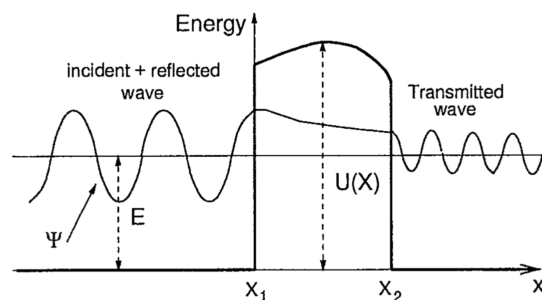


Figure 4.5 Transmission of an electron represented by the envelope function Ψ through a potential energy barrier. x_1 and x_2 are the so called classical turning points.

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]:

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

where $U(\vec{r})$ incorporates $E_{C0}(\vec{r})$. For a one dimensional barrier as that in the direction perpendicular to the Si-SiO_2 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]:

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

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 $E_{||} = (p_y^2+p_z^2)/2m^*$ are conserved along the tunneling path. Therefore, $E_{||}$ 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:

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

APPLE INC.
EXHIBIT 1008 - PAGE 0172

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

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

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

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

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:

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

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 $T \leq 1$.

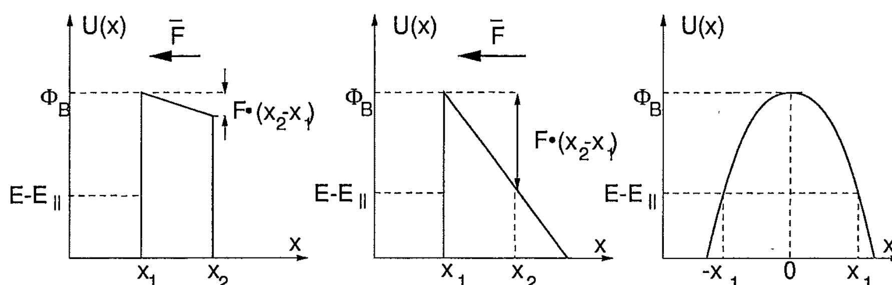


Figure 4.6 Trapezoidal, triangular and parabolic potential energy profiles.

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-SiO_2 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):

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

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

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

where F denotes the absolute value of electric field for the triangular or trapezoidal barriers, and $\xi = \Phi_B - (E - E_{||}) = \Phi_B - p_x^2/2m^* \geq 0$ 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 $\Phi_B < E - E_{||}$. For a fixed total energy E, ξ increases and T decreases exponentially as $E_{||}$ increases, because less energy and momentum are available in the tunneling direction.

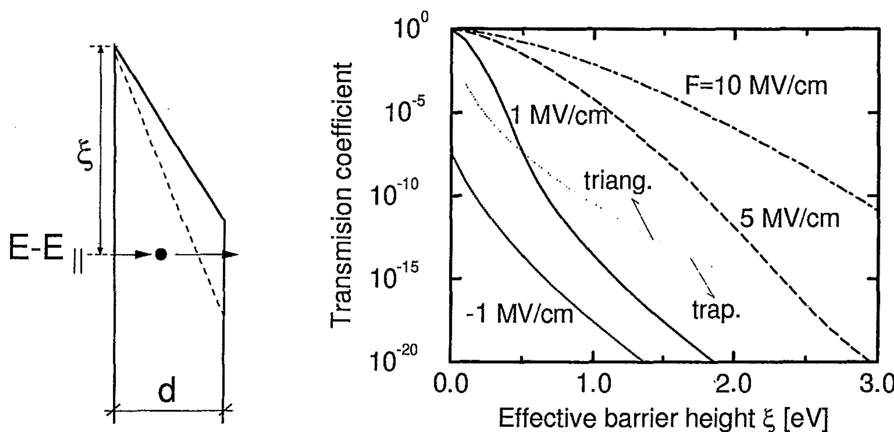


Figure 4.7 Transmission coefficient through an energy barrier as a function of the effective barrier height ξ for different field strengths and an barrier thickness $d = 5$nm. The transition between triangular and trapezoidal barrier shape is indicated by the dotted line.

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]:

$$dJ_{1\to 2} = q\frac{p_x}{m^*}Tg_1^*(E,p_x)f_1(E)\Big(1 - f_2(E)\Big)dp_x dE \quad , \qquad (4.18\ )$$

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:

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

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 $T_{1\to 2}$ is much smaller than $T_{2\to 1}$ 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

APPLE INC.
EXHIBIT 1008 - PAGE 0175

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},

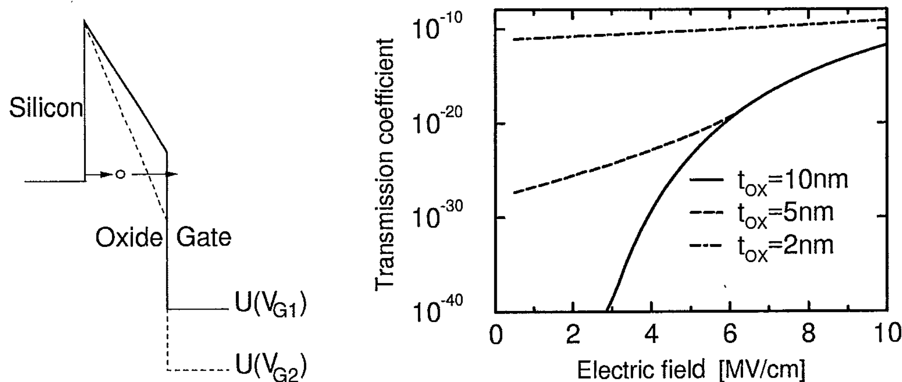


Figure 4.8 Left: Schematic representation of direct and Fowler-Nordheim tunneling conditions. Right: Transmission coefficient as a function of electric field for $\xi = 3.15$eV.

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 (< 3nm) 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 $\xi = 3.15$eV as in a MOS system. As can be seen, the tunneling probability does not depend on oxide thickness in the Fowler-Nordheim regime ($F_{OX}t_{OX} \geq \xi$). 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 ≈7–8nm [14]. For $t_{ox} \geq 7$nm, detectable currents can be observed only for $F \geq 6$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, $F \approx 10$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 ($F_{OX} = (V_G - V_{FB} - \Phi_{Si} - \Phi_P)/t_{ox}$), 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. $E - U = \hbar^2 k^2/2m^*_{ox}$ [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].

APPLE INC.
EXHIBIT 1008 - PAGE 0177

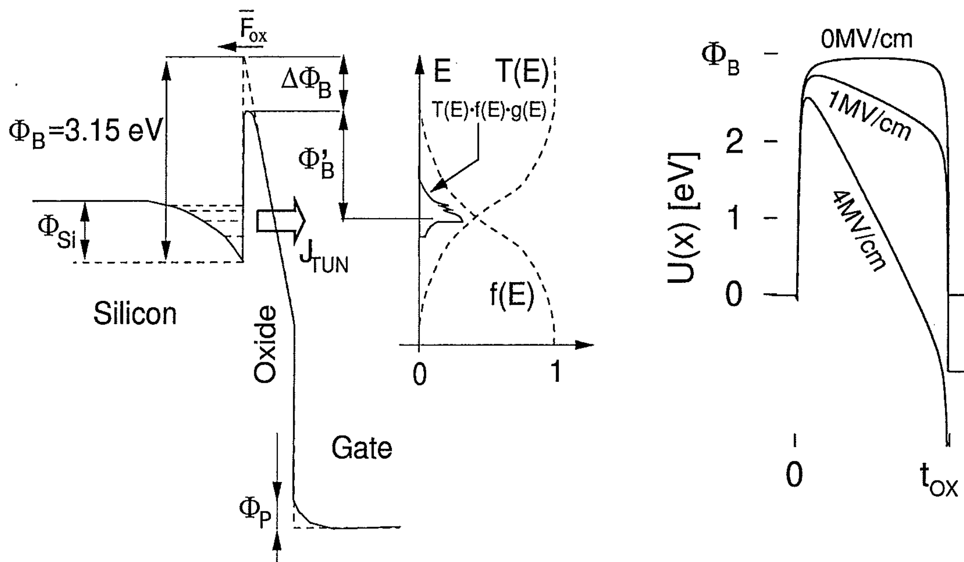


Figure 4.9 Left: Band diagram of a MOS structure in the FN regime. Dashed lines represent the quantized energy levels in the injecting electrode and the ideal barrier profile. Center: Schematic representation of the distribution function f, the transmission probability T and the product $f \cdot g \cdot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

- 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 ($\epsilon \to \infty$), 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 ($0 < x < t_{\text{ox}}$) is given by [18]:

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

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 $t_{\text{ox}} > 5\text{nm}$, as is the case in Flash technologies. By neglecting it, we can estimate the image force induced reduction of the effective barrier height as $\Delta\Phi_B = \sqrt{q^3 F_{\text{OX}}/4\pi\epsilon_\infty}$. 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 $\Phi'_B = \Phi_B - (d_0 + d_1|F_{\mathrm{OX}}|)$ where $d_0 = 0.112\mathrm{eV}$ and $d_1 = 1.69 \times 10^{-8}\mathrm{eV\ cm/V}$ 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 $T_L = 0\mathrm{K}$ (so that f_1 is a simple step function), and neglecting the reverse current from the oxide to the silicon, we get [27]:

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

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.

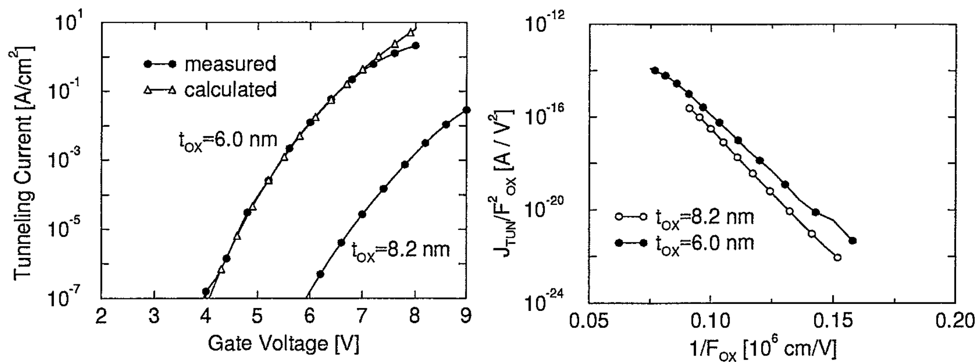


Figure 4.10 Left: Experimental tunneling current through thin gate oxides as a function of gate voltage. The current calculated with (4.21) and $\Phi'_B = 2.85\text{eV}$ is also shown for comparison. Right: Fowler-Nordheim plot of the J-V characteristic shown on the left.

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 $J_{\text{TUN}}/F^2_{\text{OX}}$ 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 ($S_{\text{FN}} = d(\log J_{\text{TUN}}/F^2_{\text{OX}})/d(1/F_{\text{OX}}) = -4\sqrt{2m^*_{\text{ox}}}\Phi'^{3/2}_B \log_{10} e$) 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].

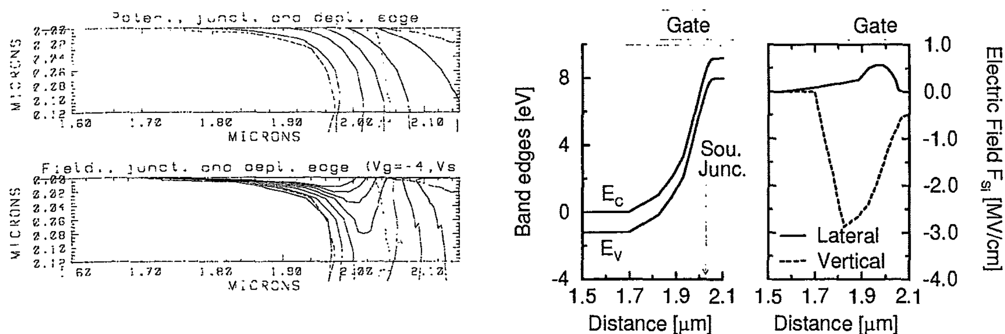


Figure 4.11 Left: simulated potential (upper plot), electric field (lower plot) and depletion region contours (dashed line) at the source side of a Flash memory cell for a typical erase bias ($V_G = -4$V, $V_S = 8$V). The Si-SiO$_2$ interface is located at $y = 0$, the gate extends for $x \geq 1.82\mu$m. Notice the deeply depleted layer in the gate-source overlap region on the left of the junction (represented by the dotted line at $x \simeq 2.05\mu$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

APPLE INC.
EXHIBIT 1008 - PAGE 0181

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.

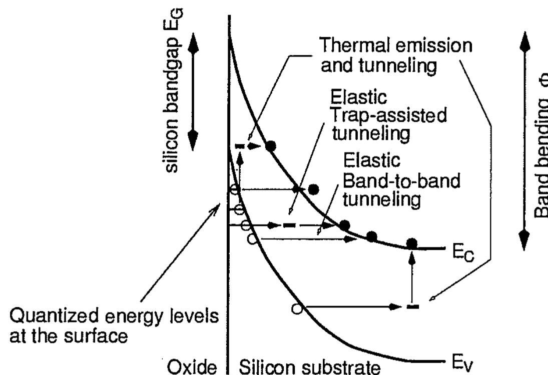


Figure 4.12 Band diagram of a MOS structure along the vertical direction through the depletion region illustrating band-to-band and trap-to-band tunneling mechanisms.

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 ($\simeq$ 1.12eV at room temperature) is much smaller than the direct one ($\simeq$ 3.5eV) 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 $\Phi \approx$ 1–1.5V 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

APPLE INC.
EXHIBIT 1008 - PAGE 0182

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

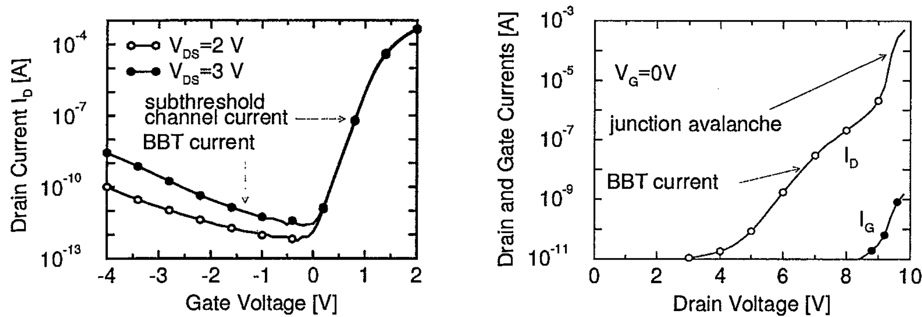


Figure 4.13 MOSFET drain and gate currents in the BBT-TBT regime. $L = 0.5\mu$m, $t_{ox} = 10$nm.

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: $F_{SI} \simeq F_{OX}\epsilon_{ox}/\epsilon_{si} \simeq V_{GD}\epsilon_{ox}/\epsilon_{si}t_{ox}$. 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

APPLE INC.
EXHIBIT 1008 - PAGE 0183

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.

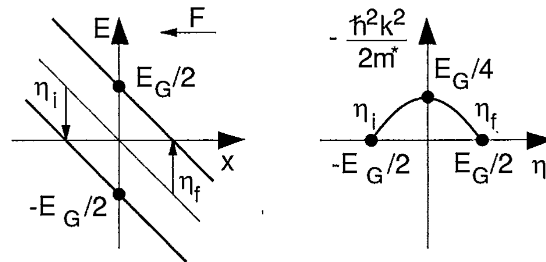


Figure 4.14 Energy band diagram and equivalent barrier height for BBT tunneling.

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-

APPLE INC.
EXHIBIT 1008 - PAGE 0184

ing to (4.18) the former is proportional to $Tg_vg_cf_v(1-f_c)$ and the latter to $Tg_cg_vf_c(1-f_v)$, 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:

$$dJ_{\mathrm{BBT}} \propto Tg_vg_c(f_v - f_c) \quad . \tag{4.22}$$

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:

$$\hbar^2k^2/2m^* = \hbar^2k_\perp^2/2m^* + \hbar^2k_{||}^2/2m^* = -\Big((E_G/2)^2 - \eta^2\Big)/E_G \quad , \tag{4.23}$$

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 $x_1 = \sqrt{(E_G/2)^2 + E_GE_{||}}/qF$ and $\xi = \Phi_B - (E - E_{||}) = E_G/4$. Hence, the transmission coefficient is [50]:

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

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:

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

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 $A_{\rm BB}$ and $B_{\rm BB}$ values indirectly confirms this statement. For example, $B_{\rm BB}$ ($\simeq$ 23MV/cm according to (4.24)) ranges between 18MV/cm [33] and 27MV/cm [39] in experiments. Much larger $B_{\rm BB}$ values ($\approx$ 36MV/cm) have been attributed to indirect BBT in the absence of traps [52].

The approximations made to derive $J_{\rm BBT}$ 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 10–15nm, 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:

$$\mathcal{F}(\vec{r}, \vec{k}, t)d\vec{k} = \frac{1}{4\pi^3} f(\vec{r}, \vec{k}, t)d\vec{k} \quad , \tag{4.26}$$

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:

$$\mathcal{F}(\vec{r}, E, t)dE = f(\vec{r}, E, t)g(E)dE \quad , \tag{4.27}$$

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:

$$n(\vec{r}) = \frac{1}{4\pi^3} \int f(\vec{r}, \vec{k})d\vec{k} \quad , \tag{4.28}$$

$$\vec{J}(\vec{r}) = -\frac{q}{4\pi^3} \int \vec{v}_g(\vec{k}) f(\vec{r}, \vec{k})d\vec{k} \quad . \tag{4.29}$$

Neglecting anisotropicity 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:

$$n(\vec{r}) = \int f(\vec{r}, E)g(E)dE \quad , \tag{4.30}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0187

$$\vec{J}(\vec{r}) = -q\int \vec{v}_g(E)f(\vec{r},E)g(E)dE \quad , \tag{4.31}$$

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:

$$\frac{d\vec{r}}{dt}\cdot\nabla_{\vec{r}}f + \frac{d\vec{k}}{dt}\cdot\nabla_{\vec{k}}f = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} \quad , \tag{4.32}$$

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 $drdk$ 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 $(\vec{r},\vec{k}')$ $((\partial f/\partial t)_{\mathrm{in}})$, and the carrier flux outgoing from state $(\vec{r},\vec{k})$ towards any other state $((\partial f/\partial t)_{\mathrm{out}})$. This balance can be expressed as:

$$\left(\frac{\partial f}{\partial t}\right)_{\mathrm{coll}} = \left(\frac{\partial f}{\partial t}\right)_{\mathrm{in}} - \left(\frac{\partial f}{\partial t}\right)_{\mathrm{out}} =$$

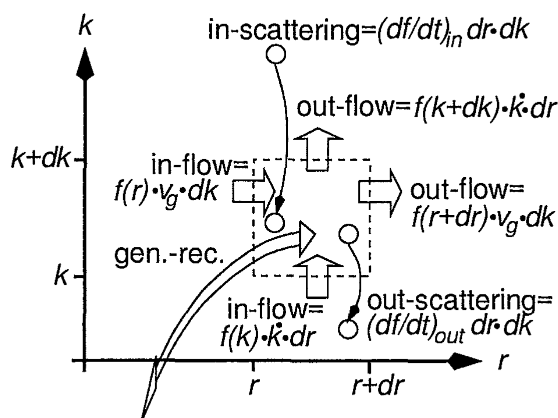


Figure 4.15 Schematic representation of particle flow in one-dimensional real and one-dimensional wave vector space according to the Boltzmann transport equation [55]. $\dot{r} = dr/dt = \vec{v}_g$ and $\dot{k} = dk/dt = -qF/\hbar$. Fluxes in real space are proportional to the surface in momentum space (dk) and vice versa. The number of particles scattered in and out of the elemental volume are proportional to the volume itself ($drdk$).

$$\begin{aligned} &= \sum_{\vec{k}'} S(\vec{r},\vec{k}',\vec{k})\ f(\vec{r},\vec{k}')(1-f(\vec{r},\vec{k})) - \\ &- \sum_{\vec{k}'} S(\vec{r},\vec{k},\vec{k}')\ f(\vec{r},\vec{k})(1-f(\vec{r},\vec{k}')) \ , \end{aligned} \tag{4.33}$$

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:

$$\left(\frac{\partial f}{\partial t}\right)_{\text{coll}} = \frac{f - f_0}{\tau_f} \ . \tag{4.34}$$

In summary, (4.32) is easily derived equating to zero the algebraic sum of the flux components depicted in Fig. 4.15.

APPLE INC.
EXHIBIT 1008 - PAGE 0189

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.

Ionized impur.	screened	elastic	isotropic
	unscreened	elastic	anisotropic
Intravalley phonons	acustic	~elastic	isotropic
	optical	inelastic	isotropic
Intervalley phonons	acustic	inelastic	isotropic
	optical	inelastic	isotropic
carrier-carrier		inelastic	anisotropic
carrier-plasmon		inelastic	anisotropic
impact ioniz.		inelastic	anisotropic

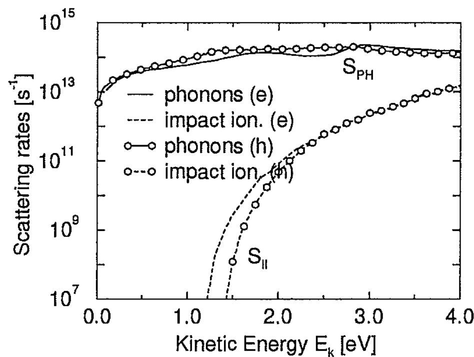


Figure 4.16 Left: Basic properties of the main electron scattering mechanisms in bulk silicon assuming a spherical parabolic band structure. Right: Phonon (S_{PH}) and impact ionization (S_{II}) scattering rates in silicon as a function of carrier energy before scattering. The non-monotonicity of the 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 $\tau_E \simeq$ 0.3–0.4ps, $\tau_p \simeq$ 0.01–0.1ps, for electrons [58], and $\tau_E \simeq$ 0.2–0.4ps [59, 60], $\tau_p \simeq$ 0.03–0.1ps 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:

$$f(E) \simeq \exp\Big(- (E - E_F)/k_B T_L \Big) , \quad (4.37)$$

APPLE INC.
EXHIBIT 1008 - PAGE 0191

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

$$\vec{v}_d(\vec{r}) = -\frac{\vec{J}}{qn} = \frac{\int \vec{v}_g(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k}}{\int f(\vec{r}, \vec{k}) d\vec{k}} \quad , \tag{4.38}$$

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

$$w(\vec{r}) = \frac{\int E_K(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k}}{\int f(\vec{r}, \vec{k}) d\vec{k}} \quad , \tag{4.39}$$

is very small ($w = 3k_BT_L/2 \approx 40$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) $\hbar d\vec{k}/dt = d\vec{p}/dt = -qvecF$ 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 computed these distributions is a statistical technique to solve the BTE, the distributions

APPLE INC.
EXHIBIT 1008 - PAGE 0192

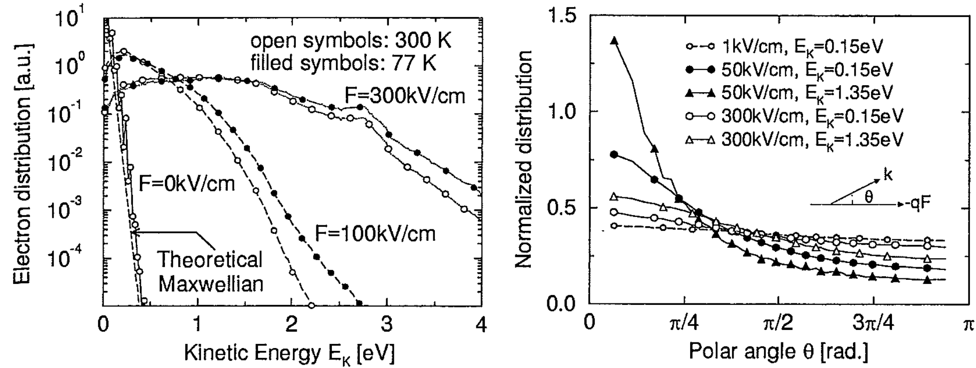


Figure 4.17 Left: Electron distributions normalized to unit concentration in homogeneous silicon slabs doped $1 \times 10^{17}\text{cm}^{-3}$ at various electric fields. A theoretical Maxwell-Boltzmann distribution at 300K is also shown for reference. Right: Distributions of carriers at given energy 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 $\approx 5 \times 10^4$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 $F \gg 5 \times 10^4$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 $F =$ 300kV/cm compared to those at 50kV/cm. Thus, v_d saturates to an almost constant value $v_s \simeq 10^7$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 $\approx 10^8$cm/s, that is much larger than the saturated velocity $v_s \simeq 10^7$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]:

$$f(\vec{r},\vec{k}) \approx \exp\Big((F_n(\vec{r})-U(\vec{r}))/k_BT_e\Big)\exp\Big(-|\hbar\vec{k}-m^*\vec{v}_d|^2/2m^*k_BT_e\Big)\ , \quad (4.40\ )$$

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:

$$w = \frac{1}{2}m^*v_d^2 + \frac{3}{2}k_BT_e\ . \quad (4.41\ )$$

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 $F = 10^5$V/cm, $w \approx 0.5$eV whereas $m^*v_d^2/2 \approx 30$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 $3k_BT_L/2$.

At high homogeneous fields $|\vec{p}| \gg |m^*\vec{v}_d|$. Hence (4.40) can be further approximated as:

$$f(\vec{r},E) \approx \exp\Big((F_n(\vec{r})-U(\vec{r}))/k_BT_e\Big)\exp\Big(-E_K/k_BT_e\Big)\ , \quad (4.42\ )$$

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.

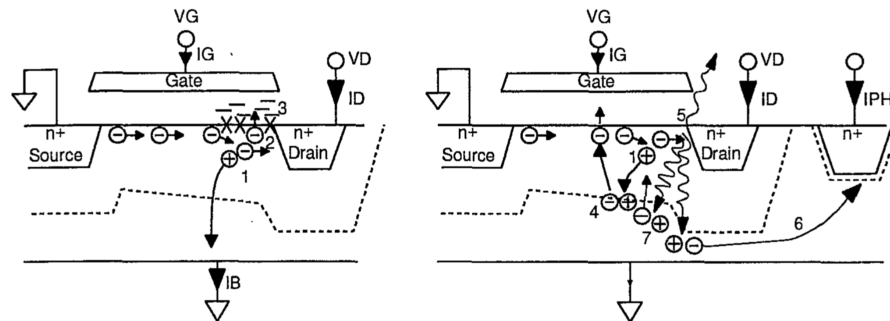


Figure 4.18 Qualitative sketch of the main hot carrier effects in an n-MOSFET.

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).

APPLE INC.
EXHIBIT 1008 - PAGE 0195

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

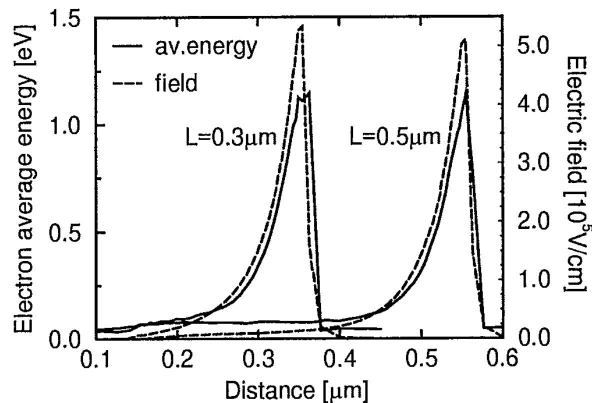


Figure 4.19 Electric field and average energy along the channel of sub-micrometer n-MOSFETs biased at $V_{GS} = V_{DS}/2 = 1.5$V, and featuring the same doping profiles and oxide thickness ($t_{ox} = 10$nm).

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

APPLE INC.
EXHIBIT 1008 - PAGE 0196

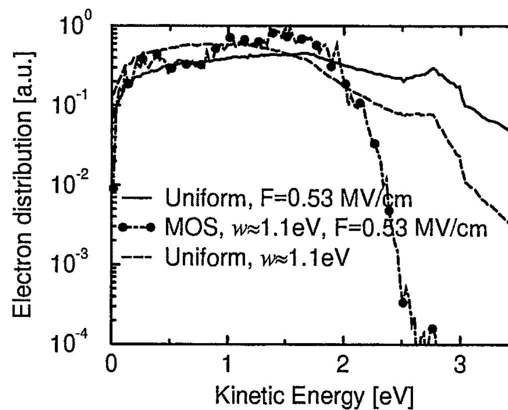


Figure 4.20 Electron distribution in a uniform field $F = 0.53$MV/cm (solid line); in a uniform field featuring $w \approx 1.1$eV (dashed line); at a point along the channel of a 0.3μm MOSFET biased at $V_{GS} = V_{DS}/2 = 1.5$V, and featuring $w \approx 1.1$eV and $F = 0.53$MV/cm (dash-dotted line). 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

APPLE INC.
EXHIBIT 1008 - PAGE 0197

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.

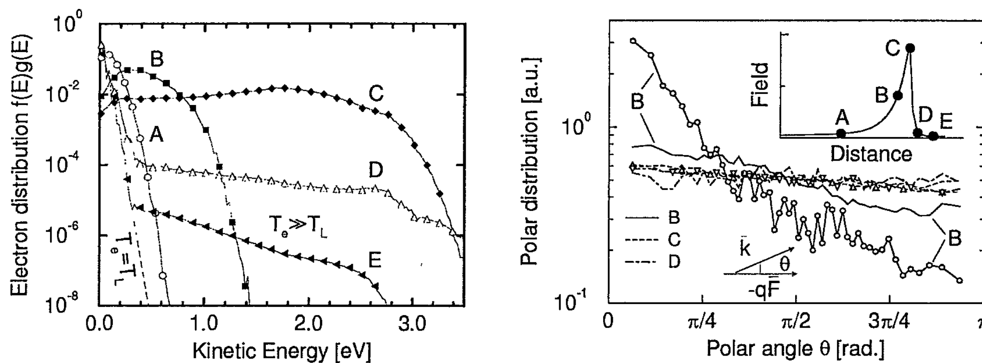


Figure 4.21 Left: Electron distributions at different positions along the channel of a MOSFET. Right: Normalized polar distributions at $E_K = 0.1$eV (lines) and $E_K = 1.5$eV (lines with symbols).

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 loose 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).

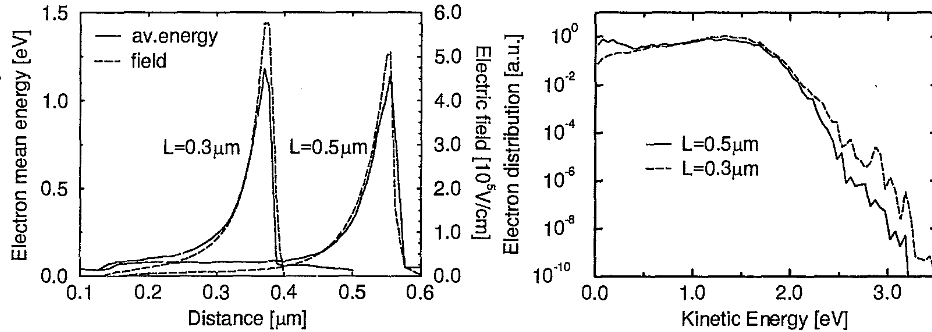


Figure 4.22 Left: Electric field and average energy along the channel of n-MOSFETs biased at $V_{GS} = V_{DS}/2 = 1.5$V. Gate length $L = 0.5\mu$m, $t_{ox} = 10$nm, junction depth $x_J = 100$nm, and $L = 0.3\mu$m, $t_{ox} = 7$nm, $x_J = 70$nm, respectively. Right: Corresponding electron energy distributions 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

APPLE INC.
EXHIBIT 1008 - PAGE 0199

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]:

$$\frac{dw}{dx} = \frac{3}{5}\, qF(x) - \frac{w(x) - w_0}{\lambda_w} \quad , \qquad (4.43\ )$$

where $w_0 = 3k_BT_L/2$, $\lambda_w = 5v_d\tau_E/3$, 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 $\lambda_w \approx$ 65–80nm for electrons [73, 74, 75] and $\lambda_w \approx$ 55–100nm 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

APPLE INC.
EXHIBIT 1008 - PAGE 0200

as:

$$\Delta w(x) = \frac{3}{5}\int_0^x qF(\xi)\exp\left(-\int_\xi^x \frac{dz}{\lambda_w(z)}\right)d\xi \ , \qquad (4.44)$$

where $w(0) = w_0 = 3k_BT_L/2$. 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 $w = 3k_BT_e/2$. 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:

$$T_e = T_L + \frac{2}{5}\frac{qF\lambda_w}{k_B} \ , \qquad (4.45)$$

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:

$$f(E) = A\exp\left(-\frac{E^\xi}{\eta(k_BT_e)^\nu}\right) \ , \qquad (4.46)$$

where T_e is usually derived from energy balance solutions under the approximation $w \simeq 3k_BT_e/2$.

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 ($V_{DS} > 5$V). For shorter devices and smaller drain biases, relevant non local effects make these models less reliable.

APPLE INC.
EXHIBIT 1008 - PAGE 0201

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 $E_G(\mathrm{Si})$. 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:

$$G_{\mathrm{II}}(\vec{r}) = \frac{1}{4\pi^3} \int S_{\mathrm{II}}(\vec{k}) f(\vec{r}, \vec{k}) d\vec{k} \ , \tag{4.47}$$

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 $S_{\mathrm{II}}(\vec{k})$ over equi-energy surfaces in $\vec{k}$ space and to express G_{II} in terms of an energy dependent scattering rate $S_{\mathrm{II}}(E)$ as:

$$G_{\mathrm{II}}(\vec{r}) = \int_{E_G}^{\infty} S_{\mathrm{II}}(E) f(\vec{r}, E) g(E) dE \ . \tag{4.48}$$

Adding the contributions of both carriers, G_{II} is often written as:

$$G_{\mathrm{II}}(\vec{r}) = \alpha_n \frac{|\vec{J}_n(\vec{r})|}{q} + \alpha_p \frac{|\vec{J}_p(\vec{r})|}{q} \ , \tag{4.49}$$

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:

$$\alpha_{n,p}(\vec{r}) = \frac{\int S_{\rm II}(\vec{k})f(\vec{r},\vec{k})d\vec{k}}{\int f(\vec{r},\vec{k})v_g(\vec{k})d\vec{k}} \quad , \tag{4.50 }$$

where $S_{\rm II}$, 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 $S_{\rm II}$ 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.

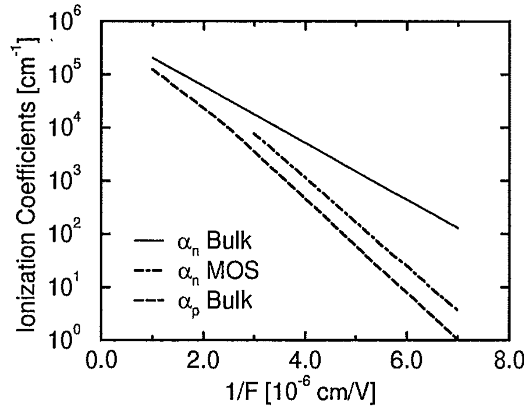


Figure 4.23 Electron and hole impact ionization coefficients at room temperature. Data from [88] (solid line), [89] (dot-dashed line), [88] (dashed line).

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]:

$$\alpha_{n,p} = \alpha_{n,p}^{\infty} \exp(-F_{n,p}^{\rm crit}/F) \quad , \tag{4.51 }$$

with $\alpha_n^{\infty} = 7.03 \times 10^5 \text{cm}^{-1}$, $F_n^{\rm crit} = 1.231\text{MV/cm}$, $\alpha_p^{\infty} = 1.582 \times 10^6\text{cm}^{-1}$, $F_p^{\rm crit} = 2.036\text{MV/cm}$, for $F < 4 \times 10^5\text{V/cm}$ and $\alpha_p^{\infty} = 6.71 \times 10^5\text{cm}^{-1}$, $F_p^{\rm crit} = 1.693\text{MV/cm}$ 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 $F_{\mathrm{EFF}} = 5(w - w_0)/3q\lambda_w$, 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:

$$I_B = q \int_V G_{\rm II}(\vec{r}) d\vec{r} = q \int_{E_G}^{\infty} S_{\rm II}(E) \left(\int_V f(\vec{r}, E) g(E) d\vec{r} \right) dE \ . \qquad (4.52)$$

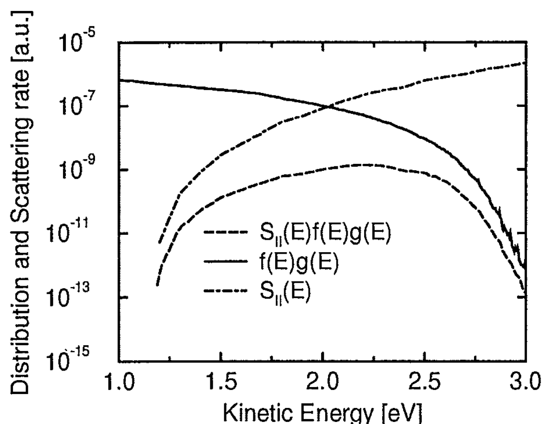


Figure 4.24 Impact ionization rate (dot-dashed line), electron distribution integrated over the device (solid line) and corresponding product (dashed line) in a $L = 0.5\mu$m MOSFET biased at $V_G = V_D/2 = 1.5$V. $t_{\rm ox} = 10$nm. $E = 0$ at the bottom of the conduction band.

Fig. 4.24 reports $f(E)g(E)$ and $S_{\rm II}(E)f(E)g(E)$ integrated over the volume of a 0.5μm MOSFET. As the result of the trade-off between the decreasing $f(E)g(E)$ and the increasing ionization rate $S_{\rm II}(E)$, the largest contribution to impact ionization comes from electrons between 2 and 2.5eV, 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 $V_{\rm GS} = V_{\rm DS}/2$, 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, $G_{\rm II}$ 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

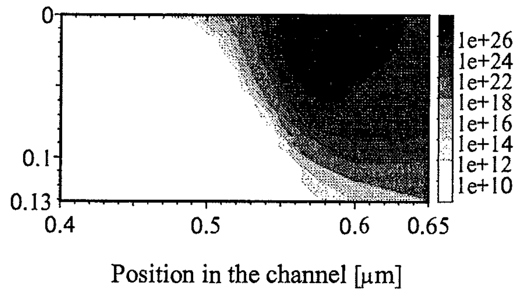


Figure 4.25 Impact ionization generation rate contours for electrons at the drain end of a $L = 0.5\mu$m MOSFET biased at $V_G = V_D/2 = 1.5$V. $t_{\mathrm{ox}} = 10$nm. The solid line represents the drain junction.

vertical field, G_{II} is maximum somewhat inside the bulk of the material and not at the interface.

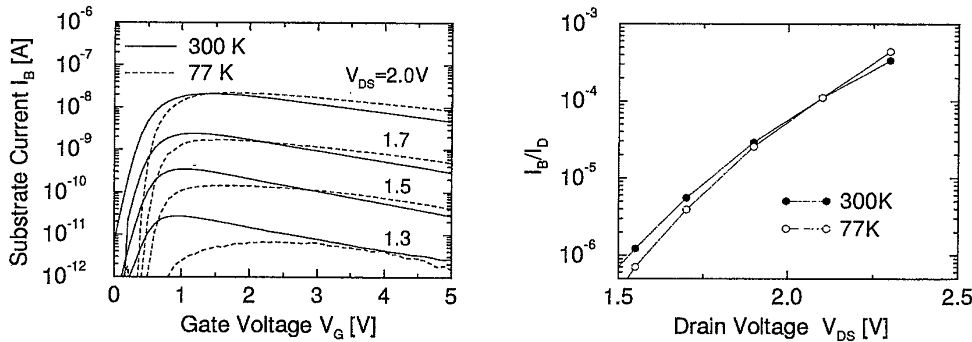


Figure 4.26 Left: Substrate current as a function of gate and drain voltages and temperature. Data from [94]. Right: Normalized substrate current as a function of drain voltage at 300 and 77K. Data from [95].

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,

APPLE INC.
EXHIBIT 1008 - PAGE 0206

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 $V_{DS} > 2.25$V). 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 that at room temperature [95, 96, 97], as observed in the right plot of Fig. 4.26 for $V_{DS} < 2$V.

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:

$$J_{\mathrm{inj},n}(x) = q \int_0^{\infty} v_\perp(x,E) f_\perp(x,E) g(E) P(x,E) dE \quad , \qquad (4.53)$$

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;

APPLE INC.
EXHIBIT 1008 - PAGE 0207

2) the electron energy and momentum distributions (because only carriers with velocity directed towards the interface contribute to $J_{\text{inj},n}$); 3) the probability of injection from Si to SiO_2 ($P(x,E)$). 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 $V_{\text{DS}} > 0$ 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 SiO_2, transport within the SiO_2 conduction band, and, possibly, tunneling through the SiO_2 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0208

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:

$$I_G = W \int_0^L \Big[\eta_n(x) J_{\mathrm{inj},n}(x) - \eta_p(x) J_{\mathrm{inj},p}(x)\Big] dx \quad , \qquad (4.54)$$

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,

APPLE INC.
EXHIBIT 1008 - PAGE 0209

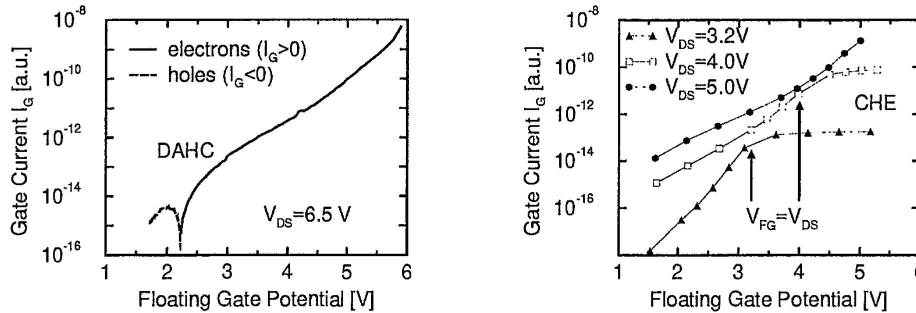


Figure 4.27 Left: Gate current of a Flash cell at high V_{DS} demonstrating electron and hole injection. $L_{EFF} = 0.4\mu$m. Data from [100]. Right: Gate current of a Flash cell as a function of floating gate and drain potentials. $L_{EFF} = 0.3\mu$m.

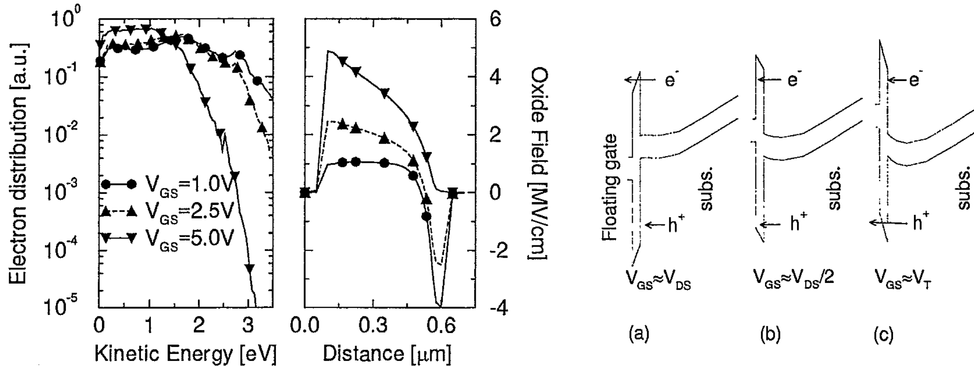


Figure 4.28 Left: Electron distributions at the point of maximum carrier heating in a 0.5μm, t_{ox} = 10nm MOSFET biased at: a) V_{GS} = 5V; b) V_{GS} = 2.5V; c) V_{GS} = 1V $\approx V_T$. V_{DS} = 5V, V_{SB} = 0V. Center: Oxide field along the interface. Right: Vertical band diagram at the drain junction. 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

APPLE INC.
EXHIBIT 1008 - PAGE 0210

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).

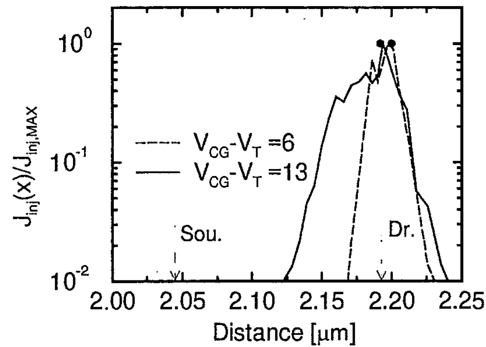


Figure 4.29 Simulated gate current density along the channel of a Flash cell normalized to its maximum value. Filled points indicate the abscissae of maximum injection.

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 $V_{\rm GS}$ is reduced below $V_{\rm DS}$, 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 $F_{\rm OX}$ (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 $V_{\rm GS}$ is further reduced, $\eta_p J_{{\rm inj},p}$ increases and $\eta_n J_{{\rm inj},n}$ decreases until the total gate current goes to zero. This happens almost at the same $V_{\rm GS}$ 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 $\eta_n J_{{\rm inj},n}$ and $\eta_p J_{{\rm inj},p}$ that balance each other.

Finally, for $V_{\rm GS}$ 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 $V_{\rm FG} < 2.2$V, 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 $\Phi_{B,e^-} \simeq 3.15$eV, 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 $V_{\rm SB}$ 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

APPLE INC.
EXHIBIT 1008 - PAGE 0212

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 ($V_{\mathrm{GS}} > V_T$), 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

APPLE INC.
EXHIBIT 1008 - PAGE 0213

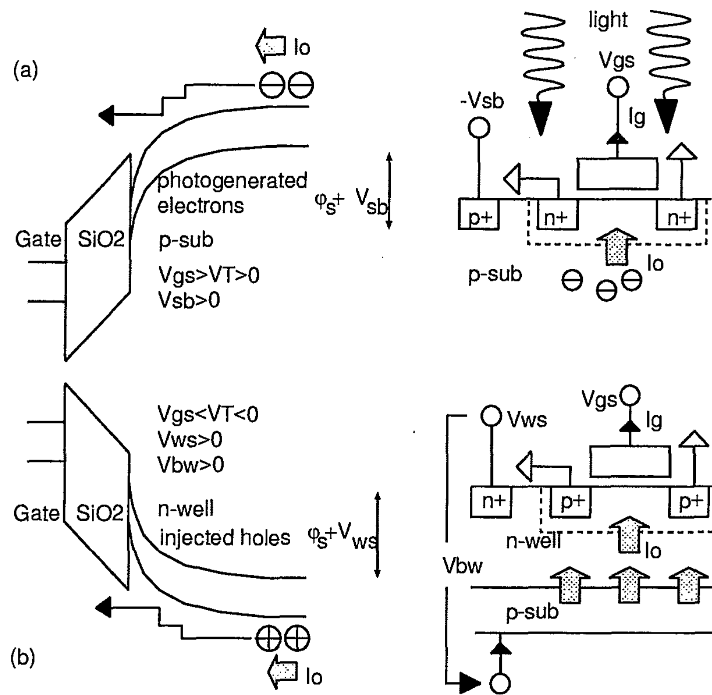


Figure 4.30 Schematic representation of the substrate hot electron (a) and hole (b) injection experiment.

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 ($F_{\rm OX} = (V_G - V_{\rm FB} - \phi_s)/t_{\rm ox}$, where $V_{\rm FB}$ is the flat band voltage) is controlled only by $V_{\rm GS}$. $V_{\rm SB}$, 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

APPLE INC.
EXHIBIT 1008 - PAGE 0214

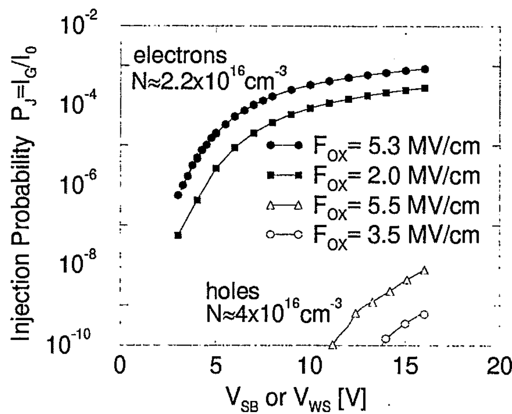


Figure 4.31 Measured homogeneous electron and hole gate currents normalized to the injected current I_G/I_0 as a function of accelerating substrate voltage and oxide field in n- and p-MOS devices featuring comparable substrate doping (N). Data from [68].

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-SiO_2 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 10μs 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

APPLE INC.
EXHIBIT 1008 - PAGE 0215

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-SiO_2 energy barrier and for impact ionization ($\Phi_{B,e^-}/q = 3.15$V and $\Phi_{B,h^+}/q \approx 4$V, $\Phi_{II}/q \simeq E_G/q \simeq 1.1$V, 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

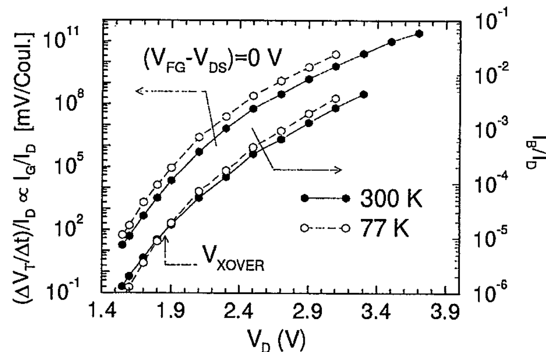


Figure 4.32 $\Delta V_T/\Delta t I_D$, proportional to I_G/I_D, and I_B/I_D of a Flash memory cell at low V_{DS}. Data from [123].

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} (≈ 1V, that is

APPLE INC.
EXHIBIT 1008 - PAGE 0216

$\ll \Phi_{B,e^-}/q$) 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 $V_{\rm DS}$ 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 ($\approx$ 1V/s) is forced on the MOS capacitor and the breakdown voltage ($V_{\rm BD}$), or field $F_{\rm BD} \approx V_{\rm BD}/t_{\rm ox}$, 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 ($<$ 5MV/cm). The latter occur at much larger fields (typically $>$ 10MV/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 ($\approx 10^{-4}$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-

APPLE INC.
EXHIBIT 1008 - PAGE 0217

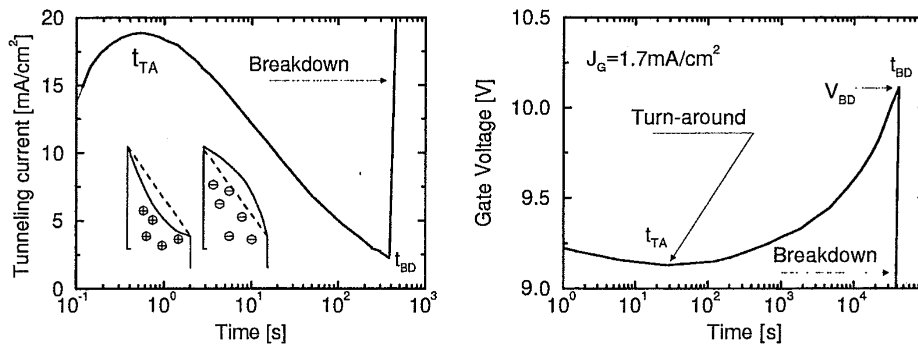


Figure 4.33 Typical gate current (left) and voltage (right) during constant voltage (current) stress experiments on MOS capacitors. The inset shows the deformation of the oxide barrier shape due to positive and negative net charges.

crease (voltage decrease), for $t < t_{\mathrm{TA}}$. Eventually, negative charges accumulate in the oxide, thickening the barrier and reducing the current (increasing the voltage), as observed for $t > t_{\mathrm{TA}}$. 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 ($t_{\mathrm{ox}} < 10$nm) where the voltage drop is insufficient to provide electrons of the required energy ($E \approx E_G(\mathrm{SiO_2}) > 8$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 $t_{\mathrm{ox}} \approx 4$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

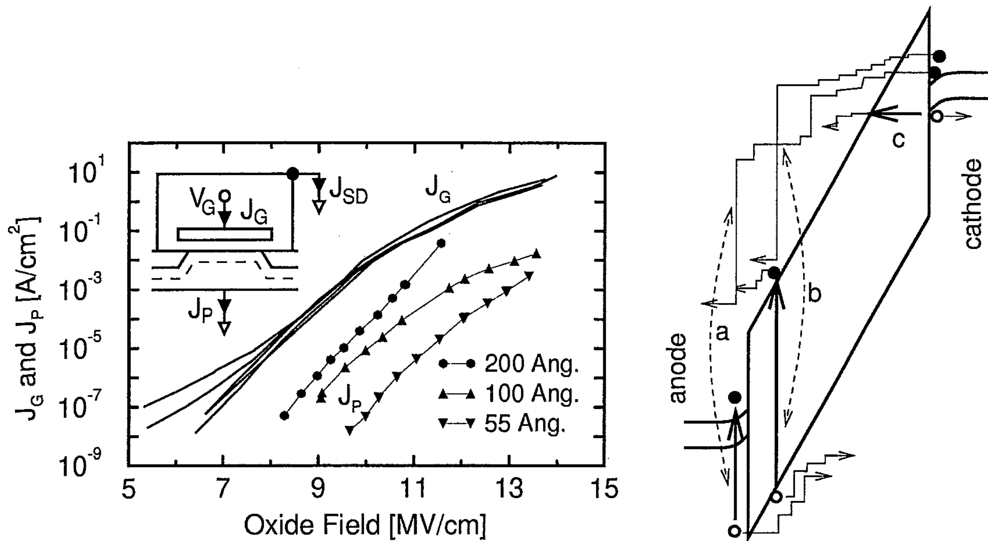


Figure 4.34 Left: density of the electron gate current (J_G) and of the hole substrate current (J_P) during tunneling experiments. Data from [131]. Right: Schematic representation of a few processes leading to the generation of holes during tunnel injection.

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

APPLE INC.
EXHIBIT 1008 - PAGE 0219

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-SiO_2 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 where 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].

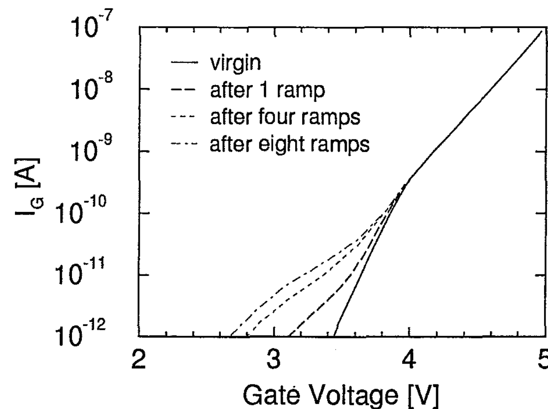


Figure 4.35 I-V characteristics of thin oxide capacitors showing low field leakage current increase after one, four and eight stress voltage ramps. Data from [147].

4.7.1.1 Stress Induced Leakage Currents (SILC). If the oxide thickness is reduced below ≈ 10nm (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 ($\simeq$ 10nm), 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].

APPLE INC.
EXHIBIT 1008 - PAGE 0221

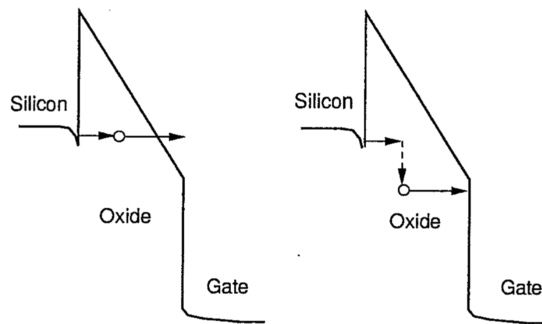


Figure 4.36 Possible SILC mechanisms in thin oxides. Left: elastic trap assisted tunneling; Right: inelastic trap assisted tunneling.

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 ($Q_{\mathrm{P,BD}}$) is reached. Therefore:

$$Q_{\mathrm{P,BD}} = \int_0^{t_{\mathrm{BD}}} J_P(t)dt \quad , \tag{4.55}$$

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 $Q_{\mathrm{P,BD}}$ 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, $Q_{\mathrm{P,BD}}$ 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]:

$$Q_{\mathrm{BD}} = \int_0^{t_{\mathrm{BD}}} J_G(t)dt \quad , \tag{4.56}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0222

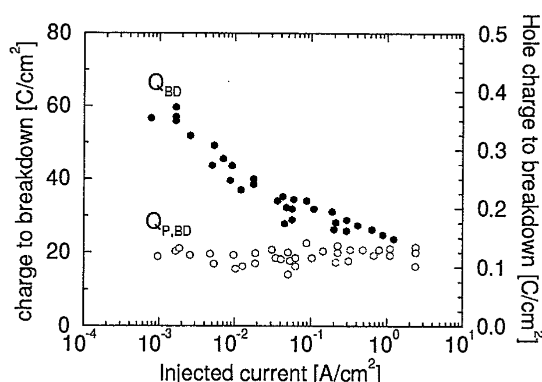


Figure 4.37 Charge to breakdown and hole charge to breakdown as a function of stress current density and stress field. $t_{ox} = 10.7$nm. Data from [131].

where J_G denotes the electron tunneling current.

$Q_{\rm BD}$ 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 $Q_{\rm BD}$ and fewer breakdown events at low field, probably due to less asperities and to a more uniform distribution of the generated stress [145]. $Q_{\rm BD}$ decreases for increasing J_G above $\approx 10^{-3}$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

APPLE INC.
EXHIBIT 1008 - PAGE 0223

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:

$$J_{SS} \propto D_{OT} J_{TUN}(\Phi_{eff}) \quad , \qquad (4.57)$$

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:

$$D_{OT}(t) \propto J^{\beta}_{stress} Q^{\alpha}_{inj}(t) \qquad (4.58)$$

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

APPLE INC.
EXHIBIT 1008 - PAGE 0224

as a function of $1/V_J$, where V_J is the erase junction voltage used as a stress acceleration factor.

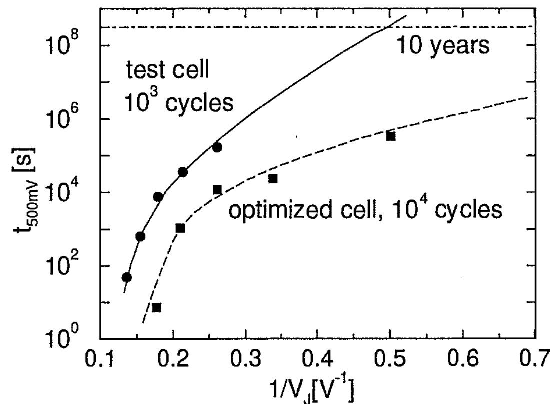


Figure 4.38 Experimental (symbols) and simulated (lines) SILC induced disturb characteristics for two flash cells. Data from [169].

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

$$t_{BD} = \frac{Q_{BD}}{J_G} = \frac{Q_{P,BD}}{J_P} \quad . \tag{4.59}$$

On the other hand:

$$J_P = \alpha_P \theta_P J_G \tag{4.60}$$

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)

APPLE INC.
EXHIBIT 1008 - PAGE 0225

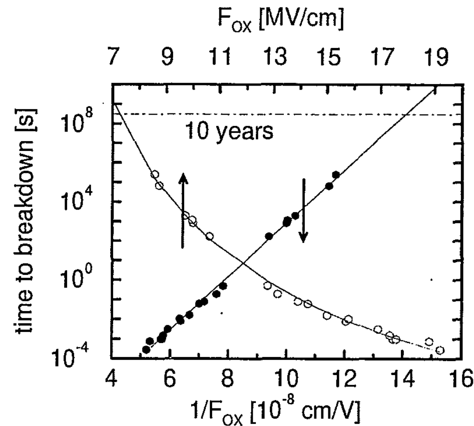


Figure 4.39 Time to breakdown data plotted as a function of stress oxide field and inverse stress oxide field. $t_{ox} = 7.9$nm. Data from [138].

that, if modeled as the electron one, is proportional to $\exp(-B\Phi'_{B,h^+}/F_{OX})$, where B is a constant, and Φ'_{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

APPLE INC.
EXHIBIT 1008 - PAGE 0226

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 payed in the characterization phase.

APPLE INC.
EXHIBIT 1008 - PAGE 0227

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 $V_{\mathrm{GS}} \simeq V_{\mathrm{DS}}/2$. 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

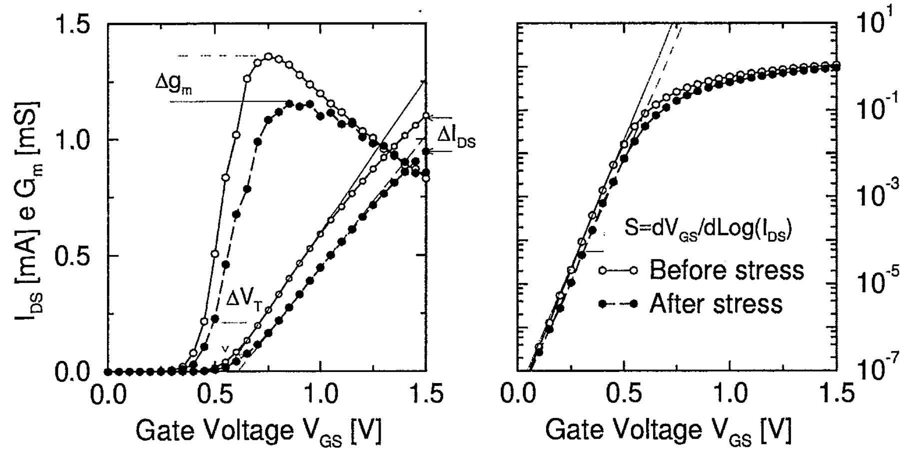


Figure 4.40 Degradation of linear MOSFET characteristics (in linear and logarithmic scales) after hot-carrier stress at $V_{GS} = V_{DS}/2$.

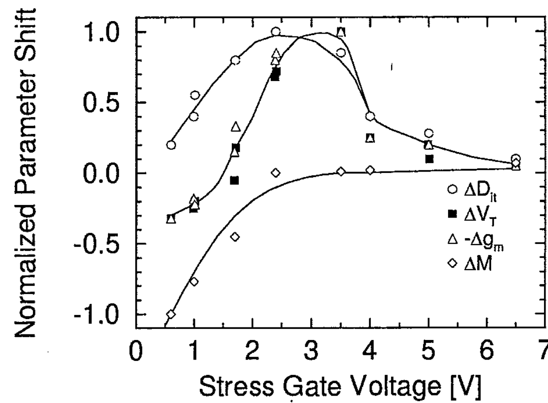


Figure 4.41 Interface state density (measured with the charge pumping technique), threshold voltage, transconductance and multiplication factor shift during stress experiments at constant drain voltage as a function of the stress gate voltage. Data from [180, 181].

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].

APPLE INC.
EXHIBIT 1008 - PAGE 0229

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 $\Phi_{Si-H} \approx 0.3$–0.5eV, an interface trap generation threshold energy $\Phi_{ITG} = \Phi_{B,e^-} + \Phi_{Si-H} \approx 3.5$–$4$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].

APPLE INC.
EXHIBIT 1008 - PAGE 0230

4.8.3 Lifetime Evaluation Models

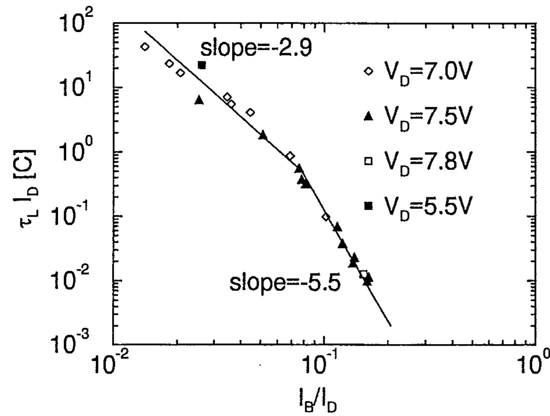


Figure 4.42 Lifetime plot versus I_B/I_D. Lifetime is defined by $D_{\rm it}$ shifts. Data from [181].

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 $\Delta D_{\rm it}$) can be described by:

$$\Delta D(t) = (D_0 t)^n \quad . \tag{4.61 }$$

Here, $\Delta D(t)$ represents the absolute parameter shift during aging and D_0 is a function of the stress conditions related to the 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 $V_{\rm GS} \approx V_{\rm DS}/2$, while $n \simeq 0.2 \div 0.3$ for ΔV_T at $V_{\rm GS} \approx V_{\rm DS}$ and $V_{\rm GS} \approx V_T$ [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 ($\Phi_{\rm IIeff}$ and $\Phi_{\rm Beff}$, 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]:

$$\frac{I_B}{I_S} = \exp(-\Phi_{\rm IIeff}/k_B T_e) \quad , \tag{4.62 }$$

APPLE INC.
EXHIBIT 1008 - PAGE 0231

$$\frac{I_G}{I_S} = C(F_{\text{OX}})\exp(-\Phi_{\text{Beff}}/k_B T_e) \quad , \tag{4.63}$$

where I_S is the channel current and $C(F_{\text{OX}})$ is a function of the oxide field that accounts for back scattering in the image force potential well. Thus,

$$\frac{I_G}{I_S} = C(F_{\text{OX}}) \left(\frac{I_B}{I_S}\right)^{\left(\frac{\Phi_{\text{Beff}}}{\Phi_{\text{IIeff}}}\right)} \quad . \tag{4.64}$$

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 $V_{\text{GS}} \simeq V_{\text{DS}}/2$), we can exploit again (4.42) and (4.61) to express D_0 and ΔD, respectively as:

$$D_0 \propto I_S \exp(-\Phi_{\text{ITG}}/k_B T_e) \quad , \tag{4.65}$$

$$\Delta D = H\left(I_S \exp\left(-\frac{\Phi_{\text{ITG}}}{k_B T_e}\right) t\right)^n \quad , \tag{4.66}$$

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:

$$\tau_L = \frac{C_\tau}{I_S}\left(\frac{I_B}{I_S}\right)^{-\frac{\Phi_{\text{ITG}}}{\Phi_{\text{IIeff}}}} \quad . \tag{4.67}$$

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 $-[\Phi_{\text{ITG}}/\Phi_{\text{IIeff}}] \approx -2.9$ is obtained for stress at $V_{\text{GS}} \simeq V_{\text{DS}}/2$ and above, which yields $\Phi_{\text{ITG}} \simeq 3.8\text{eV}$ if $\Phi_{\text{IIeff}} \simeq 1.3\text{eV}$ 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 al 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 $F_{\max} = \theta(V_{\mathrm{DS}} - V_{\mathrm{DS,SAT}})$ where $V_{\mathrm{DS,SAT}}$ 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 $1/(V_{\mathrm{DS}} - V_{\mathrm{DS,SAT}})$ 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 $V_{\mathrm{DS,SAT}}$, 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 $1/V_{\mathrm{DS}}$ [195].

Lifetime prediction at stress bias conditions different from $V_{\mathrm{GS}}/V_{\mathrm{DS}}/2$ 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0233

[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 $SiSO_2$ 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0234

[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 SiO_2". *Solid State Electronics*, **20**, p. 11.

[22] Suñè 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-SiO_2 interfaces in the WKB effective mass approximation". *Phys. Status Solidi (b)*, **132**, p. 153.

[25] Suñè 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 SiO_2". *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. Suñè J. and Riccò B. (1991) "On the determination of the Si-SiO_2 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0235

[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 Subrahmanyan 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0236

[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.

APPLE INC.
EXHIBIT 1008 - PAGE 0237

[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.

APPLE INC.
EXHIBIT 1008 - PAGE 0238

[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 SiO_2 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 $0.1\mu m$: 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0239

[79] Fukuma M. and Iizuka T. (1988) "Advanced physical models for MOSFETs". *Proc. SISDEP*, p. 57.

[80] Fiegna C., Venturi F., Melanotte M., Sangiorgi E. and Riccó 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., Streukter 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0240

[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 MOSFET". *Microelectronic Engineering*, **36**, 1-4, p. 293.

[102] Takeda E. (1984) "Hot carrier effects in submicrometer MOS VLSIs". *Proc. IEE*, **131**, p. 153.

APPLE INC.
EXHIBIT 1008 - PAGE 0241

[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 submicron 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-EPROM 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.*.

APPLE INC.
EXHIBIT 1008 - PAGE 0242

[115] Verwey J.F. (1973) "Nonavalanche injection of hot carriers into SiO_2". *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 SiO_2 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0243

[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-SiO_2 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 SiO_2-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) "SiO_2 induced substrate current and its relation to positive charge in field effect transistors". *Journal of Applied Physics*, **59**, 3, p. 824.

APPLE INC.
EXHIBIT 1008 - PAGE 0244

[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 SiO_2". *Solid State Electronics*, **21**, p. 251.

[140] Olivo P., Riccò B. and Sangiorgi E. (1983) "Electron trapping-detrapping within thin SiO_2 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 SiO_2 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 SiO_2". *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 SiO_2 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/SiO_2 interface during Fowler-Nordheim tunneling". *Journal of Vacuum Science and Technology*, **20**, 3, p. 743.

APPLE INC.
EXHIBIT 1008 - PAGE 0245

[149] Nguyen T.N., Olivo P. and Riccò B. (1987) "A new failure mode of very thin (< 50Å) 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 ultra-thin 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0246

[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., Haspeslagh L., Deferm L., Groeseneken G. and Maes H.E. (1996) "A new quantitative model to predict SILC-related disturb characteristics in Flash E^2PROM 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 $SiO_2/Si_3N_4/SiO_2$ 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0247

[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-SiO_2 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 SiO_2 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0248

[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 0.25μm 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 0.1μm 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.

APPLE INC.
EXHIBIT 1008 - PAGE 0249

5 MEMORY ARCHITECTURE AND RELATED ISSUES

Maurizio Branchetti, Giovanni Campardo, Stefano Commodaro,
Stefano Ghezzi, Andrea Ghilardelli, Carla Golla,
Ignazio Martines, Marco Maccarrone, Rino Micheloni,
Matteo Zammattio, Stefano Zanardi

STMicroelectronics
Via Olivetti 2, 20041 Agrate Brianza (Milano), Italy
carla.golla@st.com, stefano.commodaro@st.com

5.1 FLASH ARCHITECTURE: GENERAL OVERVIEW

In the Flash memory scenario, several different approaches can be found, each one with its characteristics that make each solution more suitable for a particular application. One method to classify these different approaches is to consider the "memory architecture", i.e. the way in which the array, and consequently the device, is organized.

The device architecture, therefore, is the result of the following:

- cell architecture;
- cell functionality, i.e. voltages to be applied in Read, Program and Erase operations;
- array organization that derives from the cell functionality;
- target device performance.

APPLE INC.
EXHIBIT 1008 - PAGE 0250

Another factor that should be taken into account when talking of memory architecture is the interface between the Flash and the external world: this implies, for instance, the correct protocol to provide the device with the user commands and the way the Flash internal status is reported to the external world. Flash memories are playing a key role inside the world of the system-on-chip; in this context the protocol used by the Flash to communicate with the other devices (Microprocessor, DSP, ASIC, ...) could be the criterion of choice. Anyway, although the implementations can be slightly different, the key circuital blocks of the Read and the Program/Erase paths are similar: a brief overview is done in this first paragraph of this chapter.

5.1.1 Flash Architecture Scenario

Another (and perhaps the most immediate) way to classify the Flash memories is to divide them into two families, according to their main applications:

1. "EPROM like" applications, i.e. telecom (cellular phones: GSM, DECT, ...), automotive, hard disk drives, printers, set-top box, PC BIOS, ..., in most of which Flash memories have replaced the EPROM. Both low density and high density Flash memories are required; in several systems the trend is towards an integration between Flash and other devices inside the application (other types of memory, logic gates, DSP, micro, ...). The market requirements for these types of applications are: speed, low power consumption, low supply voltage, density and number of cycles (a cycle is defined as a sequence of Program operations terminated by an Erase operation; the first cycle which produces a failure sets the maximum number of cycles allowed for a particular device).

2. Mass storage applications, i.e. miniature cards and multimedia palm-top (like PDA, PCS and portable network PC). This is a market where the Flash memories have a high potential: however, the key issue to be solved is the profitability (i.e. cost/Mbyte) compared to other alternatives (as in the case of miniature cards applications) and the performances (PDA, PCS, ... applications). The market requirements for this application are: cost, density, number of cycles, low power consumption and speed.

As it was said before, there is a different memory architecture for every application; in turn, the device organization depends on the cell array architecture, for which three implementations exist, namely: NOR, NAND and EEPROM.

It is important to remind that the choice is determined by the following issues:

- sector erase size;

APPLE INC.
EXHIBIT 1008 - PAGE 0251

- power supply requirements;
- Read performance;
- Program/Erase (P/E) performance;
- P/E cycling performance;
- complexity of process manufacturing and device size (i.e. cost).

The NOR Flash memory derives from the EPROM: it has the same "cell construction", although it differs in terms of tunnel oxide thickness and source junction. The NOR Flash memories are those most commonly used in a wide range of applications that require both performance and medium density. The whole content of this chapter is related to circuital solutions for a NOR Flash memory approach; however, most of these concepts can be extended to the other types of Flash memories.

The NAND Flash cell is similar to the NOR Flash, but the access to the matrix information is different; the cells are arranged inside the array in serial chains: the drain of one cell is connected to the source of the following cell. The name (NAND) comes from the way in which the Read operation is performed; data sensing is serial and therefore the access time is very slow: this is the reason why this architecture is not suitable for fast random access time applications.

The voltages used to Program ($\approx$ 20V on the selected gate) and Erase ($\approx$ 20V on the substrate) are very high; the availability of this high voltage inside the system can be a serious constraint and during the device lifetime some reliability problems can occur.

The Flash EEPROM cell derives from the EEPROM cell, and the difference is that its contents are alterable on a block rather than on a byte basis. The functionality is the same as that of EEPROM cell. Reliability is a main limitation because of both the high voltages used and tunnel oxide thickness; furthermore, for a memory size greater than 1Mb, device size increases significantly because repairing algorithms (like redundancy and error correction) must be implemented.

5.1.2 NOR Cell Operation and Array Organization

In this paragraph, the different operations that can be performed on a Flash memory cell (together with the correspondent operating conditions) are shown, namely Read, Program and Erase.

NOR matrix organization is considered: cells are arranged in rows (called wordlines) and columns (called bitlines): all the gates of the cells in a row are

APPLE INC.
EXHIBIT 1008 - PAGE 0252

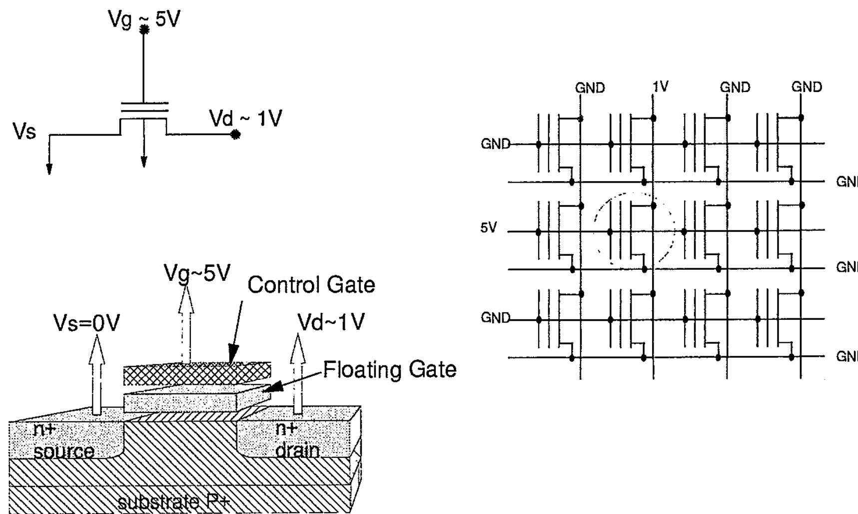


Figure 5.1 NOR Flash cell: Read operation.

connected to the same wordline, while all the drains of the cells in a column are connected to the same bitline; the sources of all the cells in the sector are connected to a common source line. This organization is repeated 8 or 16 times, thus obtaining either byte or word output.

The Read operation is the method to evaluate the content of the selected cell. The selected row is biased at $\approx 4 \div 5$V, while the selected bitline is biased at about 1V (see Fig. 5.1). Reading is done by byte or by word, therefore one cell for each output is addressed. There are several methods to identify the cell status; the most commonly used compares the current of the matrix cell with the one of a reference cell; the result of the comparison is then converted into a voltage which is fed to the output.

The Program operation is used to raise the threshold voltage of the selected cell, thus changing its state to "programmed" (i.e. logic "0"). The physical mechanism which is responsible for this effect is the so-called "channel hot electron". The gate of the selected cell is connected to a "high" voltage (i.e. much higher than V_{CC}), which can be either provided externally (through the V_{PP} pin) or generated internally (Fig. 5.2).

The drain voltage needs careful regulation to avoid dangerous conditions such as snap-back, drain turn-on, gain degradation, drain stress, ... The programming time strongly depends on the following factors:

- cell length, cell width, coupling ratio;
- temperature;

APPLE INC.
EXHIBIT 1008 - PAGE 0253

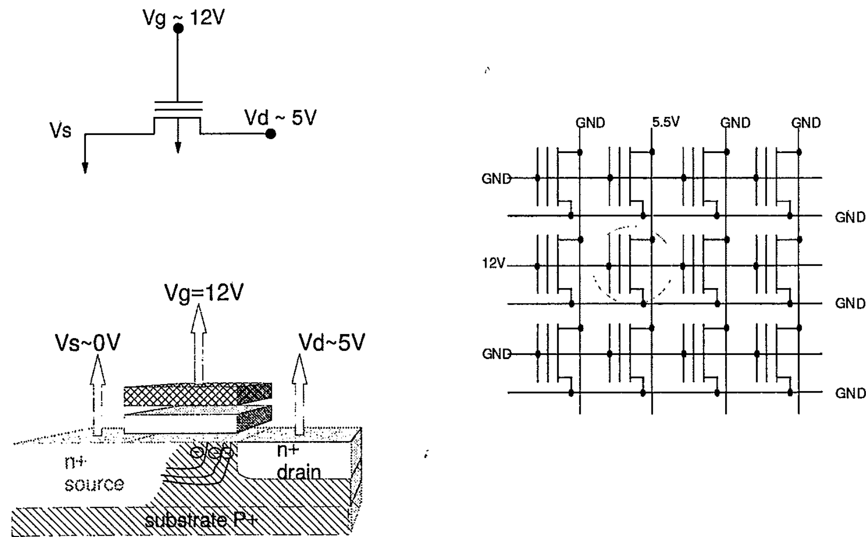


Figure 5.2 NOR Flash cell: Program operation.

- gate and drain voltages.

Like Read, Program is performed by byte or by word.

The electrical Erase operation is used to lower the threshold voltage of all the cells inside a sector, thus changing their state to "erased" (i.e. logic "1"). It can be performed exploiting the same physical phenomenon (i.e. Fowler-Nordheim tunneling) in two different ways: through the source junction or through the channel. In the first case, the source junction must be a deep junction to reduce the substrate current during erasing. Two erasing schemes can be adopted: source erase (Fig. 5.3) or negative gate erase (Fig. 5.4). Negative gate erase has the advantage of a larger margin with respect to breakdown voltage limitations (it is thus more reliable) but it presents a major drawback since it needs generating and switching negative voltage; however, in both cases pumping V_s for low voltage devices implies some circuit overhead.

Erase operation can be done using one of the following methods:

1. by applying a fixed voltage at the source terminal;

2. by forcing a constant source current by means of a current regulator (constant electric field in the tunnel oxide).

With the latter method, the erase time dependence on the applied voltage, the cell sizing (W and L) and the tunnel oxide thickness is strongly reduced. As a global result, the intrinsic endurance of a Flash array increases.

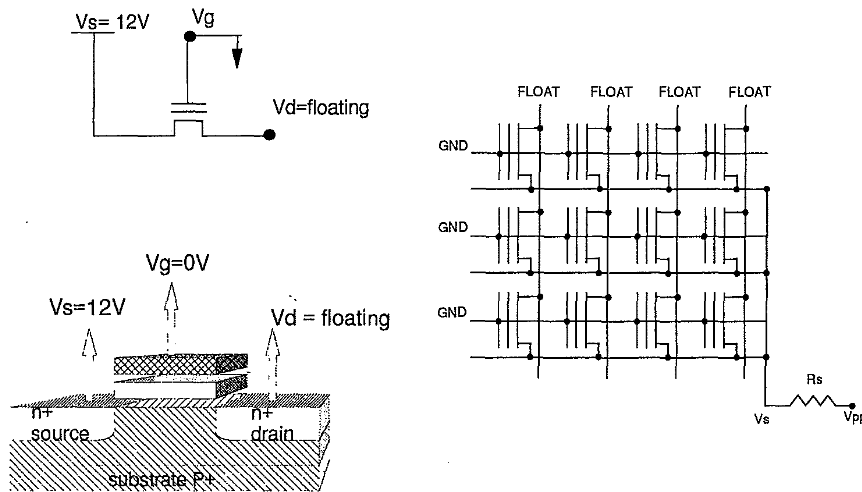


Figure 5.3 NOR Flash cell: source Erase operation.

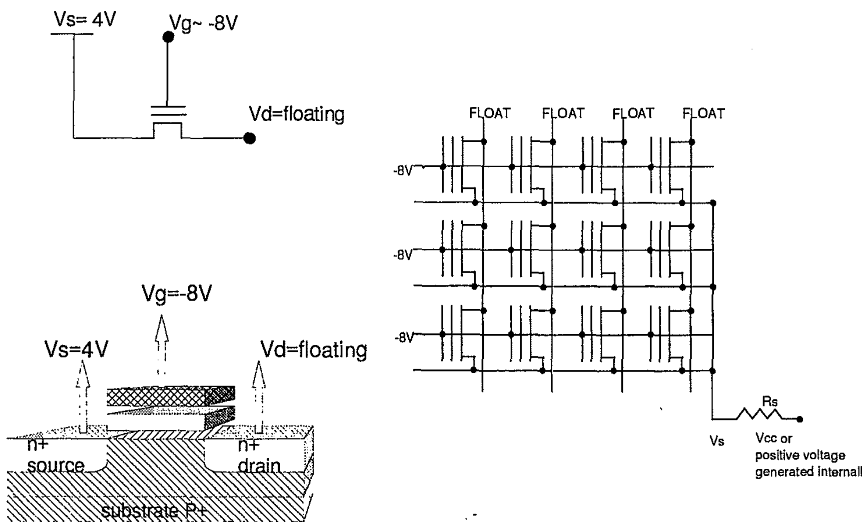


Figure 5.4 NOR Flash cell: negative gate Erase operation.

From the description of both Program and Erase operations, it is clear that a Flash memory needs voltages which could be higher than V_{CC} or lower than GND. The fact that these voltages are provided externally or generated internally accounts for another classification: a Flash device is "Double supply" if one pin is dedicated for Program/Erase (V_{PP} power supply) and another one is used as general purpose power supply V_{CC} (usually V_{PP} is forced at higher

APPLE INC.
EXHIBIT 1008 - PAGE 0255

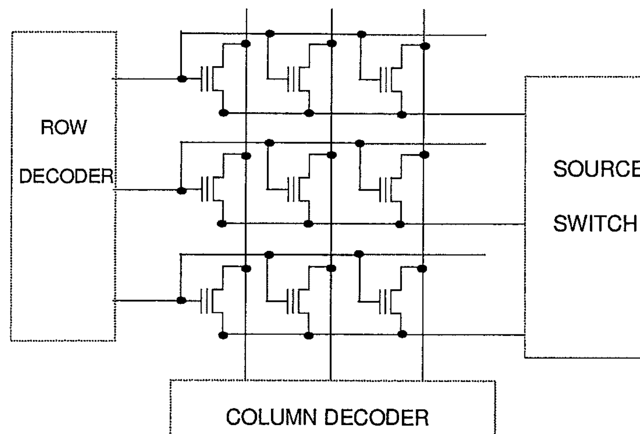


Figure 5.5 NOR Flash memory array organization.

voltage value than V_{CC}); it is called "Single supply" if only V_{CC} pin is present, and the other voltages are generated internally.

After analyzing the different operations, it is possible to get deeper inside matrix organization (see Fig. 5.5).

Through the row and the column decoders the proper cell is addressed and through the source switch the voltage necessary to erase the cells is forced. For the present generation of devices, the total number of output pins is 8 (x8 organization) or 16 (x16 organization). The total number of output pins will be 32 in the next generation. It is straightforward that the same number of columns converge to each output and that a reading circuitry is attached to it to evaluate the cell contents, as shown in Fig. 5.6: this figure represent a bulk erase Flash memory, since the source line is common for every cell.

In some cases the same device can be used in both ways (x8/x16 user-selectable), by means of the BYTE# pin; in this case, the Read operation is actually performed by 16 regardless the configuration chosen, and the selection of either the lower or the upper byte is done just before the output buffers.

Another distinction can be made as for the organization of the source lines; the first generation of Flash devices was "bulk erasable", i.e. all the array was erased simultaneously without any type of selectivity inside the matrix. Then "sector erasable" Flash memories were designed, in which the array of cells is physically divided in different sectors, each one erasable separately by means of a dedicated source switch. The sectors can be of equal or different size, like in the Boot Block architecture; in this case the so-called "boot block" has the

APPLE INC.
EXHIBIT 1008 - PAGE 0256

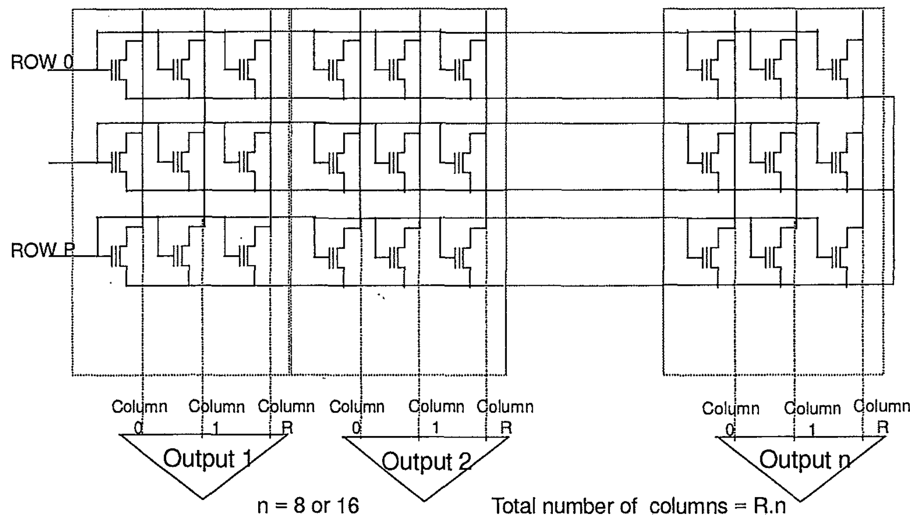


Figure 5.6 NOR bulk erase Flash memory array organization.

peculiar feature that its content can be protected/unprotected, thus preventing the user from undesired modification.

If V_{PP} pin is present, the source erase scheme is most often used; in other words the source of the block to be erased is raised and the gates are put to GND; the sectors are usually divided by columns, because cells of different sectors share the same row, but have different source lines and bitlines so that the Erase operation can be performed inside each sector separately. In Fig. 5.7 and Fig. 5.8 sectorization schemes for a Double supply Flash device are shown.

For each of these methods there are advantages and drawbacks that can be easily understood, such as the number of source switches, number of sensing circuitry, ... The sectorization of a Double supply device can also be done by rows, separating each row for different sectors and using the negative gate erase approach. In this case, when Erase is performed, all the rows of the sector must be tied at GND, independently of the rows of the other sectors.

In theory, there is no limit to sector dimension: in practice, the possibility of a much finer sectorization is limited by area/cost/performance reasons due to:

- source line separation and switching;
- complex routing due to word/byte organization;
- complex/more expensive redundancy;
- impact on P/E cycling endurance.

APPLE INC.
EXHIBIT 1008 - PAGE 0257

A Single supply Flash memory (only V_{CC} pin present on the device) is usually sectored by rows (Fig. 5.9); the divided bitline scheme is implemented, so the main bitline is common to all sectors, while the local bitlines are assigned to

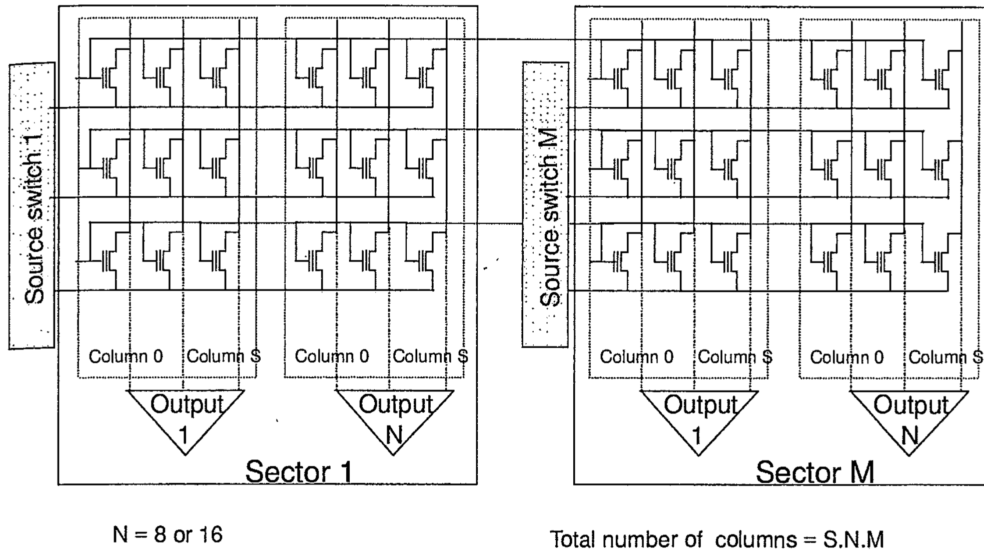


Figure 5.7 Double supply sectorized Flash memory array organization by sectors.

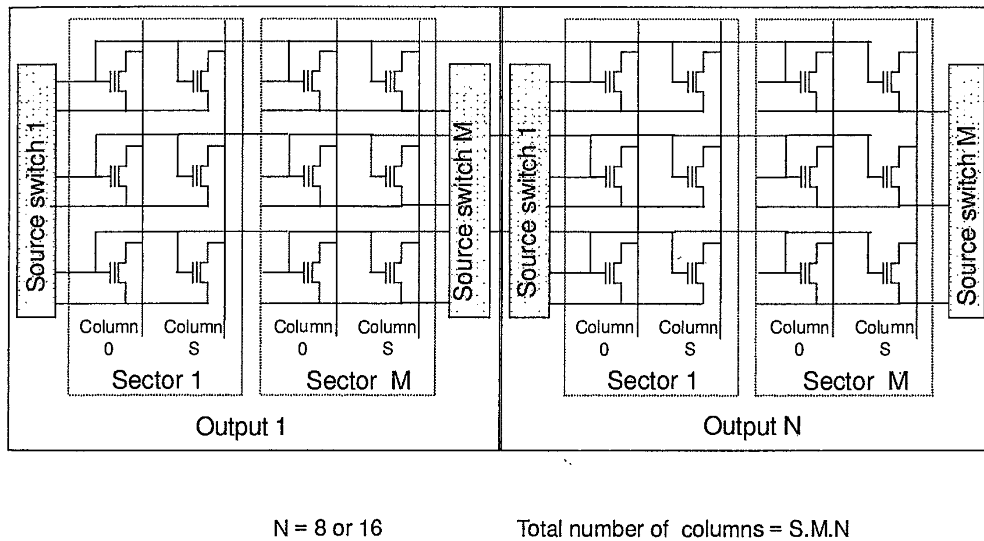


Figure 5.8 Double supply sectorized Flash memory array organization by outputs.

APPLE INC.
EXHIBIT 1008 - PAGE 0258

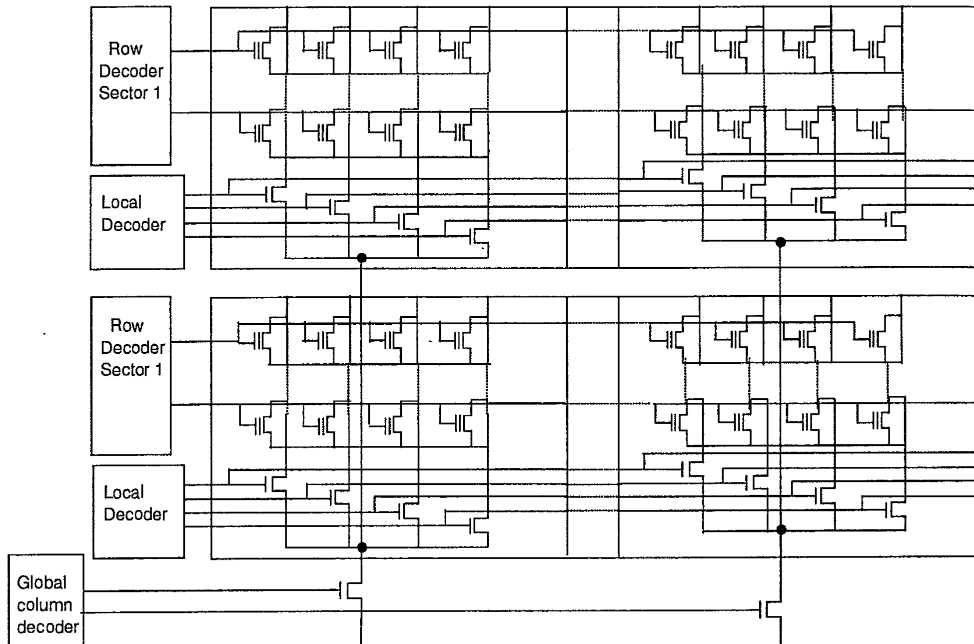


Figure 5.9 Single supply - negative gate erase Flash memory array organization.

each sector. Normally the second metal layer is used to connect the local bitlines with the main bitline. In this type of organization, in absence of a third level of metal, the key issue for the access time is the time constant of the wordline; since both the first and the second level of metal are used to realize the local and the main bitlines, there is no way to realize a strap on the wordline.

The voltages that are required for Program and Erase are generated internally starting from the V_{CC} value available. It is clear the issue that arises when the V_{CC} is low: it is from this low value of power supply that the internal voltages for Program and Erase must be generated. The increment of the voltage is obtained by using charge pump circuits in which the key element is the capacitor: these circuits are analyzed in detail in Section 5.6.

5.1.3 Flash Memory User Interface

A Flash memory can be controlled using some particular pins; typically they are:

- CHIP ENABLE (E#)
- OUTPUT ENABLE (G#)

APPLE INC.
EXHIBIT 1008 - PAGE 0259

- WRITE ENABLE (W#)
- POWER DOWN/RESET (RP#)
- WRITE PROTECT (WP#)
- BYTE#

In addition, some address pins can be used and other particular control pins are added in dedicated devices. It is worth explaining the meaning of these control pins.

The CHIP ENABLE (E#) activates the memory control logic, input buffers, decoders and sense amplifiers. E# High unselects the device and reduces the power consumption to the stand-by level. E# can also be used to control writing to the command register and to the memory array, while W# remains at Low level. Addresses and Data, which represent the command to be performed, are latched on either the falling or the rising edge of E#.

The OUTPUT ENABLE (G#) gates the output through the data buffers during a Read operation.

The WRITE ENABLE (W#) controls writing to the Command, Address and Data latches when E# is Low. Again, Addresses and Data are latched on either the falling or on the rising edge of E#.

The READY/BUSY# (R/B#) pin is an open drain output pin which is Low when a modify operation is in progress.

The POWER DOWN/RESET (RP#) pin provides a hardware reset function when it is Low: the memory will be reset after a falling edge on RP# and the reset pulse width must be large enough to avoid undesired reset caused by spikes on the RP# pin. If the hardware reset is requested during a Program or Erase operation, the device will not respond immediately, since the ongoing algorithm should not be terminated abruptly; if the hardware reset is requested during all the other operations (R/B# pin High), the device will respond after the RP# pulse. Any hardware reset during Program or Erase operation will corrupt the content of the addressed memory. Additionally, the RP# pin provides protection against undesired command writes due to invalid system bus conditions that may occur during system reset and power up/down sequences.

The WRITE PROTECT (WP#) pin provides the lock and unlock of certain sectors. Usually, when it is High sectors are unlocked, so their content can be modified.

The BYTE# pin selects the output configuration for the device: byte (x8) or word (x16) mode. When BYTE# is Low, the x8 mode is selected and the data are read or programmed on DQ0-DQ7 pins. Under this mode, DQ8-DQ15 pins are at high impedance and the lower or higher byte to be output is selected by

APPLE INC.
EXHIBIT 1008 - PAGE 0260

means of a further address pin (in some devices this function is accomplished by DQ15 pin, usually named DQ15/A_1). When BYTE# is High, x16 mode is selected and data are read or programmed on DQ0-DQ15.

By applying proper voltages on the control pins, it is possible to perform some simple operations, like:

- Read Electronic Signature;
- Boot Block Temporary Unprotection;
- Sector Protection and Temporary Unprotection;
- Output Disable;
- Stand-by;
- Power Down/Reset;
- Command Write.

More complex operation (like Program or Erase) require a complex control sequence: in the first generation of Flash memories, this task was accomplished by the external microprocessor. In the new generation of Flash Memories, complex algorithms are embedded and are activated by a proper command sequence; the minimum command set is composed by the following instructions:

- Read Electronic Signature;
- Program Set-Up;
- Erase Set-Up and Erase Confirm (if chip erase is available);
- Sector Erase Set-Up and erase Confirm, with the address of the sector(s) to be erased;
- Erase Suspend and Erase Resume (only during sector erase);
- Read Array.

Each Flash is powered-up in Read Array mode, to avoid spurious damaging of the array content, then the Command Interpreter can decode the commands and drive the device in the proper mode to execute them.

A command is made up of a sequence of binary codes given to the device through the I/O pins and the address pins with specified timing on the control pins (E# and/or W#) which act as a clock for the Command Interpreter.

APPLE INC.
EXHIBIT 1008 - PAGE 0261

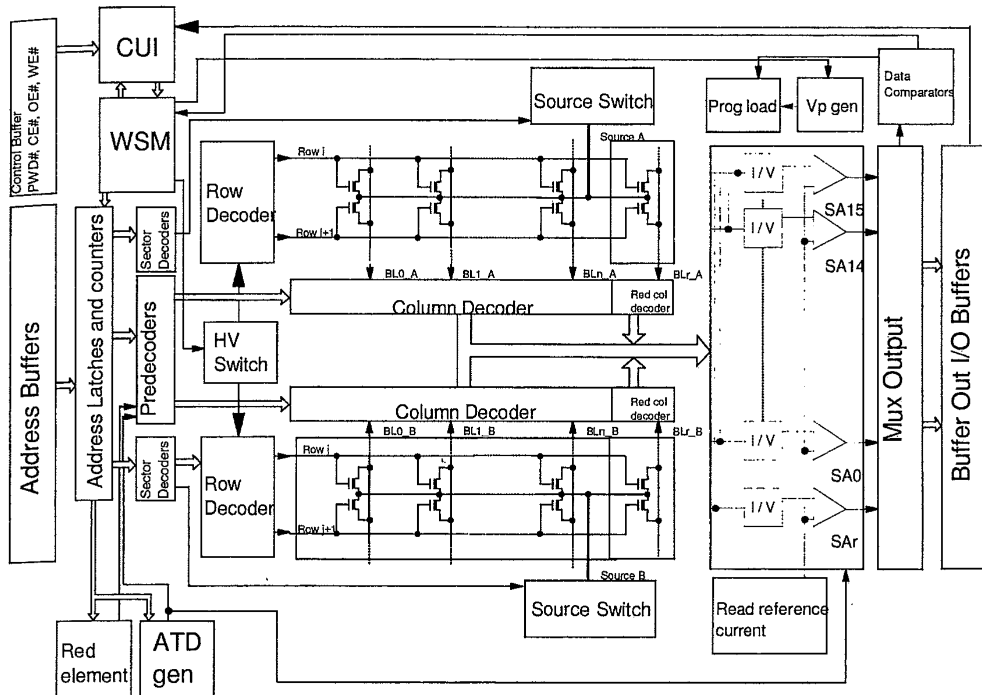


Figure 5.10 General architecture scheme.

5.1.4 Flash Memory Operations: Overview

The building blocks of a Flash device are common to all Flash memories, although the circuital implementation is strongly dependent on the technology available. In addition, it can be slightly different due to specifications requirements like operating voltages, DC and AC performances.

As it was explained in Section 5.1.2, in a Dual voltage device the high voltages necessary for Program and Erase are derived from V_{PP}, whereas in a Single Voltage Flash memory the high voltages are generated internally. This consideration should be taken into account when considering circuit implementation.

A "common" architecture for a Flash memory is shown in Fig. 5.10.

A detailed analysis of each block will be performed throughout this chapter; the following section gives a brief overview of the circuits involved in the main operations performed inside a Flash memory.

5.1.4.1 Read Path Building Blocks Description. As far as stored information is concerned, Flash memories can be divided into two categories, "standard" and "multilevel"; in the former case, every cell contains one bit of

APPLE INC.
EXHIBIT 1008 - PAGE 0262

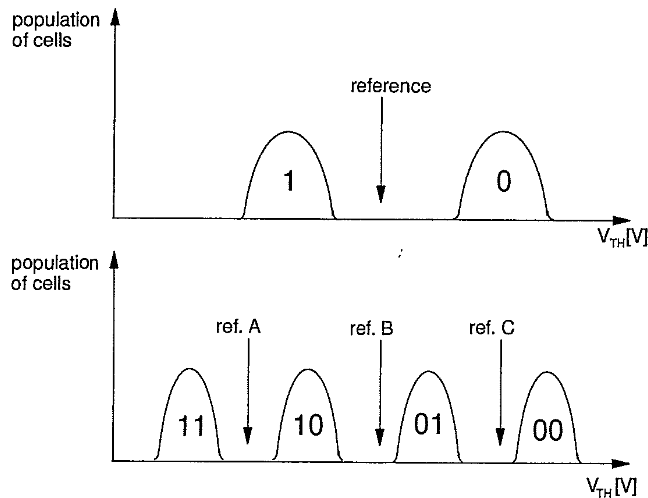


Figure 5.11 Flash cell: threshold distributions.

information, while in the latter 2^n bits of information are stored in a single cell. The main issue with a multilevel Flash memory is technological, since its development demands a great confidence in the process, because a tight distribution of the threshold of the programmed cells is needed.

In Fig. 5.11, threshold distributions for both standard and multilevel Flash memories are shown, together with the criteria used to determine the content of the cell.

No matter how many bits are stored, the complex of activities necessary to detect the cell content is referred to as Read operation; this operation is performed by a set of circuits which form the so-called Read path.

First of all, the address of the data is acquired by the input buffers: these circuits must fulfill requirements such as the noise immunity (to avoid a spurious commutation of the inputs), TTL voltage compliance and third level sensitivity (some input pins are connected not only to the usual buffer, but also to a circuit whose output goes high when the input reaches two or three times the V_{CC} value, in order to perform certain operations).

The input buffers drive the decoding circuits, which are made up of a predecoder followed by the decoder (see Section 5.2), whose purpose is to bias the selected cells with the proper read voltage; eight or sixteen cells are simultaneously selected through the column predecoder and decoder, and the proper drain voltage is forced to the cells.

In order to determine the value stored in the cells, a differential sensing is performed, comparing the matrix cells with a known reference. Being a current

APPLE INC.
EXHIBIT 1008 - PAGE 0263

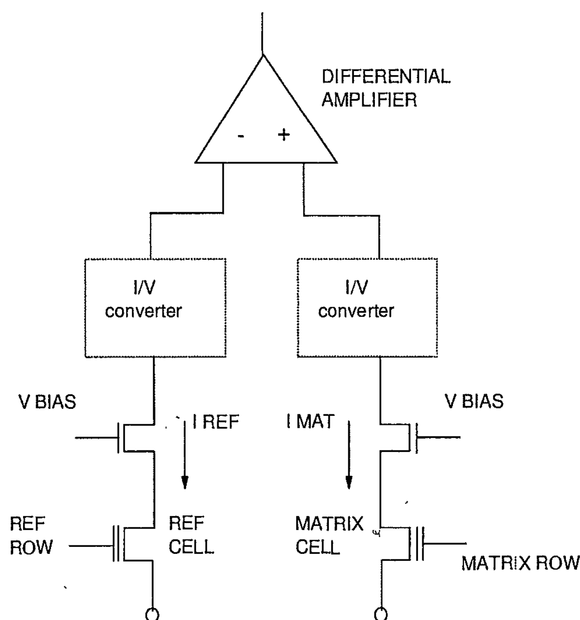


Figure 5.12 Differential sensing scheme.

comparison (a Flash cell sinks more or less current depending on its threshold voltage), a current-to-voltage conversion is performed on both the matrix and the reference branches (where transistors are used as loads), and finally the two voltages are compared by a differential amplifier (Fig. 5.12). It is worth noting that if both cells or both load transistors should vary in strength, this difference would result in a common mode voltage shift but not in a differential voltage variation, thus achieving a correct sensing.

If the cell of the matrix is programmed, it drains less current than the reference cell: the output voltage from the matrix cell is higher than that of the reference cell. If the cell of the matrix is virgin or erased, it has the same current as the reference cell: the cell can be shown virgin (more current than in the reference) if for example the load transistor for reference side is larger than the matrix load (these issues are deeply analyzed in Section 5.4).

The result of the sensing (which is a voltage) passes through a multiplexer to the output buffers, which cause the external data pin to change their state accordingly; the main problem is the noise introduced on V_{CC} and GND due to the large current required to charge and discharge the huge load capacitance of the data output pins.

As to the general issues related to the Read operation, great efforts are spent in developing solutions to reduce as much as possible the delay between

APPLE INC.
EXHIBIT 1008 - PAGE 0264

the address transition and the valid data at the output buffers; in other words, to achieve a faster access time.

5.1.4.2 Program Path Building Blocks Description. A Flash device can be programmed on a word or byte basis. As it was pointed out in Section 5.1.3, Program is a very complex and slow operation which is performed, in the actual Flash generation, by an embedded Program/Erase Controller; the user has only to provide the correct instruction, which is composed of a sequence of different commands, followed by the data to be programmed together with their address location. Then, the Program/Erase Controller automatically starts and performs the Program operation. The user or the external microcontroller can check the status of the operation using either the R/B# pin or the status on DQ0-DQ7 (the former can be issued either in "Data Polling - Toggle Bits" or "Status Register" fashion). In some Flash memories the Program can be suspended to read some other locations inside the array.

When the algorithm starts, the content of the memory location to be programmed is compared with the data: if they are valid (since to program means to change cell status from "1" to "0", and the opposite task can only be performed by an Erase operation, it is wrong to try to program a "1" onto a "0"), then a high voltage (HV switch of Fig. 5.10) is driven on the wordline of the cell to be programmed and then a program pulse is applied on the drain of the cell through the V_p generation circuit (Fig. 5.10). At the end of the program pulse, another verify is performed, reading the byte/word and comparing it with the data to be programmed (latched at the beginning of the operation). If comparison fails, another program pulse is applied to the uncompletely programmed bits. Fig. 5.13 shows a simplified flow chart and the timing of the main signals involved. The program efficiency decreases while threshold voltage increases: therefore it is possible to obtain a common value for programmed cells threshold voltages, even if the starting values are different.

5.1.4.3 Erase Path Building Blocks Description. The Erase operation is performed on all the cells inside a sector (thus sharing a common source) and can be divided in three fundamental steps:

1. Program All 0;

2. Erase;

3. Recovery of over-erased cells.

With the first step all the bits of the sector are put in the same condition as for their threshold voltages; an internal counter is used to address all the

APPLE INC.
EXHIBIT 1008 - PAGE 0265

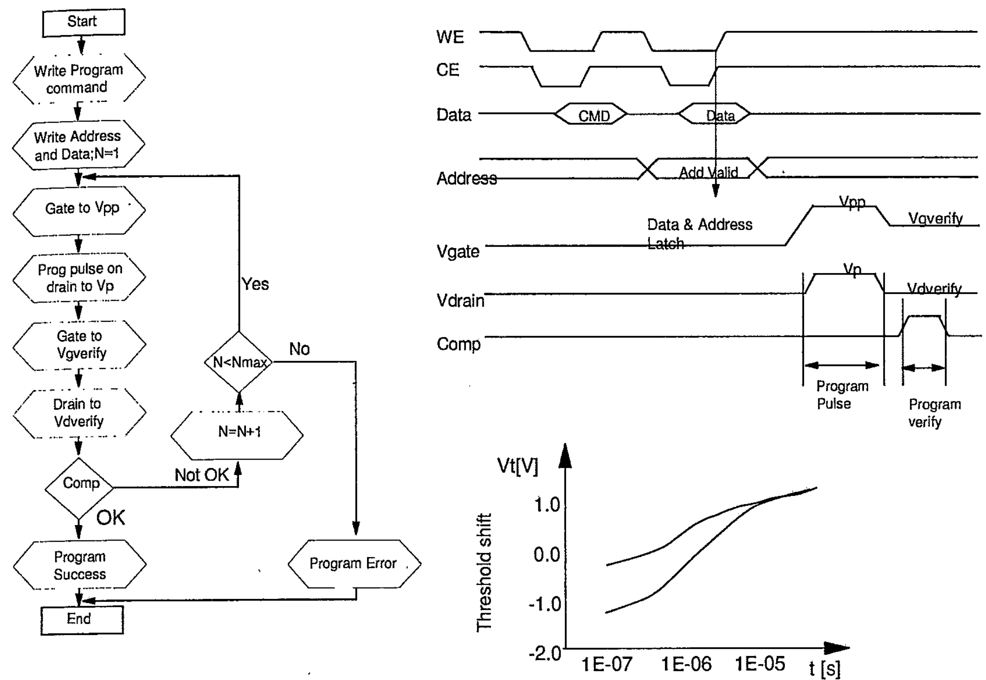


Figure 5.13 Program operation - flowchart.

words inside the sector, since Program operation is performed sequentially on every word.

The second step consists in providing a high voltage on the common source through Source switch keeping all the gates to GND and all the drain floating (this is valid both in a source and a gate erase scheme). If high voltage (V_{PP}) is not available on input pin, an internal circuit will generate it. At the end of erase pulse, a verify is performed on every word, in order to see if all the cells have been erased. If this verify fails, another erase pulse is given.

The third step performs a depletion verify to detect the over-erased bits. If present, these bits are recovered by programming them with a proper algorithm.

A simplified erase flow chart and the erase timing are shown in Fig. 5.14.

5.2 READ PATH: DECODING

As far as information storage or retrieval is concerned, it is necessary to devise a method to address the data inside the memory array. As already pointed out in Section 5.1.2, NOR matrix organization is considered: Flash cells are arranged in rows and columns selected by means of separate address decoding

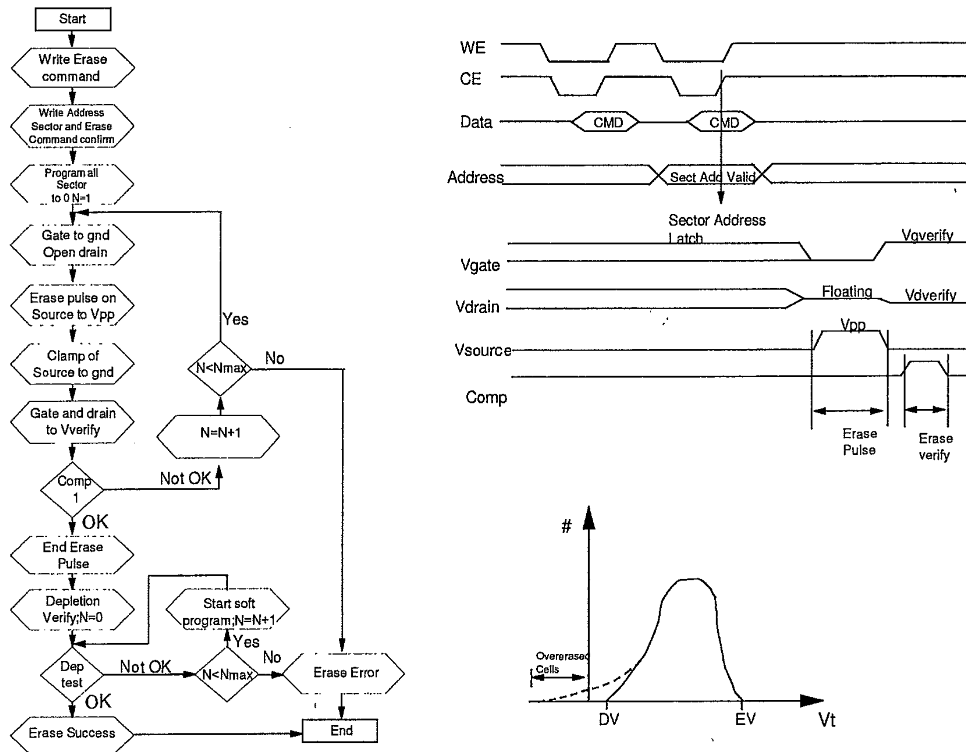


Figure 5.14 Erase operation - flowchart.

circuitry. Cells inside every couple of adjacent rows are specular, so that both source lines and drain diffusions are common every two cells. Cells control gates are realized in polysilicon, arranged in stripes called wordlines; cells drains are contacted by metal stripes called bitlines.

Under these assumptions, decoding means to select a single row and a single column in order to address the desired data inside the matrix: the externally-provided address, which is a n binary digits number, must be converted into a corresponding 2^n binary digits data, in which only one bit at a time is on. Furthermore, in the case of a non-volatile memory, decoding circuitry should also be able to handle all the different biasing voltages necessary to Erase, Program and Read a memory cell.

In the following paragraph, for sake of simplicity, the whole subject is divided in two separate parts: the "predecoding", i.e. the circuitry which selects the row and the column depending on the address, and the "decoding", i.e. the circuitry whose task is to bias both cell gates and drains.

APPLE INC.
EXHIBIT 1008 - PAGE 0267

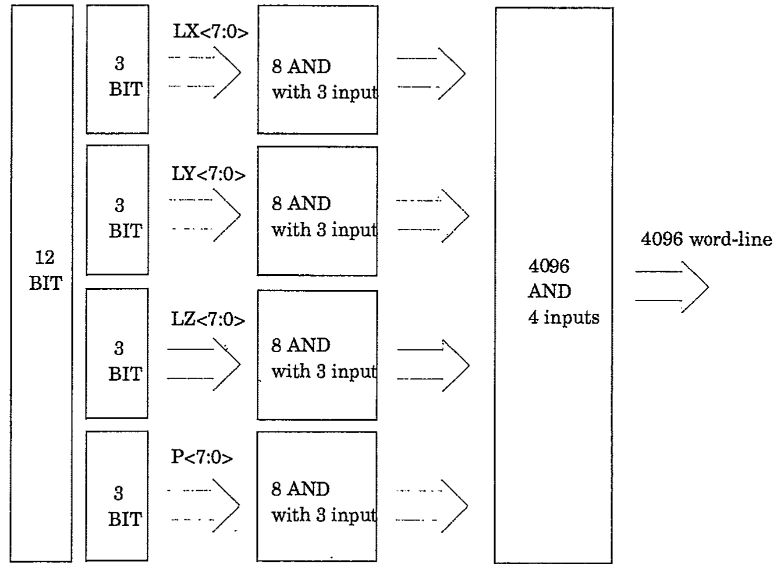


Figure 5.15 Predecoding stage.

5.2.1 Predecoding

In principle, considering a Flash memory composed of n rows, predecoding could be implemented using 2^n AND gates whose inputs are the possible combinations of the n addresses; of course such a solution is not acceptable because of the amount of area required and the layout complexity when densities of Mbits are reached. Therefore a hierarchical structure has been devised, in which the address bits are grouped into several subsets and processed separately. In Fig. 5.15 an example is shown, where the address width is 12; predecoding is realized using 32 3-inputs AND plus 2^{12} 4-inputs AND instead of 2^{12} 12-inputs AND.

5.2.2 Row Decoder

Row decoders address individual rows in a memory array according to the applied coded address. The basic scheme of row decoders may be represented by a number of inverters (one for each row) controlled by a combinational circuit, which receives the input addresses and drives the inverters so that only one at a time presents a High output (see Fig. 5.16a). The simplified arrangement described above operates correctly in Read mode, wherein both the combinational circuit and the inverters present read voltage V_{CC} as the High logic level, but not in Program mode, where the combinational circuit

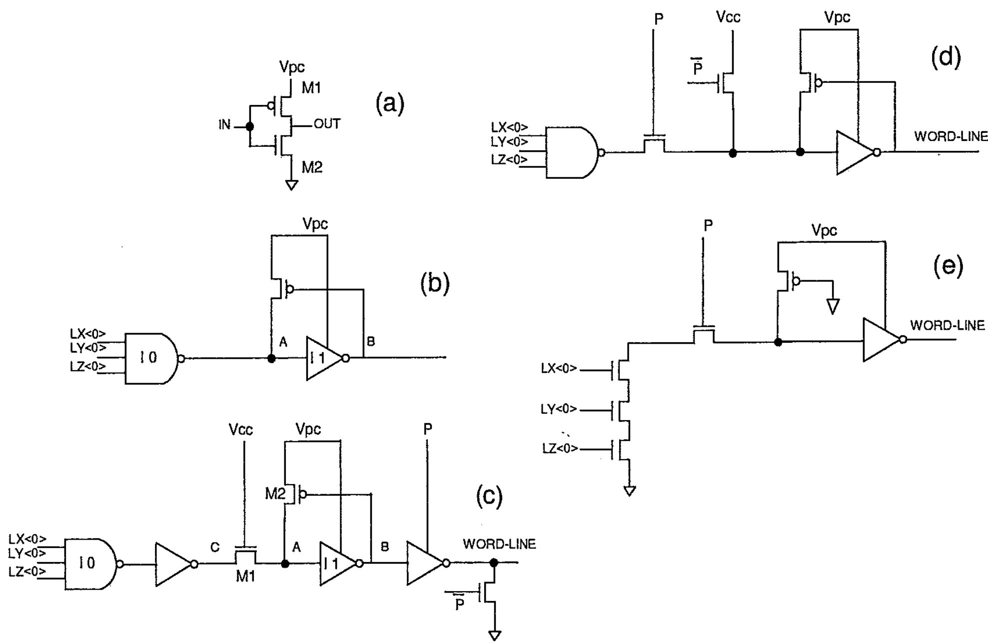


Figure 5.16 Row decoders. LX, LY, LZ and P are predecoding signals. $V_{PC} = V_{CC}$ or V_{PP}.

supplies read voltage V_{CC} as the High logic level at input IN of the non-selected row inverters, whereas supply line V_{PC} is at programming voltage $V_{PP} > V_{CC}$. Consequently, a voltage drop exists between the gate and source terminals of the pull-up transistor of the inverter, and when it reaches the threshold voltage, M1 is turned on: in such a way output OUT is prevented from reaching the zero voltage required to avoid stress on the connected cells and to ensure a correct logic level at the output.

One possible solution is to use a positive-feedback inverter with a pMOS feedback transistor connected between V_{PC} and input A, and with the gate terminal connected to output B (Fig. 5.16b). As such, when the voltage at output B falls, the feedback transistor is turned on and connects node A to the programming voltage V_{PP}, thus ensuring a complete turn-off of the pull-up transistor of the inverter I1 and a zero output voltage. The above solution, however, presents some drawbacks. First of all, layout problems arise owing to the output of the inverter being brought back; solving the problem by driving the feedback transistor with a separate signal in turn creates problems in terms of synchronization. Secondly, problems arise as regards direct biasing of the drain-bulk junction of the pMOS transistors of NAND gate I0, which

APPLE INC.
EXHIBIT 1008 - PAGE 0269

would have the source and bulk regions biased at V_{CC} and the drain regions (connected to the output) biased at V_{PP}. One possible solution is to provide an nMOS pass transistor to separate the low-voltage (predecoding) section from the high-voltage (actual decoding) section. Such a solution is shown in Fig. 5.16c wherein a 3-input AND, supplied at Read voltage V_{CC} and forming part of the combinatorial circuit for selecting the row, drives inverter I1 via an nMOS pass transistor M1 with the gate terminal biased at V_{CC}. When the node C is at V_{CC}, pass transistor M1 operates as a diode by presenting two terminals (the gate terminal and the terminal connected to C) at the same voltage, and therefore causes, between node C and node A, a voltage drop equal to its threshold voltage (plus body effect). In addition to further complicating the circuitry, the solution shown in Fig. 5.16c is also unsatisfactory in the presence of low supply voltage: the voltage drop across pass transistor M1 prevents node A from reaching the high voltage required to ensure that the pull-up transistor of I1 is turned off completely. Figs. 5.16d and 5.16e contain other commonly implemented solutions; the basic principle is the same.

Moreover, besides merely shifting the problem of undesired biasing to other parts of the circuit, a CMOS pass switch is too bulky to be accommodated in the decoding stage, which is formed within the spacing between the array rows.

Another type of decoder can be implemented; it operates correctly both in Read and Program, even in the presence of low supply voltages, and it involves no problems in terms of layout or synchronization. In the row decoder the high (programming) voltage is supplied not only to the final inverter (decoding) stage but also to the predecoding stage: to this purpose the predecoding stage presents two parallel paths, one supplied with low voltage and used in Read mode, and the other supplied with high voltage and used in Program mode (Fig. 5.17). A CMOS switch separates the two paths, and it is driven by the high voltage already available in the predecoding stage, and, being formed at predecoding level, involves none of the integration problems posed by the final decoding stage, because the predecoder is placed outside the matrix.

If the logic signal ENHV is Low, then node A is Low as well, and the pass CMOS is on, thus allowing the signal IN to get to OUT. In this situation, since the memory is in Read mode, the node V_{PC} is held at V_{CC}. To let a voltage V_{PC} (not equal to V_{CC}) reach the row decoder, ENHV is biased at V_{CC}; then the node A, through the feedback structure which consists of M1, M2, M3, M4 and I1 (level shifter), is brought at V_{PC} and switches off the pass CMOS. By means of another level shifter, the signal IN is raised at V_{PC} and is thus able to drive the inverter I2 in CMOS logic. The advantages of the circuit described are the following: it operates correctly even at low voltage, by featuring no nMOS pass transistor or other components involving a voltage drop at the terminals. It ensures correct output readings, and prevents stressing the cells

APPLE INC.
EXHIBIT 1008 - PAGE 0270

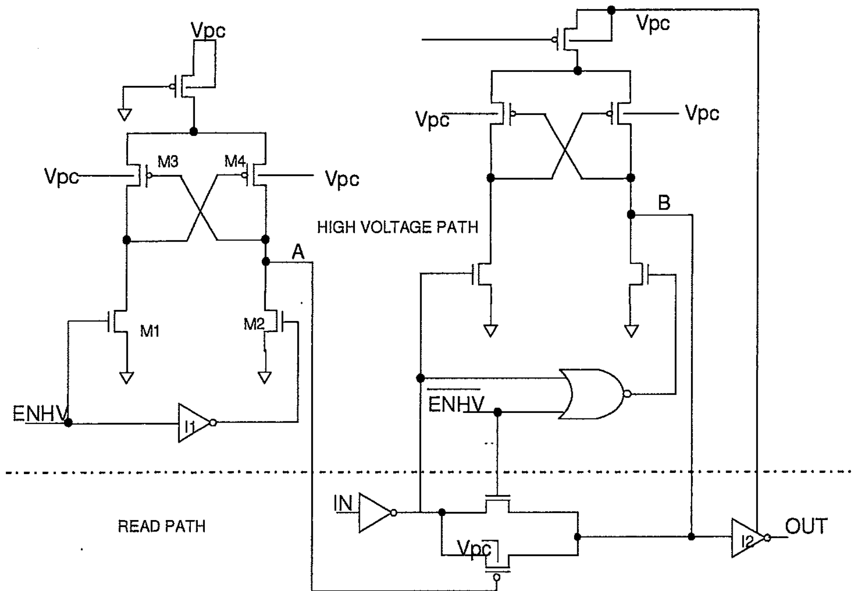


Figure 5.17 Predecoding stage with two parallel paths.

connected to the non-selected rows, by featuring supply branches ensuring that the components along the transmission paths of the row drive signals are turned off or on completely. The formation of a high-voltage path in the predecoding circuit enables the use of a CMOS pass switch, which, despite presenting no voltage drop, requires a high-voltage drive circuit for its correct operation, and is too bulky to be accommodated in the final decoding circuit. Despite the use of numerous high-voltage components in predecoding stage, the layout of the decoder is simplified since it does not require any feedback branch. Finally, by forming two separate paths in the predecoding circuit, the low-voltage path may be extremely simple (merely a connecting line) to greatly reduce access time, and the high-voltage path may be optimized by using voltage shifters, although they introduce a slight delay in signal propagation (it is a feedback structure), at no expenses in terms of read performance.

After analyzing the circuitry connecting V_{CC}-biased and V_{PP}-biased parts, it is possible to take a deeper look at the structure of the final inverter of the row decoder. The design of this circuitry is strictly related both to the type of erasing and to the available technology (these concepts will be considered again in Section 5.6.2). To electrically erase a Flash memory cell, it is necessary to apply a proper voltage bias (typically 12V) between the source and gate terminals of the cell. A "source Erase" occurs when the gate is held at GND

APPLE INC.
EXHIBIT 1008 - PAGE 0271

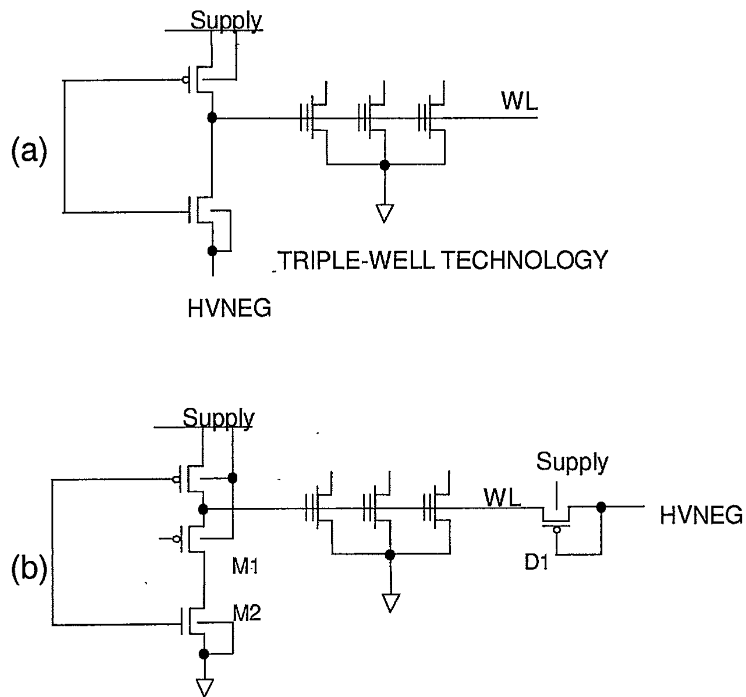


Figure 5.18 Possible row decoders for negative gate erase.

and the whole erase voltage is applied on the source of the sector. In this case the final inverter is simply a CMOS one. If "negative gate Erase" is required (i.e. the gate is negative, typically −8V, and the source positive, typically 5V) a simple CMOS inverter is no longer suitable. If a triple-well technology is available, i.e. if it is possible to realize nMOS transistors whose bulk can be biased separately from the chip substrate, then a negative voltage can be passed directly through the final inverter (Fig. 5.18a); otherwise it is necessary to isolate the nMOS in the final inverter from the wordline in order to avoid direct biasing of drain/p-well junction when negative voltage is applied to the wordline itself. In this case, the negative voltage HVNEG is supplied to the control gate of the cells through an insulation diode D1 and M2 is protected via M1 (Fig. 5.18b).

5.2.3 Column Decoder

Once the row has been selected, column decoding must be performed in order to select univocally the desired memory cells. If the matrix is m columns wide

APPLE INC.
EXHIBIT 1008 - PAGE 0272

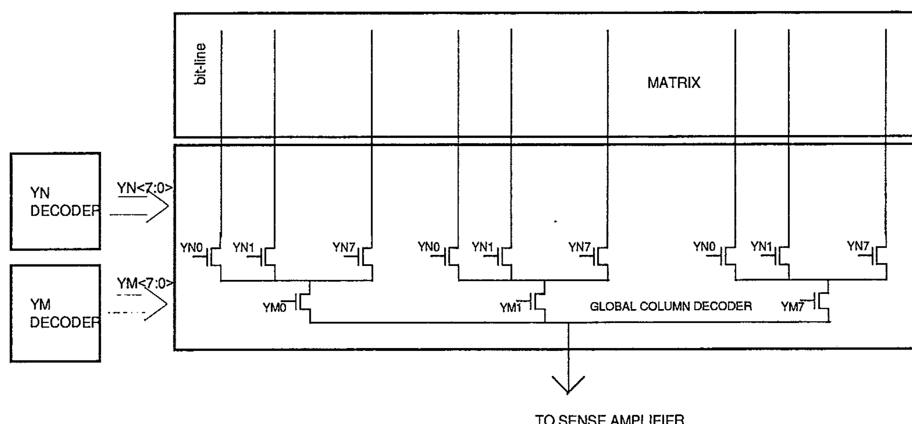


Figure 5.19 Column decoder.

and the output is composed of b bits, then it is necessary to have m/b mutually exclusive decoding signals, since every access is made on b bits at a time. The same division between predecoding and decoding applies: predecoding is a hierarchical combinational logic circuitry which forms the first level signals labeled in Fig. 5.19 as YM and YN. Decoding is composed of a nMOS transistors multiplexer, repeated for every output bit. The task of the column decoder is to pass the proper biasing voltage on the drain of the cell. YM and YN decoders are supplied with VPCY: in Read mode this voltage equals V_{CC}, while in Program mode it is raised to V_{PP}, because this circuitry is used to supply the drain of the cells to be programmed with a suitable voltage (typically $\approx$ 6V). The solutions devised to interface V_{CC} and VPCY are the same as in the row decoder.

5.2.4 Hierarchical Decoder

As outlined in Section 5.1.2, sectors in a non-volatile memory can be organized either by row or by column. In the latter case, wordlines are common to every sector and sector dimension is chosen according to the number of columns comprised in it (Fig. 5.20). A sector is defined as the set of cells which share the same source node, i.e. those cells which are erased together. Using this kind of architecture, the parasitic capacitance of the bitline is limited, and the sense amplifier greatly benefits from this solution.

Unfortunately there are also several drawbacks: first of all, the wordline, i.e. the line which connects all the control gates of the cells on the same row, is realized in polysilicon. From an electrical point of view, it can be thought as a distributed RC network: parasitic resistance depends on the polysilicon, while

APPLE INC.
EXHIBIT 1008 - PAGE 0273

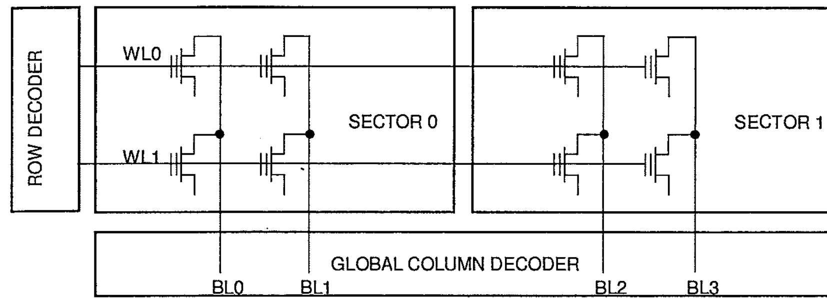


Figure 5.20 Sectors organized by column.

parasitic capacitance depends on the gate contact of the cells. Even considering a small number of cells (for instance one thousand) the time constant associated to this network is several nanoseconds. This delay directly affects access time, since it causes a delay in the correct biasing of the selected cell. Considering a two-metal technological process, one layer is used for the bitlines, while the other is used to short-circuit the wordline, thus reducing the parasitic resistance and the associated time constant.

Another problem of the shared wordline approach is that every time a cell is addressed, all the other cells on the same wordline are biased as well, thus suffering the so-called gate stress, which could affect the correct functionality of the cells during the device lifetime. Furthermore, the shared wordline approach is not the best choice in the case of negative gate erase, because the positive voltage drop is also applied to the cells of the sectors which are not to be erased. One way to overcome these problems, in particular the one related to gate stress, is to choose the row organization approach. In this case it is the bitline that is shared by all the sectors and it is the number of rows that influence the dimension of the sector. Actually it is not possible to have one single shared bitline, in order to avoid the drain stress; every sector has its set of bitlines, called local bitlines: every single bitline is connected by means of a pass transistor to another metal layer called main bitline, as shown in Fig. 5.21.

For each sector there is a set of local pass transistors, which are on only in the addressed sector, thus avoiding drain stress effects on the cells of other sectors. Because of the high integration of the memory cells, local and main bitlines are usually designed using two metal layers; if a third metal layer is not provided, it is impossible to short-circuit the wordline and the access time is greatly affected by this problem: therefore it is necessary to use a

APPLE INC.
EXHIBIT 1008 - PAGE 0274

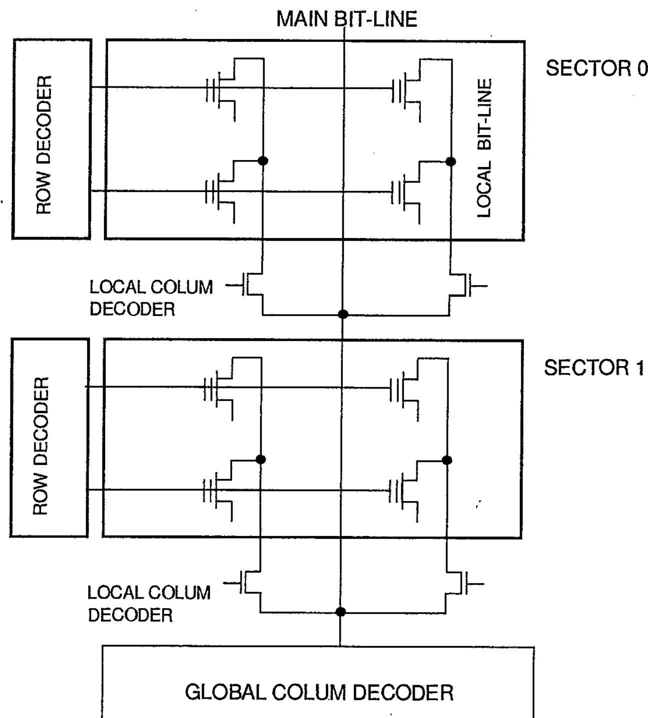


Figure 5.21 Sectors organized by row.

technology in which polysilicon has a low resistance. The row organization approach has a considerable impact on the structure of the column decoder, because it requires a local decoding stage in every sector, whose effects on the area are not negligible. On the other hand, eliminating any kind of stress on the cells greatly improves reliability.

5.2.5 Low V_{CC} Problems

As already outlined, cells of non-volatile memories are read by biasing their gate with the read voltage and by detecting the current flowing through the cell. If the cell is programmed, its threshold voltage must be higher than the read voltage, so that no current is drawn by the cell; if the cell is erased, its threshold voltage must be such as to let the current flow through; detecting the current flow provides for discriminating between programmed and erased cells. To ensure correct read operation and reliable cycling (multiple cycle operation) of the memory array, certain limits must be observed in the distribution of the threshold voltages of the cells. More specifically, currently used tech-

nologies require that the threshold of the best erased cells be above zero, and the threshold voltage of the worst erased cells be about 2.5V. The lower limit substantially arises from the need to prevent read errors caused by depleted cells (cells with a threshold voltage below zero); while the upper limit is due to the intrinsic distribution of the cell threshold according to the fabrication technology used. Since the read voltage normally coincides with the supply voltage, a supply voltage of over 3V poses no problems. A problem arises in the case of memories operating at low V_{CC}. In fact, with a supply voltage V_{CC} of 2.5V, all the cells with a threshold voltage V_{TH} close to this value conduct little or no current, so that the cell is considered programmed, thus resulting in a read error. A solution to the problem consists in boosting the read voltage, i.e. supplying the gate terminal of the cell to be read with a voltage higher than the V_{CC} which is generated by an appropriate boosting stage.

5.2.6 Boost Concept: Continuous Boost and "One-shot" Bost

A possible scheme of a boosting circuitry is shown in Fig. 5.22 together with the required timing of all the signals involved; BN is the node whose voltage is to be boosted.

At the beginning, the auxiliary boost capacitance (C_{boost}) and the parasitic one (C_{load}) are precharged to the supply voltage via the pMOS transistor. When boosting is needed, signal B switches to zero, thus raising the lower "plate" of C_{boost} to V_{CC}. At the same time M1 is switched off and node BN is isolated. This configuration produces the charge-sharing phenomenon due

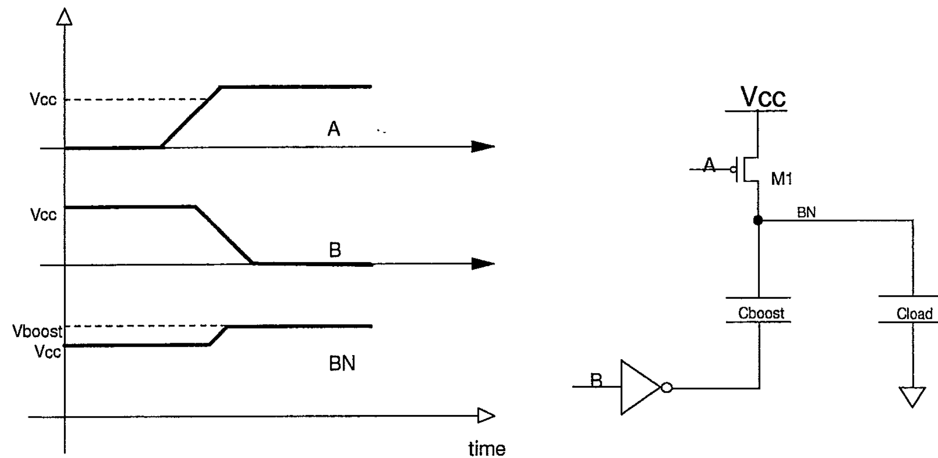


Figure 5.22 Boosting principle.

to the charge conservation on the isolated node. By imposing the Kirchhoff voltage law and the energy conservation, the final voltage on BN is

$$V_{BN} = V_{CC} + \frac{C_{boost}}{C_{boost} + C_{load}} V_{CC} \ . \qquad (5.1)$$

Two solutions based on this principle are usually implemented: continuous and "one-shot" boost. In the continuous boost solution, a timed circuit supplied with a clock signal provides for gradually charging a boost capacitor to a voltage higher than V_{CC}; the boost capacitor then provides for maintaining a common (boost) line of the memory at the desired overvoltage. The advantage of this solution lies in the small size of the boost capacitor, due to the overvoltage being reached by means of a series of small increments. For this reason, however, initial charging, and hence access to the memory when turned on or on re-entry from stand-by, is very slow. To eliminate the latter delay, a second smaller boost circuit may be used to keep the boost capacitor charged in stand-by mode, but only at the expense of increased consumption. The "one-shot" boost solution, on the other hand, employs a huge boost capacitor, which is charged by a single pulse only at predetermined times (upon address switching in Read mode, or active switching of the E# signal). While solving the problems of slow access on re-entry from stand-by (or when the memory is turned on) and increased consumption in stand-by mode, the "one-shot" solution presents other drawbacks of its own, due to the large area required for the capacitor and the necessary drive circuits.

5.2.7 A New Boost Approach: Miniboost

There is another solution, named "miniboost", to provide a read voltage higher than V_{CC}. This solution maintains the advantages and simultaneously minimizes the disadvantages of known pulsating and "one-shot" boost solutions. A possible scheme is shown in Fig. 5.23.

BN is the supply node of the row decoder; only one wordline with its driving inverter is depicted. In the absence of the clamping diode D1, the charge sharing effects would drive the voltage on node BN to a value given by (5.1). Between node BN and GND, there is also a parasitic capacitor, which represents the capacitances of the well regions of pMOS transistors, the junction capacitances, and other parasitic capacitances connected to node BN. M1 and M2 represent the final inverter of the row decoder. When transferred to the row to be read via M1, the above voltage would drive the addressed line to the high read voltage ensuring correct reading of the cell, but, on the other hand, it would cause biasing problems on the non-addressed rows: the pMOS transistor of their final inverters would have the gate terminal biased at supply voltage V_{CC} and the

APPLE INC.
EXHIBIT 1008 - PAGE 0277

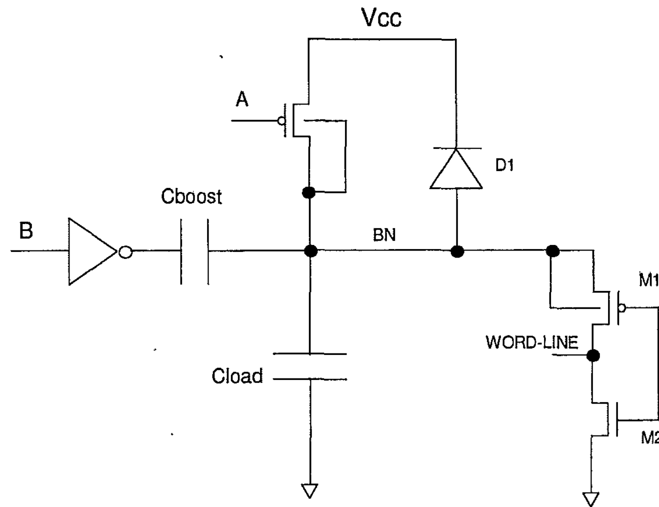


Figure 5.23 Local boosting or "miniboost".

source terminal at the higher voltage V_{BN}. As such, pMOS transistors of non-addressed rows would not definitely be turned off, thus resulting in undesired discharge of node BN, which frustrates the voltage boost function. The purpose of clamping diode, on the other hand, is precisely that of only permitting node BN to be boosted to voltage $(V_{\mathrm{CC}} + V_d)$ where V_d is the voltage drop across diode. This diode is realized by a pMOS transistor similar to transistor M1 but diode-connected, so that voltage drop V_d equals the threshold voltage V_{TH_p} of M1. As the voltage at node BN tends to exceed $(V_{\mathrm{CC}} + V_{\mathrm{TH}_p})$, it is clamped by the conducting diode, thus discharging boost capacitor in controlled manner towards GND. The voltage drop between the source and gate terminals of transistors M1 driving the non-addressed rows therefore equals the threshold voltage of diode D1, which, even if close to being turned on, remains off.

This solution, therefore, allows the read voltage of the addressed cells to be boosted at a value which, even if only slightly higher than the supply voltage, is normally sufficient to ensure correct reading of the cells, and may be increased if necessary by providing native transistors. No matter how boosting is realized, it must always be clamped, in order (i) to limit the stress on the gate of all the cells on the selected wordline and (ii) not to reduce the $V_{\mathrm{cc}_{\mathrm{max}}}$ (see Section 5.4).

At the end of a Read operation, it is necessary to restore the initial conditions (first of all to clamp BN to V_{CC}) by means of a proper signal, so that the procedure described above can be repeated for every following Read. The described boost circuit solves the slow access problems typically associated with continuous boost circuits by boosting the read voltage by means of a

APPLE INC.
EXHIBIT 1008 - PAGE 0278

single pulse. Moreover, only a small area is required, by employing a number of local boost capacitors (for each sector or half-sector), which, being small, may be integrated in unused areas of the chip or, at any rate, in such areas as to optimize the layout of the memory.

5.3 READ PATH: INPUT AND OUTPUT BUFFERS

This section describes the characteristics of the input and of the output buffers, whose task is to interface a device to the external world. In particular, for the input buffer, the specifications due to the connection to a TTL system and the noise margin are analyzed. For the output buffer the problems of power consumption, the impact on power supplies during its operation and the consequences due to a reduction of the supply voltage are discussed. Finally, the specification "High voltage tolerance" is introduced.

5.3.1 Input Buffer

CMOS technology is widely used for integrated circuits to exploit its superior features such as low power consumption, high level of integration and large noise margin. CMOS systems, however, must be interfaced with TTL or ECL systems and their interface stages must be able to translate signals into different standards; in particular input buffer stages of pure CMOS chips must be able to convert TTL voltage levels into CMOS voltage levels. The input buffer logic threshold voltage must lie between the TTL logic Low ($V_{\text{il}} = 0.8\text{V}$) and TTL logic High ($V_{\text{ih}} = 2\text{V}$) values in the whole supply voltage range. The noise margins are defined as follows (see Fig. 5.24) for a TTL to CMOS buffer:

$$NM_{\text{high}} = V_{\text{oh}} - V_{\text{ih}} = 2.4 - 2 = 0.4\text{V}\ , \tag{5.2}$$

$$NM_{\text{low}} = V_{\text{il}} - V_{\text{ol}} = 0.8 - 0.4 = 0.4\text{V}\ . \tag{5.3}$$

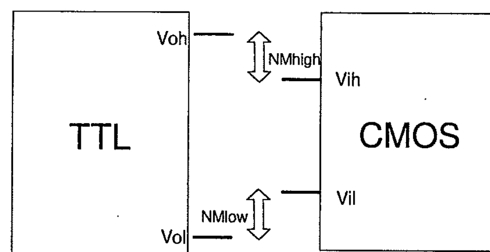


Figure 5.24 Definition of the noise margin.

A well-designed input buffer should:

APPLE INC.
EXHIBIT 1008 - PAGE 0279

- enter a high-impedance configuration to avoid consumption in stand-by;
- detect the right TTL logic levels in the supply voltage range from $V_{CC_{min}}$ to $V_{CC_{max}}$;
- ensure a noise margin as large as possible.

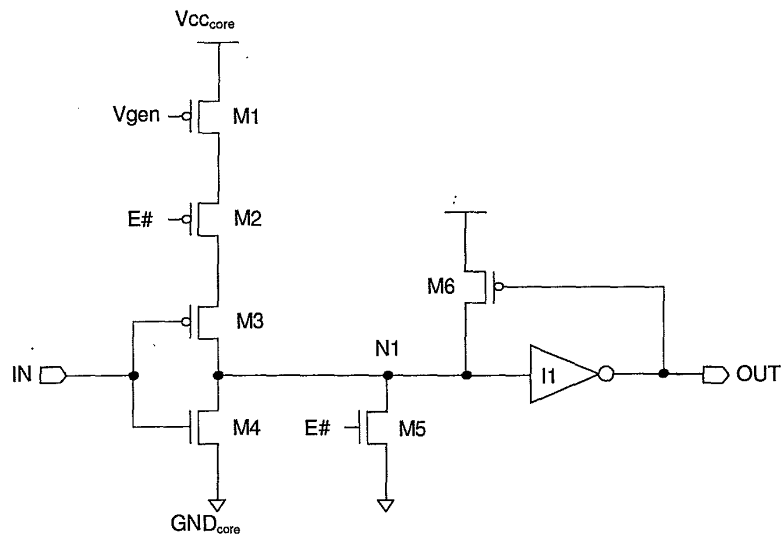


Figure 5.25 Example of input buffer.

Fig. 5.25 shows an example of input buffer. When signal E# is High, the input buffer is in three-state configuration and its output is High. The pull-up (M3) and the pull-down (M4) transistors of the input stage ensure that its logic threshold voltage lies in the range $V_{il} \div V_{ih}$ for every value of the supply voltage range. At lower V_{CC}, it is possible to increase the noise margin by making the pull-down more conductive than the pull-up; in this case the threshold voltage is closer to V_{il}. At higher V_{CC}, the threshold voltage tends to increase: the current generator (M1) limits the current through the pull-up, thus reducing this undesired effect.

To improve further the noise margin, pMOS transistor M6 is introduced as shown in Fig. 5.25. M6 produces a hysteresis of the threshold voltage of the inverter I1 (Fig. 5.26a). In Fig. 5.26b the output of the input buffer circuit with (V_{OUT}^{**}) and without (V_{OUT}^{*}) the feedback transistor M6 is shown as a function of the time variations of the voltage of node N1.

It's worth noting that during output buffer commutation, voltage variations are produced on the internal GND and V_{CC}, thus reducing the noise margin

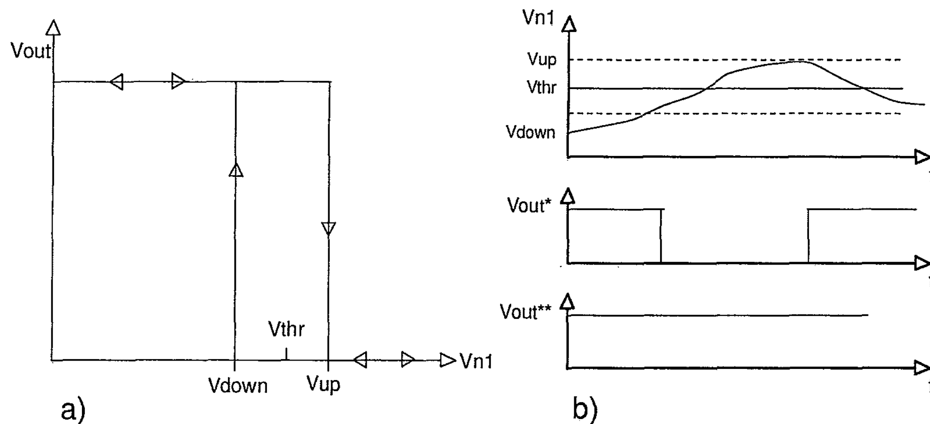


Figure 5.26 Hysteresis of the threshold voltage of the inverter I1 (a). Comparison of the output OUT without (OUT*) and with (OUT**) the feedback transistor M6 (b).

of the input buffer. For sake of simplicity, we will investigate the effect of a GND fluctuation only: if $V_{\text{in}} = V_{\text{ih}}$ a positive variation on the GND terminal reduces the overdrive on the pull-down of the input buffer ($V_{\text{gs}}(\text{nMOS}) = V_{\text{in}} - V_{\text{GND}_{\text{core}}}$); in particular, if the spike is too high in amplitude, the pull-down might even be unable to keep the output Low, and the output could be seen as a High level. In the same way if $V_{\text{in}} = V_{\text{il}}$, a negative variation on GND increases the voltage $V_{\text{gs}}(\text{nMOS})$ and the pull-up might even be unable to keep the output High and the output could be seen as Low.

5.3.2 Output Buffer

When a Read operation is performed on a Flash memory, the data read are available at the output buffers. These stages are characterized by a high current capability, so that they are able to drive interface-buses with load capacitances in the order of 100pF.

A standard output buffer is shown in Fig. 5.27, where the data to be output, which is provided by a sense amplifier stage, is stored in a the LATCH block on the rising edge of the ENABLE signal.

The pull-up and the pull-down transistors must be driven with different signals for two reasons:

1. to ensure that the final stage can be put in three-state; in this case V_{CC} and GND are forced on the gate of the pull-up and pull-down, respectively, thus turning off both transistors;

APPLE INC.
EXHIBIT 1008 - PAGE 0281

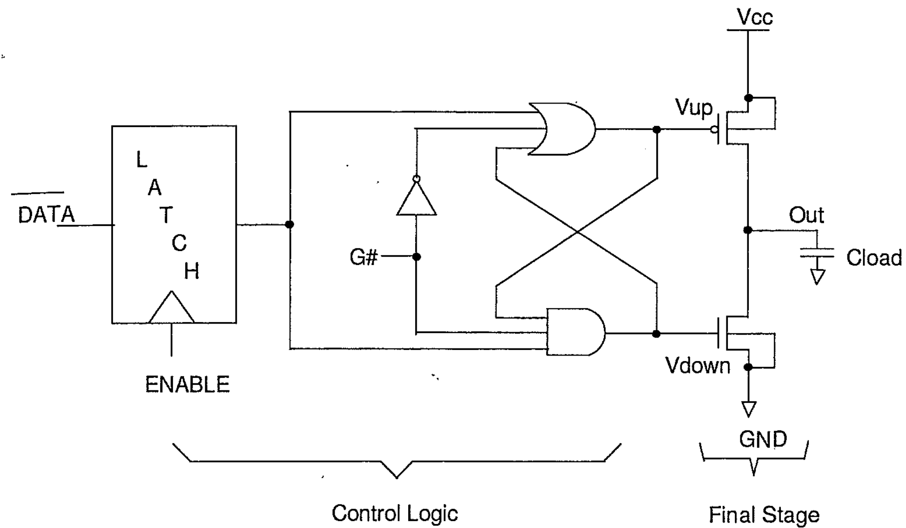


Figure 5.27 Example of output buffer.

2. to avoid the crow-bar current, which is a source of consumption, during a transition; in this case the control logic switches off the driving transistor before switching on the other one.

The output buffer is one of the key elements of the read path and therefore its performance, in particular in terms of switching time, strongly influences the memory access time that can be divided into three factors:

- decoding time (about 30% of the access time);
- reading time (about 40% of the access time);
- switching time of the output buffer (about 30% of the access time).

A major problem arises with low supply voltages operations, since it is possible to determine a relationship of inverse dependence between the supply voltage V_{CC} and the switching time of the output buffer. Such a relationship can be determined, in a simple way, by assuming that the pull-up and pull-down work in saturation. Therefore the driving transistor can be considered as an ideal current generator, so that the switching time t_s is merely the time needed to either charge or discharge a capacitor C by means of a constant current source:

$$t_s = C\,\frac{V_{CC}}{2}\,\frac{1}{K(W/L)(V_{gs}-V_{TH})^2}\,, \tag{5.4}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0282

where K is the intrinsic conductivity factor, W/L the transistor aspect ratio, and V_{gs} the gate to source voltage. In the case of capacitance discharge, by assuming for the pull-down transistor $V_{gs} = V_{CC}$, the inverse dependence between t_s and V_{CC} is obvious.

To keep the same t_s while reducing V_{CC}, the transistor aspect ratio, and therefore the area of the output buffer, must be dramatically increased. An alternative solution consists in driving the pull-up and the pull-down transistors with gate voltages lower than GND and higher than V_{CC}, respectively, by means of boost operations (the concept of boost was explained in Section 5.2.5 and 5.2.6).

Fig. 5.28 shows the basic scheme of a circuit driving the final stage with a dynamic higher than V_{CC}. During initialization (Sn1 = Sn2 = Sn3 = 1; Sp1 = Sp2 = Sp3 = 0), V_{N2} is driven to V_{CC} by Mn1 (the other terminal of C_{boostN} being at GND), while V_{P2} is driven to GND by Mp1 (the other terminal of C_{boostP} being at V_{CC}).

If the data to be transferred to the external load is Low (logic "0"), Sn1 is turned Low: V_{N1} is then switched to V_{CC} and nodes N2 and N3 (gate of the pull-down) reach a voltage higher than V_{CC}. The overvoltage V_{boostN} (i.e. the voltage on node N2) can be calculated by imposing the charge conservation

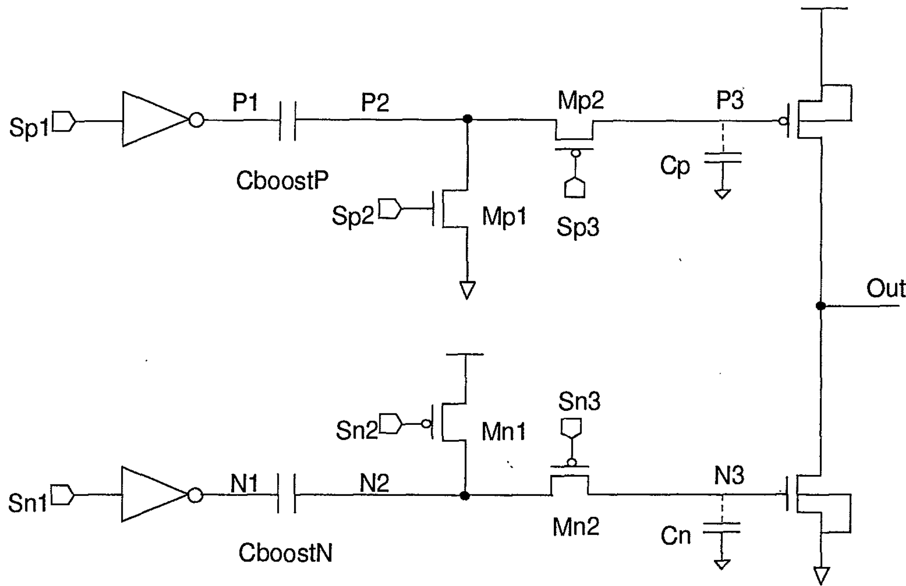


Figure 5.28 Basic scheme of a circuit driving the final stage with a dynamic higher than V_{CC}.

APPLE INC.
EXHIBIT 1008 - PAGE 0283

$Q_i = Q_f$, where Q_i and Q_f are the charge before and after the phase of boosting respectively:

$$Q_i = C_{\text{boostN}} V_{\text{CC}} \quad (5.5)$$

$$Q_f = C_{\text{boostN}} (V_{\text{boostN}} - V_{\text{CC}}) + C_n V_{\text{boostN}} \quad (5.6)$$

thus resulting in

$$V_{\text{boostN}} = 2V_{\text{CC}} \frac{1}{1 + C_n/C_{\text{boostN}}} \ . \quad (5.7)$$

Eq. (5.7) shows that the overvoltage V_{boostN} is inversely proportional to the ratio C_n/C_{boostN} and when this ratio tends to zero V_{boostN} gets to its maximum value $2V_{\text{CC}}$.

If the data to be transferred to the external load is High (logic "1"), Sp1 is turned High: V_{P1} is then switched to GND and then nodes P2 and P3 (gate of the pull-up) reach a value lower than GND. It is clear that to avoid the direct bias of the junction drain-pwell all the nMOS connected with these nodes must be in triple well technology.

By imposing the charge conservation, the following relationship for V_{boostP} (i.e. the voltage on node P2) is found:

$$V_{\text{boostP}} = V_{\text{CC}} \left(1 - \frac{2}{1 + C_p/C_{\text{boostP}}}\right) \ . \quad (5.8)$$

For the pull-up transistor, the maximum value of the overvoltage V_{boostP} is, in module, equal to V_{CC}. Fig. 5.29 shows the qualitative behavioral of the overvoltages V_{boostN} and V_{boostP} as a function of the ratio C_{boostX}/C_x. The value of the boost capacitor should be a trade-off between switching time performance of the final stage and the area occupied by the final stage itself.

5.3.3 Noise Issues

It is known that integrated electronic circuits are assembled into a package which consists of a thermosetting resin case and of embedded lead frame in order to connect the chip to the external pins. The terminals (or "pads") of the electronic circuits are connected to the package lead frame by small interconnections wires which are referred to as "bonding wires". As an example, Fig. 5.30 shows a schematic view of a semiconductor chip having a four-terminals connection to external supply pins ($V_{\text{CC}_{\text{ext}}}$ and GND_{ext}); more precisely, $V_{\text{CC}_{\text{IO}}}$ and GND_{IO} supply only the terminal stage of the output buffers, while $V_{\text{CC}_{\text{core}}}$ and GND_{core} supply the rest of the device.

The bonding wires have an inherent inductance which is denoted by L. The external capacitive load is normally in the order of about 100pf, so a large

APPLE INC.
EXHIBIT 1008 - PAGE 0284

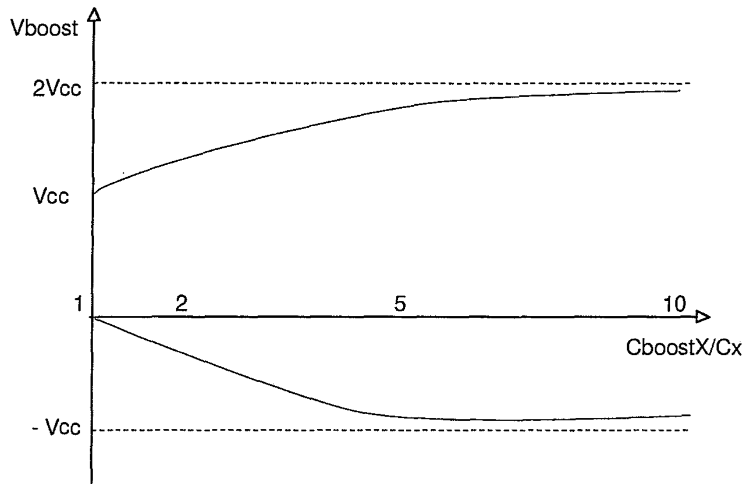


Figure 5.29 Behavior of the overvoltage V_{boostX} as a function of the ratio C_{boostX}/C_x.

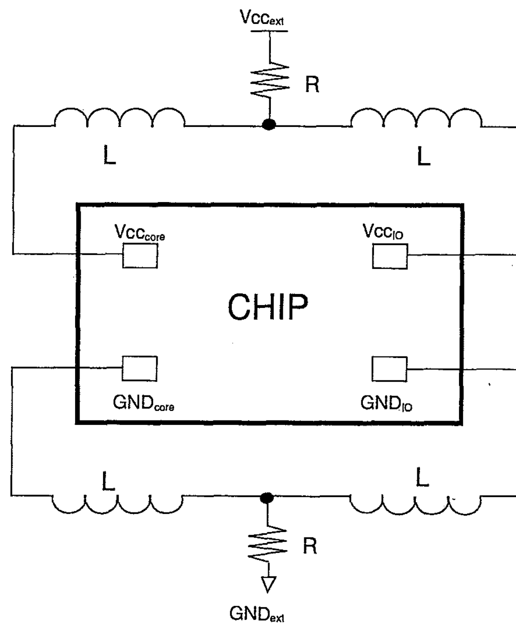


Figure 5.30 Connection between the terminals of GND and V_{CC} of a chip and the external supply pins.

current must be provided by the output stage to charge or discharge the load

APPLE INC.
EXHIBIT 1008 - PAGE 0285

in a time in the order of tens of nanoseconds. During the switching phase, either the pull-up or the pull-down transistor is turned on, so that a current flows through the active transistor to the external power supply (either GND_{IO} for the pull-down or $V_{CC_{IO}}$ for the pull-up) and a voltage drop is produced across the related bonding wire. In short, the internal supply voltage $V_{CC_{IO}}$ decreases and the GND_{IO} increases during normal operation of the circuit. The voltage drop which causes these variations is tied in with the inductance value of the bonding wires by the following relationship:

$$V = L\,\frac{di}{dL}\,. \tag{5.9 }$$

If all the N output buffers of the chip switch simultaneously towards the same value, Eq. (5.9) becomes:

$$V = NL\,\frac{di}{dt}\,. \tag{5.10 }$$

These voltage variations reduce the dynamics of the pull-up and the pull-down and produce an increase of the switching time.

In reality, with reference to Fig. 5.31, the terminals GND_{IO} and GND_{core} are connected to each other within the circuit by:

- a resistance $R_{substrate}$ of the semiconductor substrate p;
- a pair of resistances R_{ring} due to the bias rings of both the pull-down transistor of the input and the output buffers.

The voltage variations on GND_{IO}, through the series of these three resistances, is transferred to the terminal GND_{core}.

Such a situation, as depicted schematically in Fig. 5.32, may cause a false reading in a memory circuit: in fact, if an input pin (due to this voltage variation) changes its state, then a new undesired reading starts.

To avoid the variations on the GND_{core}, the pull-down can be used in a triple-well structure. As shown in Fig. 5.33, by connecting the bulkP to GND_{IO} and the nwell to $V_{CC_{IO}}$, the junctions ipwell-nwell and nwell-psubstrate are inverse-biased. In this way it is possible to prevent the propagation of the discharge current from the external load through the semiconductor substrate.

5.3.4 High Voltage Tolerance

The specification "High voltage tolerance" is required for those devices which are powered with a low supply voltage and share the DATA BUS with others devices whose supply voltage is higher. Fig. 5.34 shows the connections between

APPLE INC.
EXHIBIT 1008 - PAGE 0286

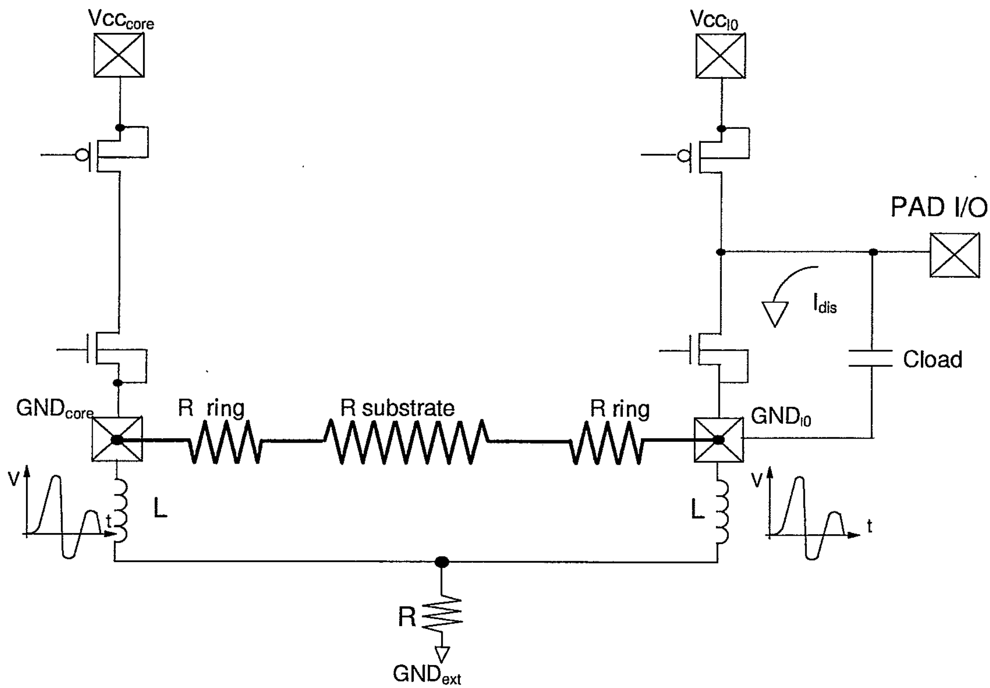


Figure 5.31 Intrinsic connections between GND_{IO} and GND_{core}.

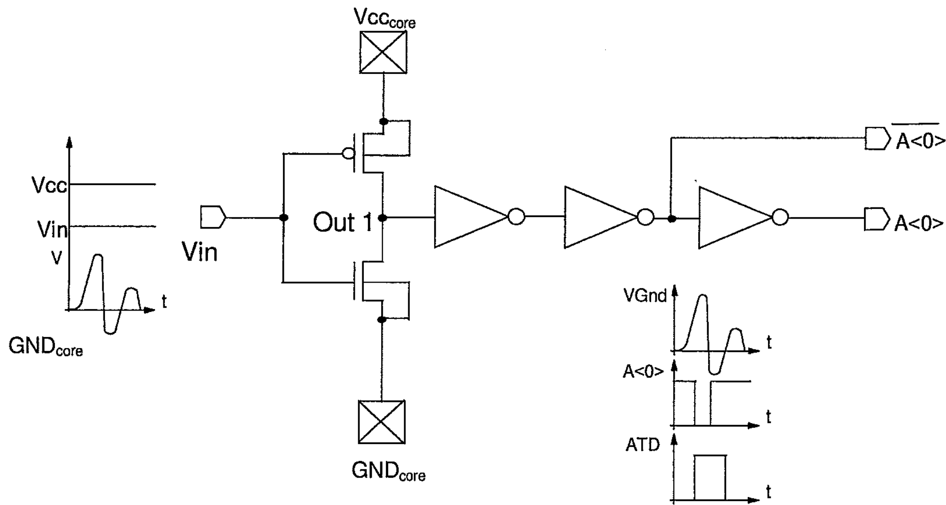


Figure 5.32 A voltage variation on GND_{core} may cause a false reading in a memory circuit.

APPLE INC.
EXHIBIT 1008 - PAGE 0287

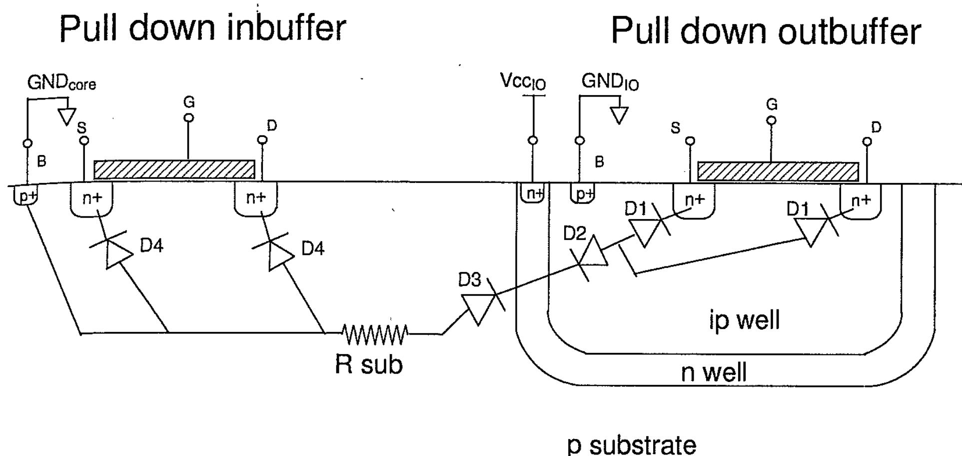


Figure 5.33 Pull-down in triple-well.

a microprocessor and devices supplied with different voltages. The microprocessor chooses the device to communicate with by forcing the appropriate E# and G# terminals, and puts the other ones in a three-state configuration.

When the device which is supplied with high voltage ($V_{CC_{high}}$) is activated, the maximum voltage which can be transferred on DATA BUS is given by:

$$V_{CC_{max}} = V_{CC_{low}} + \min\left[|V_{TH_p}|, V_\gamma\right] \tag{5.11}$$

where:
$V_{CC_{low}}$ is the lowest voltage of the devices on the board;
V_{TH_p} is the threshold voltage of the pMOS;
V_γ is the threshold voltage of the junction drain-nwell (p-n junction).

If the threshold voltage of the pMOS V_{TH_p} is lower than the threshold voltage of the junction drain-nwell V_γ, the device at $V_{CC_{high}}$ shows net power consumption; on the other hand, when $V_\gamma < |V_{TH_p}|$, there are latch-up problems for the p-n junction. It is possible to overcome these problems by using, for example, the solution shown in Fig. 5.35a.

When the device is in three-state configuration, the bulk and the gate of the pMOS are connected to the exit of the circuit SWITCH, whose transfer function and schematic are shown in Fig. 5.35b and 5.35c, respectively. It is clear that, when $V_{in} < V_{CC} - |V_{TH_p}|$, the transistor M1 is turned on and the output is V_{CC}; conversely when $V_{CC} - |V_{TH_p}| < V_{in} < V_{CC}$, the transistor M4 is turned on and the output is V_{CC}. Finally, when $V_{in} > V_{CC}$, the transistor M2 is switched on and the output follows the input, while the transistor M3 switches

APPLE INC.
EXHIBIT 1008 - PAGE 0288

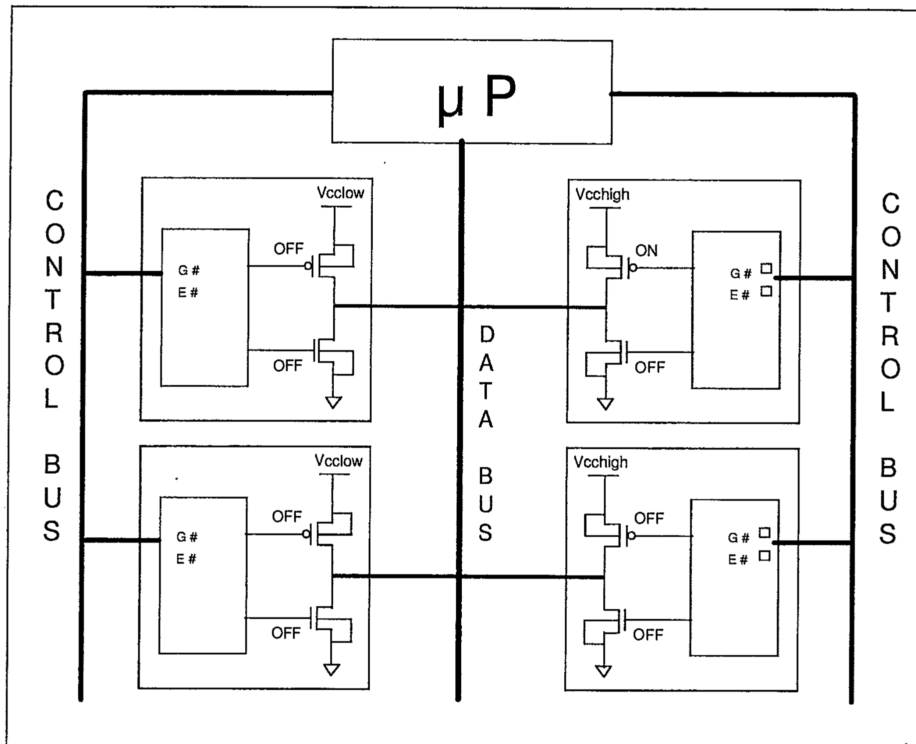


Figure 5.34 Connections between a microprocessor and devices supplied with different voltages.

off the transistor M4. During the fall time of the input transistor M6 discharges the gate of M4.

5.4 READ PATH: SENSING TECHNIQUES

This section deals with the principal architectures employed in a non-volatile memory, EPROM and Flash, to read the information stored in the cells. The survey covers the issue from the first solution up to the latest implementations for the low supply voltage design.

The block which accomplishes the reading of the cell content is called sense amplifier, and is normally divided into different sections, the first being the current-voltage converter, and the second the comparator. The attention will be focused on the converter, because the comparator is a classical circuit with enhanced commutation speed.

APPLE INC.
EXHIBIT 1008 - PAGE 0289

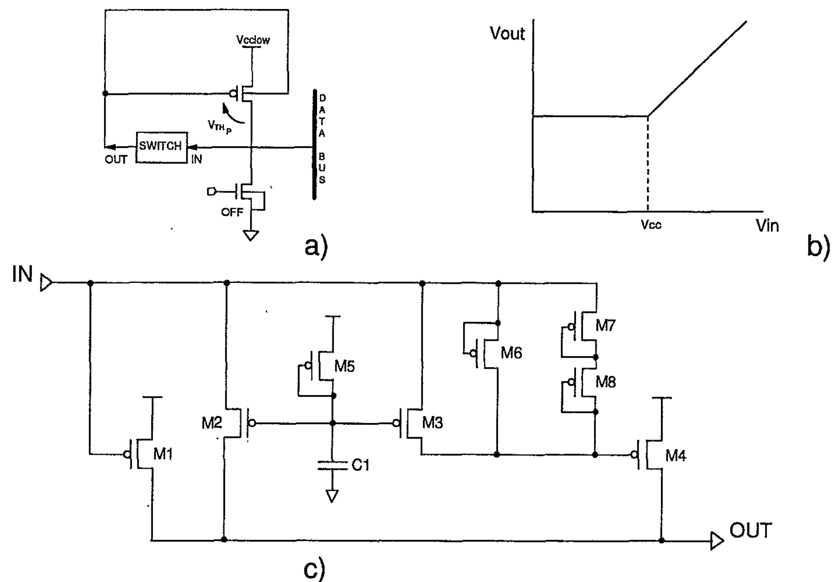


Figure 5.35 Improved high voltage tolerance using a SWITCH circuit: a) connections with the pull-up; b) transfer function; c) schematic.

The main feature of this type of converter is commutation speed as well: it ranges between 10ns and 30ns, depending on the current sunk by the cell, and it shows huge robustness to any variation in temperature, supply voltage and threshold voltage variations due to process drift.

5.4.1 Sensing Techniques: An Overview

Sensing the information stored as charge inside the floating gate of the cell is perhaps the most critical operation executed in a non-volatile memory device. During this operation, the cell source is tied to GND and the gate is driven by the row decoder to a voltage which is usually V_{CC} (or more, in case of boost: see Section 5.2.5), while the drain is connected to the supply voltage through a load (see Fig. 5.36a); C_{bitline} is the parasitic capacitance of the metal line where the drains of all the cells on the same column are connected. In Fig. 5.36b, the column decoder transistors M2 and M3 are also shown, together with M1 in cascode configuration; in the analysis, M2 and M3 are normally neglected, because their aspect ratios are large and therefore the voltage drops on their channels are not significant.

During Read operation, cell's drain voltage must be low enough to avoid stress phenomena, but high enough to allow a proper current to flow and guar-

antee a quick reading. Thanks to the presence of M1, it is possible to set this value to about 1V, which causes the cell to operate in "linear region":

$$I_{\mathrm{ds}} = k\left[\Big(V_{\mathrm{gs}} - (V_{\mathrm{TH_{UV}}} + \Delta V_{\mathrm{TH}})\Big)V_{\mathrm{ds}} - \frac{V_{\mathrm{ds}}^2}{2}\right] \tag{5.12}$$

where $V_{\mathrm{TH_{UV}}}$ is the threshold of an UV-erased non-volatile cell and ΔV_{TH} is the shift of the threshold of the cell under consideration from $V_{\mathrm{TH_{UV}}}$.

If $V_{\mathrm{ds}} = 1\mathrm{V}$ and $V_{\mathrm{gs}} = V_{\mathrm{CC}}$:

$$I_{\mathrm{ds}} = k\Big[(V_{\mathrm{CC}} - (V_{\mathrm{TH_{UV}}} + \Delta V_{\mathrm{TH}})) - 0.5\Big] \tag{5.13}$$

From Eq. (5.13) it is clear that the characteristics of different cells are parallel and the shift from the one of an UV-erased cell is determined only by the threshold shift ΔV_{TH}.

At this point, the behavior of the biasing configuration of Fig. 5.36b can be analyzed: let V_{bias} be equal to 2V, and suppose that M2 and M3 are on: then there is a peak current that charges the parasitic bitline capacitance C_{bitline}: as soon as 1V is reached, M1 is turned off (it is a negative feedback). The gate of the cell is biased at V_{CC}: if the cell is erased, e.g. with a threshold of 2V, then it starts to sink current, discharging C_{bitline} and lowering the potential of

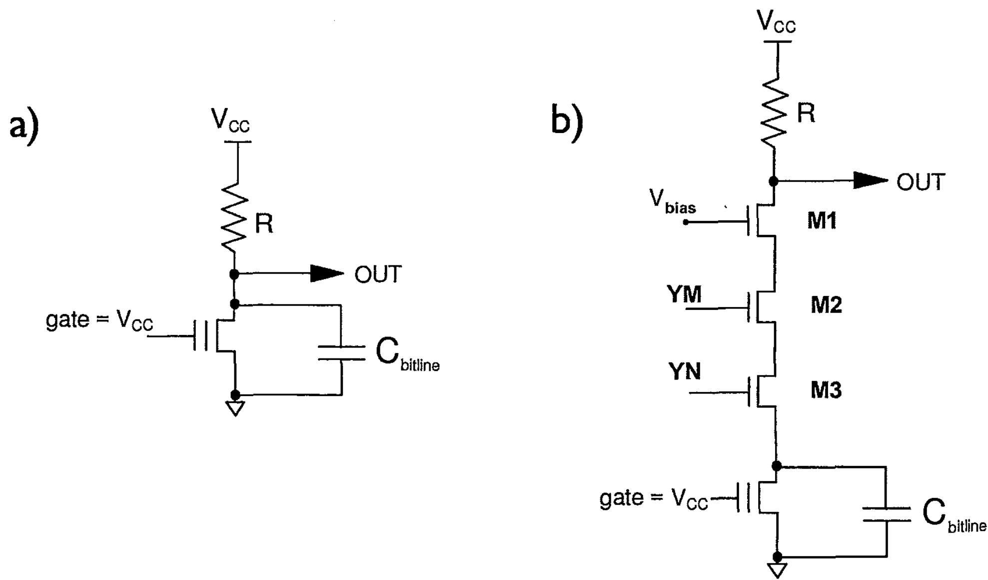


Figure 5.36 a) Simplified Read mode biasing; b) Biasing configuration for Read mode.

APPLE INC.
EXHIBIT 1008 - PAGE 0291

the node OUT; if the cell is programmed (threshold higher than V_{CC}), $C_{bitline}$ is not discharged and the node OUT remains at V_{CC}.

Replacing the resistance R with an ideal current generator, Read operation can be regarded as a comparison between the current of the generator I_{ref} and the one sunk by the cell (Fig. 5.37).

Again, two opposite conditions should be considered:

- the cell is virgin: from the plot it is clear that if $V_{CC} < V_{CC_m}$, the cell is not able to sink all the current supplied by the load and node OUT is pulled up at V_{CC}: the virgin cell is read as a programmed one; if $V_{CC} > V_{CC_m}$, cell current is always greater then load current and the node OUT is pulled down to GND: the virgin cell is read correctly.

- the cell is programmed: if $V_{CC} < V_{CC_M}$, cell current is lower than load current and the node OUT remains high (correct reading); if $V_{CC} > V_{CC_M}$, then also a programmed cell sinks a current which is greater than I_{ref} and node OUT is pulled down (incorrect reading).

Therefore the V_{CC} range for a correct sensing of the cell is $V_{CC_m} < V_{CC} < V_{CC_M}$, which is equal to the threshold voltage shift ΔV_{TH} because of the parallelism of the characteristics.

The right choice of the load current should also satisfy dynamic constraints; it should be large enough to charge $C_{bitline}$, quickly, but not so high as to prevent a virgin cell from pulling down OUT node.

Also the way of generating V_{bias} has an effect on the dynamic behavior of the circuit: if the gate of M1 is tied at a fixed value, when M2 and M3 are turned on, the V_{gs} of M1 is equal to V_{bias}, since node A is virtually at GND

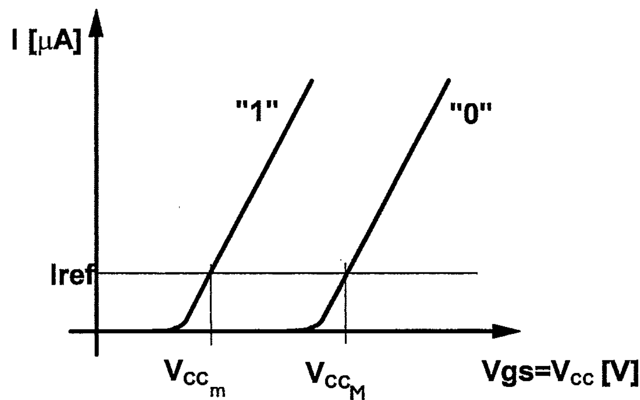


Figure 5.37 Read characteristic for a single end converter.

APPLE INC.
EXHIBIT 1008 - PAGE 0292

(see Fig. 5.38a); therefore the charge of C_{bitline} is performed through M1 with a limited V_{gs}. To overcome this problem, the V_{bias} network is replaced by an inverter as shown in Fig. 5.38b; when node A′ is at GND, V_{bias} is at V_{CC}; therefore the V_{gs} of M1 is the maximum available, and the charge of C_{bitline} is as fast as possible. By designing the inverter with the nMOS aspect ratio greater than the one of the pMOS, $(W/L)_{\text{nMOS}} \gg (W/L)_{\text{pMOS}}$, M1 can be turned off when node A′ reaches V_{TH_n}, about 1V. The main drawback of this solution is power consumption: feedback needs current to work properly.

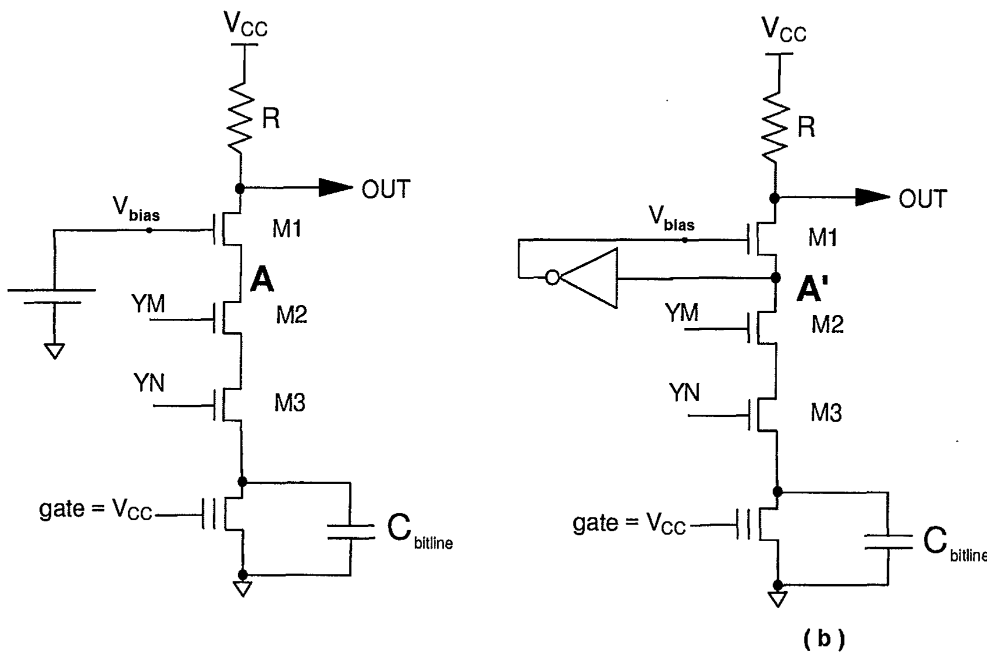


Figure 5.38 Typical cascode biasing schematics.

Fig. 5.39 shows the connection of node OUT to the simplest buffer (i.e. an inverter) to produce CMOS level at the OUT′; the buffer is necessary to strengthen the signal, thus allowing its propagation over long distance.

At this point, it is worth noting that C_{bitline} is very large (some picofarads), while the contribution of C_{OUT} is much smaller (tens of femtofarads); therefore a small variation of the voltage of node A means a little modification of the charge within C_{bitline}, but it produces a great change in the OUT voltage (see Fig. 5.40).

The main drawback of the use of the "inverter" approach is that the threshold voltage shift that the cell should perform when programmed must be large to be significant, while process trends are moving towards reduction of both

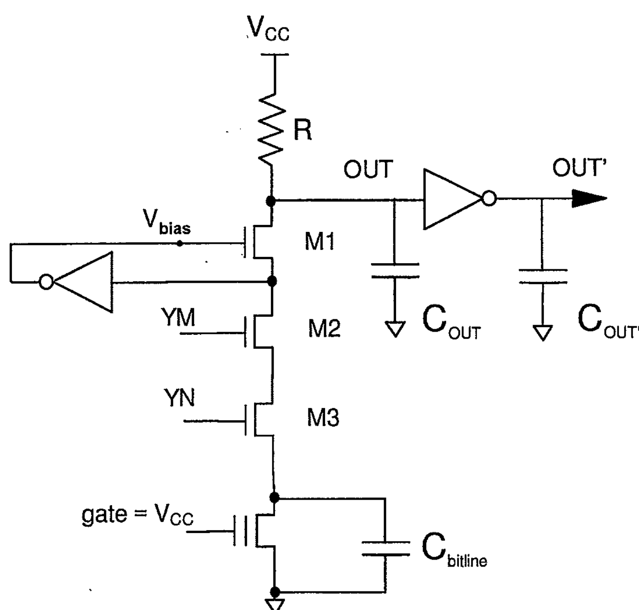


Figure 5.39 Cascode stage.

programming time and voltages to decrease stress on cell oxides. To guarantee a reliable read operation, differential sensing structures should be implemented.

5.4.2 Differential Sensing Technique

Fig. 5.41 shows the basic scheme of a differential sensing architecture: where MM is the matrix cell to be read, and MR is a reference cell whose threshold voltage is known.

If an EPROM device is considered, MR is a virgin cell (i.e. UV-erased) whose threshold, $V_{\mathrm{TH_{UV}}}$, is equal to 2V, while MM is a matrix cell whose threshold is a function of the charge within the floating gate; both cells are biased by the same signal and have the same parasitic elements.

Typical cell threshold distribution for an EPROM device is depicted in Fig. 5.42; it is very narrow for the erased cells and much larger for the programmed ones; it is worth noting that, while $V_{\mathrm{TH_{UV}}}$ is the center of the distribution, $V_{\mathrm{TH_{PG}}}$ is defined as the threshold of the worst programmed cell.

A typical $V_{\mathrm{TH_{PG}}}$ value is $V_{\mathrm{TH_{UV}}}$ plus 3V to guarantee a good separation between the distributions; in this case the Read operation is more reliable, because there exists a range of V_{gs} where virgin cells sink current while the

APPLE INC.
EXHIBIT 1008 - PAGE 0294

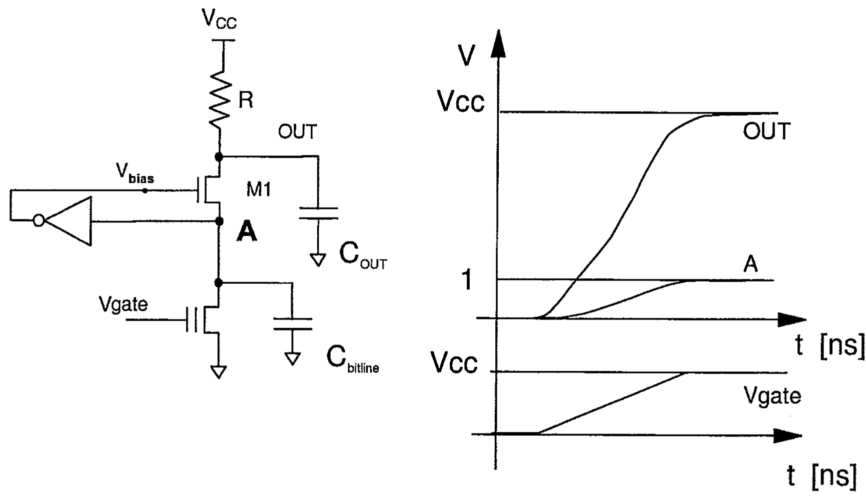


Figure 5.40 Cascode operation.

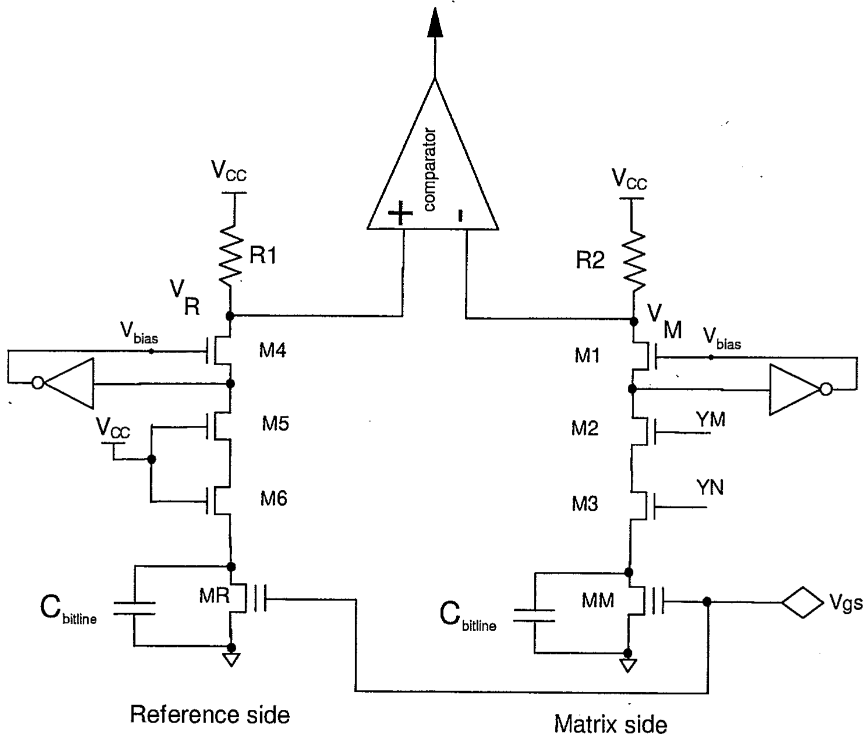


Figure 5.41 Differential architecture.

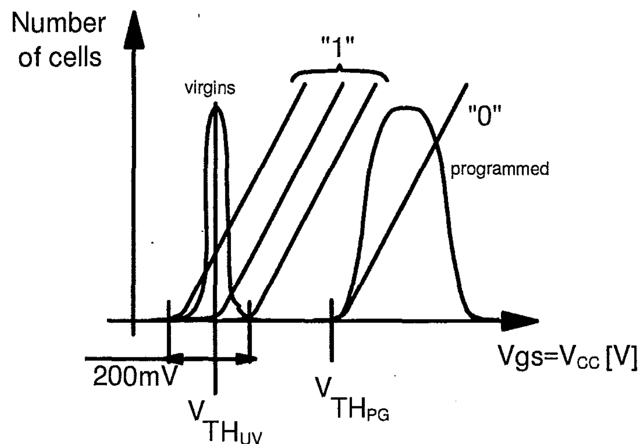


Figure 5.42 EPROM threshold voltage cell distribution.

programmed ones do not; in Fig. 5.42 the characteristics of different cells are shown, under the assumption that there is no distribution gain.

Consider again Fig. 5.41: the two branches of the current-voltage converter are identical, both for electrical properties and layout disposition; M5 and M6 are necessary to balance transistors M2 and M3 of the column decoder. For the time being, R1 and R2 act as loads: later on they will be replaced by active loads.

If MM is programmed, no current is sunk in the matrix side and V_M is at V_{CC}, while MR (being a virgin cell) is on and V_R is pulled down: therefore a voltage difference exists, and the comparator switches its output. The problem with the structure of Fig. 5.41 arises when MM is virgin: of course, V_R and V_M have the same potential and the comparator is not able to decide what type of information it is reading. The parameters that can be changed to obtain a correct behavior in every conditions are the value of the loads, R1 and R2; no matter how these values are chosen, V_R node should always be between V_{M_V} (the potential due to a virgin matrix cell) and V_{M_P} (the potential due to a programmed matrix cell), and it should always have the same value, chosen according to dynamic considerations, independently of the matrix side cell. The relation is therefore the following:

$$V_{M_V} < V_R < V_{M_P} \ . \tag{5.14}$$

Provided that the upper limit is intrinsically satisfied, there are two ways to obtain the lower one (see Fig. 5.43): (i) decreasing R1, thus pulling up node V_R or (ii) increasing R2, thus pulling down the node V_{M_V}. In both cases, if

APPLE INC.
EXHIBIT 1008 - PAGE 0296

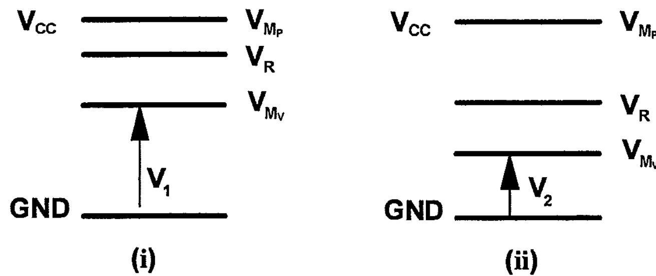


Figure 5.43 Different solutions to separate output nodes of voltage converter.

the two branches sink the same current, node V_R is higher than V_{M_V} because voltage drop on R1 is lower than on R2 (R1<R2).

The optimal solution is the former, since smaller load values mean higher current to charge C_{bitline} quickly.

Fig. 5.44 shows a first schematic with load resistances replaced by pMOS transistors in diode configuration: reference side has two transistors, matrix side has one transistor (R2 = 2R1).

In Fig. 5.45, both reference and matrix characteristic are drawn. The curves are plotted under the assumption that MR and MM have the same dimensions; reference characteristic starts at the same point of the one of the virgin cell, but with half angular coefficient (it is important to remember that this is not the real situation, since it is not possible to have a reference cell half of the matrix cell, but it is an easy way to obtain a legible plot). According to the plot, a cell is virgin if its characteristic lies above the one of the reference, while it is programmed if it is below.

By simple inspection, however, it is clear that this statement is not completely true because of the crossing between reference and programmed cell characteristics; the conclusion is still valid if a constraint on the maximum value of the device supply voltage is put. Supposing that the characteristics are straight lines, the following relation holds:

$$V_{\text{CC}_{\text{max}}} = V_{\text{TH}_{\text{UV}}} + \frac{n}{n-1}\,\Delta V_{\text{TH}} \quad (5.15)$$

$$V_{\text{CC}_{\text{min}}} = V_{\text{TH}_{\text{UV}}} \quad (5.16)$$

with n equal to the ratio between the dimension of the loads in the reference and in the matrix side.

By choosing different values of n, it is possible to modify the current differences between the reference and the matrix sides thus giving more margin either to the virgin or to the programmed cell, as needed.

APPLE INC.
EXHIBIT 1008 - PAGE 0297

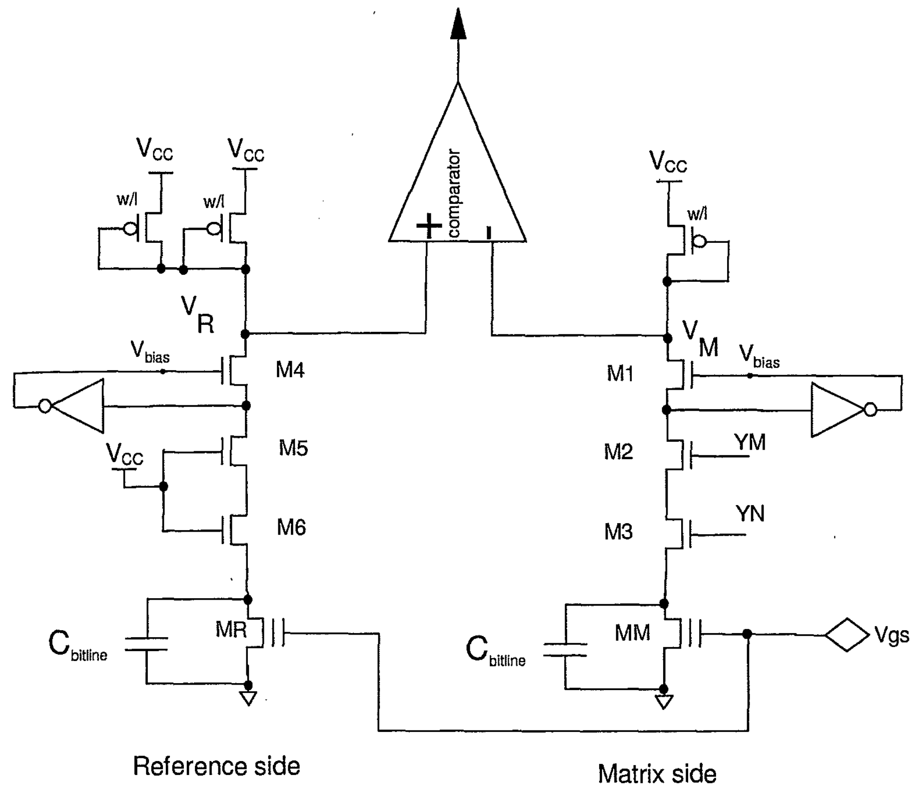


Figure 5.44 Active loads for differential sensing solution.

5.4.3 *Differential Sensing Technique with Offset Current*

The main problem for a current-voltage converter with unbalanced loads is that $V_{CC_{max}}$ fixes a maximum supply voltage for the device operation; in fact, if V_{CC} becomes higher than $V_{CC_{max}}$, the read circuitry misinterprets a programmed cell as virgin. For a typical device, $V_{CC_{max}}$ is equal to 7V and the V_{CC} range is $4.5 \div 5.5$V (and therefore it could seem that $V_{CC_{max}}$ is high enough); nevertheless, a margin is necessary to satisfy all the parameter variations such as temperature, V_{CC}, frequency, process, ..., and to avoid problems due to the charge loss during lifetime which result in a reduction of the cell's working window.

A first solution to overcome the $V_{CC_{max}}$ limit in a non-volatile memory was the introduction of a "parallel reference", i.e. a reference whose characteristic is parallel to that of the matrix cells. This choice led to the situation of Fig. 5.46, that shows the characteristics of a programmed and of a virgin cell, the latter being coincident with the one of the reference cell.

APPLE INC.
EXHIBIT 1008 - PAGE 0298

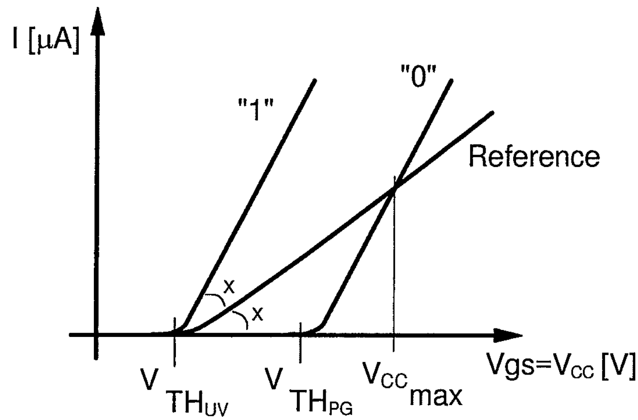


Figure 5.45 Read with unbalanced loads.

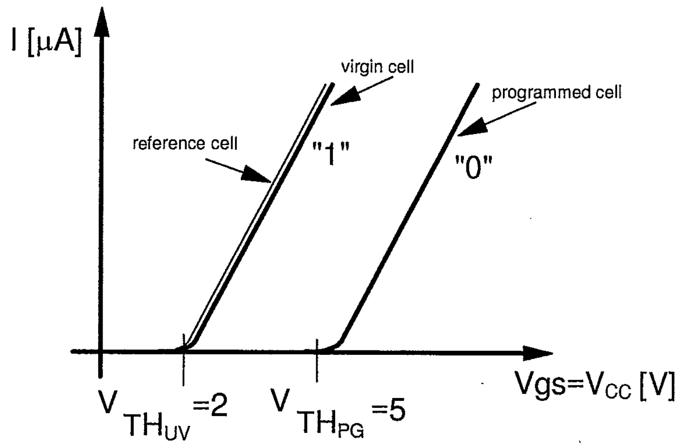


Figure 5.46 Parallel characteristics approach.

By shifting rigidly the two characteristics upwards, the plot of Fig. 5.47 is obtained: the characteristic of the virgin cell is completely on the left side of the reference one, while that of the programmed cell only shares a common part at low V_{CC}, but all the rest is on the right side.

Therefore for $V_{CC} > V_{CC_{min}}$ the reference characteristic is always separated from those of the matrix; it is important to note that this type of solution solves the $V_{CC_{max}}$ problem but raises $V_{CC_{min}}$.

The $V_{CC_{min}}$ value (see Fig. 5.47) is related to I_{off}, which is $20 \div 30\mu A$ typical; the latter choice is a good compromise between the $V_{CC_{min}}$ and the separation

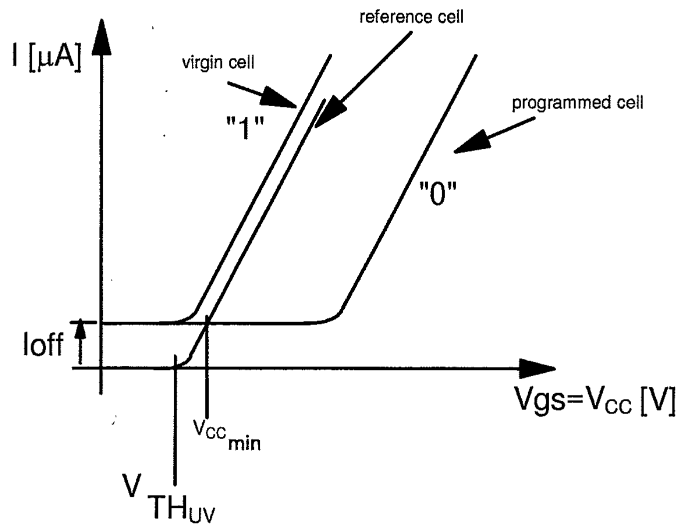


Figure 5.47 Characteristic separation using offset current.

of reference and virgin cell curves. In fact Fig. 5.47 is modified into Fig. 5.48, because the I_{off} current is generated by a circuit supplied by the same V_{CC}, and therefore it could not be present if $V_{\text{CC}} = 0$.

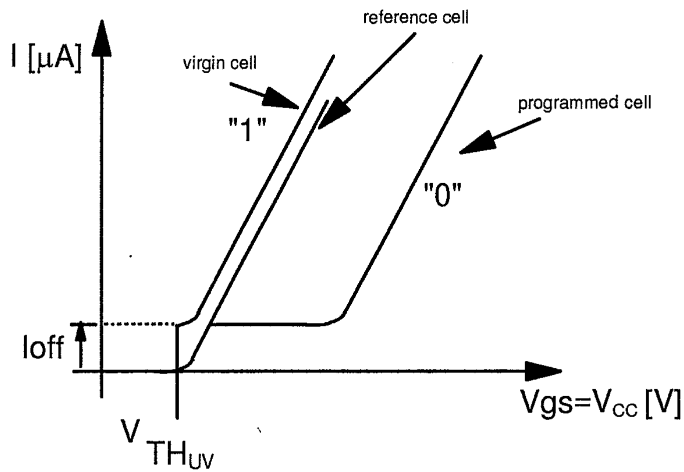


Figure 5.48 Real shape of characteristic versus V_{CC}.

APPLE INC.
EXHIBIT 1008 - PAGE 0300

Fig. 5.49 shows the circuitry used to obtain the characteristic displayed in Fig. 5.48, while Fig. 5.50 shows the circuit to produce the offset current using nMOS device and EPROM cells.

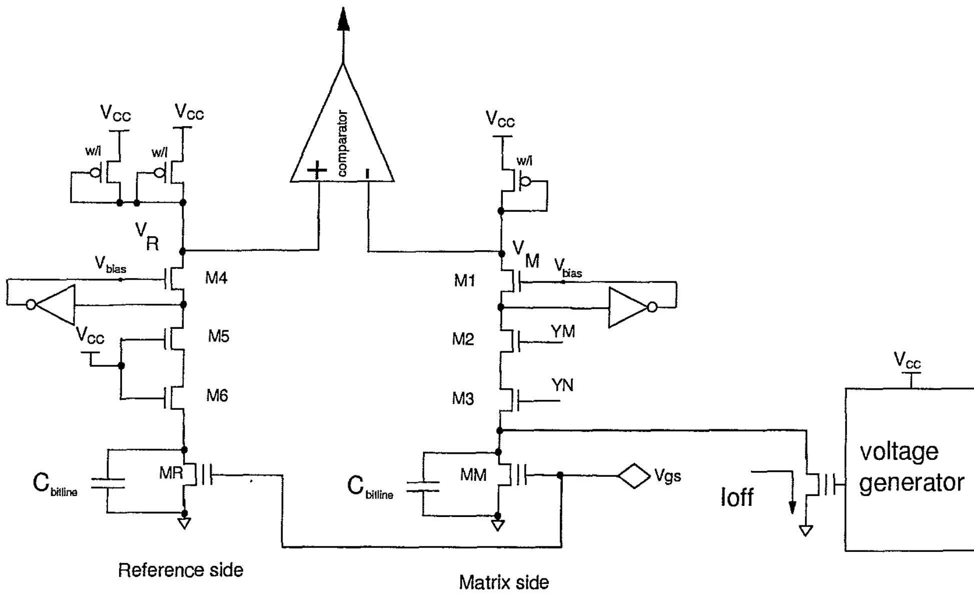


Figure 5.49 Sense amplifier with current offset configuration.

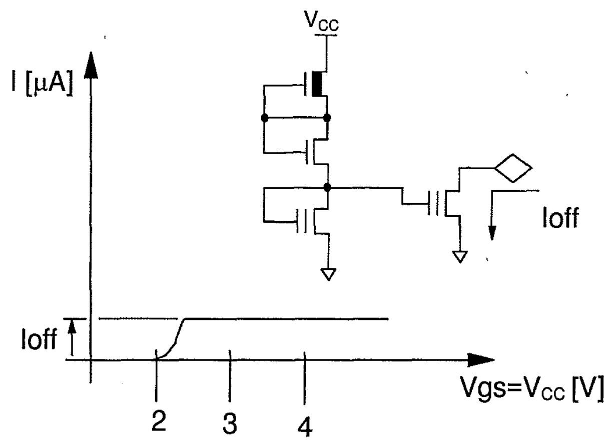


Figure 5.50 Offset current generation for EPROM device.

APPLE INC.
EXHIBIT 1008 - PAGE 0301

For this type of current-voltage converter, the ideal value of $V_{CC_{max}}$ is infinity. For each positive value of ΔV_{TH}, $V_{CC_{max}}$ versus threshold voltage shift is shown in Fig. 5.51 for the two different types of converters analyzed.

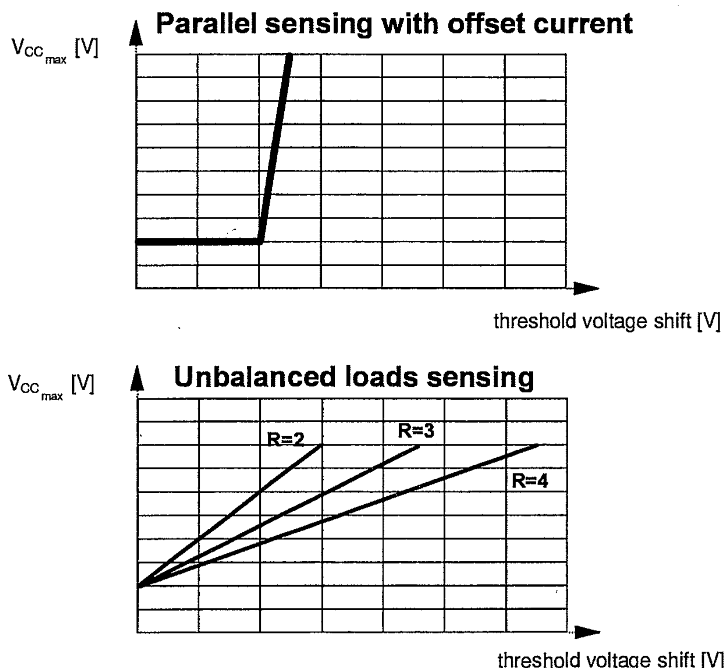


Figure 5.51 $V_{CC_{max}}$ versus threshold voltage shift.

5.4.4 *Differential Semi-Parallel Sensing Technique*

The converter previously described was adopted in many nMOS devices with good results, paying particular attention to the shrink influence on the offset current value. In the case of CMOS non-volatile memories, a new type of converter was designed to produce the "ideal reference characteristic" for an EPROM device. The aims were to shift the value of $V_{CC_{min}}$ down to $V_{TH_{UV}}$; and the plot of Fig. 5.52 was the designer's goal.

Reference characteristic is a composition of those previously examined: unbalanced loads approach is used in the range of V_{CC} between $V_{TH_{UV}}$ and V_S (which is a "safe" value to guarantee a good separation between reference curve and virgin curve), parallel approach is used everywhere else.

First of all, it is necessary to obtain a characteristic like that of I_X shown in Fig. 5.53; then it is possible to generate a new characteristic I_r by subtracting

APPLE INC.
EXHIBIT 1008 - PAGE 0302

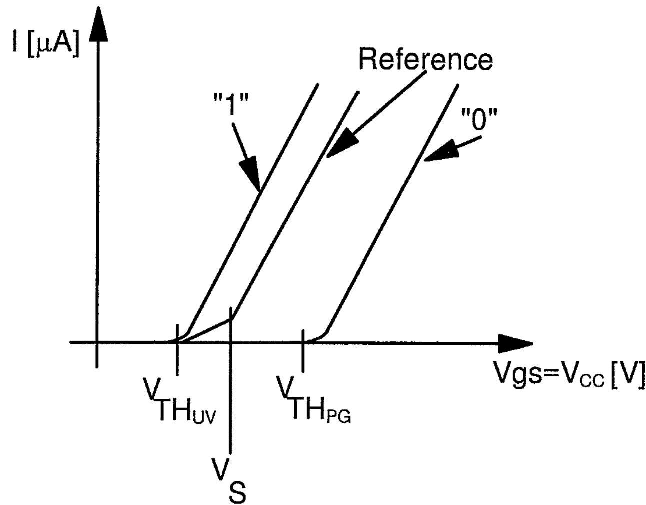


Figure 5.52 The ideal reference characteristic for EPROM devices.

the I_X curve from the I_V curve (Fig. 5.54). I_X is obtained as

$$I_X = I_{V_{\mathrm{TH_{UV}}}} - \frac{I_{V_S}}{n} \tag{5.17}$$

where factor n has to reproduce the situation of unbalanced loads.

Fig. 5.55 shows the I_X current generator: M6 is a virgin cell, while M1 and M2, (properly dimensioned) drive the M3 virgin cell with a voltage ($\approx$ 1V) equal to $V_{\mathrm{CC}} - V_{\mathrm{TH_{PG}}}$: this is a way to obtain a "virtual threshold" of M3 which is $\approx$ 1V higher than the one of M6. This trick was useful for an EPROM device where writing a reference cell was not feasible; for a Flash memory, the same characteristic is obtained using reference cell programmed at EWS. Transistors M4 and M5 form a mirror configuration; reducing the aspect ratio of M5 by a factor n with respect to that of M4, the mirrored current is also reduced by the same factor (see Fig. 5.56).

Kirchhoff's current law on node A gives:

$$I_X = I_V - \frac{I_3}{n} \tag{5.18}$$

The complete scheme of the differential semi-parallel sense amplifier is depicted in Fig. 5.57; I_r is obtained subtracting I_X from the UV-erased cell (i.e. MR) current. V_M and V_R versus V_{CC} are depicted in Fig. 5.58: they are always separated and the minimum value of V_{CC} is $V_{\mathrm{TH_{UV}}}$, while a V_{CC} maximum value does not exist. For an EPROM device this type of sense is ideal.

APPLE INC.
EXHIBIT 1008 - PAGE 0303

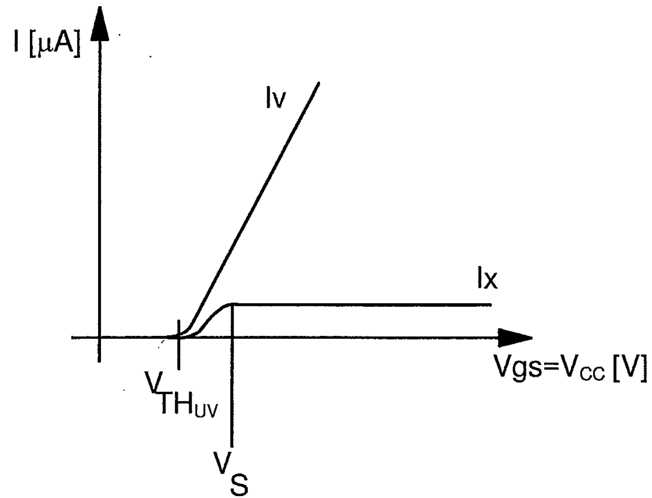


Figure 5.53 Offset current generation.

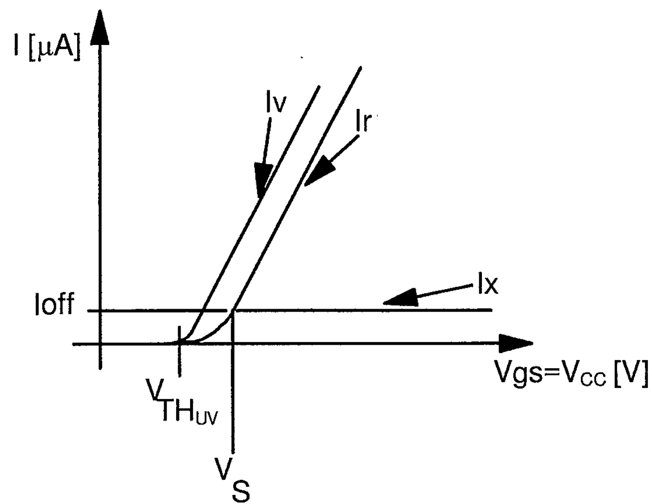


Figure 5.54 I_r current generation.

5.4.5 Reading Speed-up Techniques

After analyzing the circuits to read a non-volatile cell, it is useful to inspect some techniques to improve both their speed and reliability.

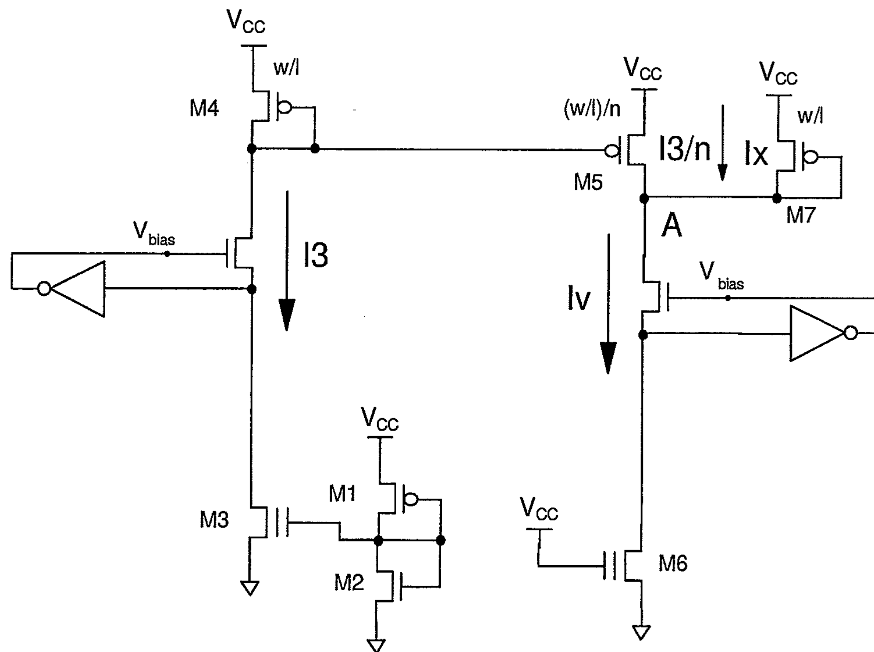


Figure 5.55 I_X current circuit generation.

The first one is "node equalization", i.e. a way of reducing voltage variations of certain nodes in order to speed up commutations. Fig. 5.59 shows an unbalanced loads converter with two matrix cells: MM1, virgin and decoded by YN1, and MM2, programmed and decoded by YN2.

Fig. 5.60 shows the time diagram of two consecutive read sequences. Suppose that the first read is on MM1 and the second on MM2, and that, for sake of simplicity, V_R has a constant value, then it is clear that V_M has a considerable voltage swing, which means a long transient time. Conversely, to enhance the time responsiveness of the circuit, all the nodes should make small movements around their biasing values: therefore Fig. 5.59 is modified into Fig. 5.61, where a natural transistor (i.e. a transistor whose threshold voltage, ≈ 0.1V, is lower than the one of a normal implanted enhancement nMOS) M1 has been added; its task is to short-circuit nodes V_R and V_M when ATD is High (ATD is an acronym for "Address Transition Detector", i.e. a circuit able to generate a pulse, used as a clock, when the state of either E# or one or more address pins changes). During this phase, V_M and V_R are both clamped at a proper value, for example $V_{CC}/2$: when ATD pulse stops, node V_R and V_M start swinging to reach their final values according to the matrix cell contents, with

APPLE INC.
EXHIBIT 1008 - PAGE 0305

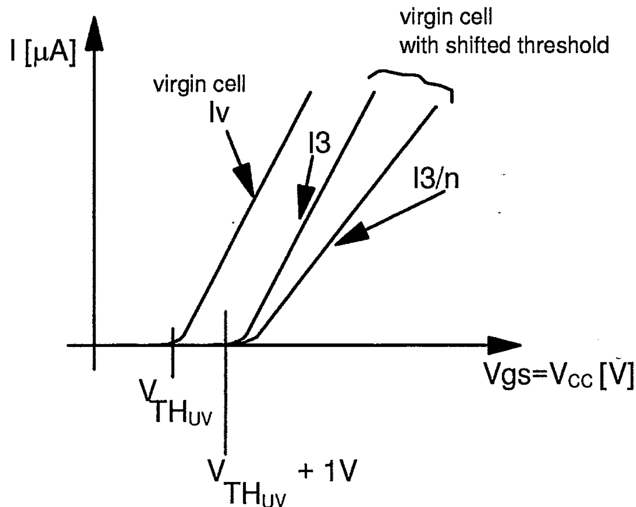


Figure 5.56 Shifted threshold voltage cell characteristic.

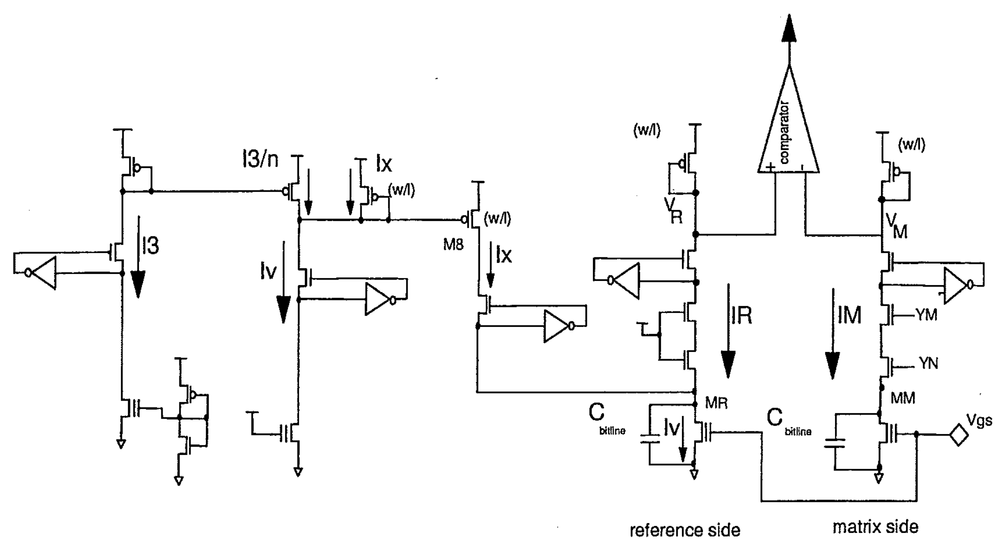


Figure 5.57 Complete schematic of differential semi-parallel sense amplifier.

a voltage excursion lower than in the previous solution. Normally equalization technique also involves other nodes, increasing the speed (and, unfortunately, the complexity) of the circuitry.

APPLE INC.
EXHIBIT 1008 - PAGE 0306

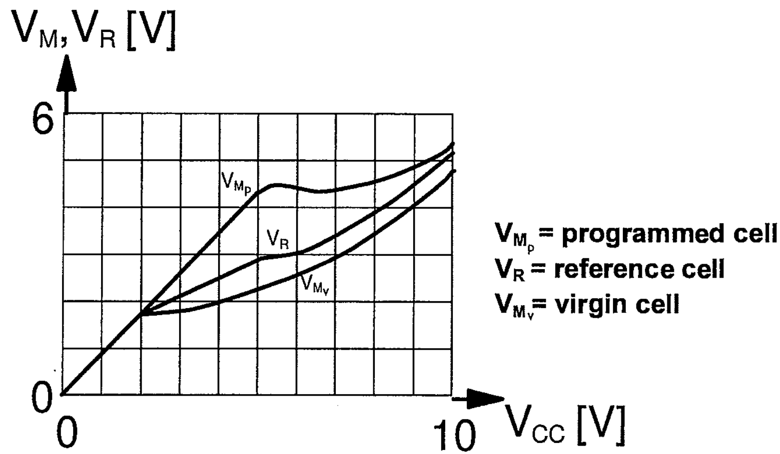


Figure 5.58 Current-voltage semi-parallel converter output node.

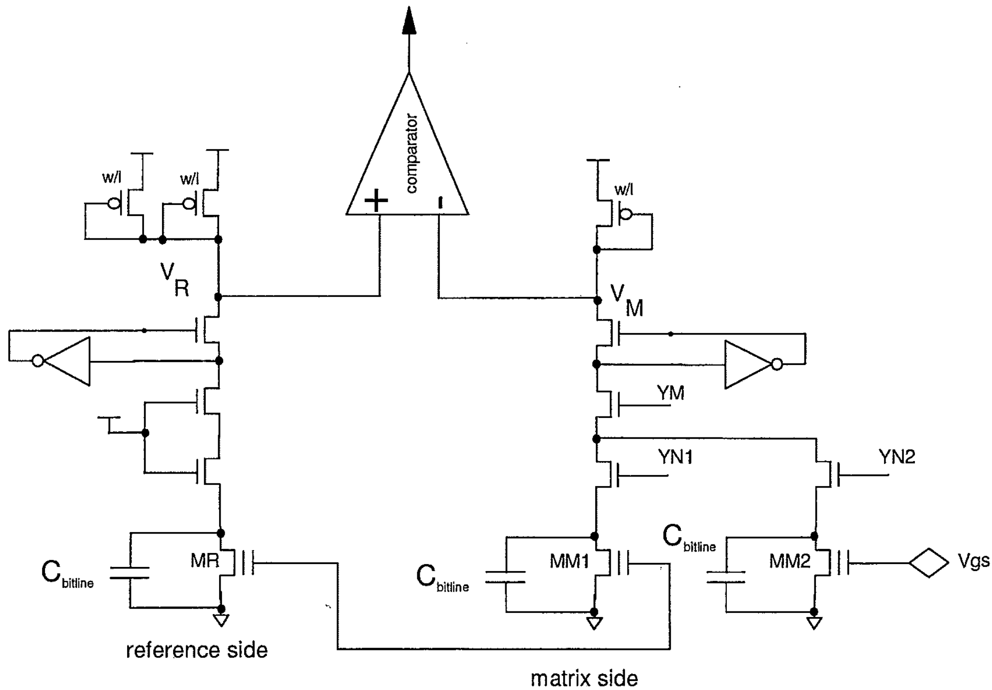


Figure 5.59 Schematic to study consecutive reading of different cells.

Together with equalization, the "precharge" technique is used to speed up Read operations; to precharge a node means to use time between different

APPLE INC.
EXHIBIT 1008 - PAGE 0307

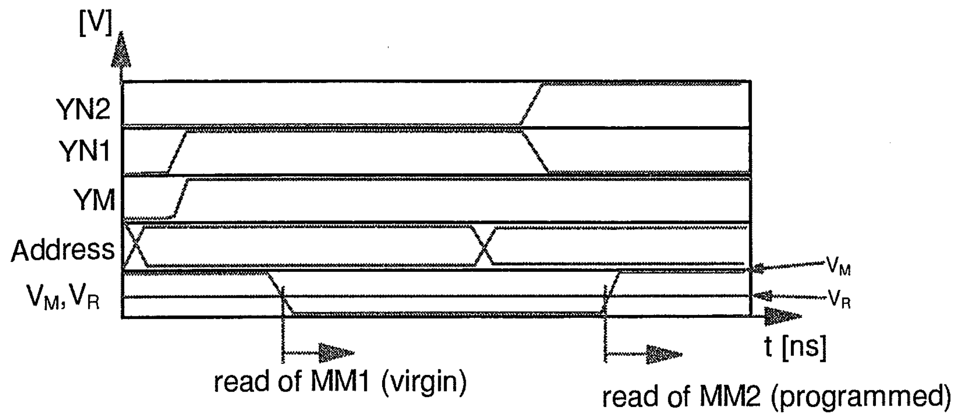


Figure 5.60 Waveform of consecutive reading of two cells located on different columns.

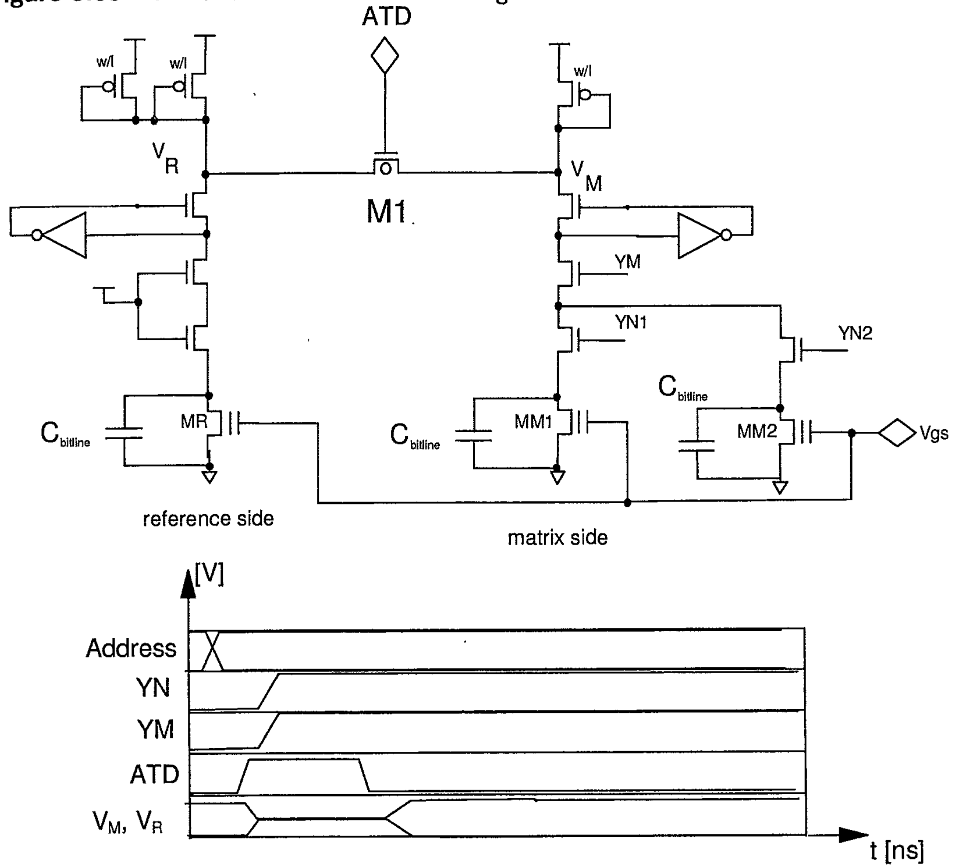


Figure 5.61 Equalization of output nodes converter.

phases to prepare the parasitic capacitances: Fig. 5.62 shows the precharge transistor for an unbalanced load current-voltage converter: M6 and M7 are switched on during ATD-active phase to share stray capacitances charge: equalization, by means of M1÷M5, is simultaneous to precharge, while the reading is performed only by the loads.

Many other "tricks" can be devised to speed up the converter: as shown in Fig. 5.63, it is possible to add two nMOS in diode configuration, M8 and M9, one opposite to the other; the effect is that nodes V_M and V_R cannot be separated by more than two V_{TH_n}, thus decreasing the voltage swing and saving time. The drawback for this type of solution is the deterioration in noise margin.

Another way to improve reading is to consider a different arrangement of the pMOS load transistors. Instead of connecting them both in a diode configuration, as it was in the previous schemes, it is possible to realize a mirror configuration, as shown in Fig. 5.64. The advantage is an improvement in the dynamic of V_M: the lower limit (reading of a virgin cell) is lowered from 1V down to GND, while the upper bound (reading of a programmed cell) is raised from $V_{CC} - V_{TH_p}$ to V_{CC}.

It is important to choose properly the dimension of both M1 and M3 (the equalization transistors, as shown in Fig. 5.65): if they were too big, then the stray capacitances between V_M and V_R and their gates (driven by the equalization signal) become significant: Fig. 5.66 shows a simulation, in which the coupling between source-gate and drain-gate for M1 and M3 is evident.

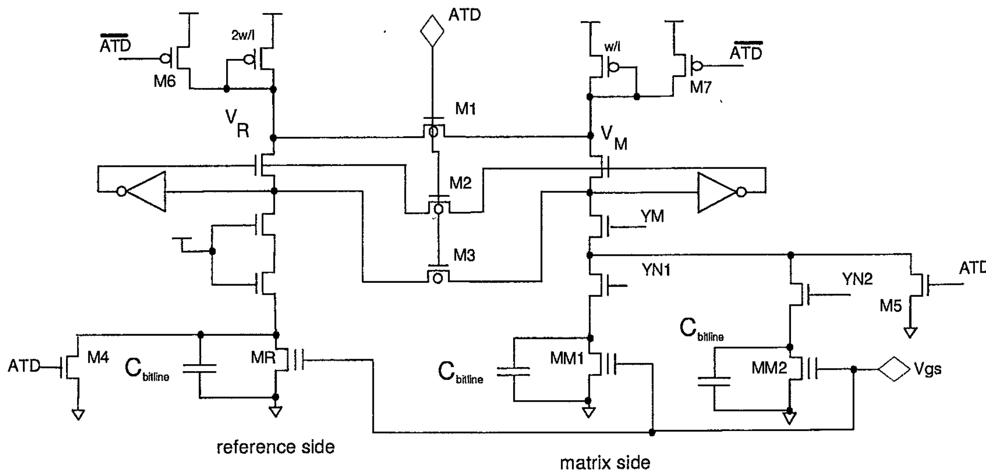


Figure 5.62 V_M and V_R precharge nodes.

APPLE INC.
EXHIBIT 1008 - PAGE 0309

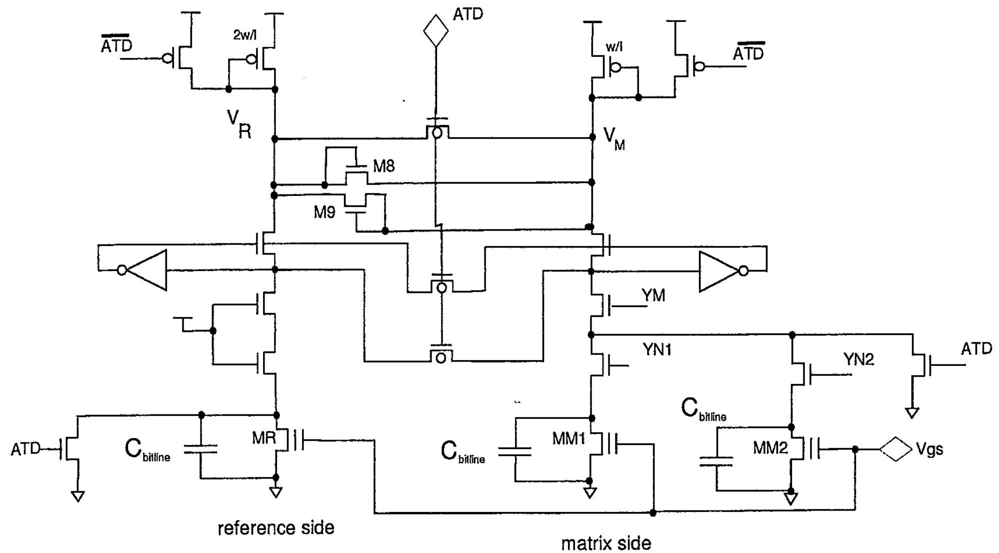


Figure 5.63 Converter outputs nodes clamping.

Fig. 5.67 shows a different solution which uses two additional transistors, MA and MB, whose on-channel resistance is about 2kΩ; CK1 has a rising edge coincident with the CK signal but stays High for a longer time; when MA and MB are still on, and M1 and M3 are already off, the current through MA and MB generates a potential drop between V_M and V_R. The behavior of the circuit in the case of a virgin and of a programmed cell is shown in Fig. 5.68a and Fig. 5.68b respectively.

If the cell is virgin (Fig. 5.68a), then

$$I_1 = I_V + I_X \quad (5.19)$$
$$I_2 = I_V - I_X \quad (5.20)$$

and then

$$2I_V = I_1 + I_2 \quad (5.21)$$

Using the unbalanced load sense amplifier

$$I_1 = 2I_2 \quad (5.22)$$

therefore

$$I_X = \frac{1}{3} I_V \quad (5.23)$$

APPLE INC.
EXHIBIT 1008 - PAGE 0310

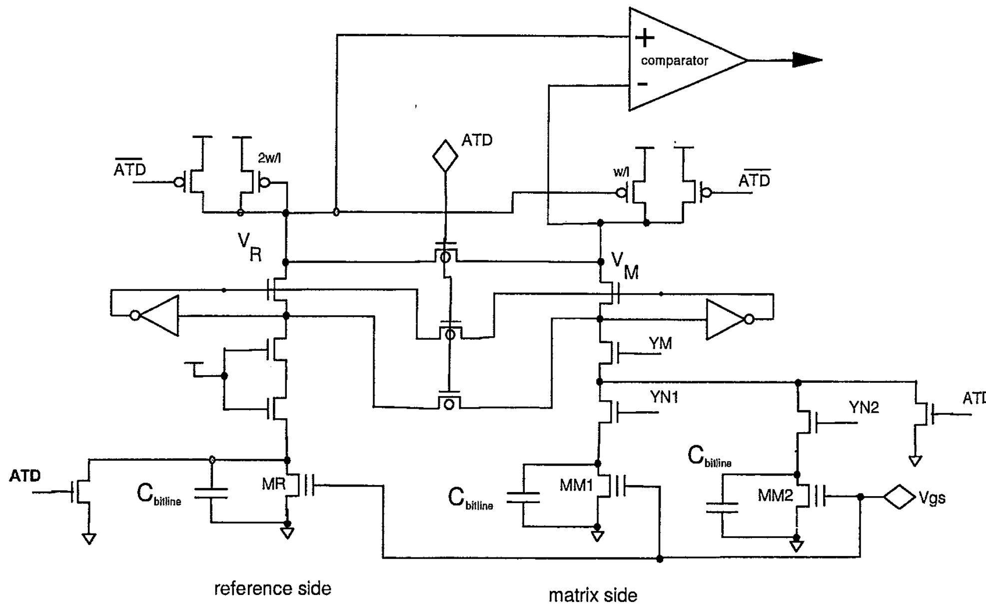


Figure 5.64 Differential read with loads in mirror configuration.

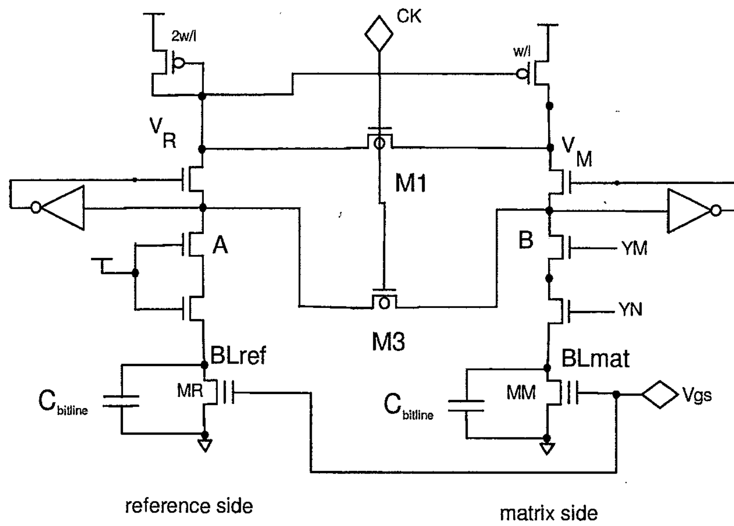


Figure 5.65 Differential reading with equalized load in mirror configuration.

APPLE INC.
EXHIBIT 1008 - PAGE 0311

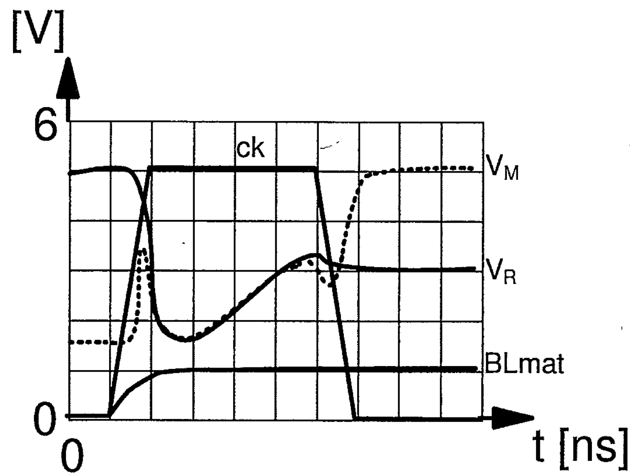


Figure 5.66 Clock interaction on output node.

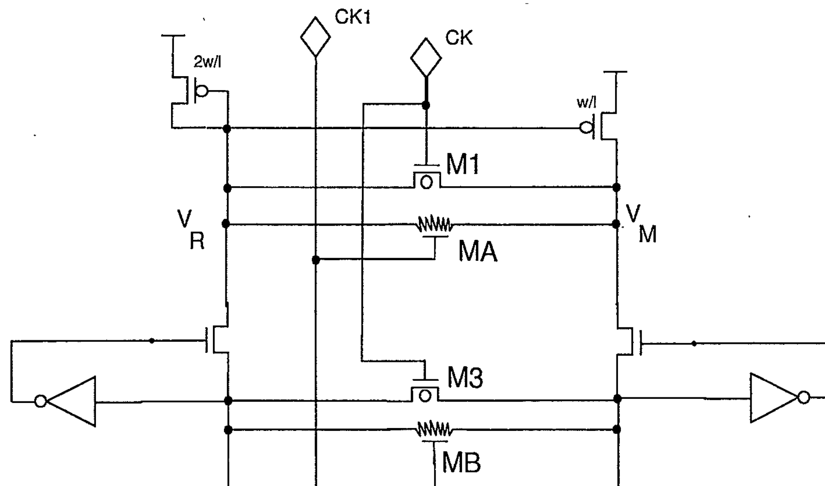


Figure 5.67 Equalization scheme using also resistive transistor.

$$V_R - V_M = \frac{R}{3} I_V \qquad (5.24)$$

If the cell is programmed (Fig. 5.68b), then

$$I_1 = I_V - I_X \qquad (5.25)$$

$$I_2 = I_X + I_S \qquad (5.26)$$

$$I_1 = 2I_2 \qquad (5.27)$$

APPLE INC.
EXHIBIT 1008 - PAGE 0312

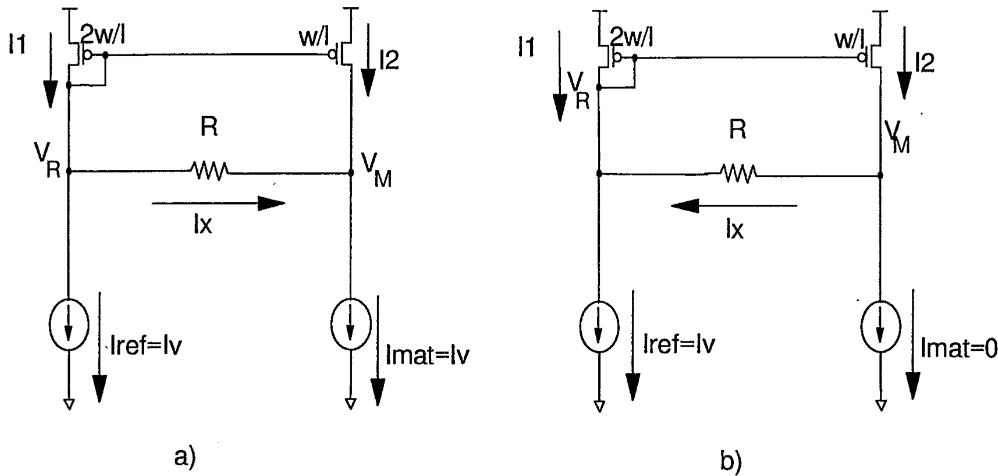


Figure 5.68 Principle schemes for resistive equalization.

Defining $k = I_S/I_V$, i.e the ratio between the current of a virgin and a programmed cell, and assuming $0 < k < 1/2$, then

$$I_X = (1/3)(1-2k)\, I_V \qquad (5.28)$$

$$V_M - V_R = (1/3)(1-2k)\, I_V R \qquad (5.29)$$

Depending on the state of the cell, current I_X flows in opposite direction; if the cell is virgin, the reference side "helps" the matrix side to charge the capacitance, while if the cell is programmed the opposite situation occurs. From Eq. (5.29) it is clear that it is possible to separate V_M and V_R nodes of the known quantity even during the ATD phase, provided that the decoder has already activated the cell.

5.4.6 From EPROM to Flash

All the concepts presented in the previous paragraphs are also valid to design a sense amplifier for Flash memories, so that different approaches can be used (unbalanced, parallel or semi-parallel loads). However there are new issues, related to the electrical erase, which must be taken into account. Fig. 5.69 shows a typical distribution of the erased cells threshold and the threshold value of an UV-erased cell ($V_{TH_{UV}} = 2V$). The former should not exceed the lower limit of 0.5V, and therefore is placed between 0.5V and 2.5V. Under these conditions it is clear that if the worst erased cell ($V_{TH} = 2.5V$) is read with $V_{CC} = 2.7V$ (minimum value for a typical low voltage device), then its

APPLE INC.
EXHIBIT 1008 - PAGE 0313

overdrive is only 200mV, i.e. only $1 \div 5\mu A$ of current: it means that the sense amplifier resolution must be really fine.

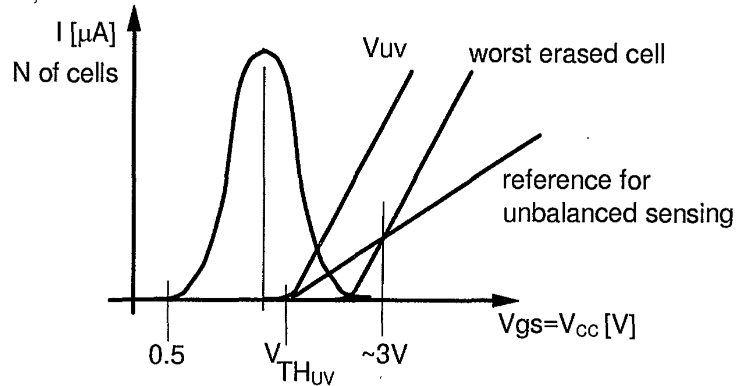


Figure 5.69 Erased Flash cells threshold distribution.

On the other hand, another problem arises when there are erased cells whose $V_{TH} < 0$ (i.e. "depleted" cells): this undesired situation is depicted in Fig. 5.70. Suppose that C1 (programmed, $V_{TH} = 6V$) is the addressed cell; other cells share the same bitline: C3 (virgin, $V_{TH} = 2V$) and C2 (depleted, $V_{TH} < 0$); obviously the "traditional" sense amplifier is not able to understand whether the current contribution comes from C1 or from C2. Usually the problem of depleted bits is solved by means of the so-called "Soft Programming": at the end of the Erase algorithm, depleted cells are detected and are slightly re-programmed until their threshold goes above zero. In the next paragraph a different approach is presented.

5.4.7 Reading Flash Memories with Depleted Bits

Instead of recovering depleted bits, it is possible to read correctly despite their presence on the selected bitline, provided that matrix architecture is changed as shown in Fig. 5.71: another "bitline" (BLP, made up of nMOS pass P1 ... Pn) is placed inside the sector: these transistors are connected to bitline, wordline and source in the same way as the Flash cells. If A is the addressed cell, WL2 is High and P3 is on, thus connecting BLP line with the sources of all the cells whose wordline is either WL2 or WL3.

If BLP is biased at GND, then the above-mentioned sources are at GND as well, while those of the other cells are floating, since their Pi transistors are off.

Unfortunately, if a NOR matrix architecture is chosen (i.e. all the cells share the same source), a depleted cell (e.g. cell B or C), can give its unwanted

APPLE INC.
EXHIBIT 1008 - PAGE 0314

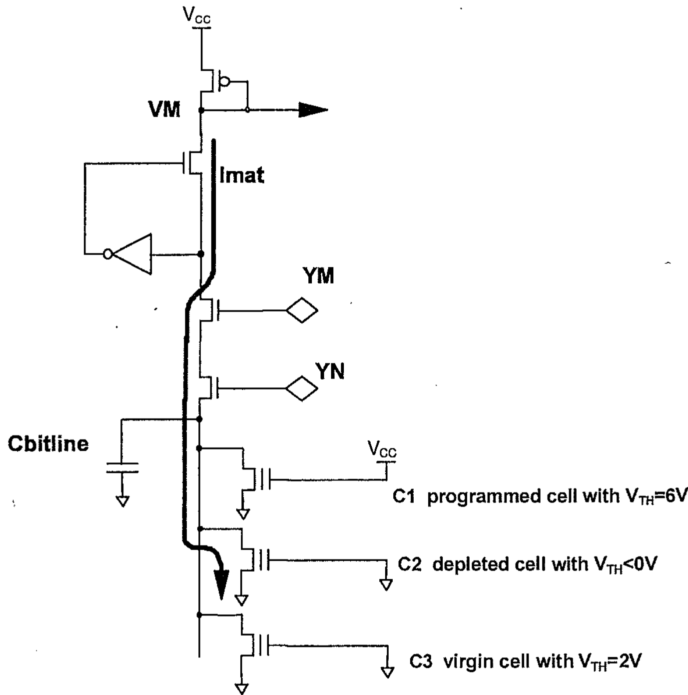


Figure 5.70 Typical read fail induced on sense amplifier due to a depleted cell.

current contribution even if its pass (P4, in this case) is off, because source node (which is at GND) is common to all the cells with gate driven by either WL2 or WL3.

To overcome this problem (i.e. to switch off the non-addressed cells, depleted or not), body effect can be exploited: the value of the threshold is modified as follows:

$$V_{\mathrm{TH}} = V_{\mathrm{TH}_0} + \Delta V_{\mathrm{TH}_{\mathrm{body}}} \quad (5.30)$$

$$\Delta V_{\mathrm{TH}_{\mathrm{body}}} = \gamma\left(\sqrt{2|\Phi_p| + |V_{\mathrm{sb}}|} - \sqrt{|\Phi_p|}\right) \quad (5.31)$$

where Φ_p is the Fermi potential, V_{sb} the source-bulk voltage, V_{TH_0} the threshold voltage if $V_{\mathrm{sb}} = 0$ and γ the body effect coefficient.

Biasing is as follows:

- BLP is driven at V_{CC} instead of GND;
- non-addressed bitlines are tied at V_{CC};
- the addressed bitline is tied at $V_{\mathrm{CC}} - 1.5\mathrm{V}$;

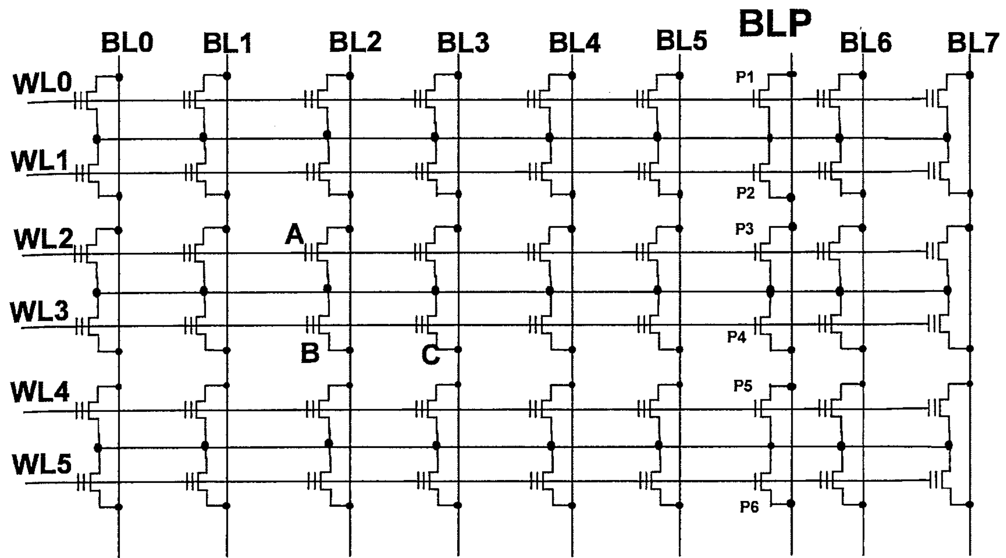


Figure 5.71 Flash NOR modified architecture.

- non-addressed wordlines are tied at GND;
- the addressed wordline is boosted (see Section 5.2.5).

Therefore for all the non-addressed cells (i.e $V_g = 0\text{V}$) on the selected bitline (i.e. $V_s = V_{\text{CC}} - 1.5\text{V}$)

$$V_{\text{gs}} = V_g - V_s = -(V_{\text{CC}} - 1.5\text{V}) \tag{5.32 }$$

This equation also applies if one of those cells is depleted: because of the particular biasing scheme, its V_{sb} is higher than zero, and therefore its threshold (which is undesirably negative) can be raised, thus making its switching-on more difficult. Therefore, to turn on a depleted cell whose threshold is V_{T0}, its V_{gs} must be

$$V_g - V_s > V_{\text{TH}} = V_{\text{TH}_0} + \Delta V_{\text{TH}_{\text{body}}} \; . \tag{5.33 }$$

Another advantage of the above mentioned solution is that it is possible to reverse cell source and drain, thus devising different approaches to perform both Program and Erase.

APPLE INC.
EXHIBIT 1008 - PAGE 0316

5.4.8 Low Voltage Flash Read

The main problem when reading at low voltage is the amount of cell current available. Supposing that the worst erased cell has a threshold equal to 2.5V, the current sunk by the cell could range from 1μA to 5μA at 2.7V, which is the minimum value of V_{CC} for a low voltage device. On the other hand, fast reading requires that reference characteristic is separated from matrix cell characteristic: in this case, however, the differences between reference and matrix currents could be so small to imply very long switching time when sensing.

For a quick movement of all the nodes, it is mandatory that both an erased cell and the reference cell sink large currents, in order to discharge the bitline capacitance and to read a programmed cell, respectively.

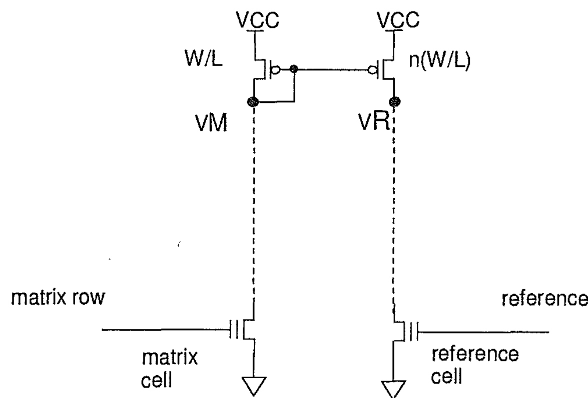


Figure 5.72 Mirror connection to amplify cell current.

It is necessary to design a new type of converter, which is capable of amplifying cell current; such a circuit is shown in Fig. 5.72: the matrix cell current is mirrored to the reference branch and it is amplified by the mirroring factor n, thus overcoming the problem of poor current when low voltage operation is performed.

In Fig. 5.73 characteristics for a programmed matrix cell and for the reference cell are shown, while Fig. 5.74 shows nodes V_M and V_R when $V_{TH_{PG}} > 6$V.

In Fig. 5.74, two different conditions should be investigated: $V_{CC} < V_{TH_{PG}}$ and $V_{CC} > V_{TH_{PG}}$; in the first case nodes V_M and V_R are tied together at $V_{CC} - V_{TH_{PG}}$ until V_{CC} is high enough to turn the reference cell on. Subsequently, V_R is forced to 0V by the reference, while V_M voltage keeps on rising as the matrix cell does not draw any current. When $V_{CC} > V_{TH_{PG}}$ the matrix cell starts conducting, so that node V_M is forced downward, and V_R is consequently driven upward due to the mirror amplification.

APPLE INC.
EXHIBIT 1008 - PAGE 0317

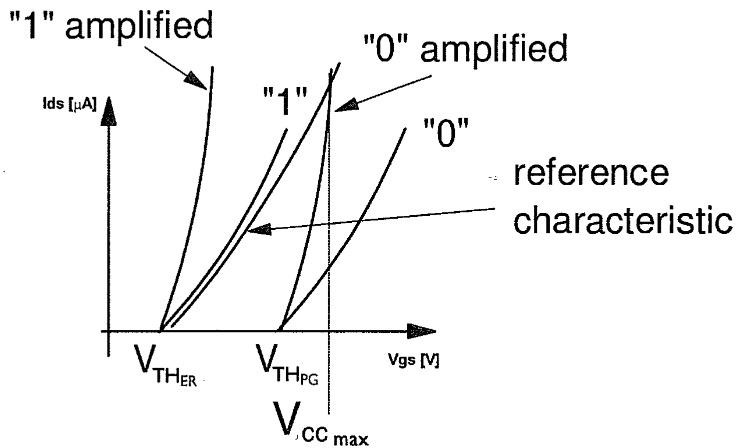


Figure 5.73 $V_{CC_{max}}$ problem for the amplified converter.

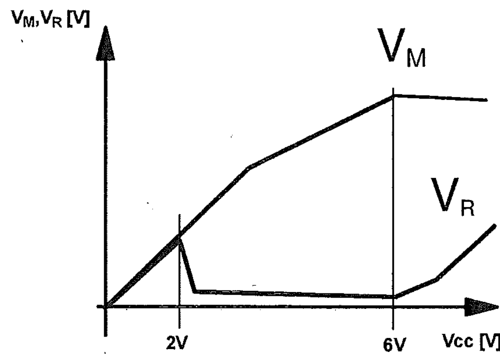


Figure 5.74 Output nodes for a programmed cell with threshold greater than 6V.

However, if the threshold shift for a programmed cell were small, there could exist a value of supply voltage where the two curves V_R and V_M have an intersection, thus fixing a V_{CC} maximum value (Fig. 5.75). On the other hand for an erased cell the two nodes are always divided, as in Fig. 5.76.

The most important thing to do is to design a reference characteristic able not to cross erased and programmed characteristics within the entire supply voltage range; Fig. 5.77 shows cell and reference characteristics linearized under the hypothesis that drain voltage is equal to 1V: $V_{TH_{REF}}$, $V_{TH_{ER}}$ and $V_{TH_{PG}}$ are the thresholds of the reference, of the worst erased and of the worst programmed cells respectively.

Supposing to apply a boost to the row of V_b volt, the characteristics are shifted towards left; if g is the transconductance of the cell and n is the ampli-

APPLE INC.
EXHIBIT 1008 - PAGE 0318

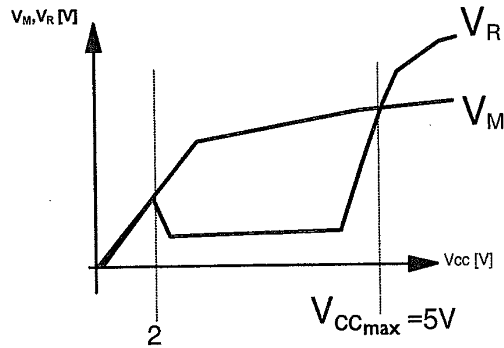


Figure 5.75 Output nodes for a programmed cell with a low voltage threshold shift.

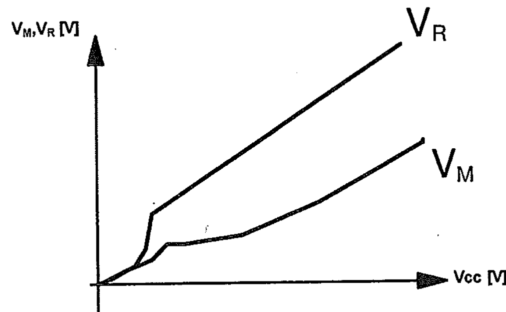


Figure 5.76 Output nodes for an erased cell.

fication factor of the matrix current, then

$$I_{\mathrm{ds}} = g(V - V_{\mathrm{TH}}) \tag{5.34}$$

$$V_{\mathrm{CC_{min}}} = \frac{nV_{\mathrm{TH_{ER}}} - nV_b - V_{\mathrm{TH_{REF}}}}{n-1} \tag{5.35}$$

$$V_{\mathrm{CC_{max}}} = \frac{nV_{\mathrm{TH_{PG}}} - nV_b - V_{\mathrm{TH_{REF}}}}{n-1} \tag{5.36}$$

If the slope of the reference is changed at a proper value V_s, then $V_{\mathrm{CC_{max}}}$ problem no longer occurs, as shown in Fig. 5.78; Fig. 5.79 shows a possible scheme to generate reference current through an appropriate mirroring scheme.

Care must be taken when designing this converter because a reading fail could be induced by bitline leakage that can be amplified by the mirror connection: a programmed cell could be disguised as erased.

APPLE INC.
EXHIBIT 1008 - PAGE 0319

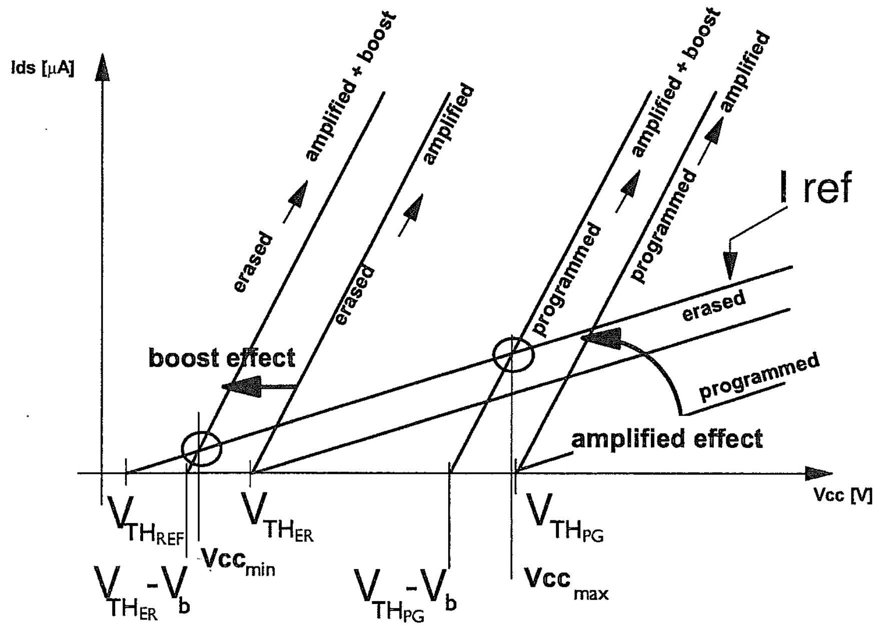


Figure 5.77 Analysis of amplified converter in (V_{CC}, I_{ds}) plane.

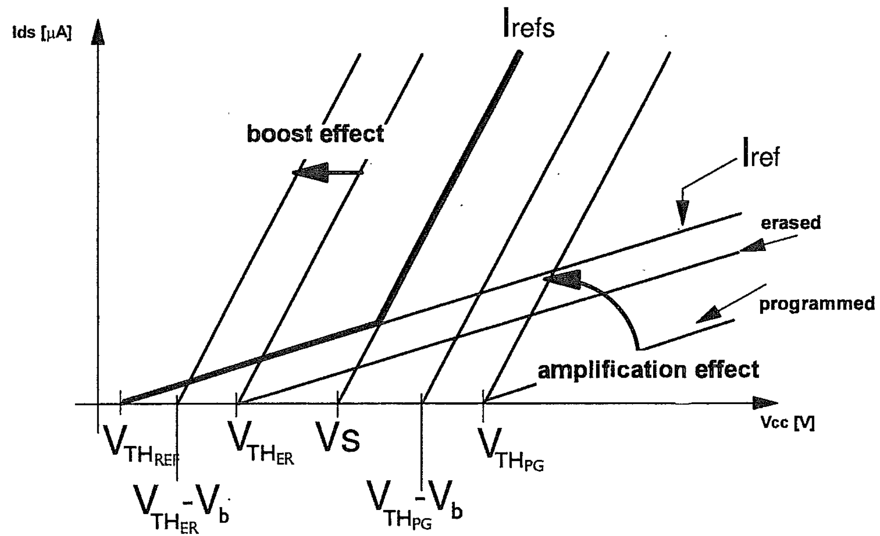


Figure 5.78 $V_{CC_{max}}$ problem solved for the amplified converter.

APPLE INC.
EXHIBIT 1008 - PAGE 0320

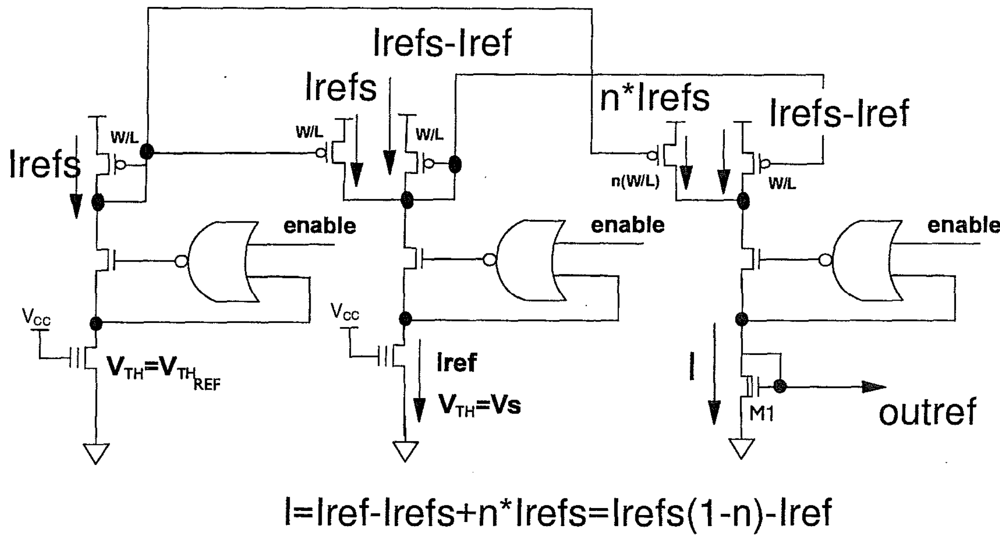


Figure 5.79 Scheme for current reference generation.

5.4.9 Reference Problems

As shown in Fig. 5.79, reference current is generated by a reference cell which is usually located in a small matrix outside the data matrix; this current is carried around the chip by means of a nMOS transistor in diode configuration.

In order to to understand and better appreciate this solution, the global structure of the reference should be investigated further.

In a EPROM device, every output has a dedicated column where all the cells are reference cells; the addressing of a row automatically turns on also the reference cell for the related output. This solution guarantees a good timing correlation between matrix and reference cells, because the row is the same, as well as the loads on the two converter branches due to the architectural choice. The drawback of using such an approach is that the number of reference cells is too big: for example, for 1 Megabit Flash with 8 outputs, organized as 1024 rows by 1024 columns, the number of reference cells is 8192, 1 column for each output and one cell for each row.

In the first generation of Flash memories, when the possibility of choosing the threshold of the reference cell had not been introduced yet, UV-erased cells were used and the organization was as in EPROM, with a reference column every output. These Flash featured source Erase, i.e. gate at zero volt and source at 12V, and the sources of the reference cells were separated from the sources of the matrix cells.

APPLE INC.
EXHIBIT 1008 - PAGE 0321

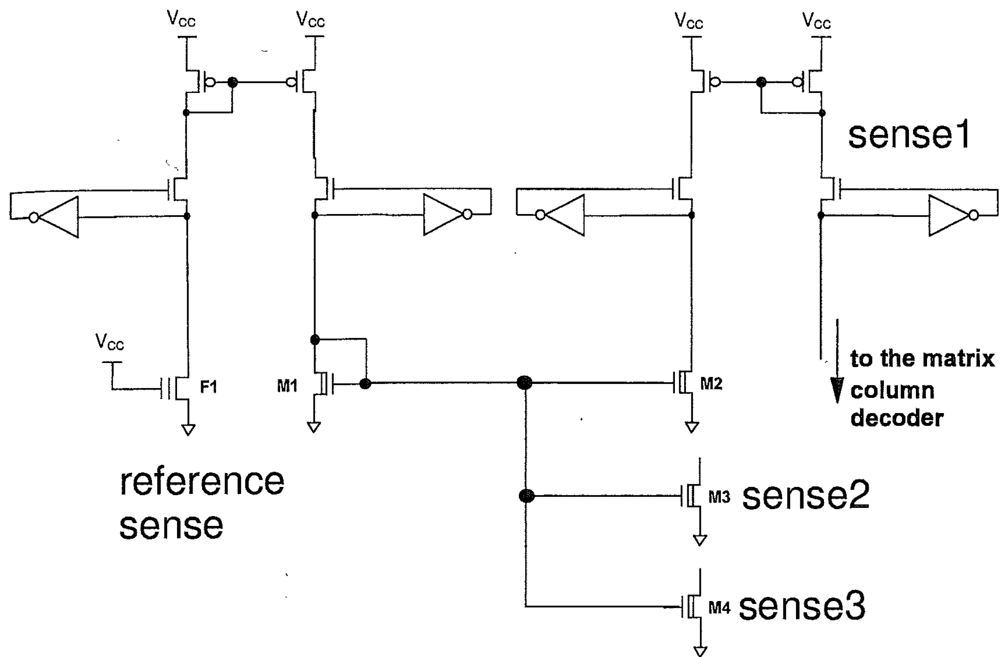


Figure 5.80 Current reference cell F1 is mirrored inside different branches of sense amplifier.

Introducing the new Flash generation featuring negative voltage Erase, it was mandatory to remove reference cells from the matrix, to avoid stress which could modify the threshold of the reference cells themselves.

In this case the problem of carrying around the chip the current generated by the reference circuitry arises and it is easy to understand that (see Fig. 5.79) it is easier to use a natural nMOS M1 to mirror reference current in all other reference sides (Fig. 5.80) (really the circuit shown in Fig. 5.80 is more complex, because the configuration used to mirror F1 current in M1 would work like a perfect mirror only if the two pull-down transistors F1 and M1 will had the same electrical characteristics: this is not true, and circuit stops working properly at $V_{CC} \approx 4V$; this limit can be eliminated using a circuit featuring current compensation).

A brief excursus through principal current-voltage converter used to read Flash non-volatile memory current cell was discussed. Many details were omitted to give the readers only the perception of the fundamental problems involved in this type of design.

APPLE INC.
EXHIBIT 1008 - PAGE 0322

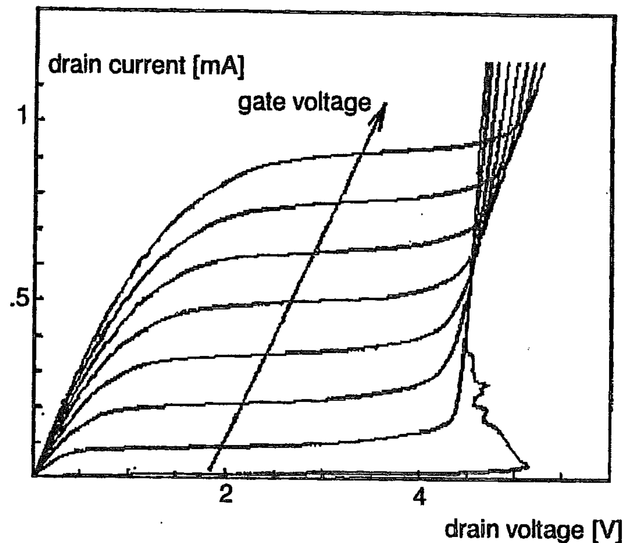


Figure 5.81 Snap-Back triggering vs. gate voltage.

5.5 PROGRAM OPERATION CIRCUITRY

In Section 5.1 it was pointed out that some voltages (greater than V_{CC}) are needed to modify the state of a Flash memory cell. More in details, to program a Flash cell using the Channel Hot Electron effect, two different positive voltages have to be applied to gate and drain terminals, while the source is tied to GND. Moreover, these voltages are greater than the conventional power supply level (5 or 3V) and a regulation is also required to avoid spurious phenomena which can damage the memory cell or reduce its reliability after several Program/Erase cycles are executed. Furthermore, other voltage levels are necessary to perform the Program Verify operation, i.e. a check of the cell state to ensure that its threshold has properly shifted after the program pulse. This Section analyze the circuitry devoted to the generation and the regulation of these programming voltages.

5.5.1 Cell Programming Voltages: Optimum Choice

The choice of the optimum programming voltages is often a hard challenge, and several considerations drive the levels and timings definition. First of all during the Program phase the operating point of the memory cell must be kept within its "Working Window", which defines the allowed range for the drain voltage versus the effective cell length, taking into account the program efficiency, the Snap-Back and the Drain Stress effects.

APPLE INC.
EXHIBIT 1008 - PAGE 0323

The gate voltage also affects some important parameters like the program efficiency, the drain current and the Snap-Back triggering point. The dependence of Snap-Back triggering on the gate voltage can be found in Fig. 5.81. From this plot it is easy to see that a critical gate voltage exists for which the Snap-Back turn on takes place at the minimum drain voltage value. Consequently, it is important to raise the gate of the memory cell before the activation of the program load, which provides the drain voltage. This means that, during the gate voltage ramp-up, the drain program voltage is not applied to the cell, thus avoiding any risk of Snap-Back. Fig. 5.82 shows a typical timing diagram for gate/drain voltages and the drain current during programming. The Program Verify phase is also shown in this figure.

As far as Drain Stress is concerned, it is important to consider both the voltage value applied to the drain and for how long it is applied.

Since the onset of the previous described parasitic effects (Snap-Back and Drain Stress) is strongly dependent on the programming voltages, the gate and drain voltages are usually regulated by a devoted circuitry.

In conclusion, the main constraints that drive the optimum programming voltage choice are the following:

drain voltage:

- program efficiency (i.e. Program time);
- Drain Stress effect on the others cells of the same bitline;
- Snap-Back effect;
- drain current;

gate voltage:

- program efficiency;
- Snap-Back effect;
- drain current.

5.5.2 *Typical Program Path*

Fig. 5.83 displays the block diagram for the analog program signals in a Flash memory device. It is supposed that the appropriate instruction has been applied to the device, so that the Program/Erase Controller is activated, and it is therefore developing its algorithm; furthermore it is assumed that the cell address, and the data to be programmed have been stored in the input latches. The Program algorithm starts with a check (or "Program Verify") of the actual cell contents, and if the latter are not equal to those of the input latches,

APPLE INC.
EXHIBIT 1008 - PAGE 0324

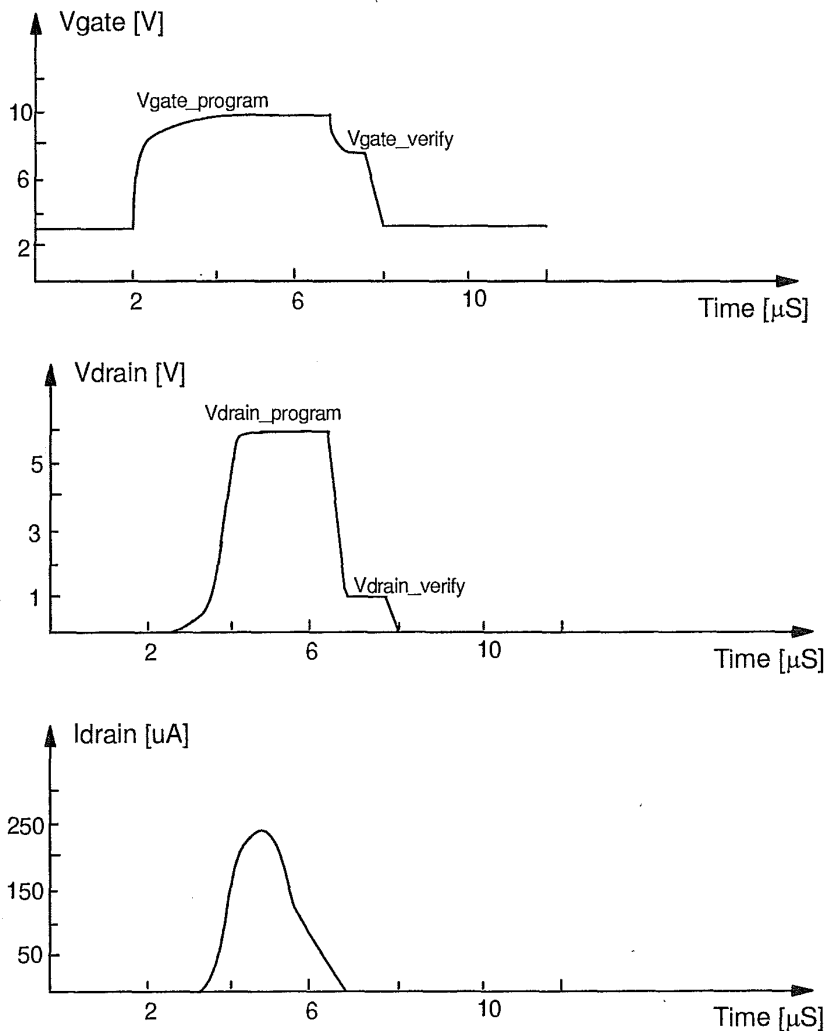


Figure 5.82 Programming voltages and current: timing.

high-voltage signals are forced to the gates and drains of the selected memory elements for a given period of time, and then another Program Verify is performed to check the results. Again, if the comparison between the latched and memory data is successful, the algorithm will terminate; otherwise the above procedure will be repeated until the number of program attempts exceeds an allowed maximum limit, in which case the algorithm terminates with a fail.

In this paragraph, the circuits to provide the Flash cell with regulated high voltage signals are analyzed; in the following, VPCX is the voltage to be applied

APPLE INC.
EXHIBIT 1008 - PAGE 0325

to the cell gates through the row decoder, VPCY is that for the column decoder, and VPD is that for the cell drains through the program load stages.

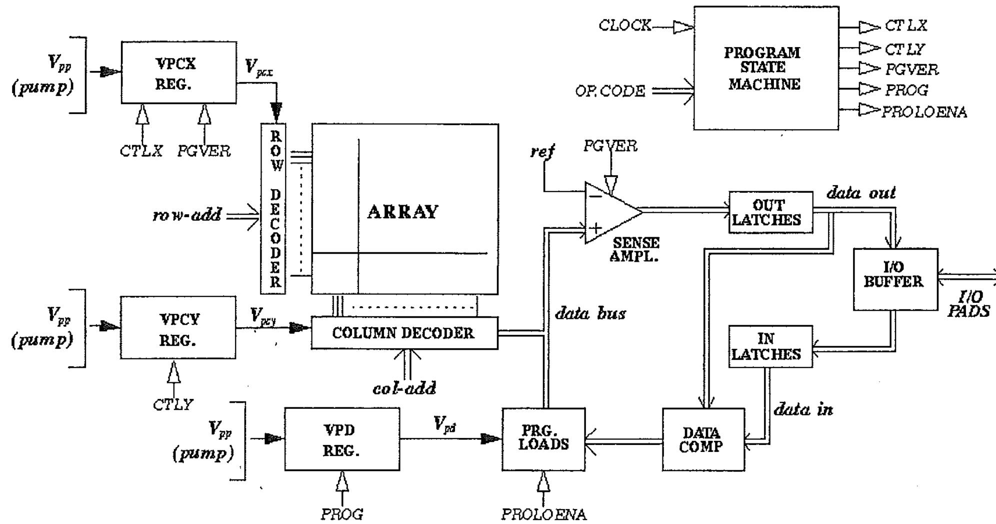


Figure 5.83 Block diagram for program.

5.5.3 *Drain Voltage Regulation: Principles and Basic Circuits*

It is worth remembering that the above voltages are derived from an external pin named V_{PP} in "Double supply" devices, while in "Single supply" ones they are obtained from the power supply V_{CC} by means of charge pump circuits (which are explained in Section 5.6.2). For the time being, the former type of device is considered.

From Fig. 5.84 it is clear that there are two ways of forcing a suitable voltage to the drains of the cells:

1. "regulation through the column decoder" (left-hand circuit in Fig. 5.84), i.e. a tight control of the gate of one of the column decoder pass transistors. In this case, the full V_{PP} voltage is applied to the program loads with no control, while a regulation circuit for the YN column decoder pass transistor limits the voltage on the cell drain to the value:

$$V_d = V_{reg} - V_{gs} - V_1 \tag{5.37}$$

where V_1 is the voltage drop across the final YO pass of the decoder and the local bitline. The regulation is set on the YN and not on the YO pass transistor for practical reasons: in fact the YO pass gates are

APPLE INC.
EXHIBIT 1008 - PAGE 0326

physically distributed throughout the matrix, while the YN and YM ones are located at the bottom end. The achievements of this solution are that no current is provided by the regulator and the voltage drop V_1 is due to the local bitline drop only, while its drawbacks are that V_{gs} in Eq. (5.37) depends (non-linearly) on the program current of the cell, and the column decoder width is to be enhanced in order to reduce the voltage drops V_{gs} and V_1. An example of this kind of regulation is presented in Fig. 5.85. If the currents in the circuit sides are equal, all the transistors work in the saturation region and if homologous transistors (M5 and M2, M6 and M7) are given the same aspect ratios, it can be shown that the drain voltage of the cell is merely a function of both V_{PP} and the resistors (see Fig. 5.85 for the calculations). It's worth noting that the solution of Fig. 5.85 does not make use of feedback, so control can't be exercised on temperature, cell ageing or process spreading.

2. "regulation through the program loads" (right-hand circuit in Fig. 5.84): in this case a regulated VPD voltage is applied to the data bus through the program loads, while the voltage that is forced to the gates of the column decoder pass transistors is high enough to keep them in triode region, thus limiting the voltage drop across them. The drain voltage is

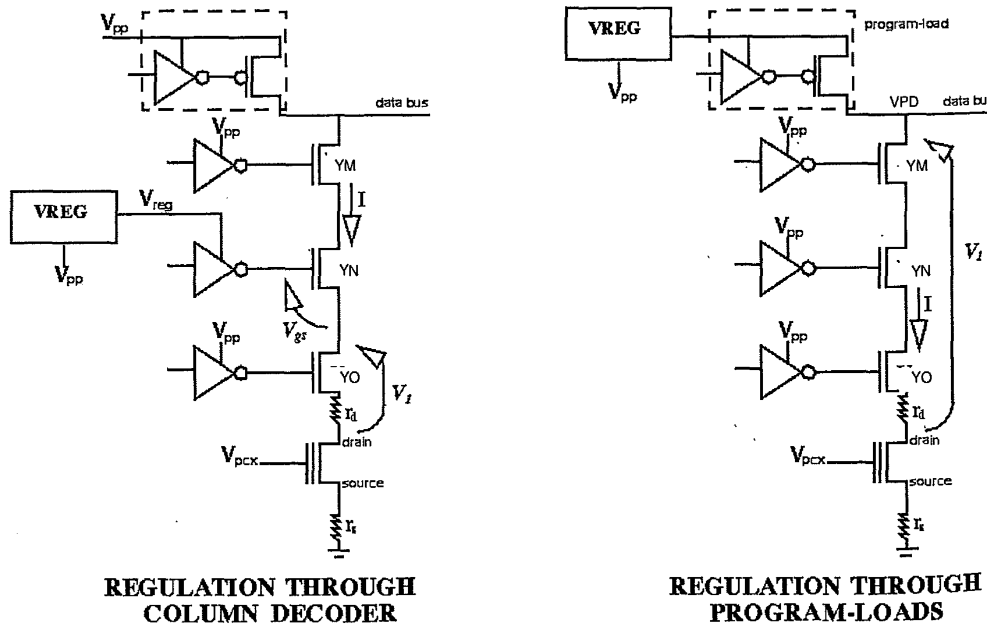


Figure 5.84 Schemes for drain regulation.

APPLE INC.
EXHIBIT 1008 - PAGE 0327

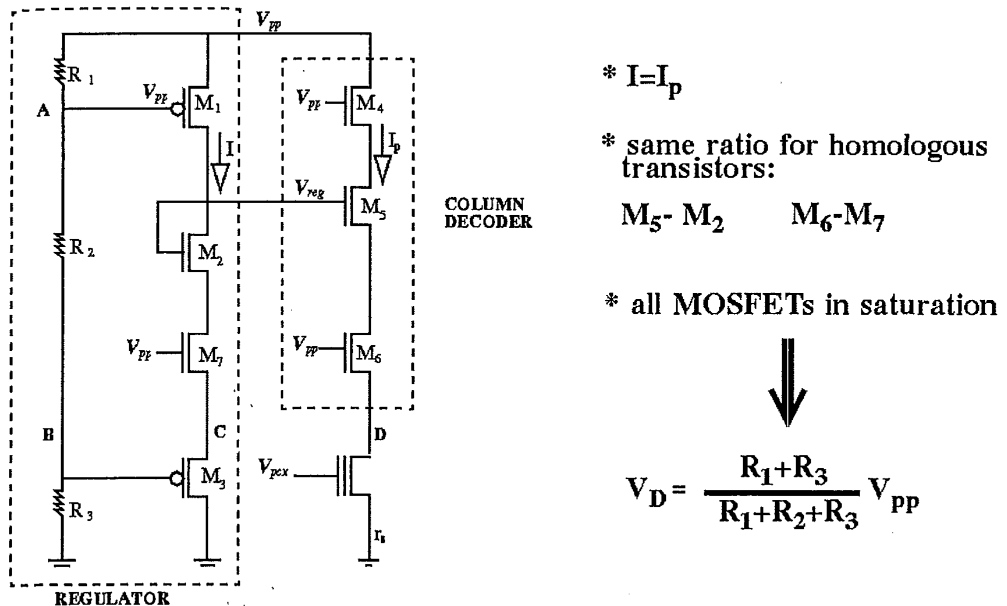


Figure 5.85 Regulation through column decoder.

then equal to:

$$V_d = \text{VPD} - V_1 \tag{5.38 }$$

The striking advantages of this solution are that a more compact column decoder can be adopted, and it will be demonstrated that an active compensation for the actual value of V_1 is achievable, while its drawbacks are that the regulator itself must provide the cells with the program current, and the regulator circuit is more complicated than the preceding scheme. An example of this kind is shown in Fig. 5.86, which makes use of an operational amplifier, a dummy decoder (i.e. a column decoder with no cells attached to it) and a circuit which mirrors the program current Ic into the dummy decoder. In Fig. 5.86 the program load stages are not indicated, as the voltage drop across them can be neglected. Again, by suitably choosing the mirror factor k, the voltage drops across the real and dummy decoders ΔV_d and ΔV_c can compensate for each other, thus the drain voltage is made independent of temperature, current and ageing:

$$V_d = \frac{R_2}{R_1 + R_2} V_{\text{PP}} \tag{5.39 }$$

From Eq. (5.39) it is clear that V_d is only a function of the regulator resistors.

If Single supply devices are considered, any high voltage must be derived from charge pump circuits. Therefore it is clear that, while the regulation scheme of Fig. 5.85 can be easily adapted to this situation, that of Fig. 5.86 cannot, as the voltage drop across the current mirror reduces the available VPD for the decoder by at least one volt due to the voltage drop ΔV across the current mirror. Consequently, a typical configuration for a drain regulation through the program loads is that of Fig. 5.87, which makes use of an operational amplifier with a voltage divider, and sets the VPD voltage from a band-gap reference voltage V_{BG} according to the relationship:

$$\mathrm{VPD} = \frac{R_1 + R_2}{R_2} V_{\mathrm{BG}} \tag{5.40}$$

Consequently, the VPD voltage is actively stabilized by the operational stage, while the drain voltage, which differs from the latter for the column decoder voltage drop, is not.

5.5.4 Gate Voltage Regulation Fundamentals

Every Flash cell inside the memory array receives the gate voltage from the row decoder through a wordline; in Fig. 5.88 the connection of the row decoder to the row address bus, to the array wordlines and to a dedicated supply VPCX is shown. This circuitry allows to transfer the VPCX supply to the memory cell which has been selected by the row address.

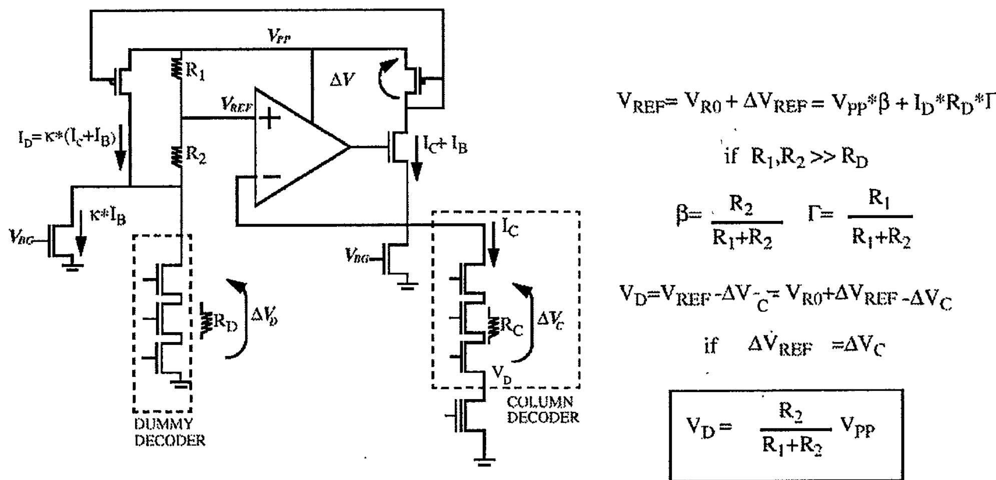


Figure 5.86 Regulation through the program loads.

APPLE INC.
EXHIBIT 1008 - PAGE 0329

Each memory operation (Read, Program, Verify, ...) requires a different VPCX voltage as shown in Tab. 5.1.

Table 5.1 Different VPCX values.

Operation	*VPCX level*
Read	V_{CC}
Program Pulse	V_{PP} (10÷12V)
Program Verify	V_{PV} (≈ 7V)
Erase Pulse	< V_{CC}
Erase Verify	V_{EV} (≈ 3V)

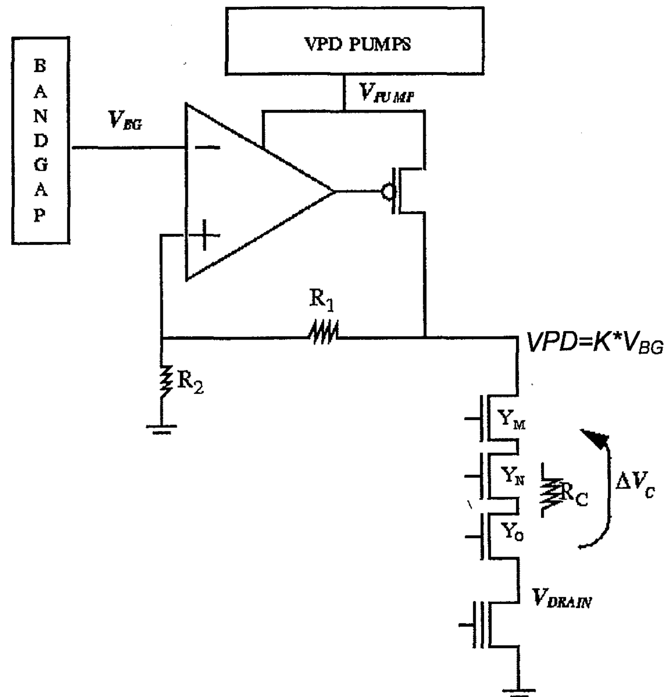


Figure 5.87 Regulation through the program loads for Single supply devices.

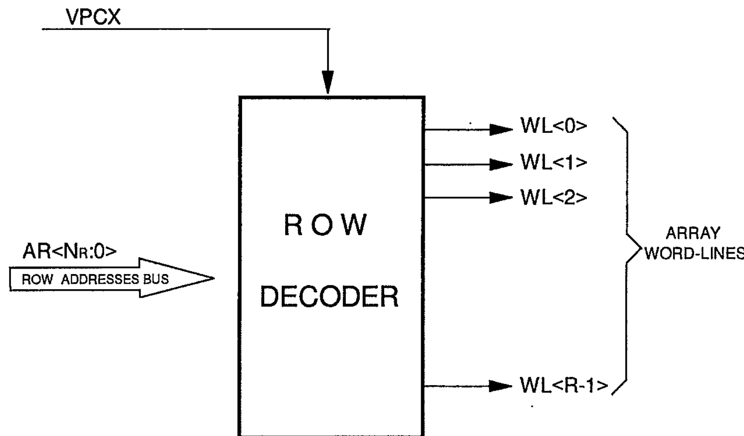


Figure 5.88 Row Decoder block diagram.

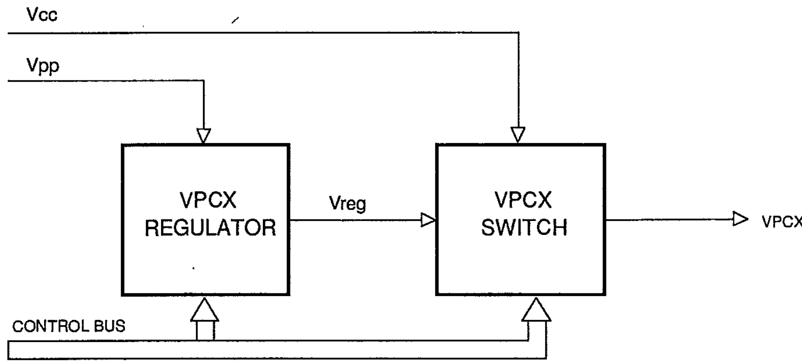


Figure 5.89 Gate voltage regulation: block diagram.

To generate all these voltage levels a circuitry according to the block diagram of Fig. 5.89 is used, which consists of a voltage regulator and a high voltage (HV) switch. The regulator provides all the VPCX levels that are different from V_{CC}, while the HV switch connects the row decoder supply to V_{CC} or V_{reg} (the output of the gate voltage regulator). The operation mode of both circuits is defined by a control bus, driven by the Program/Erase Controller.

A first simple regulator architecture, which is mainly used in the Dual supply Flash memories, is shown in Fig. 5.90. In this case a well-defined level voltage is the program voltage (V_{PP}) provided by the user on the V_{PP} pin itself. The device specifications usually require a V_{PP} level of 12 Volt with only a 5% of tolerance, so this supply can be used to derive a reference voltage by means of a simple divider. The various reference voltages are then obtained by a

APPLE INC.
EXHIBIT 1008 - PAGE 0331

set of taps (M2-M5 in Fig. 5.90) which are selected using the analog switches according to some bits of the control bus. The digital signals of the control bus are translated through some level shifters to allow the complete turn-on of the analog switches. A typical realization of these level shifters is shown in Fig. 5.91.

In this schematic M1, M2, M3 and M4 form a bistable circuit which is driven by the input digital signal IN and its complement generated by the inverter I1. If IN is at the logic level High, M3 and M2 are on, the output OUT is connected to V_{sup} and M1, M4 are off. Since there is not any DC path between V_{sup} and GND, the level shifter maintains this state without any current consumption. When IN goes Low, M3 and M2 are turned off and M1, M4 are switched on, connecting the output to GND. The power consumption, apart from the time interval required to accomplish the transition between the two steady states, is once again zero.

An operational amplifier in voltage-follower configuration is used to decouple the reference voltage generator from the high voltage switch and the row decoder. It is interesting to note that the regulator does not provide high current values since it is connected only to the memory cells gate through the row decoder. Furthermore, the internal algorithms are programmed in such a way that all the address transitions (which can produce a crow-bar current peak) take place when VPCX = V_{CC}.

Therefore the regulator must be designed in such a way that its output current guarantees the desired settling time of the VPCX voltage. As in any feedback circuit, the stability of the regulator is guaranteed by a proper compensation circuitry.

Conversely, in a Single voltage Flash memory only the V_{CC} supply is available, but normally its value is less precise than the V_{PP} one. For this reason an internal reference voltage generator (a band-gap reference, for instance) is added inside a typical Single voltage device. The new gate regulator architecture is shown in Fig. 5.92.

Since the band-gap output voltage is lower than the regulated one, the operational amplifier is now configured as a non-inverting amplifier with programmable gain. The supply named V_{PP} in Fig. 5.92 is greater than V_{CC} and is provided by a charge pump circuit, thus the design of the regulator should take great care to limit its current consumption. To reduce this contribution, the voltage divider can be realized using MOS capacitors.

At this point, the HV switch, i.e. the circuit (directly driven by the gate voltage regulator) which supplies the row decoder inside a Flash memory, can be considered; it is a circuit which is able to select two or more supply levels, even in the case of voltages greater than the conventional supply rail level (i.e. V_{CC}).

APPLE INC.
EXHIBIT 1008 - PAGE 0332

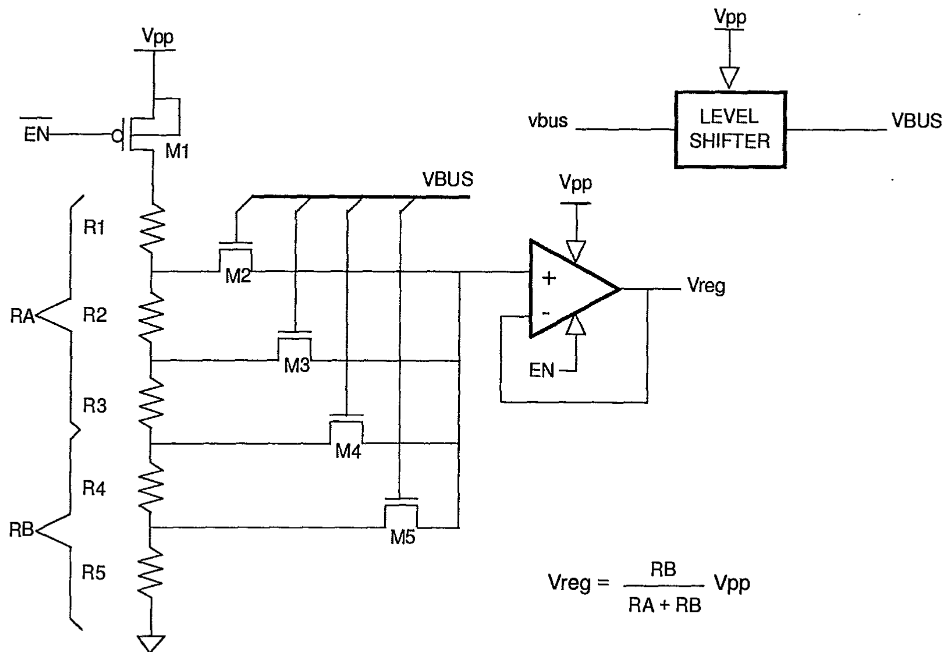


Figure 5.90 A first implementation of the gate voltage regulator.

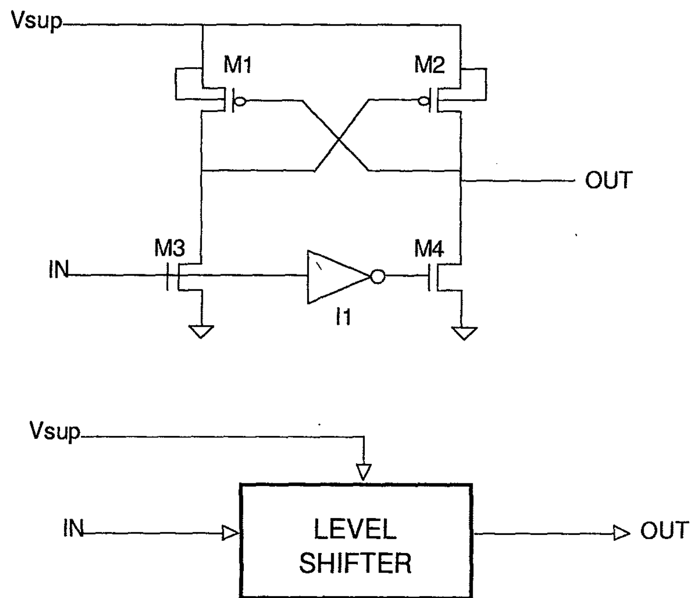


Figure 5.91 Detailed schematic diagram of the level shifter.

APPLE INC.
EXHIBIT 1008 - PAGE 0333

This circuit does not have any power consumption; this means that the input voltage levels have to be completely transferred to the output. In a CMOS process the simplest solution is based on a pMOS-based HV switch: the switching elements are pMOS, their gates being driven by two level shifters which are supplied by the two input voltages (see Fig. 5.93).

The two pMOS transistors are properly driven but, unfortunately, the drain junction of the pMOS connected to the lower supply is directly biased when the higher supply is transferred to the output.

Obviously, this is not admitted in a CMOS circuit and a proper countermeasure has to be adopted. The schematic in Fig. 5.94 provides a good solution for this problem. Each branch consists of two pMOS having opposite n-wells. If the gates are properly selected by signals having the right voltage swing, this switch can manage all the possible supply levels combinations ($V_{CC} > V_{reg}$, $V_{CC} < V_{reg}$, $V_{CC} = V_{reg}$) without any drawback.

If, for instance, V_{CC} is lower than V_{reg} and VPCX has to be connected to V_{CC}, the left branch is turned on by bringing to GND the two enable signals ENAVCCHV# and ENAVCC#; the node A is then at V_{CC} and all the junctions are correctly biased. The right branch is turned off with ENAVREGHV# =

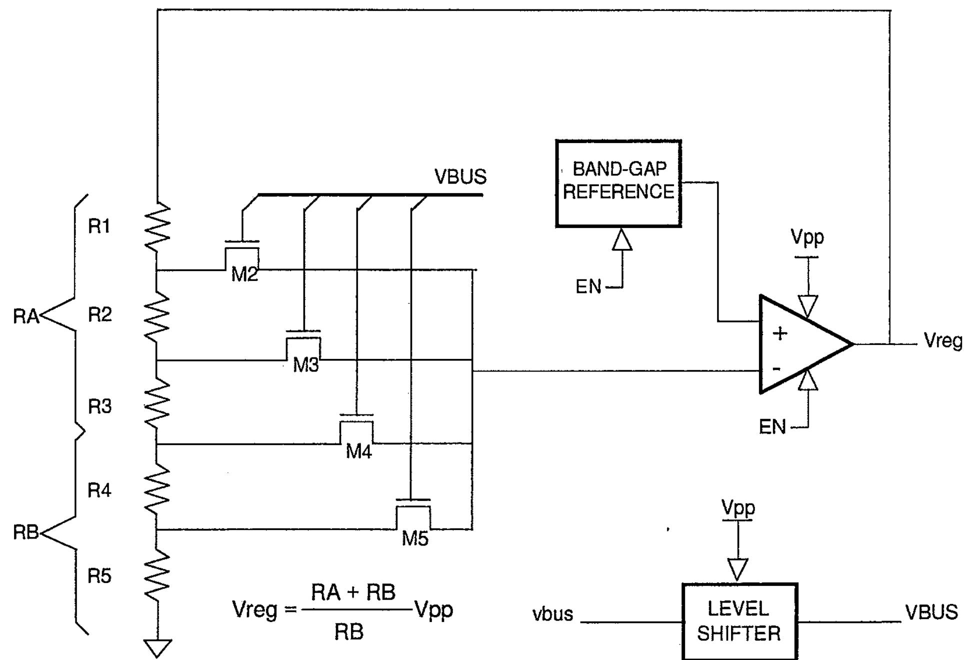


Figure 5.92 Gate voltage regulator with band-gap reference.

APPLE INC.
EXHIBIT 1008 - PAGE 0334

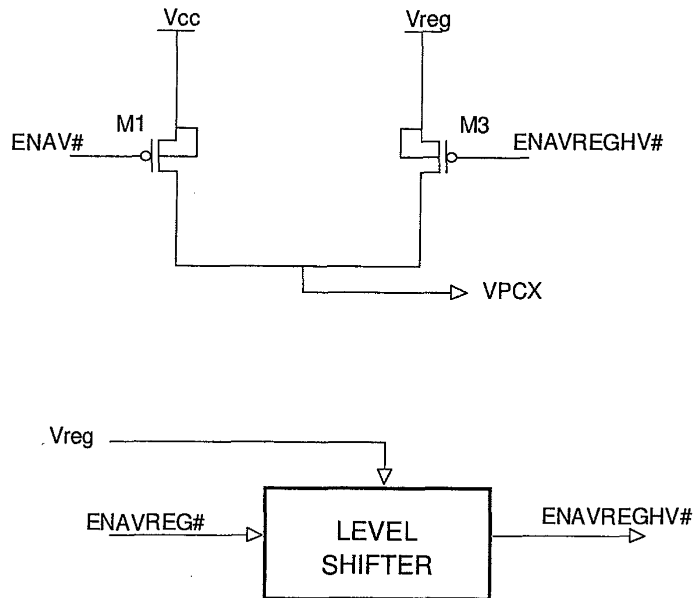


Figure 5.93 Basic HV switch.

V_{reg} and ENAVREG# = V_{CC}. M3 and M4 are both switched off, since they have $V_{\text{gs}} = 0$; this means that node B is floating and source-well and drain-well junctions of M3 and M4 are correctly biased.

If VPCX has to be set to V_{reg}, the right branch of the switch is turned on setting ENAVREGHV# = GND and ENAVREG# = GND, while the left branch is interrupted with ENAVCC# = V_{CC} and ENAVCCHV# = V_{reg}. Also in this case the source and drain junctions of all the transistors are well biased and node A is floating.

A slightly more complicated version of this HV switch is shown in Fig. 5.95. In this realization a double pair of nMOS transistors is used to keep the nodes A and B always biased. When V_{CC} is passed to VPCX, the node B has a voltage equal to $V_{\text{CC}} - V_{\text{TH}_n}$ (if V_{reg} is lower than V_{CC} and M7 is on) or equal to V_{CC} (if V_{reg} is greater than V_{CC} and M8 is on). When V_{reg} is passed to VPCX, the node A is at V_{CC} if $V_{\text{reg}} > V_{\text{CC}}$ or at $V_{\text{CC}} - V_{\text{TH}_n}$ if $V_{\text{reg}} < V_{\text{CC}}$. The two protection resistors are introduced to avoid the possibility of breakdown of the n+/psubstrate drain junctions of M5 and M7 during an Electrostatic Discharge (ESD) occurring on the supply pins.

Since the HV Switch provides two supplies, it is mandatory to prevent the two branches being simultaneously on; this critical condition, which can occur when the HV Switch changes its state, can be avoided with a proper signal

APPLE INC.
EXHIBIT 1008 - PAGE 0335

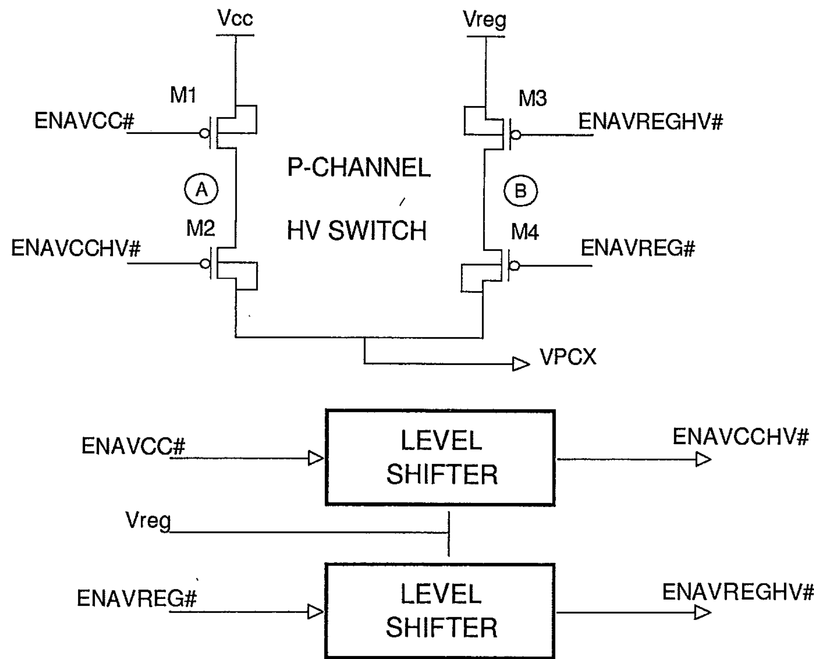


Figure 5.94 Enhanced HV switch.

timings, as shown in Fig. 5.96: one branch is turned off before turning on the other one by means of non-overlapping control signals.

5.6 ERASE OPERATION CIRCUITRY

Electrical Erase is a global operation for a Flash memory that is applied to entire sectors of the memory array. The methodology of this operation is the essential difference existing between EEPROM and Flash: the former can be erased cell by cell owing to the presence of the selection transistor, while in the latter (being high density its aim) a common source is shared by all the cells of the same array portion. The partition of the memory matrix in sectors of particular size allows to erase only certain portions leaving the others unaltered. For this reason every sector has its own internal source line and its own circuitry used to switch this line. In order to perform the electrical Erase effectively, it is mandatory to provide a high electric field between the source and the floating gate of the cell, causing the extraction of the negative charge from the floating gate by means of Fowler-Nordheim tunneling; this task is accomplished applying a voltage difference greater than 10V between source and control gate.

APPLE INC.
EXHIBIT 1008 - PAGE 0336

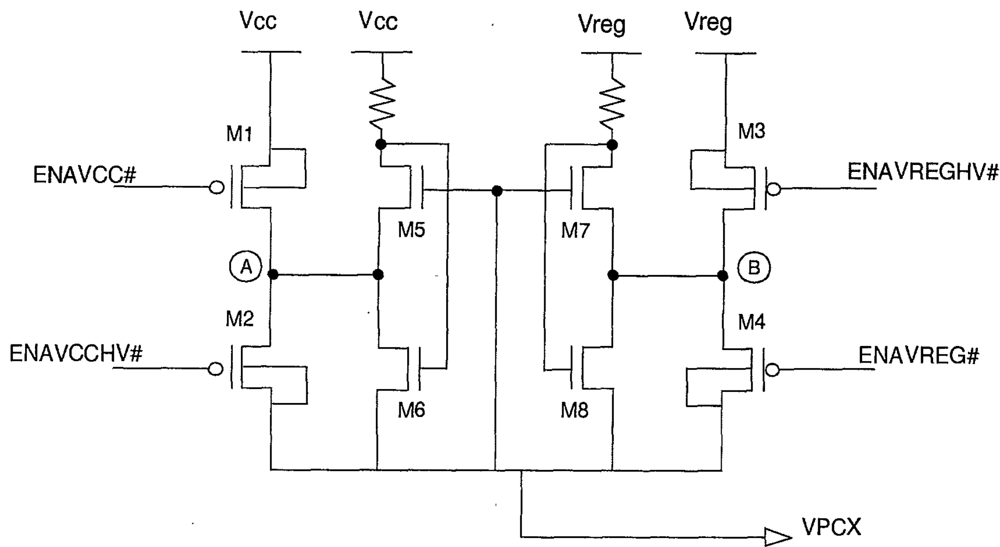


Figure 5.95 HV switch with anti-ESD and biasing nets.

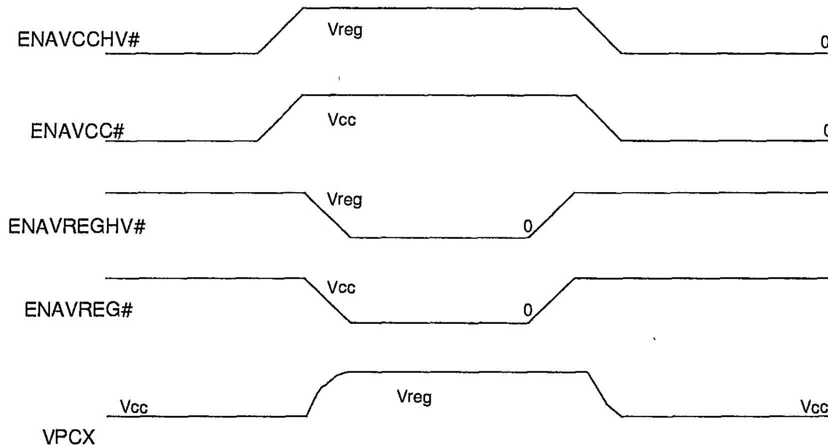


Figure 5.96 HV switch timing of the control signals.

Depending on the kind of device taken into consideration, either Double or Single supply (see Section 5.1), different circuit have to be designed; in the next paragraphs the criteria behind this choice will be analyzed, as well as the impact on the circuitry topology.

APPLE INC.
EXHIBIT 1008 - PAGE 0337

5.6.1 Double Supply Voltage Approach

If an externally-provided high voltage is available, it is immediate to connect the source line of the sector to be erased at V_{PP} and the gates of the corresponding cells at GND. This design choice leads to a reduction in the complexity of the high voltage internal lines and allows to use the highest voltage drop without any other internal manipulation (partition or regulation).

The erasing internal high voltage is provided through source line by means of pulses of about 10ms of length. At the end of each pulse the threshold voltage of the cells is verified, to stop the operation if all the cells are under a desired value. Further details about the Erase algorithm can be found in Section 5.7.4. Two undesired phenomena occur in the cell source-substrate junction because of its high reverse biasing: the band to band tunneling effect and the avalanche breakdown. The former is due to a great folding of the conduction and valence band at the periphery of the n+ source region; some electrons cross the gap between the two bands and are captured by the terminal at the highest potential, i.e. the source contact. The latter (the avalanche breakdown) can occur when a high reverse voltage drop is applied to the source-substrate junction. These two undesired effects have the common property of producing a current from the source line (from the V_{PP} external source) to GND which is steady throughout the whole erase pulse.

5.6.1.1 Source Erase Circuitry. The source line of the sector under Erase must be switched from the GND level to the high level for all the erase pulse length. This line must be clamped to GND during Read and Program operation. The circuit which accomplishes this task is commonly called "source switch". A principle scheme is shown in Fig. 5.97. During the reading operation the ERASE1_N, SLOWD and ERASE2_N signals are at V_{CC}, therefore transistors MN1, MN2 and MN3 are on; ERASE signal is Low, in order to keep the output of the HV switch at V_{PP} (a complete description of the behavior of this circuit can be found in Section 5.5.4), thus switching off MP1.

The source line shows a huge parasitic capacitance towards GND, labeled C_{sect} in Fig. 5.97. This capacitance is due to different contributions: the perimeter and area capacitance of each Flash cell source-substrate junction and the coupling capacitance of the source line towards each node at zero potential (e.g. the wordlines that connect the gate of the cells).

An estimated value for C_{sect} is about 2nF/Mbit, for the most recent geometry of Flash array. During the erase pulse, the undesired presence of the band to band and breakdown current can be modeled by introducing a reverse biased Zener diode (DZ in Fig. 5.97). In order to let the circuit self-limit the amount of current, a resistive element is put between the source line and V_{PP}. A simple

way to do this is to design MP1 with a resistive aspect ratio. The V_{ds} voltage drop of MP1 does not allow the source line to reach a full V_{PP}, and it limits the source current of an entire sector to $10 \div 20$mA (this value can grow for large sector size and high operating temperature).

5.6.1.2 Slow Discharge of Critical Nodes. Discharge of the source line must be performed by introducing some precautions. First of all, the ground noise due to the discharge of the C_{sect} parasitic capacitor is significant and it can introduce spurious commutations in the internal logic networks. To limit this phenomenon it is necessary to reduce the amount of discharge current in the source line. In practice a slow discharge is fundamental. In Fig. 5.97 the slow discharge transistor is MN2. The gate signal SLOWD (see Fig. 5.98) has a slow rise front to delay the activation of MN2 at the beginning of the discharge. In the meantime, MN1 and MN3 are switched off by lowering ERASE1_N and ERASE2_N.

The voltage level reached during the erase pulse by the source line is about 1 volt below V_{PP} (because of the voltage drop on MP1). To maintain the discharge current in the range of some mA, the discharge time (T_{disch} in Fig. 5.98) must lie between 10 and 50ms. Of course this time is a function of the size of the erased sector.

Another reason for a slow rise of SLOWD is to avoid a snapback effect for MN2. This undesired condition, as shown in Fig. 5.81, takes place when high drain and gate voltage are applied; a way to avoid snapback primer window for NM2 is to decrease its drain voltage (the source line) while its gate voltage (SLOWD) is rising. In addition, some design rules must be observed to reduce

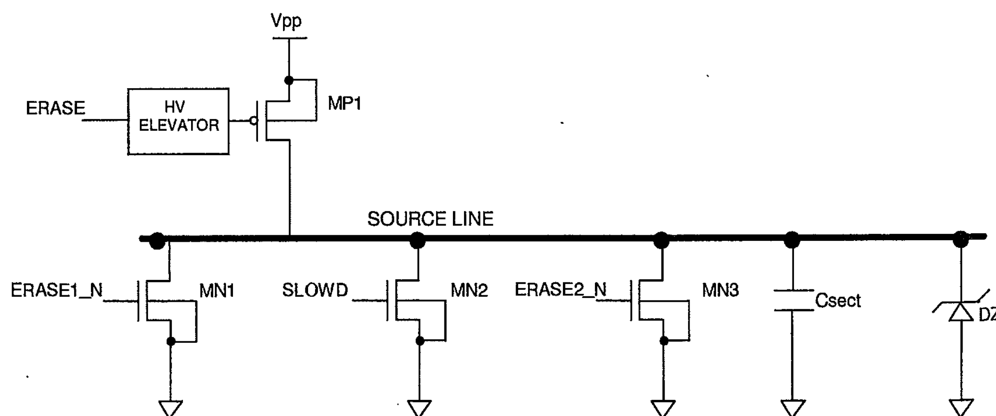


Figure 5.97 Source line elements.

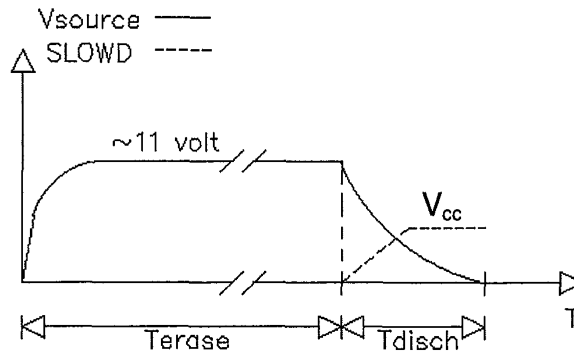


Figure 5.98 Source discharge timing.

the snapback effect for this transistor, as ensuring a non minimum channel length.

The capacitive coupling between wordline and source line is another effect which should be taken into account when discharge is considered. During the erase pulse, the coupling capacitance C_{ws} (see Fig. 5.99) is charged at the source line erase voltage; when this line is discharged, the other terminals, i.e. the wordlines, are therefore driven at negative voltage. This effect causes the forward biasing of drain-substrate junction of the wordline pull-down drivers. Once again, a slow discharge is suitable to reduce the forward current in that junction.

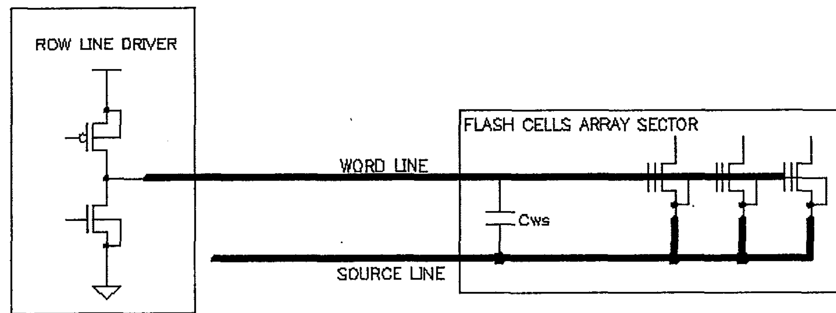


Figure 5.99 Coupling capacitance presence in Flash cells array.

5.6.2 Single Supply Voltage Approach

In Single supply devices, both the voltages higher than V_{CC} and those lower than GND (required for the programming and erasing of the Flash cell) must

be generated internally. This is accomplished by means of the so-called "charge pumps".

One of the drawbacks of this approach is that charge pumps have little driving current capability; this entails that the current sunk from them must be limited as much as possible. Because of this, the Erase operation is performed with a negative voltage (≈ -8V) on the gate, which allows for a lower voltage on the source (≈ 5V). Note that the electric field applied between the control gate and the source of the cells is the same as in the Double supply voltage approach, in which the absolute values of the biasing of these terminals are different, but their difference is the same in both cases. This implies a much lower source current, because the component due to the source junction breakdown is not longer present (only the band-to-band tunneling current still exists).

5.6.2.1 Charge Pumping. The principle of charge pumping is based on the storing of charge with the proper sign at the desired node. Two kinds of pumps exist:

positive pump, that generates voltages higher than the power supply, by storing positive charge in the desired node;

negative pump, that generates voltages lower than GND, by removing positive charge from the desired node.

In both positive and negative pumps, the element used to transfer charge from one node to another is the capacitor, because this component has the property of keeping the voltage drop across its plates unchanged as far as small time intervals are considered.

The basic scheme of a positive charge pump is shown in Fig. 5.100: it includes several serially-connected stages, each consisting of a capacitor and a switch. Two phases are needed, "a" and "b" in the figure, pulsing in phase opposition between GND and the V_{CC}, driving the capacitors as shown. When the phase "a" is High and "b" is Low (T1; Fig. 5.101), the switches S2 and S4 are kept closed, while the others are open. Positive charge is transferred from stage 1 to stage 2, and from stage 3 to stage 4. When the phase "a" is Low and "b" is High (T2; Fig. 5.102), positive charge is transferred from V_{CC} to stage 1, from stage 2 to stage 3 and from stage 4 to the output. Positive charge is accumulated at the output node, so that its voltage increases. During a charge transfer, the voltages of stage i and stage $i+1$ equalize. If the capacitors have the same value, the final voltage after the equalization will be the average value of the two initial voltages.

With no current delivered at the output, after an initial transition phase, the voltage gained at each stage will be equal to V_{CC}. Obviously, with the intuitive changes, the same concepts apply for the negative pump.

APPLE INC.
EXHIBIT 1008 - PAGE 0341

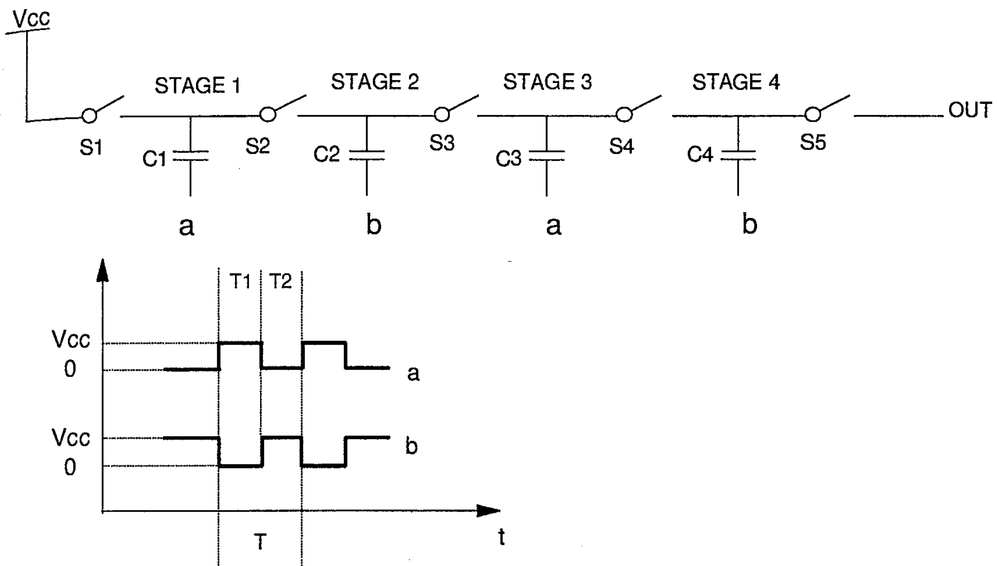


Figure 5.100 Positive charge pump: basic scheme.

The fundamental analytical relations concerning the charge pumps are the following:

- Capacitors driving voltage $\in (0, V_{CC})$;
- Voltage gained through n stages:
  $V_{\max,\text{pos}} = (n+1)V_{CC}$ (positive pump: V_{CC} is the initial voltage),
  $|V_{\max,\text{neg}}| = -nV_{CC}$ (negative pump: GND is the initial voltage);
- Charge transferred from one stage to another at every period of the clock:
  $Q = C\Delta V \Longrightarrow Q_{\max} = CV_{CC}$;
- Current supplied to the output (T = period):
  $I = Q/T \Longrightarrow I_{\max} = CV_{CC}/T$;
- Output resistance:
  $R_{\text{pos}} = V_{\max,\text{pos}}/I_{\max} = (n+1)T/C$
  $R_{\text{neg}} = |V_{\max,\text{neg}}|/I_{\max} = nT/C$

The main design issue concerning any charge pump is the way the switches are realized. Of course, the switch should be either fully open or completely closed, which is not trivial at all. The main problem is that MOSFETs suffer from body effect, and different solutions have been found, depending on the technology available.

APPLE INC.
EXHIBIT 1008 - PAGE 0342

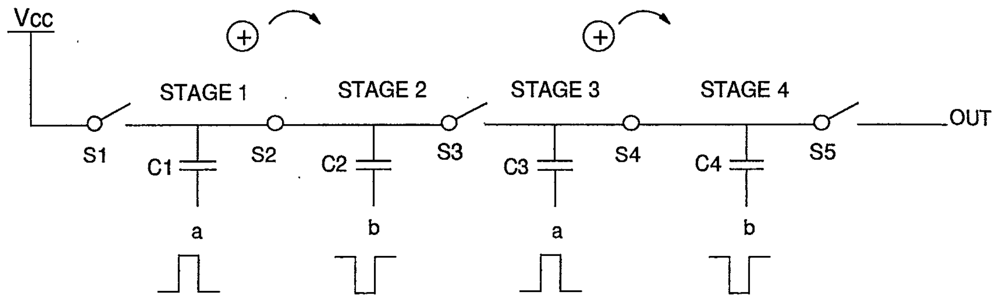


Figure 5.101 Positive charge pump: during T1.

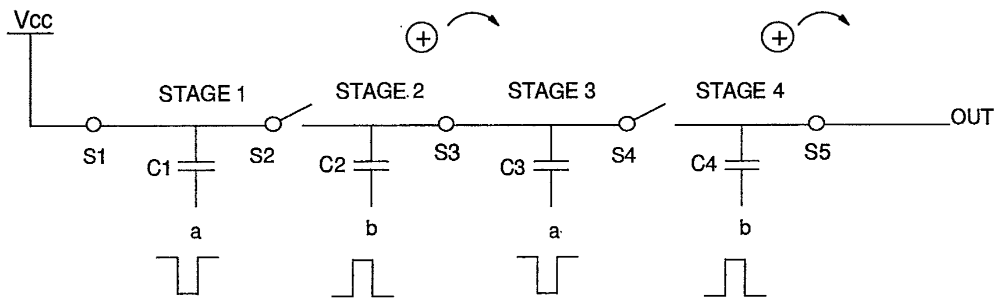


Figure 5.102 Positive charge pump: during T2.

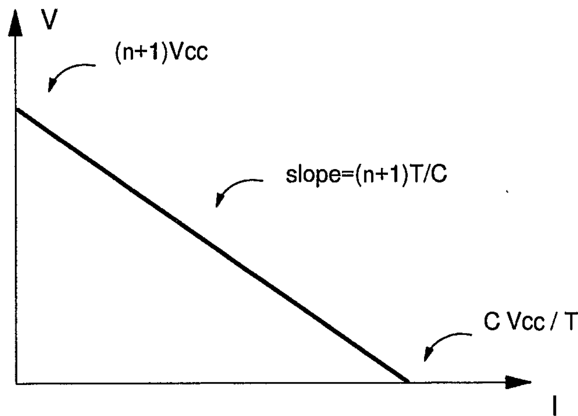


Figure 5.103 I/V characteristic.

Another concern regarding charge pumps is the area they require. As shown in Fig. 5.103, the higher the current required at the output, the lower the reachable output voltage; both voltage and current depend on area, though

in different ways: once V_{CC} is given, the output voltage only depends (substantially) on the number of capacitors (i.e. stages) n (and more stages means greater area), while the current is related to the dimension of the capacitors used. This brings to the use of many capacitors with large area, and this in turns reflects negatively on the overall layout area.

Last, but not least, the issue of the power consumption must be faced. The inverters driving the boosting capacitors consume a lot of power and this turns back negatively both on the total power consumption and on the noise generated internally on the GND and V_{CC} nodes.

Because of the limited capability of the pumps in providing current, all the circuits connected to them must absolutely satisfy two requirements:

1. they should sink as little current as possible (for instance, no crow-bar current is allowed);

2. they should introduce parasitic capacitances as little as possible, in order to avoid the slow down of the transient time in which the node is driven from its initial voltage to the desired one.

5.6.2.2 Voltage Regulators. The purpose of the voltage regulators is to limit and regulate the voltage supplied by the charge pump. Usually these two goals are realized by two different circuits: a limiter and a fine regulator (see Fig. 5.104), both exploiting the feedback concept.

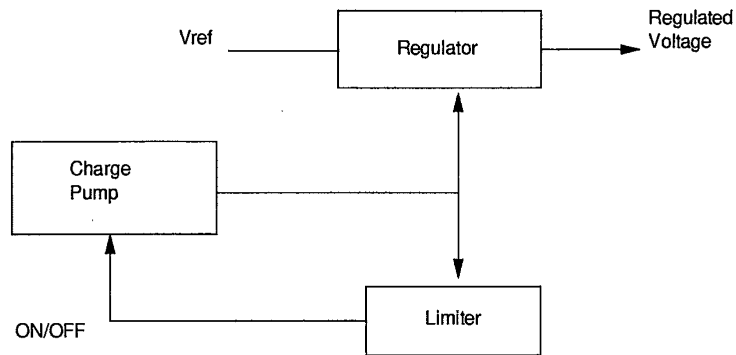


Figure 5.104 Regulator and limiter.

The limiter samples the output of the charge pump and, if this becomes too high (in module), disables the pump, by turning off its phases (clock). This prevents damages caused by too high voltages. The regulator, sometimes supplied by the output of the charge pump, compares the value of a reference voltage – which must be as precise as possible – with the one to be regulated. Usually,

APPLE INC.
EXHIBIT 1008 - PAGE 0344

the output stage of a regulator consists of a pull-up and a pull-down. If the regulated voltage is higher than the desired one, then the pull-down is turned on and the pull-up is turned off; the opposite happens if the regulated voltage is lower. A typical negative voltage regulator is shown in Fig. 5.105. HVNEG is the voltage to be regulated; it is sampled by means of a capacitive voltage divider and compared with GND. The use of a capacitive voltage divider, instead of a more usual resistive one, avoids current consumption from the pumped node, as previously justified. Of course, the values of the capacitors, together with the V_{ref} value, determine the regulated voltage.

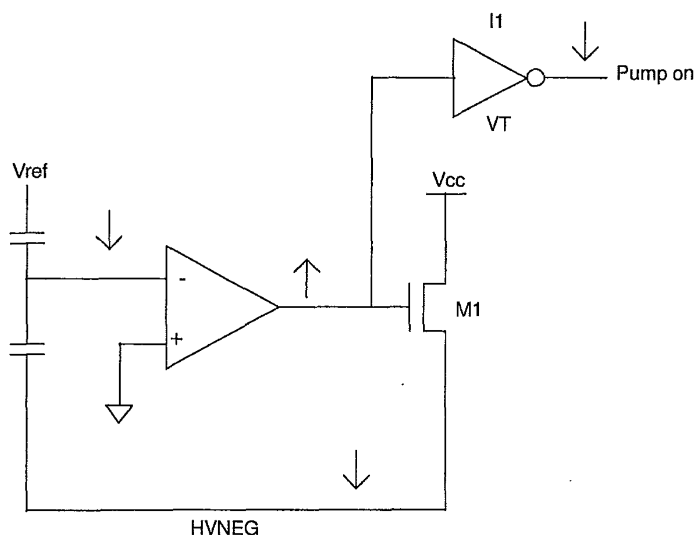


Figure 5.105 Negative regulator.

If the negative voltage goes too low (high in absolute value) the operational amplifier turns on the nMOS, which provides current to the regulated node HVNEG, thus rising its value. If this is not sufficient (that is if the output of the operational amplifier goes high, beyond a certain limit VT) the limiter turns off the pump. Alternatively, M1 may be omitted, renouncing to a precise regulation.

5.6.2.3 Source Switch. For the Erase operation, the negative voltage must be applied through to the wordlines (cells' gate). Basically, two approaches are used:

1. wordline diodes (no triple-well technology),

2. normal row decoder (triple-well technology).

APPLE INC.
EXHIBIT 1008 - PAGE 0345

Section 5.2.2 has already dealt with this issue: if the latter approach can be chosen, then it is clear that the source of that pull-down must be switched between GND and a negative voltage, and this requires a specific (small) circuitry, not present in the former implementation.

With the exception of the different voltage applied to the source, no particular differences occur between the Single and the Dual voltage approach. The main difference consists in the fact that for the former, being the source voltage and current provided by means of pumps (thus with limited current capability), no elements are necessary to limit the current sunk by the cells.

Of course, the considerations analyzed in Section 5.6.1 about the slow discharge of critical nodes still hold. However, in the case of Single voltage devices, the fact that the Erase operation is performed with a negative voltage on the gate "helps" the line discharge. The parasitic capacitance, introduced by the gate of the cells of the sector being erased, is of the same order of magnitude of the parasitic capacitance of the source of the same cells (about 1nF). These must be both discharged to GND after the Erase operation, and both are huge. Nevertheless, discharging the source – positively charged – implies a positive current towards GND, while discharging the gate – negatively charged – entails a negative current towards GND. These two currents compensate each other, so that the discharge operation may be faster, as the GND bouncing is much more limited. Nonetheless, a certain control must be accomplished on this discharge: the two currents must flow contemporarily and – more or less – with the same amount, so needing a constant current discharging, and none of them should be too large, in order to avoid electromigration and the Joule effect. It is known that an Erase operation, performed by imposing a constant current (instead of a constant voltage as usual) to the source, allows a constant Erase time, independently of the geometrical differences of the cells, and even of the number of cycles the cells have been subjected to. In this case, the structure of the source switch becomes a bit more complicated, with respect to the one depicted in Fig. 5.97, because a current generator must be introduced (Fig. 5.106). This is accomplished by transistors M1 to M4, which realize two current mirrors, mirroring a suitable ratio of I_{ref}. The voltage of the sector source is imposed by the sources of the cells.

5.7 CONTROL LOGIC AND EMBEDDED ALGORITHMS

Among other ways, the different operation modes of a Flash memory could be classified according to their duration: this parameter is highly variable, being about 100ns for Read, 10μs for Program and several seconds for Erase. In most of the applications, however, the microcontroller is not dedicated to the Flash only, but it has to interact with several other devices; therefore it is not possible

APPLE INC.
EXHIBIT 1008 - PAGE 0346

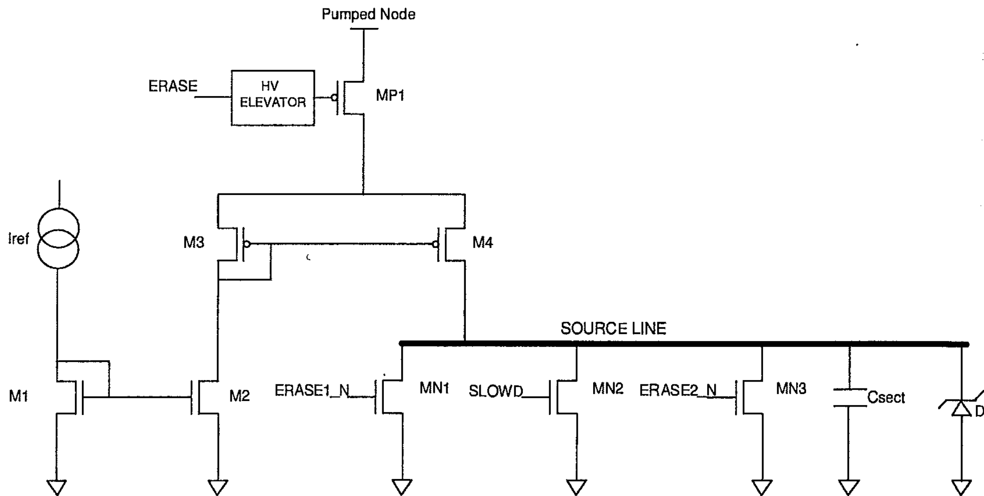


Figure 5.106 Constant current source switch.

to keep it busy in a single (and complex) operation (e.g. handling a complete Erase cycle), because it would mean introducing 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 (i.e. Program and Erase) should be designed embedded entirely inside the Flash memory, thus delivering the microcontroller from the handling of such a time-consuming task. Another advantage of this approach is the simplicity of the interface: all the timings required to perform the operation correctly (setting of different voltages, counting of the program or the erase pulse, verify and so on) are transparent to the user, since it is sufficient to provide the operation code (and the data to be programmed or the sector to be erased) and check, when the microcontroller is not busy in other tasks, the current status of the memory.

In designing Flash memories, this approach has not been followed since the very beginning: by making a comparison between the percentage of embedded logic (i.e. excluding the conventional one, such as row and column decoders, output multiplexers and so on) in different devices, provided that a typical first generation Flash sets the reference value to 0%, it is possible to see the constant increment of this parameter: from 15% for the second generation of Flash memory up to the 40% for a state-of-the-art present device. This growth is strictly related to the implementation of new solutions inside the design: for instance, a Flash memory in which high voltages are provided externally (Dual

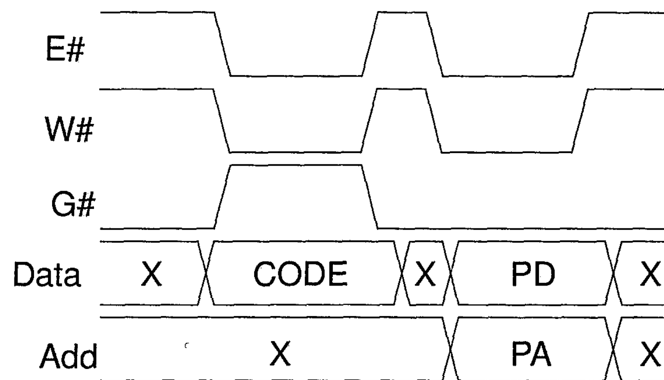
voltage device) requires fewer controls than another one in which these voltages are generated internally (Single voltage device), because of the complex timings to be handled inside the high voltage generation circuitry. In this latter case the advantage of having an embedded algorithm is therefore more evident.

5.7.1 Logic Architecture

As already pointed out in Section 5.1.3, a Flash memory can be controlled by means of its pins: analyzing the device pinout, they can be grouped into three categories (apart from supply and GND pins): address, data and control. Typically the control signals are used to activate the device and to obtain some simple operations like "Read Electronic Signature", "Output Disable" and so on. When used in conjunction with some data pins (usually the 8 least significant), they allow the microprocessor or the microcontroller (since a Flash memory can work either with a microprocessor or a microcontroller, in the following these two devices will be cited indifferently) to write a command into the user interface so that the Flash memory can perform a complex operation. Several protocols exist to write a command and Fig. 5.107 gives an example. In this figure both data and addresses are latched on the rising edge of the control signal W# (Write Enable). Anyway, the latching phase could be controlled by E# (Chip Enable) or by any combination of the two signals or in other ways. When latching occurs, the data and/or the address represents a particular command; a group of commands decoded by the user interface defines an instruction.

The two main instructions are "Program a byte (or a word)" and "Erase a sector". Other instructions exist, for example "Read Electronic Signature", "Read/Clear Status Register", ..., but the most complex operations are those related to either Program or Erase. Therefore, when reading a specification for a Flash memory, instructions like the following can be found: Program byte (word) to let the device perform a Program of 8 (16) bits at any memory location; Chip Erase, to let the device execute (sequentially or simultaneously) the Erase of all the sectors; Sector(s) Erase, to let the device execute the Erase of a (list of) sector(s); Sector Erase Suspend/Resume: since the Erase is a very long operation (compared to the microprocessor execution time) which can be performed on a subset of sectors, there is the possibility to suspend this action to either Read or Program into a sector not to be erased. As it was said before, all the complex operations are executed by the Flash with no help from the external controller; moreover, the memory is able to inform the microcontroller about the status and the correctness of each operation. The status register, or something similar, has the task to record information during the execution of the embedded algorithm so that the microcontroller can read them. It's clear

APPLE INC.
EXHIBIT 1008 - PAGE 0348



CODE: program command code
PD: data to program
PA: address where program

Figure 5.107 Example of Command Write.

that the control logic in a Flash memory has a very defined task to perform, so that the logic circuitry must be really well timed and the analog and digital parts must be very well bound together.

Fig. 5.108 shows a very simple block diagram of a Flash memory architecture; the simplicity lies in the fact that not all the connections are drawn and that only the main blocks are represented, but this scratch is sufficient to see all the fundamental circuits inside a Flash memory. The concepts which drive the design are also presented; starting from this point any development is allowed, but the basis are always the same. For example, Program operation was and will be mentioned, but no words are spent about its real implementation. A last comment about Fig. 5.108 is that the rounded boxes are digital and all the other could be not fully analog; in fact some circuits must have their main properties as analogic, but they can also implement a logic function or their output can drive other logic parts.

In Fig. 5.108, the already mentioned Command Interpreter is shown; its task is to latch the code and to understand if the command sequence provided is correct; if this holds, a group of signals are asserted so that the device is prepared to execute the instruction decoded. This block could be very simple or very complex because of the protocol chosen to write the command and to answer to the microcontroller; generally the choice is a trade-off between code safety and time saving. Furthermore the Command Interpreter itself could manage the simple operations required by some instructions or it could be only a code translator for other active circuits; for sure, in case of either Program or

APPLE INC.
EXHIBIT 1008 - PAGE 0349

Erase, the control is given to the Program/Erase Controller already mentioned in Section 5.1.

Before analyzing its structure and behavior, it is better to consider the other circuits involved in control logic. First of all the address input buffers and the data input/output buffers. The former are connected to the Command Interpreter (if the instruction protocol requires that the commands are built using also the address bits), and to the row and column predecoders to select a memory location in the matrix. The latter are used to input the code of a command (usually only the 8 least significant bits are used) and to write the data to be programmed in case of Program operation; they are used to output the electronic signature, the status register and the content of an addressed memory location. To perform their task, these buffers are connected to the Command Interpreter, the Program/Erase Controller, the sense amplifiers and the program loads.

In the following, the main operations performed inside a Flash memory are considered (namely "Read", "Program" and "Erase"), wandering through their "paths".

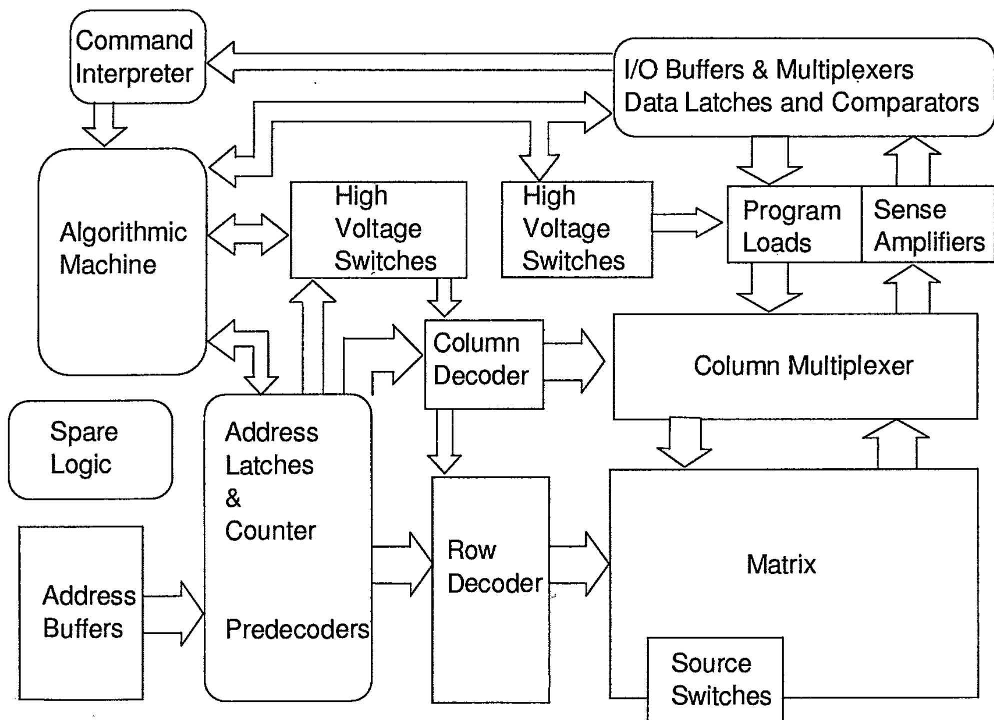


Figure 5.108 Block Diagram.

The Read path includes the address input buffers, row and column predecoders, row decoder, column multiplexer, sense amplifiers and data output buffers. The tasks of the input buffer are to set a filter between external and internal signals and to restore the CMOS levels in case of TTL input signals. The data output buffers indicates a memory location which is recognized as the cross between a row and 8 (or 16) columns. There are row and column decoders, typically built on two levels, which select the row and the columns. After that a memory location has been identified, the sense amplifiers read the cells' current and give the result to the data output buffers which work as an interface between the internal circuitry and the external data bus. "Read" is the default operation for a Flash memory (and for any memory device) and it does not need control logic; this means that once the address of a memory location is given, the entire operation is self-timed (actually, a block is missing: the Address Transition Detector, which activates the the read procedure when there is a transition on at least one address bit) and the only control is done by G# (Output Enable), which enables the data output buffers. The circuits mentioned are built also respecting the requirements for the other operations; in fact they are not completely dedicated to Read, but they also allow to Program and to Erase in the meaning explained below.

The two Write operations (both Program and Erase are considered Write operations, since they alter the memory content) start in same way by issuing the proper code sequence to build the desired instruction. As it was already said, this phase is managed by the Command Interpreter; its inputs are provided by the control signal buffers and by the data input buffers to build each instruction; other inputs are the feedback from the internal circuitry. The output signals command the latching of the address from the address input buffers, and of the data from the data input buffers, only in Write. After that an instruction is decoded, the control of the whole device passes to the Program/Erase Controller (which is an Algorithmic State Machine) until the end of the Write (Program or Erase) operation. There are several specifications on the behavior of the Command Interpreter during this phase, but the concept is that it cannot stop a Write operation until the Program/Erase Controller gives its acknowledge.

Actually, it is an over-simplification to reduce the Program/Erase Controller to an Algorithmic State Machine, since it also comprises: an oscillator, a time counter, a cycle counter, an end-count logic, and a status register. The Oscillator provides the clock to the Program/Erase controller; in some implementations there is a phase generator which provides non-overlapping phases. The Counters and the end-count logic guarantee the time steps needed to the algorithm (e.g. for wait states and pulse widths) and the number of attempts allowed in Program and in Erase, respectively. The status register informs the

microprocessor about the condition of the device during a Write operation and about the result at the end.

If a Program instruction is provided to the device, the blocks involved, in addition to decoders and array, are: the high voltage switches, the high voltage generators together with the program loads, (to provide the voltages to the gate and the drain of the selected cells); the sense amplifiers and the data comparator, to verify the programmed data. If an Erase instruction is fetch, the blocks involved, in addition to decoders and array, are: the source switches, to provide the proper voltage to the addressed sector(s); the address counter, the sense amplifiers and the data comparators, to verify the erased sector(s). The address counter scans byte by byte (or word by word) the whole addressed sector(s) to verify the content after the erase pulse. All the other circuits mentioned are typically analog (see Section 5.5), but they need a logic control, no matter how they are built. In fact their outputs must be properly timed or selected so that the array receives the correct voltages at the stated time or their outputs are used as feedback to the Program/Erase Controller. An example could be the output of the sense amplifiers after a Program Verify: they are compared with the data latched and the result is read by the Program/Erase Controller which decides if another program pulse is needed. The logic control in this case has to enable the sense amplifiers whenever they have to read and compare. The last logic part not yet mentioned are the Input/Output multiplexers. The Flash memory can be used in conjunction with devices which communicate either by 8 or 16 bits, so that either a byte or a word organization is required. There is the possibility to switch dynamically between the two organizations using a control signal (usually called BYTE#). When the device is used by word, all the 16 data I/O buffers are activated and the multiplexers are transparent, so that they are directly connected to the 16 internal data lines. When the byte organization is selected, only the 8 low order data I/O buffers are used and the multiplexer connects these data to the internal low or high byte depending on a further address pin

5.7.2 Embedded Algorithms

As to the hardware implementation of an algorithm, different methodologies can be followed, in order to obtain a control unit able to perform the correct sequence required by the algorithm itself. The main possible approaches are the following:

- Sequencer (pseudo-Microcontroller);
- Finite State Machine.

APPLE INC.
EXHIBIT 1008 - PAGE 0352

5.7.2.1 Sequencer (Pseudo-Microcontroller). Since the external microcontroller cannot be devoted entirely to the Flash memory, the simplest solution is to have a similar device inside the Flash. The embedded logic consists of two separate parts:

1. a ROM, in which the control program is stored as a sequence of instructions (whose set is usually smaller and simpler than the one of a conventional microcontroller), comprising decoding and sensing;

2. an instruction decoder, whose task is to fetch an instruction after another, decode it and drive various control signals accordingly (e.g. to set a particular voltage, to set the width of a pulse and so on).

Should the control flow be modified, it is sufficient to alter the content of the ROM (modifying few layers in the process flow); on the other hand the behavior (and the number) of the instructions is strictly fixed and cannot be modified unless the decoder is entirely re-designed.

5.7.2.2 Finite State Machine. In this case the control program and the decoder are not separated: the State Machine flow is performed step by step and in every different state a proper set of control signals is driven according to the actual status of the sequence. The degree of flexibility depends on the hardware implementation of the Finite State Machine:

- Registers + combinational logic: even the smallest modification implies a complete re-layout of the block. The advantage is that such a solution can be implemented using an automatic synthesis tool.

- Registers + PLA (Programmable Logic Array): this solution requires a more complex handling of the clock (since the PLA must be realized in a precharge and evaluation fashion), but the contents of the PLA (i.e. the flow) can be easily modified changing few layers in the process flow. The drawback is that the layout of this block is "monolithic" and it is difficult to be fitted inside an automatically-placed environment.

The flows hereby described refer to a generic single Supply Flash memory; the same concepts apply in the case of Dual voltage devices, apart from the issues related to the necessity of charge pumping.

5.7.3 Program Flow

Upon successful decoding of a Program instruction, the device has stored both the data to be programmed and the destination address. The Program algorithm works as follows (see Fig. 5.109):

APPLE INC.
EXHIBIT 1008 - PAGE 0353

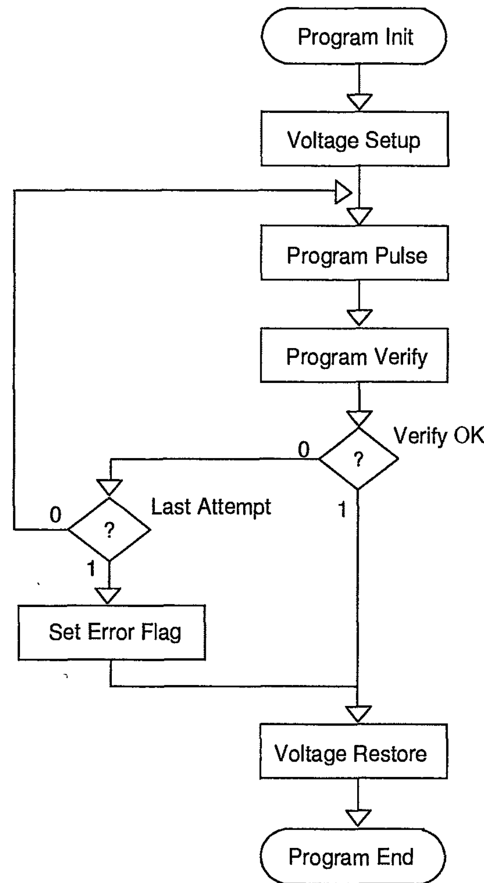


Figure 5.109 Program Flow Diagram.

Program Init: start-up procedures (e.g. counter reset) are performed.

Voltage Setup: charge pumps are enabled in order to raise the voltage from V_{CC} to program voltage.

Program Pulse: the program voltages are provided to the selected cells for the desired amount of time (typically some microseconds). At the end of the pulse, the charge pumps get ready to switch to the voltage needed in the Verify phase.

Program Verify: this procedure is a Read operation at the selected address performed biasing the gate of the Flash cells at a voltage much higher than the one usually provided in normal Read. If the data read are equal

to those to be programmed, it means (i) that the threshold voltage of the Flash cells has risen from the virgin (or erased) value to the programmed one and (ii) that, since this "Read" has given a correct result with a higher voltage, the threshold gap is enough to be detected even in a conventional Read operation. In this case the algorithm has completed successfully: the charge pumps are switched off and the control is passed back to the Command Interpreter. If Verify fails, either there are other attempts available (whose number depends on the value stored in a specific counter) or the Program algorithm ends (again switching off charge pumps) producing a failure signal which can be detected by the user.

5.7.4 Erase Flow

For sake of simplicity, the Erase of a single sector is presented: the same considerations and algorithm apply if multiple sector or chip Erase is performed. As shown in Fig. 5.110, the flow is composed of the following steps:

Erase Init: start-up procedures (e.g. counter reset) are performed.

Program All 0: during this step all the cells in the sector with low threshold (logic "1") are programmed (i.e. they change their state to "0") so that their characteristics are as similar as possible before the erase pulse. Provided that it is applied on every cell of the sector, the algorithm is exactly the one used to write a single word (see Section 5.7.3).

Erase Pulse: the whole sector is biased with the erase voltages for the proper time (some milliseconds): during this period, cells are erased by means of Fowler-Nordheim Tunneling effect.

Erase Verify: this procedure is a Read operation at the selected address performed biasing the gate of the Flash cells at a voltage much lower than the one usually provided in normal Read. If the data read are "1" (the state of the erased cell), it means (i) that the threshold voltage of the Flash cells is set to the erased value and (ii) that, since this "Read" has given a correct result with a lower voltage, the threshold gap is enough to be detected even in a conventional Read operation. If Verify fails, either there are other attempts available (whose number depends on the value stored in a specific counter) or the Erase algorithm ends (again switching off charge pumps) producing a failure signal which can be detected by the user. If Verify has given the desired result on every cell of the sector, then it is possible to perform the following step.

Depletion Verify and Soft Program: theoretically speaking, after the erase pulse all the cells in the sector should share the same characteristic; un-

APPLE INC.
EXHIBIT 1008 - PAGE 0355

fortunately, there will be a spread in the distribution because of both the non-ideality of the matrix and the different initial threshold value of the cells. The already mentioned Erase Verify is used to re-iterate the Erase Pulse until all the cells are properly erased: the problem is that if a cell were already well erased, a further pulse could shift its threshold below zero, so that it would draw current even if not biased (deplete cell). By means of Depletion Verify it is possible to detect the presence of depleted cells and slightly re-program them, in order to raise their threshold above zero. If the operation is performed within the assigned number of attempts, then Erase has been successfully carried out: otherwise, a fail signal is raised to inform the user.

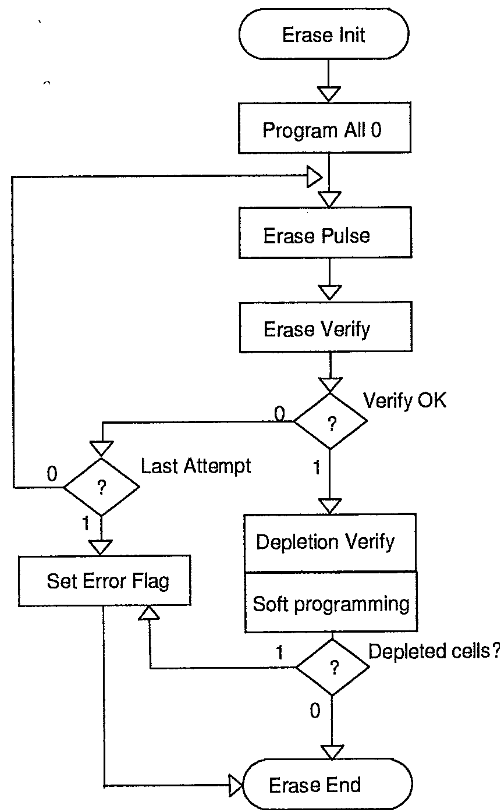


Figure 5.110 Erase Flow Diagram.

APPLE INC.
EXHIBIT 1008 - PAGE 0356

5.7.5 Erase Suspend - Erase Resume

As said above, a typical Erase cycle can last some seconds. There are cases in which the memory itself must be accessed in the meanwhile, thus interrupting the algorithm. In order to perform this task, an Erase Suspend command should be issued by the microcontroller: during the suspension it will be allowed either to Read from or to Program to the device. The main issue that should be taken into consideration when implementing such a feature is the necessity of saving the present status of the device, in order to know exactly where the flow has been interrupted. Therefore an adequate system of storage of the main parameters should be designed; furthermore it is essential to reduce power consumption to the standard Read value, so that all the internal pumps should be switched off, thus lowering the internal voltages to V_{CC}. Finally the device is available for Read or Program.

Once the "interrupt" has been resolved, the Erase operation is allowed to restart. First of all, it is necessary to restore the prior status, so that the algorithm knows exactly what has already been done. Afterwards, it is possible to let the charge pumps restore the voltages peculiar to the suspended phase, and the algorithm is resumed and executed until either it ends or another Erase Suspend command is issued.

5.7.6 Testability Issues

Early in the process life the yield can be poor, but the yield is high in production when the process is fixed; at that moment the device functionality becomes important. These few words point out the main tasks of a testing flow for an integrated device. First of all, by testing the devices the process problems must be discovered and solved to speed up the development and to reduce the time to market. Secondly all the features and the parameters written in the specifications must be tested to ensure the complete functionality over the conditions and for the years specified. Finally future developments can be discovered by a proper testing on actual devices. Focusing the issues on the Flash memories, it is not so easy to fit a generic testing flow to this kind of devices because there are several complex features asked to a Flash memory by both the suppliers and the users. Here the main needs are pointed out: a device must have an area as small as possible, because of the cost the silicon; the power consumption must be the lowest, to allow the use of low voltage batteries; the access time must be the fastest, to interface high speed microprocessors; the testing time, in production, must be the shortest, to reduce the production time. If all these needs were collected and each device implemented them internally, the result would be very large devices and very long testing time, which would lead to no salable parts ... As usual, a trade off must be found to

APPLE INC.
EXHIBIT 1008 - PAGE 0357

satisfy both the suppliers and the customers. Here also the designers have to be mentioned because their job is in the middle between what the supplier wants, i.e. all the possible tests in little area, and what the customer needs, i.e. each device fully operating inside the specifications. These considerations drive us to build the circuitry for test modes without penalize area and performances. It is clear that each single circuit would be designed so that the testing features would be combined to the standard ones, if it is possible and needed, adding only the strictly necessary transistors: fully testing circuits must be as few as possible. This choice will result in a little area, but sometimes it is impossible to be realized; therefore dedicated control signals are introduced. Often added signals cause a slowing down of previously fast circuits, so that it is preferred to add testing capability on the circuitry that does not need to be fast, but it could be not yet enough. Another possibility is to use some circuits fully dedicated to the testing: in user mode the standard circuitry is used, but the test mode switches on a dedicated one. Finally the design will match the specifications following as well as possible the previous considerations. Now let's try to give some details. It was shown that the first generation of Flash memories were EPROM like; this means that quite the whole internal circuitry was analogic: if a memory is regarded as a device to store data, no other circuits are needed. The element to be tested are the cells (as basic memory elements), the whole array (as group of cells organized by rows and columns) and the input/output circuits. First, it is extremely important to have a figure of cells' behavior (reading, programming and erasing characteristics); these measurements are useful to monitor the process and to fix the circuitry. The results of this phase are collected in such a way a statistical study can be performed. The standard, or production, testing aims to ensure the functionality of the whole device. In this case the circuits involved are decoders, high voltage switches for program and erase, sense amplifiers, input and output buffers. Every failure that should occur is considered as an alarm bell: other characterization of the cells are performed and the causes are investigated using particular test modes (each project has many test circuits used only in very peculiar situations). We have seen that the device became complex and can be also dedicated to some applications. The practical result is that pure logic circuitry is built on same silicon of an initially "pure analogic" device. All the circuitry works around the array and its main task is to read, to program and to erase the cells, but there are some fully digital circuits principally encharged to manage the user interface and the embedded algorithms. Their own functionality can be tested apart from the matrix one, because they can work also stand-alone; therefore the first question to answer is whether the well-known testing concepts can be directly applied to them. The answer is positive regarding the concepts themselves, but it is negative if we look at their realizations. The reason of

APPLE INC.
EXHIBIT 1008 - PAGE 0358

this statement is a complex mixture of simple arguments; each single reason is not so important to state the answer, but all of them together are very strong. The considerations about the area turn out again: the classic concepts of testing (SCAN, BIST, ...) require to latch each node to be analyzed in the network. Unfortunately there are only few latches on the internal signals, so the introduction of the missing latches results in a big amount of area. Another argument is about the purpose of the testing. Here the complete functional test is not needed: the circuitry is designed to perform some operations under fixed conditions. The production testing must ensure the perfect functionality as the specifications state, even if the design could operate potentially in a wider range. The last point is about the connection between digital and analogic parts. It is true that the digital circuitry can work separately from the analogic one, but actually the two parts work together and it is not realistic to perform two separate testing flows. This implies that first of all the way to activate the test mode must be unique; then the digital circuitry is tested on its capability to drive the analogic one; if one of the two parts fails, the testing flow should be stopped. Once that the purpose of the testing is clear, some words can be spent about the test mode activation in a Flash memory. The way to enter the test mode is obviously unknown to the user because otherwise many design choices about the circuits could be discovered. Furthermore the protocol to activate the testing must be complex enough to prevent the user from accidentally entering it, but not so difficult to slow down the testing time (in fact during the testing flow the test mode entry sequence is repeated many times).

5.8 REDUNDANCY AND ERROR CORRECTION CODES

Aim of the following analysis is to establish a relationship between the process yield for a Flash memory and the number of redundancy elements which have been inserted into such a device.

5.8.1 The Yield

The production of memories must deal with the problems of the process yield. It is clear that the term "yield" indicates the number of devices which, over a given production lot, operate according to the specifications. In other words, it denotes the fraction of devices which work properly at time zero.

The verification of the yield for a complex device as the Flash memory is a problem of primary importance which has consequences on the destiny of a design. This implies that appropriate measures must be taken in every step of the design flow in order to ensure a yield raise. In particular, it is fundamental that the design takes into account the process failures and adopts circuitry solutions in order to reduce their negative effects. To this aim, it is necessary

APPLE INC.
EXHIBIT 1008 - PAGE 0359

to have complete information about the most frequent defects which takes place on the chip. For example, in the case of memories, statistical studies show that the reduction of yield is mainly due to the matrix; that's obvious because in this kind of devices the latter occupies a relevant part of the chip area. The most common types of defects herein found can be divided into two categories:

1. adjacent rows whose metals are short-circuited;

2. memory cells whose behavior is incorrect.

In order to overcome the first kind of failure, designers employ supplementary rows ("row redundancy") within the matrix, while in the other case a "column redundancy" is used. In both situations the idea is to substitute the defected element with a redundancy one, whatever the practical implement might be. The drawback of this strategy is that the larger is the device, the more complex will be the realization to manage the redundancy, as the selection of a redundancy element is made on the basis of a comparison of the external address cell with the contents of appropriate registers in which the failed element address are stored.

The statistical model which is most adopted to describe a yield estimation makes use of a "Poissonian" approximation whose expression takes the form:

$$Y_{\text{tot}} = \mathrm{e}^{-AD} \tag{5.41}$$

where Y_{tot} is the total yield, A the device area, and D the number of defects per unit area.

The "prime yield" Y_0 is the yield which is obtainable without any intervention; in the present analysis it will be considered an independent variable used to evaluate the redundancy system performance. The first operation to make is to decompose the prime yield into the Y_p and Y_{m0} components, referring to the peripheral and matrix yield, respectively.

This step is necessary as redundancy will be effective only on the latter one, since the introduction of redundant control circuitry is not cost effective.

The correct way to do this splitting-up is by assuming that the total amount of defects must be constant; by indicating as A_p and A_m the area of the two portions the device has been divided into, and D_p and D_m the corresponding defectiveness, we can write:

$$A_m D_m + A_p D_p = AD \tag{5.42}$$

By substituting these relationship into Eq. (5.41) we obtain:

$$Y_0 = Y_p Y_{\text{m0}} = \mathrm{e}^{(-A_p D_p)}\ \mathrm{e}^{(-A_m D_m)} \tag{5.43}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0360

It is now opportune to introduce the following two adimensional parameters

$$k = \frac{A_p}{A_m} \qquad R = \frac{D_p}{D_m} \tag{5.44}$$

so that we can now obtain the Y_{m0} and Y_p yields in the following form:

$$Y_{\mathrm{m0}}(Y_0, k, R) = Y_0^{\frac{1}{1+kR}} \tag{5.45}$$

$$Y_p(Y_0, k, R) = Y_0^{1-\frac{1}{1+kR}} \tag{5.46}$$

We are particularly interested in the Y_{m0} yield, because we can correlate it with the probability p of a cell fail. We will assume that fail events on distinct cells are independent, and we will disregard any fail cause which is not ascribed to the cells.

On the basis of these approximations, the relative matrix yield can be expressed in the form:

$$Y_{\mathrm{m0}} = (1-p)^{N_{\mathrm{cells}}} \implies p(Y_{\mathrm{m0}}, N_{\mathrm{cells}}) = 1 - Y_{\mathrm{m0}}^{\frac{1}{N_{\mathrm{cells}}}} \tag{5.47}$$

In the following paragraph we will employ this relationship to estimate the impact of redundancy systems on the yield.

5.8.2 Static Redundancy

First of all, we will refer to column redundancy, as the row redundancy is not employed, as explained later on.

Let p be the error probability on a single bit of a block, i.e. the probability to have a fail on a cell. Furthermore, let's assume that:

- N_c number of columns in one block,
- N_r number of rows in one block,
- N_{red} number of redundant columns in one block.

We can now calculate the error probability in one column; as the latter consists of N_r bit, the probability can be expressed as:

$$p_0(p) = 1 - (1-p)^{N_r} \tag{5.48}$$

If we suppose that m columns, $0 \le m \le N_c$, have at least one error each, we obtain:

$$P_c(m,p) = \binom{N_c}{m} \Big[p_0(p)\Big]^m \Big[1 - p_0(p)\Big]^{(N_c-m)} \tag{5.49}$$

APPLE INC.
EXHIBIT 1008 - PAGE 0361

Similarly we can obtain the number of having j redundancy columns without errors, where $0 \leq j \leq N_{\mathrm{red}}$:

$$P_{\mathrm{red}}(j,p) = \binom{N_{\mathrm{red}}}{j}\Big[1 - p_0(p)\Big]^j \Big[p_0(p)\Big]^{(N_{\mathrm{red}}-j)} \qquad (5.50)$$

We must now identify the fail event which can be corrected. We point out that we can recover a sector with failed columns whenever the number of functioning redundancy elements exceeds that of the matrix failed ones; so the probability that we can recover one sector can be given by:

$$P_{\mathrm{tot}}(p) = \sum_{m=0}^{N_{\mathrm{red}}} P_c(m,p) \sum_{j=m}^{N_{\mathrm{red}}} P_{\mathrm{red}}(j,p) \qquad (5.51)$$

This is the correct expression for a given block; now, if the whole memory is organized in N_b blocks, the total recovery probability is calculated as the intersection of the single events. Again, if we assume that defects on different sectors are independent of each other, we obtain:

$$Y_{\mathrm{tot}}(p) = \Big[P_{\mathrm{tot}}(p)\Big]^{N_b} \qquad (5.52)$$

The use of the symbol Y_{tot} in the previous equation is justified by the fact that the latter can be interpreted as the ratio of the number of recoverable chips to the total number of them. The relationship explicitly links Y_0 to Y_{tot} by means of the intermediate parameters Y_{m0} and Y_p, and enables to investigate quantitatively the improvement which can be achieved on yield by using redundancy.

5.8.3 Wafer Yield

The plots which depict Y_{tot} as a function of the prime yield Y_0 show that the final yield saturates; in other words, when the prime yield is sufficiently high, the introduction of further redundancy elements marginally enhances the final yield.

In addition to these considerations, it's worth noting that while the chip geometry is rectangular, that of the wafer is circular. The increase in the device area over a certain limit can reduce the number of components which are obtainable from a given wafer, therefore it is necessary to compare the yield increase factor due to redundancy usage versus the decrease of the number of chips per wafer, in order to decide on the redundancy amount.

APPLE INC.
EXHIBIT 1008 - PAGE 0362

In particular, the following relationship gives the number of chips of area A_0 which are obtained from one wafer whose "usable radius" is r_o:

$$N(A_0, \lambda) = \frac{\pi r_o^2}{A_0} - \frac{2\pi r_o \sqrt{\lambda}}{\sqrt{A_0(\lambda^2+1)}} \tag{5.53}$$

where λ is the chip aspect ratio.

From $N(A_0, \lambda)$, one can derive the following two quantities:

1. N_{sr}: the number of functioning devices when no yield enhancement countermeasure has been adopted

$$N_{sr} = Y_0 N(A_0, \lambda) \tag{5.54}$$

2. N_{cr}: the number of functioning chips when column redundancy is employed

$$N_{cr} = Y_{tot} N(A_0 + \Delta A_0, \lambda) \tag{5.55}$$

We point out that the use of redundancy enlarges not only the matrix area, but also the peripheral area, because of the further control circuitry which must be introduced inside the chip.

5.8.4 *A Real Case*

To apply our previous considerations we will take as an example an 8Mbit Flash memory which consists of 16 sectors, each of them 512kbit-wide.

Every sector is made of 2048 columns, which are organized in groups of 4 columns (we will refer to as "local bitlines"), so as to form 512 "main bitlines".

An advantage of this architecture is that it keeps adjacent Metal-2 for main bitlines as distant as the width of four cells, thus relaxing the criticity over such a layer.

Since the Metal-1 and Metal-2 layers are used for the columns, rows will be realized in polysilicon; so, while this choice will affect the row switching times, it will allow designers to avoid row redundancy because the probability of short circuit between two adjacent polysilicon rows is very low for actual standard of integrated technologies.

Each local bitline is addressed by means of 11 bits: 7 of them are necessary to identify the columns which belongs to the same word, while the others are used to pick out the desired column within the selected word. Obviously, one

APPLE INC.
EXHIBIT 1008 - PAGE 0363

main bitline is addressed by means of 9 bits.

$$\text{row addresses} \quad [A_{18} \cdots A_{15}] \cup [A_{14} \cdots A_{12}] \cup [A_{11} \cdots A_{10}] \cup [A_9 \cdots A_7] \tag{5.56}$$

$$\text{column addresses} \quad [A_6 \cdots A_4] \cup [A_3 \cdots A_2] \cup [A_1 \cdots A_0] \tag{5.57}$$

Within the device, the failed column addresses are stored in non-volatile registers called "UPROM" (i.e. Unerasable Programmable ROM); one of them (the "guard" UPROM) will specify whether the UPROM information is valid or not. Therefore the total number of UPROMs (N_{UPROM}) will be:

$$N_{\text{UPROM}} = 12 N_{\text{red}} N_b \quad \text{local bit lines} \tag{5.58}$$

$$N_{\text{UPROM}} = 10 N_{\text{red}} N_b \quad \text{main bit lines} \tag{5.59}$$

where N_b is the number of blocks the matrix is decomposed in with respect to the redundancy elements. It's worth noting that the latter could not coincide with the number of sectors, as redundancy is not necessarily related to each physical sector.

The area increase for peripheral circuitry takes the form:

$$\Delta A_p = S_f \; N_{\text{UPROM}} \; A_{\text{UPROM}} \tag{5.60}$$

where S_f is the shrink factor for the device, and A_{UPROM} is the silicon area occupied by one UPROM cell and its corresponding control logic.

The increase in the matrix area can be easily calculated considering that the redundancy columns are the same as the others:

$$\Delta A_m = \frac{N_{\text{red}}}{N_c} A_m \tag{5.61}$$

thus we can finally obtain the dependence of N_{sr} and N_{cr} on the number of redundancy columns to investigate the recovery possibilities.

For the case of the 8Mbit memory, we have implemented a redundancy system which can substitute 8 failed local bitlines and 40 more columns organized in 4 groups of 10 columns each; this solution proved to be a reasonable compromise between flexibility and area occupation, and makes use of one redundancy-dedicated sense amplifier, in order not to penalize the redundancy access time with respect to that of the ordinary cells. The drawback is that we will not be able to correct more than one fail occurring in the same word.

Fig. 5.111 and Fig. 5.112 show the circuits which connect the redundant columns to the sense amplifier. The signals named Y_{Oi} correspond to the decoding of $A_1 - A_0$ address bits, while Y_{Mri} and Y_{Ori} are generated according to the comparison of the invoked cell address bits and the UPROM contents.

APPLE INC.
EXHIBIT 1008 - PAGE 0364

In the case of main bitline redundancy, Y_{MRj} only are necessary for selection, while Y_{Ork} are necessary for a local bitline substitution.

The array RDC[3:0] in Fig. 5.113 and Fig. 5.114 specifies which column within the word must be replaced.

5.8.5 Error Correction Codes

One possible alternative to the static redundancy is the use of error correction codes.

At present, this opportunity is purely theoretical, as the large majority of the devices in the market do not implement this technique. Furthermore, wherever error correction codes have been exploited, the choice has been driven by reliability considerations rather than by expected benefits on the yield.

However, the two alternatives of static redundancy and error correction codes can be compared to shed light on their features and limits.

In general terms, we can define as "code" any application between two finite collections, the domain (or "source alphabet"), and the co-domain (or "code alphabet") respectively.

The basic idea behind any code is the concept of "information redundancy", i.e. the fact that the code is able to transfer a larger quantity of information than it is strictly required; the main object of an information coding is thus to introduce a certain amount of "redundancy". In order to describe quantitatively the notion of "code" redundancy, it is necessary to introduce the concept of distance between two words, i.e. the number of positions the two words differ from.

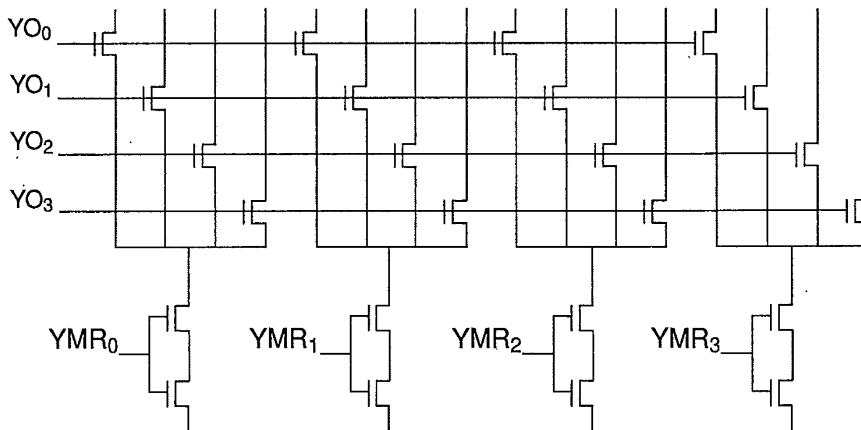


Figure 5.111 Redundancy columns - 1.

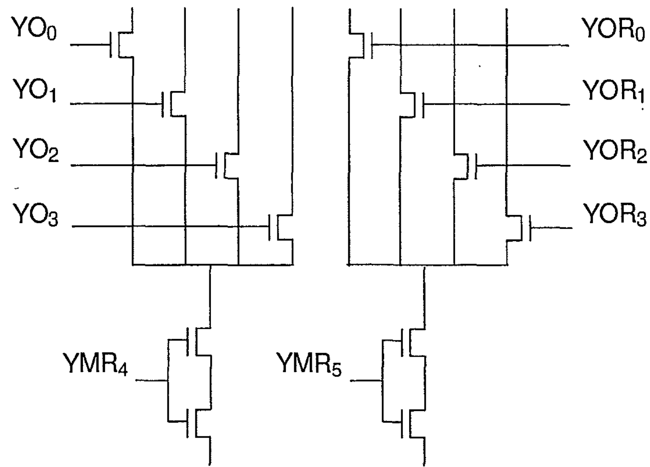


Figure 5.112 Redundancy columns - 2.

According to this definition, it is clear that the more elements two words vary from each other, the more redundant a code will be.

As an example, the Hamming codes are among the simplest ones for single error correction. The relationships among their basic parameters are:

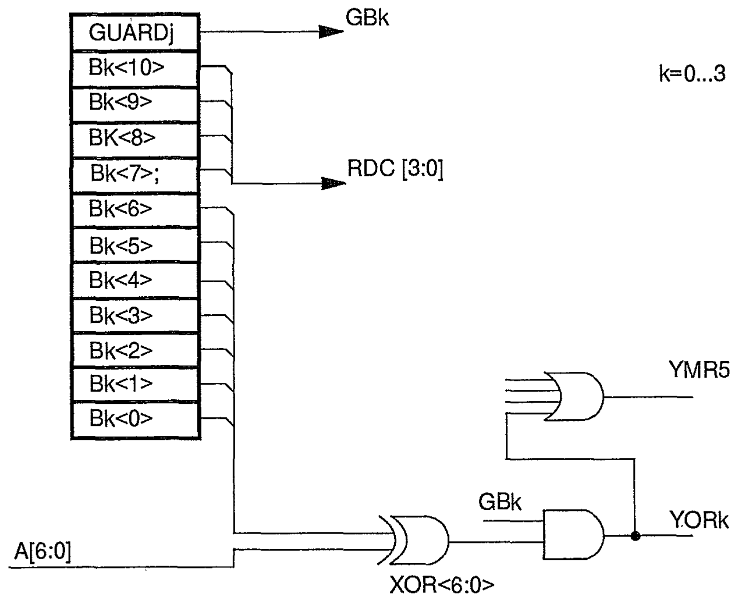


Figure 5.113 Selection signals for the local bitlines.

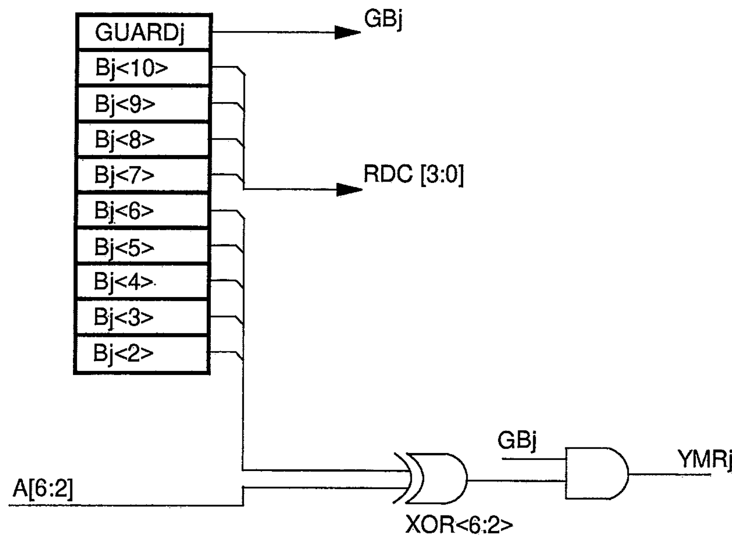


Figure 5.114 Selection signals for the main bitlines.

$n = 2^m - 1$	length of the code word,
$k = 2^m - m - 1$	information bits,
m	redundancy bits.

The minimum number of redundancy bit in a code in order to correct t errors must comply with the so called "lower limit of Hamming":

$$P_{\text{tot}} \geq \sum_{i=0}^{t} \binom{n}{i} . \tag{5.62 }$$

The above inequality can be explained as follows: the left-hand side represents the number of possible different configurations of the parity bits, and it must be greater than the right-hand side quantity which is the amount of error events which alter up to t bits. In the case with $t = 1$ the relationship reduces to:

$$2^m = 1 + \binom{n}{1} = n + 1 \tag{5.63 }$$

which was previously indicated.

It's worth noting that these codes introduce the strictly necessary (and sufficient) redundancy bits, thus they can be referred to as "perfect" codes.

The evaluation of the error correction capability of a given code is to estimate the residual error probability, i.e. the probability that errors exceeds the code

APPLE INC.
EXHIBIT 1008 - PAGE 0367

recovery power. For example, let's suppose that our information channel can damage each of the input bits with equal probability p, and that it has no memory of previous events (we talk of "symmetric binary channel"); we can state that the probability of j errors occurring on a n-bit word is given by:

$$P(j \text{ errors}) = p^j (1-p)^{(n-j)} \qquad (5.64)$$

The probability that the number j of errors exceeds the code recovery capability t can be expressed by the following inequality:

$$P(E) \leq \sum_{j=t+1}^{n} \binom{n}{j} p^j (1-p)^{(n-j)} \qquad (5.65)$$

The inequality sign is used here because any code, apart from the perfect ones, with distance $2t+1$ can correct some error whose distance overcomes t.

At this point, it is interesting to compare the alternatives of column redundancy and code redundancy on the ground of unrecovery probability; in Fig. 5.115 we plot the curves for

1. a static redundancy implementation with four additional columns for 1Mbit (2^{10} rows and 2^{10} column) block;
2. a shortened Hamming error correction code, with 2^7 information bits and 2^3 parity bits.

We now point out that the Hamming code solution uses more area than the static one: in fact in the former case we have 8 dedicated bit every 136, which corresponds to 5.88% area increase, while in the latter the area increase is merely the 0.39% of the entire matrix. Another comparison can be done in case of equal area occupation. In Fig. 5.116 we compare the Hamming code solution of the previous case and a 60-columns static one.

Although we have not so far considered the problems of the practical implementation of the two systems, Fig. 5.116 shows that the Hamming correction code is effective for p which exceed the static solution limit.

APPLE INC.
EXHIBIT 1008 - PAGE 0368

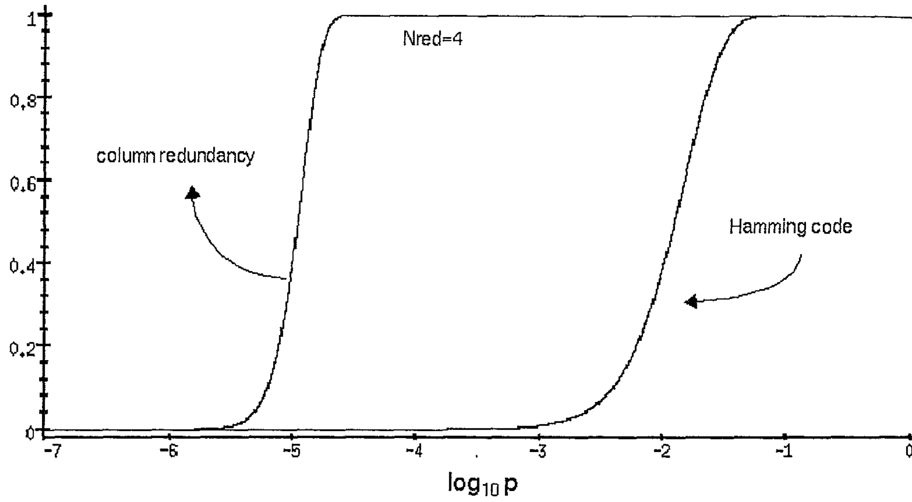


Figure 5.115 Unrecoverable events curves.

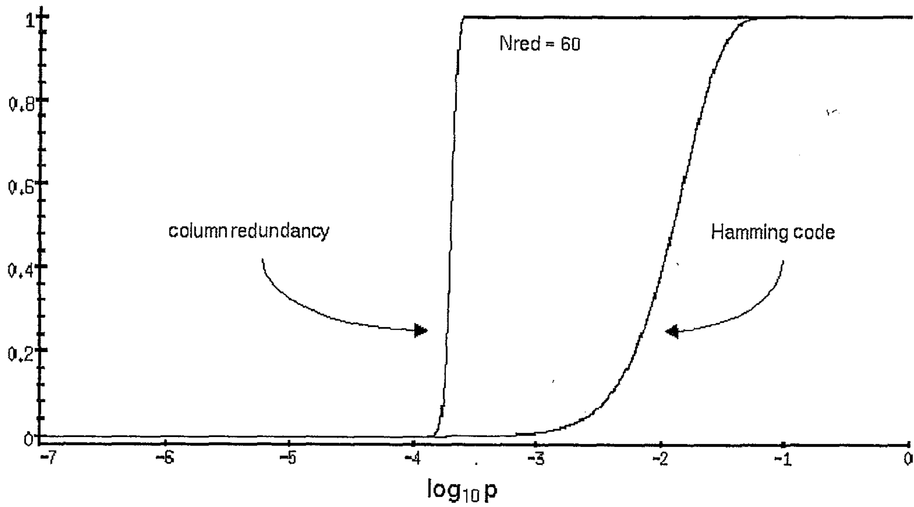


Figure 5.116 Unrecoverable events curves.

APPLE INC.
EXHIBIT 1008 - PAGE 0369

6 MULTILEVEL FLASH MEMORIES

Guido Torelli[1], Massimo Lanzoni[2],
Alessandro Manstretta[3], Bruno Riccò[2]

[1]Department of Electronics, University of Pavia
Via Ferrata 1, 27100 Pavia, Italy
torelli@ele.unipv.it

[2]DEIS, University of Bologna
Viale Risorgimento 2, 40136 Bologna, Italy
mlanzoni@deis.unibo.it, bricco@deis.unibo.it

[3]STMicroelectronics, Memory Product Group
Via Olivetti 2, 20041 Agrate Brianza (Milano), Italy
alessandro.manstretta@st.com

Portions reprinted, with permission, from Proceedings of the IEEE, 1998.

Abstract: The threshold voltage of a Flash memory transistor depends analogically on the charge amount stored on its floating gate, and if a number m of different threshold levels can be reliably programmed and sensed in a memory cell, this can store $n = \log_2 m$ bits, thereby overcoming the traditional one-to-one correspondence between cell count and memory capacity. Multilevel (ML) storage is therefore very attractive, as it allows cost-per-bit reduction for any given fabrication technology. However, to implement ML Flash memories more severe requirements have to be met as compared with the traditional bilevel approach in terms of cell sensing and writing, as well as of reliability.

This chapter reviews the fundamental issues specific of ML Flash memories, with particular regard to the case of stand-alone memories, where the large

APPLE INC.
EXHIBIT 1008 - PAGE 0370

array size justifies the overhead required to reliably implement the ML concept. The basic features of 4-level multimegabit Flash prototypes presented so far in the literature are illustrated as significant examples of implementation.

6.1 INTRODUCTION

6.1.1 *The Multilevel Approach*

Data storage density, i.e., the number of bits that can be stored per unit area, and cost-per-bit are essential driving factors for the development of any kind of memory device, and Flash memories do not represent an exception to this respect.

As both these factors are strictly dependent on the physical sizes of the memory cells, so far the development of new generations of Flash memories has entirely relied on the geometry scaling characterizing advanced microelectronics. This approach, however, can be accompanied and enhanced by the Multilevel (ML) Non Volatile (NV) data storage concept, that is in principle very simple and intuitive and can be applied to the same technology and devices of conventional (i.e., bilevel) memories. In bilevel memories, as described elsewhere in this book, the cells can have only two different values of threshold voltage V_T (or, better, two distributions of V_T values - Fig. 6.1a), and during reading it is only necessary to sense whether or not the addressed transistor is conductive. This is generally done by comparing the current, I_{CELL}, flowing through the memory transistor (often also referred to as sense transistor) biased with predetermined drain-to-source and gate-to-source voltages with that of a reference transistor under the same bias conditions, either directly (current-mode sensing) or after a current-to-voltage conversion (voltage-mode sensing).

However, the threshold voltage of a Flash memory transistor depends analogically on the amount of charge, Q_{FG}, stored on the floating gate (FG): therefore, V_T, and hence I_{CELL}, can be changed over a large range of values by simply varying Q_{FG} (Fig. 6.1b). Thus, if the reading (sensing) circuitry can resolve a difference ΔI_{CELL} in the current of differently programmed transistors, in principle each cell can be made to store n bits, with $n = \log_2\{[(I_{max} - I_{min})/\Delta I_{CELL}] + 1\}$, where I_{max} and I_{min} represent the values of I_{CELL} flowing through the most conductive and through the least conductive cell (i.e., through the cell with the lowest and the highest V_T), respectively. Generally, I_{min} is set equal to zero, and therefore $n = \log_2[(I_{max}/\Delta I_{CELL}) + 1]$.

In the extreme case of negligibly small values of ΔI_{CELL}, the cell could operate as a fully analog device able to store an arbitrarily large amount of data (indeed, the use of analog storage capabilities of NV floating-gate transistors has been proposed for non-destructive trimming purposes [12, 35, 36, 46] as well as for analog computation and neural networks [20, 29, 30, 31, 50]). However,

APPLE INC.
EXHIBIT 1008 - PAGE 0371

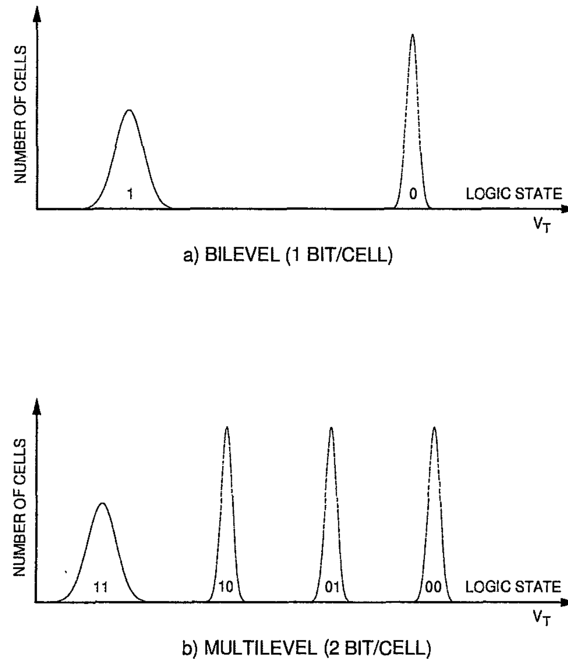


Figure 6.1 Conceptual representation of threshold voltage distributions for a bilevel (a) and for a 4-level (b) Flash memory.

for digital applications sensing would be critically difficult and prone to errors, and reliability problems would become prohibitively critical. On the contrary, in the traditional bilevel case ($\Delta I_{\mathrm{CELL}} = I_{\max}$), we have $n = 1$, but sensing is easier than for higher values of n. In practice, a number of intermediate cases can be considered, and the question arises of which value of n leads to the most cost-effective use of silicon area. Naturally, the answer to this question depends not only on technology and memory architecture, but also on programming and reading schemes, as well as on cell reliability.

In practice, the ML approach allows a cost-per-bit equivalent to future generations of fabrication technologies using the present one, thereby leading to an acceleration in the rate of cost-per-bit reduction [2, 22]. A key requirement of the ML approach is minimization of performance degradation with respect to bilevel memories: in practice, the adoption of ML storage techniques should be transparent to the users. In addition, to achieve significant cost reduction, the overhead to implement the ML concept must be reasonably small.

A few prototypes of Flash memories storing two bits per cell (i.e., 4-level charge storage) have been presented in the literature [3, 26, 27, 40, 41], while similar results are also at reach of EPROMs and full-featured EEPROMs. The

APPLE INC.
EXHIBIT 1008 - PAGE 0372

three-bit/cell solution is considered feasible, though significant research is still needed for it to become an industrial reality. Instead, the possibility to go beyond this limit for digital applications still needs careful investigation [8, 28] and, at present, it is generally looked at with scepticism.

In fact, 256-level storage has already been implemented in EEPROM technology for voice recording [58], nevertheless this is practically considered to be an analog case with special features, since the application is largely tolerant of errors. Also, 6-bit/cell Flash storage has been demonstrated for applications such as still cameras and speech recorders [45] where, however, the required performance is less severe than in stand-alone memories. In this chapter, we will no longer consider this kind of devices, as we will focus our attention on fully digital ML storage. Moreover, we will specifically refer to stand-alone multimegabit Flash memories, where the very large size of the array totally justifies the circuit overhead required to implement the ML approach reliably.

6.1.2 Basic Issues for ML Storage

From the device point of view, ML NV charge storage poses three main problems, namely: 1) precise reading, i.e., the capacity to recognize (slightly) different cell conductivities in a short time (or, equivalently, the ability of detecting the cell current I_{CELL} with sufficiently high accuracy and high speed); 2) accurate writing, i.e., placing just the amount of charge on the floating gate of the cell transistor required to obtain the target value of V_{T}; 3) reliability, in particular the cell capability to ensure adequate spacing between adjacent stored levels for a sufficiently long time interval under specified operating/storage conditions. Moreover, both program and read disturbs must be carefully considered. All these requirements become obviously more severe as the distance between different levels (measured in terms of either stored charge or cell conductivity) decreases with increasing number of stored bits per cell. (In this chapter, the data change operations performed on the NV memory cells will be denoted according to the following convention: *erase* is the operation of changing the data state in a block of memory cells (either the whole array, or a sector or even a byte, depending on the memory organization, i.e., it is an *unselective* data change operation); *program* indicates the operation of changing the data state in memory cells on a "bit-by-bit" basis (i.e., a *selective* data change operation); *write* indicates one or the other of these operations indifferently. Thus, when a new content is to be stored (or written) in a memory block, this is first erased, and its cells are then programmed to the respective desired states.)

As for reading, the operation can substantially be seen as an analog-to-digital (A/D) conversion of the signal produced by the selected cell, and the key issue concerns the trade-off between speed and accuracy: as the latter is increased

APPLE INC.
EXHIBIT 1008 - PAGE 0373

(i.e., as smaller values of ΔI_{CELL} can be safely recognized), a larger number of bits can be stored on a single cell, but the sensing circuitry becomes more expensive and globally slower. In general, the required A/D conversion can be done in parallel or sequentially for higher speed and smaller area occupation, respectively, while intermediate solutions offer interesting trade-offs between these conflicting aspects.

A gate voltage larger than in conventional bilevel sensing is desirable to allow for a sufficiently large threshold window (or, equivalently, to provide sufficiently large current signals to sensing circuits), however the issue of read disturbs must be carefully considered, also taking oxide degradation due to program/erase cycling into account.

The problem of accurate charge placement is generally tackled by means of cell-by-cell Program & Verify (P&V) approaches (Section 6.4): the programming operation is divided into a number of partial steps and the cell is sensed after each step to determine whether or not the target V_{T} is achieved, so as to continue programming if this is not the case (as each cell is independently controlled during programming, this technique allows simultaneous programming of a whole byte or even a number of bytes). This procedure ensures that the target V_{T} is reached (with the accuracy allowed by the quantization inherent in the use of finite programming steps), but can be very long and must be controlled by on-chip logic circuitry, thus leading to a non negligible area overhead. The alternative approach is based on self-controlling techniques able to automatically stop programming when the target V_{T} is reached. These procedures are less developed and looked at with smaller confidence than P&V schemes: thus they have not yet been used in commercial products, but could provide faster and simpler operations and deserve careful consideration [33, 37].

As far as writing mechanisms are concerned, Channel Hot-Electron (CHE) injection and Fowler-Nordheim (FN) tunneling present different characteristics, advantages and drawbacks, which make them somewhat complementary and suitable for different types of architectures and applications [18]. The main advantages of CHE injection over FN tunneling as a programming mechanism are higher speed and lower electric fields in the gate oxide, resulting in better endurance, lower applied voltages, less circuit overhead and less disturbs in the memory array. High programming speed and less disturbs are particularly attractive features for ML storage, as a larger number of programming pulses than in the bilevel case is required to achieve high accuracy in the obtained threshold voltage. The higher programming speed allows for a larger number of program pulses in any given time interval, which leads to better controlled V_{T} distributions when adopting P&V techniques. On the other hand, FN tunneling requires much lower power consumption, which allows large programming parallelism to increase overall program throughput, and makes it easier to generate

APPLE INC.
EXHIBIT 1008 - PAGE 0374

the required high programming voltages by using on-chip charge-pump voltage multipliers [16, 17, 56]. FN tunneling also exhibits stronger process sensitivity [18] and is affected by the so-called "erratic bit" problem (Chapter 7).

In ML Flash memories, enhanced sensitivity to program disturbs is expected because of the reduced margins between adjacent states and/or the larger used V_T window (which increases with the number of stored levels for the same separation among them), and the increased program time (because the higher accuracy required in charge placing makes the time needed for the operation to be completed successfully longer). In NOR-based memories using CHE programming, the worst case for the bit-line disturb occurs when programming a cell to the highest V_T level because of the large V_T shift required with respect to the neutral state: the programming drain voltage is the same as in conventional Flash memories, but the programming time is longer. On the contrary, word-line disturb should not be substantially affected by the increase in program time, at least when using a staircase gate voltage ramp programming algorithm (Section 6.4), since during most of the time the word-line voltage is rather low and only the last few pulses will effectively contribute to the disturb. Optimized cell design for low drain voltage programming and divided bit-line organization are probably required to guarantee sufficient program disturb immunity.

With regard to reliability, no systematic study on ML Flash memories is yet available, thus all predictions are made essentially extrapolating the knowledge acquired with conventional bilevel memories. No specific failure mechanism is foreseen for ML Flash memories, although ML storage is obviously more critical than its conventional bilevel counterpart, due to the need for a larger V_T window (to keep adequate spacing among the stored levels) and/or reduced spacing between adjacent levels (to limit the increase in the V_T window). All the failure modes described in Chapter 7 for bilevel Flash memories must be carefully considered, including issues such as data retention, oxide defects, program and read disturbs, performance degradation induced by program/erase cycling, and stress-induced leakage current (SILC).

In essence, an increased V_T window leads to larger charge transfer through the oxide and, in practice, to higher operating voltages, thereby worsening problems related to charge trapping within the oxide (with possible impact on endurance) and excessive oxide leakage (with consequent data retention degradation). On the other hand, for any given number of stored levels, decreasing the separation between adjacent levels helps to limit the phenomena mentioned above, but makes the memory more sensitive to their effects. In particular, for instance, if the same V_T window is used to allocate m levels, the data retention in principle degrades as $m - 1$ (as the charge difference between adjacent levels becomes smaller by the same factor). As, in practice, ML Flash memories

APPLE INC.
EXHIBIT 1008 - PAGE 0375

require both higher V_T windows and smaller level separation, all the aspects mentioned above must be taken into account. Consequently, reliability margins decrease from more than 1V for 1 bit/cell memories to values in the order of few hundreds mV for the case of 2 bits per cell, and even less for 3 or 4 bit/cell storage.

For example, we can express the crucial problem of data retention in terms of number of stored electrons. With a FG capacitance in the 10^{-15}F range, a threshold voltage window of a few V requires an overall variation of about 30,000 electrons in the FG. If this variation is split, for instance, in 4 levels, the states of the sense transistors differ from one another by about 10,000 electrons, and this difference should be maintained for more than 10 years. This implies that less than about 1,000 electrons should leak out from (or into) the FG in one year, i.e., that the (average) leakage current through the oxide insulating the FG should be smaller than 10^{-24}A (or, about 10^{-16}A/cm^2). Of course, this problem is significantly aggravated if the number of levels programmed in a cell is increased and/or the total V_T window is narrowed.

The reliability problems can (and must) be partially alleviated with the use of redundancy and error correction codes (ECCs) [54, 55] (the latter, in particular, for applications demanding a large number of program/erase cycles), but the very small numbers given above clearly indicate the need of outstanding insulating properties. From this point of view, SILC and/or oxide time-dependent breakdown are expected to pose serious problems.

In conclusion, the successful realization of ML Flash memories and the number of bits that can be stored in a single cell depend on the solutions given to the main problems mentioned above (precise cell sensing, accurate charge placement, and reliability) [9, 15, 19, 61]. Such solutions, however, strongly depend on the physical mechanism used for cell writing as well as on memory architectures, which are fundamental issues determining the complexity, performance and cost-effectiveness of memory chips. The realization of ML Flash memories represents therefore a unique global problem composed of many interacting aspects, that must be concurrently considered in the conception and development of real industrial products. In the following of this chapter, we will focus our attention on basic items specific of ML Flash memories, namely array architectures, sensing and programming. The reader should refer to Chapter 7 for details on reliability issues which, although posing more severe constraints, as mentioned above, are not expected to introduce new basic problems from device physics and failure mechanism standpoints (in that chapter, issues of ML storage reliability are also addressed).

APPLE INC.
EXHIBIT 1008 - PAGE 0376

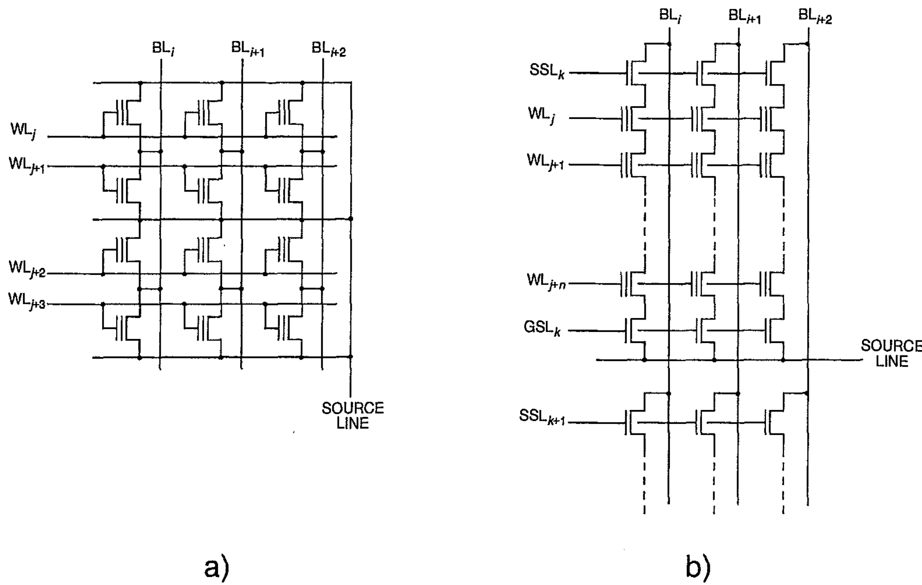


Figure 6.2 Schematic representation of NOR (a) and NAND (b) memory array architectures.

6.2 ARRAY ARCHITECTURES FOR MULTILEVEL FLASH MEMORIES

As for the case of their bilevel counterparts, two basic array architectures are available for ML Flash memories, namely the (common ground) NOR and the NAND ones (Fig. 6.2), which obviously maintain their inherent advantages and disadvantages.

So, the NOR array allows higher sensing speed, as the selected cell is directly connected between the bit-line and ground. This is by far the preferred architecture in industrial products, and therefore also gets benefits from a very strong experience in many years of development and production. However, the need for one bit-line contact hole every two cells poses a limit to the obtainable integration density. On the other hand, the NAND array achieves higher integration density, but offers intrinsically lower speed due to the number of transistors connected in series between each bit-line and ground, and is therefore particularly suited for mass storage applications, where page-mode operation makes the sensing time of individual cells not a serious concern.

As far as writing is concerned, the NOR architecture is suitable for both CHE and FN programming, while the NAND architecture only supports FN programming, due to the series-connection of the memory cells within a string. All cells in a NOR array must absolutely have a positive threshold voltage

(higher than a predetermined minimum value), to prevent bit-line leakage due to unselected cells during reading. Obviously, this feature is not required for a NAND array. Finally, the NAND architecture is affected by larger read disturbs, due to the high voltages applied to unselected word-lines during reading, and suffers from inter-bit-line capacitive coupling noise, due to simultaneous activation of adjacent bit-lines for page parallel sensing [53]. More details about the performance of these basic architectures for ML storage are provided in the following subsections.

In principle, the ML concept can be coupled to any kind of architecture and/or writing mechanism, although in this section, for conciseness, the discussion will be limited to the combination of solutions used in the prototypes reported so far in the literature [3, 26, 27, 40, 41], all implementing a 4-level (i.e., 2 bit/cell) ML scheme but differing substantially as far as programming mechanisms, architectures and implementations are concerned (see Tab. 6.2 at the end of this chapter for a performance summary).

6.2.1 NOR Architecture with CHE Programming

The common ground NOR architecture, in conjunction with CHE programming and source-side FN erasing, is the straightforward extension of the industry standard for bilevel Flash memories, and in fact this approach was followed in the first 2-bit/cell Flash memory prototype presented in the literature [3]. The typical V_T distribution of this memory is illustrated in Fig. 6.3a, where the threshold voltages of the reference cells used for verification after erasure (erase-verify), for program-verify and for reading (V_{EV}, V_{PVi} and V_{Ri}, $i = 1$ to 3, respectively) are also shown. The lowest V_T level derives from unselectively erasing a whole memory block: the dispersion of cell characteristics gives rise to a V_T distribution as large as 1.5V or more [60]. Since in conventional NOR architectures all the cells must have a sufficiently positive V_T, the large width of this level greatly decreases the threshold voltage window allowed for the programmed states. Several methods have been proposed to narrow the erased V_T distribution [25, 42, 59]. Nevertheless, in order to prevent false results, erase-verify can be carried out only when bit-line leakage is negligible.

The highest V_T level can be placed above the read voltage, as it can be detected by sensing zero current through the selected cell during both program-verify and reading. However, the read gate voltage (V_{GR}) must be higher than the maximum V_T value of the third level (01) by a given amount (in the range of 0.5 to 1V), to allow the two highest V_T levels (01 and 00) to be discriminated.

Reliability aspects (namely, read disturb of erased cells in the addressed word-line) and design considerations (need for limiting silicon area and power

APPLE INC.
EXHIBIT 1008 - PAGE 0378

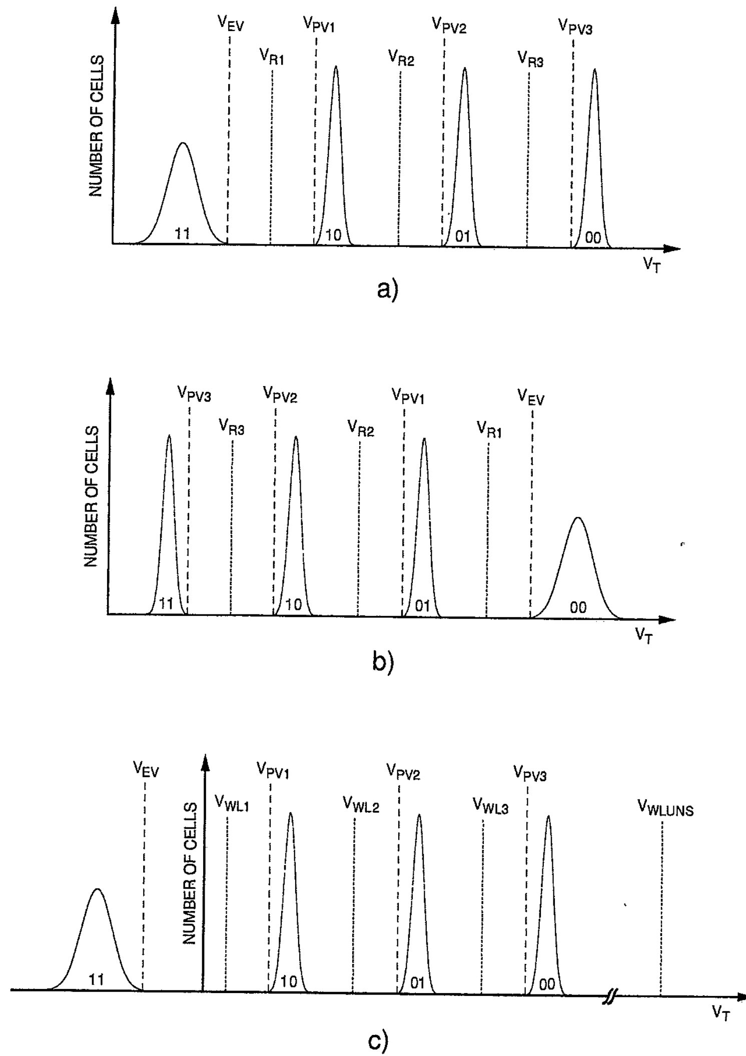


Figure 6.3 Conceptual representation of the threshold voltage distributions for: a) a NOR-array 4-level memory using CHE programming; b) a NOR-array 4-level memory using FN programming; c) a NAND-array 4-level memory (FN programming). The read gate voltages used in [27] ($V_{WL1}, V_{WL2}, V_{WL3}, V_{WLUNS}$) are also shown.

APPLE INC.
EXHIBIT 1008 - PAGE 0379

consumption overhead due to the voltage multipliers providing read voltages higher than V_{DD}) suggest a maximum value around 6V for V_{GR} [9].

The V_T distribution width of the three programmed levels can be reduced by using suitable algorithms based on the bit-by-bit P&V approach (Section 6.4), with the drawback of longer program times. The program time in [3] is given as 40μs: to achieve a program throughput of 0.1MB/s, 4 bytes are programmed in parallel.

6.2.2 *NOR Architecture with FN Programming*

As recalled above, the NOR architecture can also support FN tunneling as the programming mechanism. This is the case in the NOR ML Flash memory presented in [41], where programming is achieved by extracting electrons from the FG by drain-side FN tunneling, and erasing is performed by (unselectively) injecting electrons into the FG by channel FN tunneling.

The resulting V_T distribution in the case of 4-level cells is shown in Fig. 6.3b, where the erased state corresponds to the highest V_T level (logic state = 00). The width of the erased state is again rather large, however this does not affect the V_T window available for the other states, as it can be placed above the read voltage, where overerasing causes no problem. The lowest V_T state (11) is obtained as a result of a programming operation carried out with a bit-by-bit P&V approach: thus its V_T distribution can be made narrow, thereby optimizing the use of the allowed V_T window.

To make program time acceptable with FN tunneling, high electric fields are required, which results in more severe disturb problems. The effects of these disturbs are reduced by using segmented bit-lines and word-lines. In the prototype [41], 64 cells are programmed in parallel to achieve a program throughput of 0.16MB/s. More details on the programming method used in this device are given in Section 6.4.1.

6.2.3 *NAND Architecture*

A ML NAND architecture, exploiting FN tunneling for both programming and erasing, has been proposed in [27]. The typical V_T distribution obtained for this kind of array in the case of 4-level cells is schematically shown in Fig. 6.3c. The lowest V_T level (logic state 11) is due to channel FN erasing. The use of negative threshold voltages, possible in a NAND architecture, allows a better use of the V_T window. Moreover, the large distribution width of the erased (i.e., lowest-V_T) state is not a concern. During reading, all unselected cells in a selected string must be turned on, featuring a suitably low drain-to-source resistance. This is obtained by driving their word-lines with an adequate gate voltage (V_{WLUNS}), which however cannot be too high to prevent excessive read

APPLE INC.
EXHIBIT 1008 - PAGE 0380

disturbs. This requirement also sets an upper limit to the value of the highest programmed V_T (logic state 00).

A ML Flash memory based on a NAND array is affected by the so-called Background Pattern Dependency (BDP): the threshold voltage of a cell depends on the states of the other cells of the same string, because of the dependence of the series resistance of these cells on their programmed states. More specifically, the threshold voltage of a cell during program-verify can be different than during reading, as a result of the subsequent programming of the other cells, and hence the effective V_T distributions become wider. This effect can be minimized using a predetermined programming sequence which starts from the word-line closest to the ground contact of the string and successively moves towards the bit-line contact (verification and read-out of any cell are therefore performed with the same source resistance). Also, the voltage of unselected word-lines, V_{WLUNS}, is boosted to 6V to reduce BDP, even though this leads to some increase in read disturbs. In addition, the series source resistance of the Array Ground Line (AGL) causes an increase in the cell source voltage during verification and read-out, referred to as AGL bouncing. Even though AGL resistance is reduced through metal ground lines which strap the AGL every 32 bit-lines, this effect also causes an equivalent threshold voltage shift in the programmed cells. In [27], AGL bouncing is minimized by using a very small sense current ($\sim 1\mu$A), although this leads to increased sensing time (in conjunction with the use of a serial sensing approach, this results in very slow sensing speed: in fact, this prototype has been proposed for serial-access mass storage applications). With these methods, the residual broadening of a programmed level due both BPD and AGL can be limited within 100mV.

Program disturbs, also a serious concern in a NAND architecture due to the high program voltages used, can be reduced by means of a Local Self Boosting (LSB) technique, which operates through a capacitive coupling mechanism in the NAND string. Thanks to this approach, when the target V_T of a cell has been reached, further programming is inhibited, while other cells in the same word-line are being programmed to higher V_T values. High programming parallelism is allowed by FN tunneling: for instance, in [51] as many as 2K cells are programmed in parallel to increase program throughput. In the page organization of this architecture, each sense/write circuit is shared by two adjacent bit-lines, so that only either the even or the odd bit-lines are activated simultaneously. This minimizes interbitline capacitive coupling noise and threshold V_T distribution broadening.

Although more complex to design, the NAND architecture presents the substantial advantage of a very small cell size. For instance, in a 0.4-μm process the effective area is $1.47\mu m^2$ for a NOR cell [41] and only $1.1\mu m^2$ for a NAND cell [27]. This advantage results in higher storage density, provided the overhead

APPLE INC.
EXHIBIT 1008 - PAGE 0381

circuitry necessary to realize the NAND architecture is negligible. Moreover, the insensitivity of this structure to cell over-erasure allows a larger V_T window for programmed levels which, of course, results in larger separation between programmed states.

As mentioned above, however, in practice the NAND architecture requires complex design as it needs page programming (to achieve high throughput), high read-out voltages, suitable programming sequences (to eliminate BPD), low sensing current (to reduce AGL) and suitable LSB technique (to minimize program disturbs).

To solve the problems of the series resistance and BPD due to the series-connected cells in a string, a particular technique has also been proposed [1], where each cell includes a transfer transistor connected in parallel with the FG transistor that, when unselected, can be by-passed so as to minimize the cell series resistance. To save silicon area, the transfer device is located at the sidewall of the shallow trench isolation region, at the cost of increased fabrication complexity.

As the above solutions increase not only the chip area but also the complexity of the design and/or fabrication process, the NAND architecture seems to be the least attractive for ML storage applications [18].

6.3 MULTILEVEL SENSING

In ML Flash memories, cell reading (or sensing) is a very critical operation as it plays a dominant role in determining the number of bits that can be effectively stored in a single cell.

From a functional point of view, cell reading consists of two basic steps: signal production and signal recognition. The former step has the fundamental goal of producing tight and distinguishable ML signals (two basic sensing methodology are available, i.e., current-mode and voltage-mode sensing). The second step deals with the circuitry needed for reliably detecting the produced signal, aiming at achieving the best trade-off between complexity (hence, area occupation and power consumption) and speed. From this point of view, a fundamental choice is that between few (possibly only one) sense amplifiers to be used sequentially for determining all the bits stored in the selected cell (low complexity but low speed) and a (simpler) sensing circuit for each programmable level (i.e., many sense amplifiers to be used in parallel) for fast reading (but higher complexity). The convenience of one scheme compared to the other, as well as that of possible intermediate solutions, naturally depend on the application and must be accurately evaluated.

As for silicon area and power consumption, since in 2^n-level memories each cell stores n bits, k/n memory cells must be sensed simultaneously to read k bits

APPLE INC.
EXHIBIT 1008 - PAGE 0382

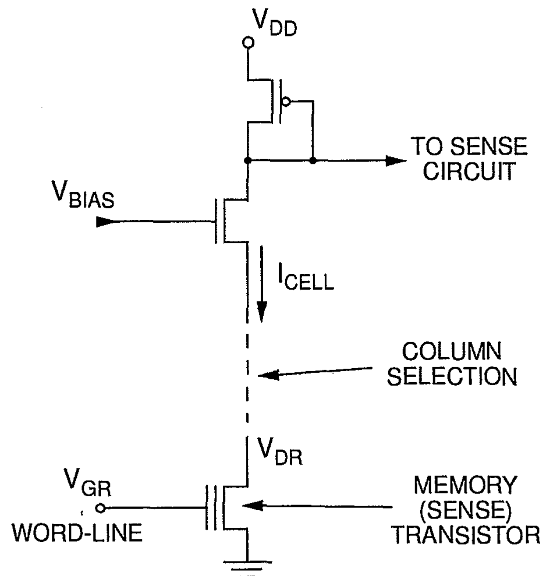


Figure 6.4 Simplified schematic diagram of the circuit used to produce the signal to be detected in memory sensing.

in parallel, and hence k/n sensing blocks are required. As a consequence, on the one hand a large value of n increases the complexity of a single sensing block, but on the other it leads to a smaller number of sensing blocks for any given data read-out parallelism. For example, to read an 8-bit word, 8 sensing blocks are needed in a conventional NOR-based bilevel memory, while 4 blocks are required in a four-level memory. The sensing block count is as low as 2 in the case of a device with 16-level cells (4 bits per cell).

6.3.1 Signal Production and Recognition

Cell reading can be done by current sensing, i.e., by directly looking at the cell current, or by voltage sensing, i.e., by detecting the voltage drop produced by the cell current across a predetermined load. In both cases, the operation is conveniently performed by comparing the cell signal with adequate references produced by cells identical to those to be read but programmed at suitable decision levels. This general scheme, used in conventional bilevel memories, is also adopted in ML memories that, however, compared to bilevel memories, have much more stringent requirements, becoming more severe with increasing number of bits per cell.

APPLE INC.
EXHIBIT 1008 - PAGE 0383

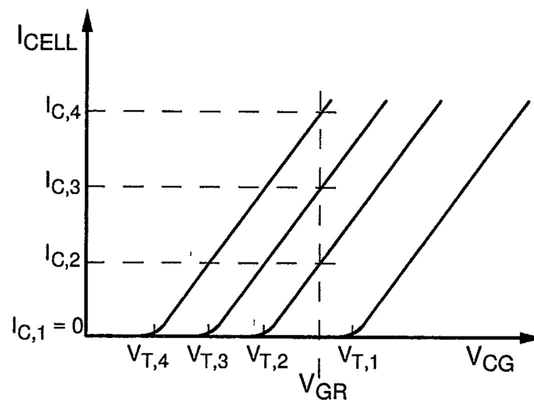


Figure 6.5 I-V characteristics of a Flash memory cell for different values of the programmed threshold voltage. V_{CG} is the control-gate voltage, V_{GR} is the value of V_{CG} during sensing.

Because of the crucial importance of sensing operation for the memory functionality, it is necessary to keep the cell current distribution corresponding to each programmed state as narrow as possible, and therefore all possible causes of current dispersion must be minimized even when, as in the case of voltage sensing, we are ultimately interested in producing different voltage levels. This is obtained by biasing the cell gate and drain electrodes with respective well defined, fixed voltages [43]. In practice, the gate read voltage (V_{GR}) is applied to the addressed word-line, while the drain read voltage (V_{DR}) is obtained by suitably driving the selected bit-lines, as shown in the simplified circuit diagram of Fig. 6.4 [21]. The resulting relationship between the programmed V_T and the read cell current (I_{CELL}) is schematically depicted in Fig. 6.5, where typical I/V characteristics of a cell are plotted.

The choice of V_{GR} and V_{DR} is a key point to reach an acceptable trade-off between reliability and design considerations. On the one hand, in fact, these voltages can not be too high in order to minimize read disturbs, namely variations in the FG charge during normal reading of the same cell or of neighbouring ones; on the other, they cannot be too low to avoid the need for sensing excessively small signals. Naturally, the programmed V_T distributions give rise to current distributions, that should be adequately spaced for safe detection. With state-of-the-art technologies, typical values for V_{DR} and V_{GR} are about 1V and 6V, respectively [9]. The required gate voltage should be generated on chip to minimize cost overheads at the system level. To this end, charge-pump based voltage multipliers can be used.

APPLE INC.
EXHIBIT 1008 - PAGE 0384

Extending the approach generally adopted for conventional bilevel Flash memories, the most straightforward sensing technique consists of comparing the current of the selected cell with that of identical reference cells biased in the same read conditions but programmed at suitable threshold voltages so as to provide adequate decision levels. In practice, to read a cell capable of storing $m = 2^n$ levels, we need $m-1$ references, which must be placed midway between adjacent programmed levels. This approach ensures optimal tracking versus process spreads and environmental conditions (in particular, temperature and supply voltage).

Signal recognition in ML memories is obviously more complex than in the bilevel case since, in practice, it implies an A/D conversion, that involves a small number of bits, but must be fast and realized with circuitry requiring minimum area overhead and power consumption. With regard to speed, it should be pointed out that sensing time is only a part of the total access time, where the dominant contribution generally comes from decoding/addressing operations and data output transfer: thus a reasonably low increase in sensing time is not dramatic for the overall performance.

As mentioned above, to carry out the comparison between the cell content and the reference, both current-mode and voltage-mode sensing techniques can be adopted. With the first approach, the cell and the reference currents are applied to the inputs of respective current differential amplifiers, which sense the input current difference and drive cascaded stages so as to provide an output digital voltage. With the second method, the cell and the reference currents are first converted into voltage signals, which are then applied to sense amplifiers capable to detect and amplify the input voltage difference. In general, current-mode techniques seem very attractive for low-voltage analog applications, especially in the presence of large capacitive loads, such as very long bit-lines, and when using submicron devices, which can provide modest small-signal voltage gain [5, 47]. Designing sense amplifiers for ML Flash memories follows similar criteria as in the case of bilevel sensing. However, more stringent requirements have to be met in terms of sensitivity and speed, as much smaller differential input signals are available.

6.3.2 Sensing Schemes

Several architectural approaches can be used for ML sensing. Basically, three sensing methodologies have been proposed: parallel-, serial- and mixed serial-parallel-sensing, the last one being suitable for a large number of programmed levels.

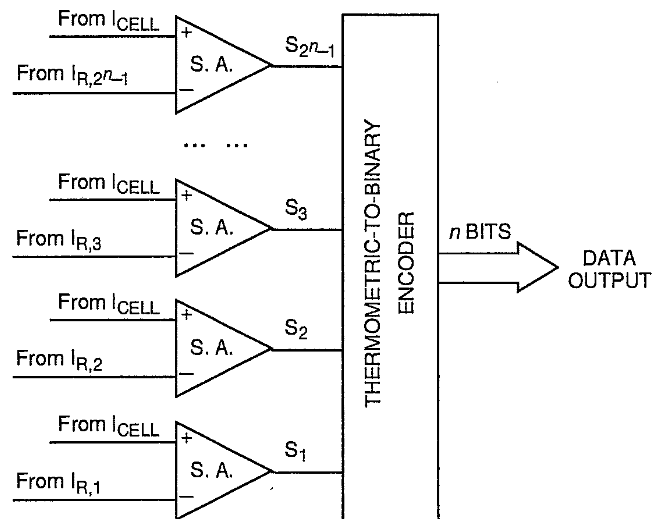


Figure 6.6 Parallel sensing architecture.

Parallel Sensing. Parallel sensing [6, 10, 19] follows a flash conversion approach: the current of the selected cell is compared simultaneously with $2^n - 1$ reference currents (Fig. 6.6). The thermometric code delivered by the bank of $2^n - 1$ sense amplifiers is converted into a binary code by a simple digital encoder.

A single comparison step is required to carry out a complete sensing, which allows for very high sensing speed. However, the number of sense amplifiers increases exponentially with n, with a corresponding increase in silicon area and power consumption. Moreover, careful attention must be paid in distributing the information of the cell contents to all sense amplifiers. In particular, suitable accuracy is required and kickback effects due to the fast switching of the amplifiers [44] must be prevented, especially when fast structures based on regenerative feedback are used.

A particular case of the parallel approach uses a "level identifying" technique based on small-area read-out circuits combined with a "winner-take-all" discriminator [38, 39]. This approach, which has however been proposed for non-verified embedded memories, determines the array cell content by sensing the minimum Euclidean distance between cell and reference currents.

Serial Sensing. The basic principle of serial sensing is to sequentially compare the cell current with a reference current which is varied at each comparison step according to a predetermined law. As a single comparison is carried out at each step, a single comparator is required. This minimizes area occupation

APPLE INC.
EXHIBIT 1008 - PAGE 0386

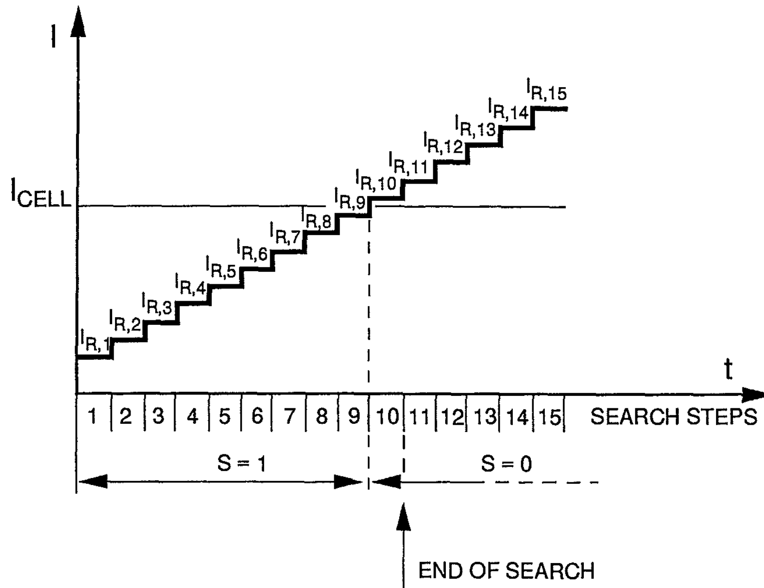


Figure 6.7 Search sequence for the sequential-serial sensing technique (16-level cells). $I_{\mathrm{R},i}$ represents the i-th reference current; S is the sense amplifier output.

and power consumption. However, additional circuitry is required to control a complete read operation, that is more complex than in the case of parallel sensing. The serial approach also eliminates any problem of distributing the cell current to many sense amplifiers as well as kickback effects.

Two different serial sensing approaches have been proposed, i.e., sequential-serial and dichotomic-serial sensing. The sequential-serial technique can be derived from the sensing method presented for ML DRAMs in [24]. The cell current is successively compared with increasing (decreasing) reference currents starting from the lowest (or the highest) one. Sensing is stopped when the reference current becomes larger (smaller) than the cell current (Fig. 6.7). The basic disadvantage of this approach is that the sensing time can be very large: in the worst-case, $2^n - 1$ sensing steps must be performed. Moreover, sensing time depends on the level stored in the selected cell.

For ML Flash memories, the sequential serial scheme has been implemented by successively applying different gate voltages to the selected word-line (V_{WL1}, V_{WL2}, V_{WL3} in Fig. 6.3c) [27], and comparing the obtained cell current with a given fixed reference. The main advantage is that multiple reference circuits are not required. This approach is particularly suitable for page-mode applications, where the simplicity of the page buffer is a key factor. Sensing speed

APPLE INC.
EXHIBIT 1008 - PAGE 0387

is inherently lower as compared to fixed-gate biasing sensing methods due to settling time requirements for the varying read voltage.

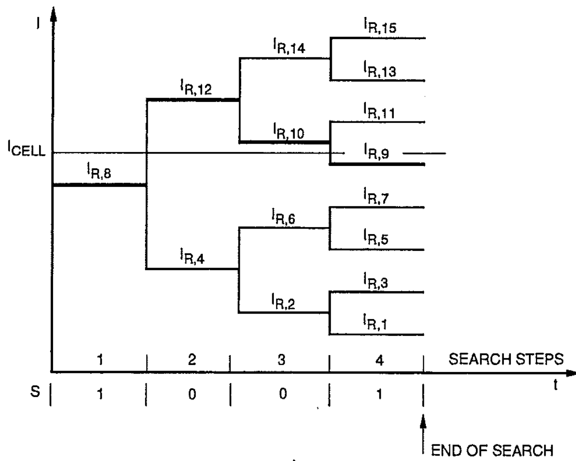
The dichotomic-serial (or binary-search) sensing technique differs from the sequential serial method in the law used to vary the reference current [3, 7, 41], that follows a successive-approximation conversion concept. The current range allowed is divided into two equal subranges (binary division, or "dichotomy"). A comparison with a first reference current, allocated in the middle of the entire range, detects in which sub-range the cell current lies. The detected sub-range is again divided into two equal parts, and a new comparison step determines to which part the cell current belongs. This procedure is iterated until the finest range containing the cell current has been detected (Fig. 6.8a). A single sense amplifier is successively used for all comparisons, while a successive-approximation register (SAR) stores the result of each comparison and controls the selection of the reference current used at each search step (Fig. 6.8b).

The number of steps required to complete a read operation is equal to the number, n, of stored bits per cell. The sensing time is higher with respect to the parallel sensing approach, as it increases linearly with n. On the other hand, the dichotomic-serial technique can provide more efficient area occupation, especially in the case of memories with more than 2 bits per cell. Indeed, while the complexity of this solution increases sub-linearly with n, that of the parallel-sensing circuitry shows a substantially exponential dependence on the same parameter. In addition, some parts of the logic circuitry implementing the dichotomic technique (e.g., timing signal generators) can be shared by all sensing blocks.

Parallel- and dichotomic-serial- sensing schemes seem suitable for different application fields [11]. In 4-level memories, the parallel approach seems to be more attractive because of its low sensing time and reasonable area and power consumption overhead. On the other hand, parallel sensing is less suited to memories with a larger number of bits per cell, because circuit complexity can become unacceptable. A niche product application very attractive for the dichotomic-serial approach is the field of page-mode Flash memories, as the increase in sensing time does not affect read-out throughput.

Mixed Serial-Parallel Sensing. For NV memories storing more than 2 bits per cell, the high sensing time inherent in the dichotomic-serial approach can represent a serious concern, and the ensuing increase in data access time can be unacceptable for many applications. On the other hand, the circuit complexity of the parallel sensing technique also constitutes a severe drawback. For this case, a mixed technique [8] has been proposed, which is a combination of the parallel and the dichotomic-serial approaches (Fig. 6.9).

APPLE INC.
EXHIBIT 1008 - PAGE 0388



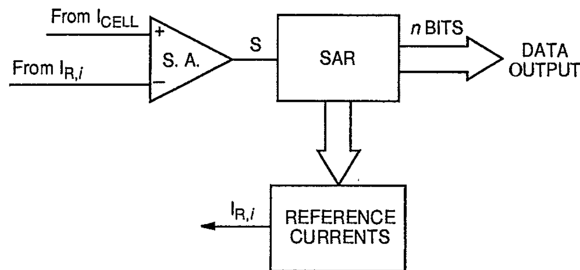
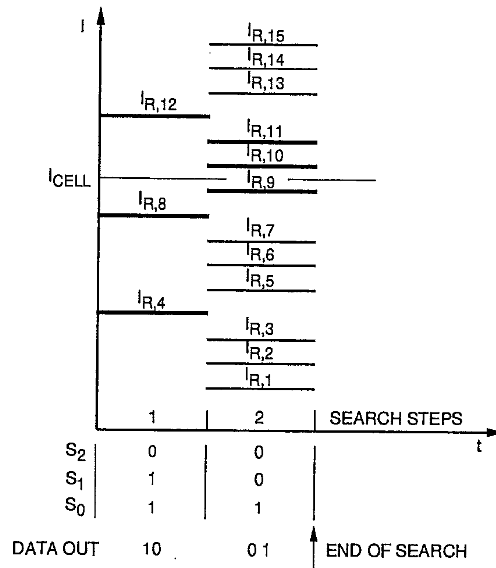


Figure 6.8 Dichotomic-serial sensing technique: a) sensing sequence for 16-level cells ($I_{R,i}$ and S are defined as in Fig. 6.7); b) architecture.

A serial search is performed following the dichotomic algorithm, however each search step is carried out with a parallel approach. For instance, in the case of 16-level cells, reading is achieved in two successive steps, each performing three parallel comparisons. The first step carries out a 2-bit coarse conversion of the cell content, and the second gives the 2 fine bits. In general, assuming that the same number of bits (z) is detected at each step, sensing an n-bit cell is carried out in n/z steps. The total comparator count is $k = 2^z - 1$.

The mixed technique provides a reasonable trade-off between the performance of the parallel and dichotomic-serial approaches in terms of sensing speed and circuit complexity.

APPLE INC.
EXHIBIT 1008 - PAGE 0389



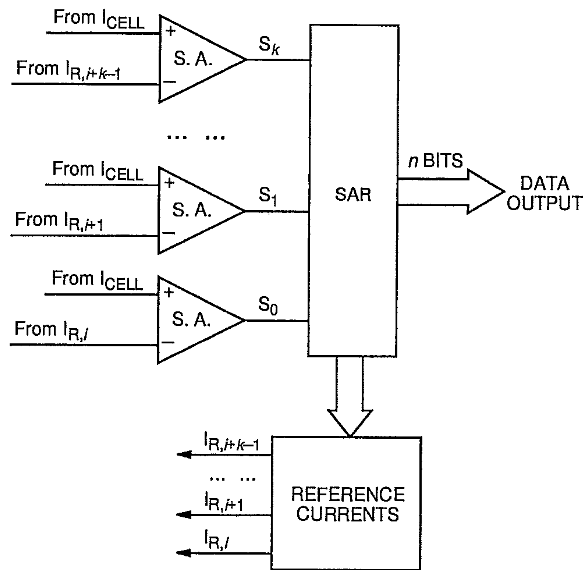


Figure 6.9 Mixed parallel-serial sensing technique: a) sensing sequence for 16-level cells ($I_{R,i}$ is defined as in Fig. 6.7; S_j represents the output of the j-th sense amplifier); b) architecture ($k = 2^z - 1$ is the number of sense amplifiers).

APPLE INC.
EXHIBIT 1008 - PAGE 0390

A summary of the main features of the different sensing approaches is provided in Tab. 6.1.

Table 6.1 Performance summary of different sensing schemes (n bits per cell).

	sensing steps	sense amplifiers	kickback compensation
Parallel	1	$2^n - 1$	YES
Serial-Dichotomic	n	1	NO
Mixed ($z$bits per step)	n/z	$2^z - 1$	YES

6.4 MULTILEVEL PROGRAMMING

The basic goal of ML programming schemes is to implement suitable algorithms and on-chip circuitry able to achieve adequately narrow and spaced V_T distributions in a reasonable time. Obviously, to allocate more than two threshold levels within a predetermined V_T window, much more stringent requirements than in the case of conventional bilevel memories must be met. In particular, the distribution width of each programmed level becomes more critical, and this generates the need for an accurate control of the programmed V_T of FG transistors, that finds applications even beyond the field of digital NV memories [36].

For a population of nominally identical cells, accurate V_T programming requires precise control of charge transfer into/from the FG, and hence of the oxide current I_{OX}, for any given program time. Since, on the other hand, for any given cell I_{OX} (due to either CHE injection or FN tunneling) depends only on the cells bias conditions, in principle precise V_T values could be achieved by accurately controlling applied voltages and program times (the specific values of these parameters depending, however, on process spreads).

If the cell bias is kept fixed during programming, as electrons are moved through the oxide, the FG potential changes and I_{OX} decreases. Thus, charge transfer becomes progressively less efficient. Fig. 6.10 shows the program curves (ΔV_T and I_{OX} vs. time) typical of Flash memory cells programmed by means of CHE injection, but a qualitatively similar behavior is also found in the case of FN tunneling.

For the same drain voltage and pulse duration, a higher program gate voltage (V_{GP}) results in a stronger oxide current, and hence in a larger threshold voltage shift ΔV_T. For any given cell, after an initial transient, a linear relationship exists between the applied V_{GP} and the obtained ΔV_T, evaluated with respect

APPLE INC.
EXHIBIT 1008 - PAGE 0391

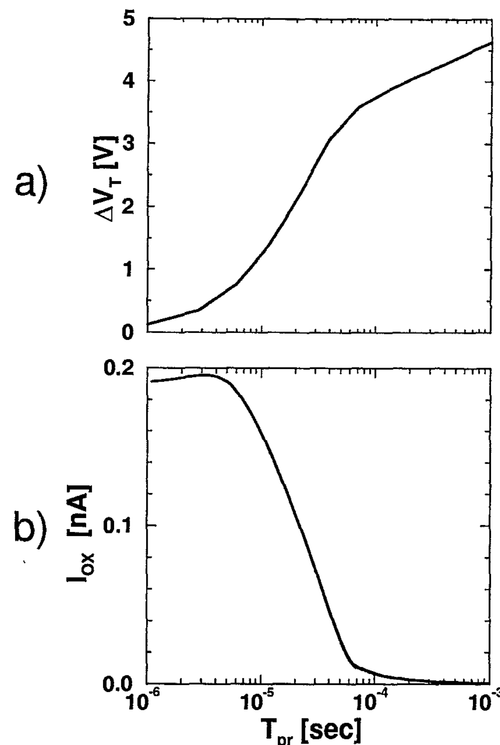


Figure 6.10 Typical programming curves of Flash memories, showing the threshold voltage shift (a) and the estimated injection current (b) as a function of time.

to the threshold voltage of the UV-erased cell [9, 15]. However, the absolute value of ΔV_T for a given value of V_{GP} strongly depends on process parameters, and the convergence of V_T to its target value is too slow: thus, in practice, this simple method cannot provide adequately reproducible and narrow V_T distributions. Consequently, some adjustment in the program conditions must be done to account for the real characteristics of individual cells.

This result can be achieved in different ways. A first type of solution consists of using self-convergent programming mechanisms providing a well defined final state, independently of the cell actual characteristics. From this point of view, CHE injection offers some (theoretical) possibilities [59], whose real viability, however, is still to be investigated.

The other technique, universally used to obtain the necessary precision, is that of verifying the result of programming, that is continued until the target V_T is actually reached. Such an approach, however, can have different

APPLE INC.
EXHIBIT 1008 - PAGE 0392

implementations according to whether the verification is done under reading conditions or during programming itself.

In the former case, which is a straightforward extension of the P&V approach generally followed in bilevel NV memories, as mentioned in the Introduction, programming is divided in a number of partial steps, and at the end of each of them the cell is read with the same circuitry that will be used for normal (sensing) operation, in order to determine whether or not V_T has reached the target values. If this is not the case, another programming step is performed and the whole procedure is repeated until successful completion. In applying this scheme, the obvious choice is a cell-by-cell P&V approach, that is traditionally used with both NOR and NAND architectures [32, 53, 57] and is also compatible with parallel programming. With cell-by-cell P&V procedures, as the threshold voltages of different cells are controlled individually, endurance issues due to the V_T window closure are not a major concern, provided that sufficiently long program times are allowed, because the target value of V_T can always be reached, although with a variable number of program steps. In fact, the achievable accuracy ideally depends only on the quantization error inherent to the use of finite program steps, although if these are made small (for high precision) program time can become excessively long. Beside that concerning endurance, P&V programming offers further reliability advantages due to the fact that only the minimum required charge flows through the oxide.

Alternatively, program verification can be done during a unique program operation, which automatically stops when the target V_T is reached. As compared with standard P&V schemes, this type of methods offers a better trade-off between accuracy and program time, while maintaining all the advantages in terms of reliability, and may require simpler control circuitry (and hence smaller area overhead). However, cell sensing is made under program conditions, and therefore significant differences can occur between verification and normal cell reading, with obvious effects on the effective accuracy of the whole operation. A particularly interesting case of this category is the self-converging technique, whereby the oxide current extinguishes spontaneously when the charge on the floating gate reaches a given value.

The rest of this section is dedicated to an overview a number of specific program methods proposed for ML Flash memories, which follow P&V and self-controlled approaches, respectively.

6.4.1 Program-and-Verify Approaches

Staircase Gate Voltage Ramp Programming. A P&V technique suitable for ML Flash memories using CHE injection exploits the above mentioned linear relationship between ΔV_T and V_{GP} for any given cell by using a staircase

APPLE INC.
EXHIBIT 1008 - PAGE 0393

waveform for the gate voltage, that is increased at each program step by a fixed amount ΔV_{GP}. Fig. 6.11 clearly shows that, for any channel length, after a small number of initial steps, the V_T increase after each program pulse, $\delta V_{T,p}$, is equal to the program gate voltage increase ΔV_{GP}. Similar results are obtained when cell width or oxide thickness spreads are considered. Indeed, it can be demonstrated [9] that the one-to-one correspondence between $\delta V_{T,p}$ and ΔV_{GP} after the first steps, is not affected by process variations, so that this method allows accurate cell programming by using a reasonable number of pulses.

The above concept is also valid in the case of FN programming. For this reason, staircase voltage ramp programming [57] has also been proposed for the case of ML Flash memories making use of FN tunneling [14, 23, 27, 48, 51, 54]. The program curves obtained in [27] are shown in Fig. 6.12.

P&V staircase programming should theoretically give V_T distribution widths not larger than ΔV_{GP} independently of the number of cells programmed to that state. Indeed, neglecting errors due to sense amplifier accuracy or voltage fluctuations, the last program pulse applied to a cell will cause its threshold voltage to be shifted above the decision level used for verification by an amount at most as large as ΔV_{GP}.

Using small incremental steps for V_{GP}, V_T distributions adequate for a large number of V_T levels can be achieved. However, this also results in increased program time, and optimum trade-offs between the required program time and the number of usable levels and/or suitable architectural solutions have to be found. In this respect, NAND multi-page architectures [52] have been proposed to increase program throughput.

Drain Voltage Programming. When programming by means of drain-side FN tunneling, the oxide field F_{OX} (hence also I_{OX}) can be varied by changing V_{GP} and/or the drain voltage V_{DP}. In particular, for parallel ML programming a technique called Drain-voltage Controlled Multilevel Programming (DCMP) [41] has been proposed, with which target V_T values are achieved by applying suitable drain voltages V_{DP}, while keeping V_{GP} at an adequate constant value (Fig. 6.13). Memory cells belonging to the selected word-line can be programmed to different levels in the same program period by providing the required different V_{DP} voltages to each bit-line: $m-1$ different bit-line voltages are used for simultaneously programming $m-1$ different levels. Each program step has a fixed time duration. Fig. 6.13 shows that a V_T difference in the range of 1.5V between the central values of adjacent states is obtained after a 100-μs program time by setting V_{GP} equal to -9V and V_{DP} in the range from 0V to 6V. A P&V approach using a Parallel Multilevel Verify (PMV)

APPLE INC.
EXHIBIT 1008 - PAGE 0394

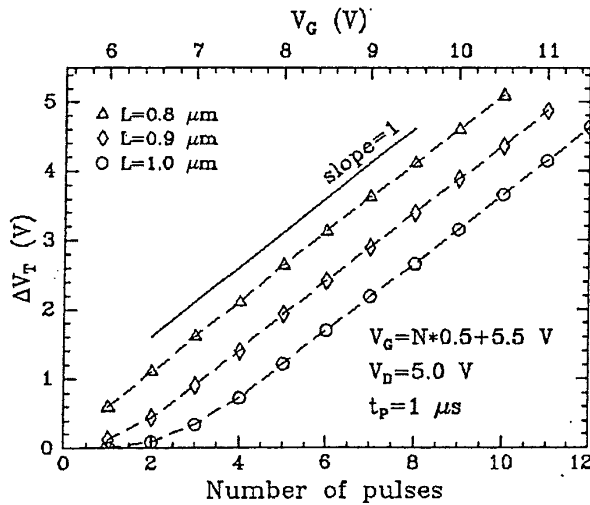


Figure 6.11 Threshold shift as a function of the number of programming pulses when using an adaptive staircase gate voltage algorithm for CHE programming (the gate voltage step is 0.5V; the program gate voltage at each step is shown in the upper axis) [9]. Three different cell channel lengths were used.

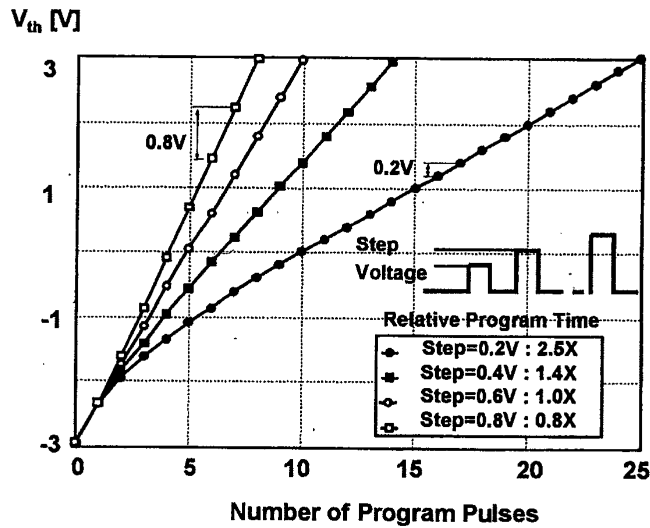


Figure 6.12 Threshold shift (V_{th}) as a function of the number of programming pulses when using an adaptive staircase gate voltage algorithm for FN programming [27]. Three different staircase steps were used.

APPLE INC.
EXHIBIT 1008 - PAGE 0395

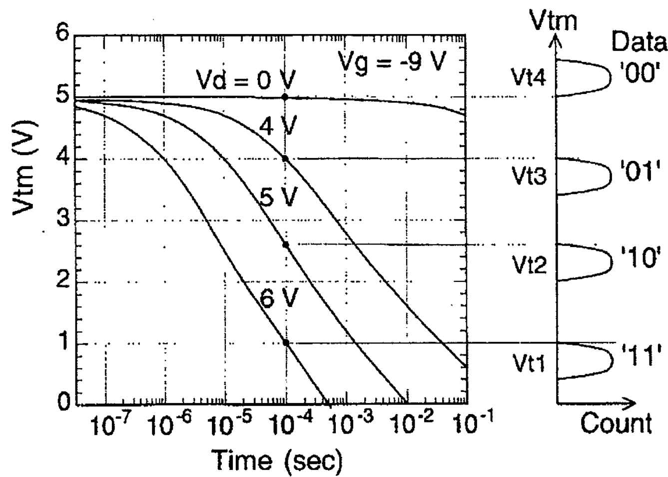


Figure 6.13 Programming curves obtained using the Drain-voltage Controlled Multilevel Programming technique [41].

scheme is adopted in [41] to achieve the required V_T control of the cells being programmed.

6.4.2 *Self-Controlled Approaches*

As mentioned above, an attractive way to speed up programming and eliminate the need for iterative P&V sequences, thus simplifying additional circuitry (and hence area overhead), consists in controlling the cell V_T while this is being programmed. This concept can be exploited with both CHE and FN programming and has been investigated in pioneering studies. It has not yet been applied in real devices but could become important in the future, especially for embedded Flash memories, where simplicity and speed can be decisive factors.

Drain Current Monitoring. When using CHE programming, for fixed values of V_{GP} and V_{DP}, a one-to-one relationship exists between V_T and the drain current I_{DP} for any given cell. Thus, such a current can be monitored during programming to determine when the target V_T has actually been reached. The accuracy of the whole scheme depends on the characteristics (offsets, parameter dispersion, etc.) and speed of the circuitry used to monitor I_{DP}, however it should be taken into account that verification is performed in different conditions than normal reading. With accurate design and cumulative expertise, acceptable precision should be achieved. So far, however, this method, has been

APPLE INC.
EXHIBIT 1008 - PAGE 0396

implemented only in a split-gate bilevel Flash memory [13], that has provided interesting results, but not (yet) the accuracy needed for ML applications.

Self-Converging Programming. The principle of self-converging approaches is to exploit a mechanism where the oxide current I_{OX} extinguishes spontaneously when a given threshold voltage $V_{T,FIN}$ is reached, thereby automatically completing the programming procedure. If $V_{T,FIN}$ is made dependent only on very few (ideally one) control parameters such as reference voltages and independent of process and environment parameters (provided that the used voltages are sufficiently high as to activate the required programming mechanism), this technique can achieve a precise control of the cell V_T with no need for a sequence of program-verify steps.

As an example of self-converging techniques, in this section we will describe the solution of [4, 49], that has been proposed to obtain low-voltage programming of Flash memory cells, and can also be attractive for ML programming. It exploits inherent characteristics of CHE injection. As mentioned above, when using CHE injection with fixed drain and gate bias, the injection current decreases in time (as the FG voltage lowers while the stored charge progressively increases) and, if enough time is allowed, I_{OX} eventually vanishes taking the programming operation to a natural limit. Since, of course, all the physical processes of interest are directly controlled only by the FG, such a limit corresponds to a specific value (V_{LIM}) of the FG potential, essentially depending only on technological parameters.

During programming, as the floating-gate voltage V_{FG} lowers, the oxide field F_{OX} driving the electrons toward the FG decreases, until it (normally) changes sign in the region of maximum carrier heating. When this occurs, F_{OX} becomes repulsive for electrons and increases the barrier height for electron injection, that gets progressively more difficult. This process, however, has complementary effects on the energetic holes (generated by impact ionization and heated by the field) which necessarily accompany hot electrons. Thus hole injection probability (much lower than that of electrons at the beginning of programming) becomes progressively larger until hole and electron injection become equal to each other and I_{OX} goes to zero. The role of hot holes in making programming to reach its final limit is important because it helps to make V_{LIM} well defined, as conceptually shown in Fig. 6.14.

Since V_{LIM} does not depend on the gate voltage V_{GP}, if programming is left to extinguish spontaneously, the charge on the FG would be $Q_{FG} = (V_{GP} - V_{LIM})C_{PP}$, where C_{PP} represents the capacitance between the control gate CG and FG. Thus, Q_{FG} varies (linearly) with the control-gate voltage V_{GP}. If different values of V_{GP} are used, this method could be exploited for ML programming (although so far it had not been proposed for this purpose).

APPLE INC.
EXHIBIT 1008 - PAGE 0397

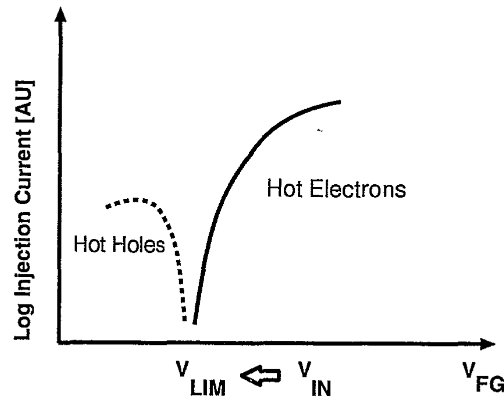


Figure 6.14 Conceptual representation of injection currents (due to hot electrons and hot holes) as a function of the voltage of the transistor floating gate, V_{FG}.

In practice, however, several drawbacks make the actual applicability of such a method seriously doubtful: 1) this programming mechanism is very slow (since I_{OX} becomes vanishingly small when V_{FG} approaches V_{LIM}); 2) it leads to serious reliability problems (since the simultaneous injection of electrons and holes represents the most critical situation for oxide reliability); 3) its accuracy in terms of V_{T} distributions is limited by the dispersion of V_{LIM} values due to technology. The first two of these items are particularly important and difficult to overcome, at least for mainstream applications, were speed and reliability represent primary targets.

A self-controlled programming method has also been proposed for the FN tunneling mechanism, which is able to stop automatically when the cell threshold voltage reaches the target value [33, 34, 35]. This technique allows conceptually simple implementation, and provides a linear dependence of the final V_{T} on the fixed voltage applied to the control gate during programming. Since this method has been proposed as particularly suited to full-featured EEPROMs, and needs further investigation to assess its suitability to Flash memories, it will not be described here.

6.5 CONCLUSIONS

The ML storage concept allows the traditional one-to-one correspondence between cell count and memory size to be overcome. When coupled to technology scaling down, this approach can significantly accelerate the reduction in the memory cost-per-bit. However, optimum trade-offs must be searched between the conflicting aspects of density (in terms of stored bits per cell) and noise im-

APPLE INC.
EXHIBIT 1008 - PAGE 0398

munity (in a general sense, including sensing and programming accuracy as well as reliability aspects), aiming at the best use of silicon area, of course depending on issues such as technology, cell structure, array architecture, algorithms, circuits and reliability.

Table 6.2 Performance summary of ML Flash experimental prototypes (2 bit/cell).

	[3]	[41]	[27]
Memory Size	32Mbit	64Mbit	128Mbit
Array Structure	NOR	NOR	NAND
Program Method	CHE	Drain-side FN	Channel FN
Erase Method	Source-sideFN	Channel FN	Channel FN
Process	0.6μm 2-Well 2-Metal 2-Poly	0.4μm 3-Well 2-Metal 2-Poly	0.4μm 3 Well 1-Metal 2-Poly
Cell Size	3.6μm^2	1.47μm^2 (*)	1.1μm^2 (*)
Die Size	152mm^2	98mm^2	117mm^2
Supply Voltage	5V	3.3V	3.3V
Programming Scheme	Pulsed Verified	Drain Voltage Verified	Staircase Gate Voltage Verified
Program Time	40μs	100μs	900μs
Program Throughput	0.10MByte/s	0.16MByte/s	0.50MByte/s
Erase Time	not available	1ms per block	6ms per block
Erase Block	128KBytes	1Kb,16Kb,64Kb	16KBytes
Sensing Scheme	Dichotomic	Dichotomic	Sequential-Serial
Access Time	120ns	80ns	25μs (random page) 25ns (burst)
Read Throughput	16MByte/s	25MByte/s	14MByte/s

(*) Effective cell size, also considering select transistors.

The practical feasibility of 4-level NV storage with present CMOS fabrication technology has been demonstrated by experimental prototypes as well as

APPLE INC.
EXHIBIT 1008 - PAGE 0399

by technology, reliability and design considerations. Three 4-level multimegabit Flash experimental chips have been presented so far in the literature. A comparison performance between these prototypes is shown in Tab. 6.2. It can be seen that the device performance pays some penalty with respect to conventional Flash memories (for instance in terms of access and program time), however the main goal of increased integration density has been achieved. Moreover, future improvements are expected. The 5-V 64-Mbit 2-bit/cell commercial device referred to in [2, 19] is based on the prototype [3], however it follows a parallel sensing approach. This chip is only 5% larger than the 32-Mbit bilevel cell device in 0.4-μm fabrication technology, and features 150-ns access time and 0.16 MByte/s program throughput.

Reliability is an important constraint: even though no specific failure mechanism is foreseen for ML Flash memories, ML storage is more critical than the conventional bilevel storage due to the larger threshold window and/or the reduced spacing between adjacent levels. The practical exploitation of the ML storage concept for mass production of Flash memories will probably require some error correction technique, mainly to cope with the oxide degradation induced by program/erase cycling, even though for 2 bit/cell storage this does not seem to be necessarily true [2].

Storage of more than 2 bits per cell is in principle feasible, however new technology and design solutions are needed to achieve the performance and reliability targets required for digital applications. Obtaining narrower V_T distributions and increasing the read gate voltage is the way, but this leads to increased program time and larger read disturbs, respectively. Oxide reliability must be improved, especially from endurance standpoint. New sensing schemes and circuits must be designed to allow fast and correct sensing. Improved program algorithms must be developed to ensure the required accuracy in charge placement with reasonable program throughput.

Therefore, many research efforts are presently being devoted to study the limits and the practical convenience of the ML storage approach, to be able to fully exploit the potential advantages of reduced cost-per-bit.

References

[1] Aritome S., Takeuchi Y., Sato S., Watanabe H., Shimizu K., Hemink G. and Shirota R. (1995) "A novel side-wall transfer-transistor cell (SWATT cell) for multi-level NAND EEPROMs", *IEDM 1995 Tech. Dig.*, pp. 275–278.

[2] Atwood G., Fazio A., Mills D. and Reaves B. (1997) "Intel StrataFlash™ memory technology overview", *Intel Technology Journal*, Q4.

[3] 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", *1995 IEEE ISSCC Dig. Tech. Papers*, pp. 132–133, 351.

[4] Bergemont A., Chi M. and Haggag H. (1996) "Low voltage NVG: A new high performance 3V/5V Flash technology for portable computing and telecommunications applications", *IEEE Trans. Electron Devices*, **43**, 9, pp. 1510–1517.

[5] Blalock T.N. and Jaeger R.C. (1991) "A high-speed clamped bit-line current-mode sense amplifier", *IEEE J. Solid-State Circuits*, **26**, 4, pp. 542–548.

[6] Bleiker A. and Melchior H. (1987) "A four-state EEPROM using floating-gate memory cells", *IEEE J. Solid-State Circuits*, **22**, 3, pp. 460–463.

[7] Calligaro C., Daniele V., Gastaldi R., Manstretta A. and Torelli G. (1995) "A new serial sensing approach for multistorage non-volatile memories", *Proc. 1995 IEEE Int. Workshop on Memory Technology, Design and Testing*, pp. 21–26.

[8] Calligaro C., Manstretta A., Rolandi P. and Torelli G. (1996) "Mixed sensing architecture for 64Mbit 16-level-cell non-volatile memories", *Proc. 1996 IEEE Int. Conf. Innovative Systems in Silicon*, pp. 133–140.

[9] Calligaro C., Manstretta A., Modelli A. and Torelli G. (1996) "Technological and design constraints for multilevel Flash memories", *Proc. 3rd IEEE Int. Conf. on Electronics, Circuits and Systems*, pp. 1003–1008.

[10] Calligaro C., Gastaldi R., Manstretta A. and Torelli G. (1997) "A high-speed parallel sensing scheme for multi-level non-volatile memories", *Proc. 1997 IEEE Int. Workshop on Memory Technology, Design and Testing*, pp. 96–99.

[11] Calligaro C., Manstretta A., Pierin A. and Torelli G. (1997) "Comparative analysis of sensing schemes for multilevel non-volatile memories", *Proc. 1997 IEEE Int. Conf. Innovative Systems in Silicon*, pp. 266–273.

[12] Carley L.R. (1989) "Trimming analog circuits using floating-gate analog MOS memory", *IEEE J. Solid-State Circuits*, **24**, 6, pp. 1569–1575.

[13] Cernea R., Lee D.J., Mofidi M., Chang E.Y., Chien W.-Y., Goh L., Fong Y., Yuan J.H., Samachisa G., Guterman D.C., Mehrotra S., Sato K.,

APPLE INC.
EXHIBIT 1008 - PAGE 0401

Onishi H., Ueda K., Noro F., Miyamoto K., Morita M., Umeda K. and Kubo K. (1995) "A 34Mb 3.3V serial Flash EEPROM for solid-state disk applications", *1995 IEEE ISSCC Dig. Tech. Papers*, pp. 126–127, 350.

[14] Choi Y.-J., Suh K.-D., Koh Y.-N., Park J.-W., Lee K.-J., Cho Y.-J. and Suh B.-H. (1996) "A high speed program scheme for multi-level NAND Flash memory", *1996 Symp. VLSI Circuits Dig. Tech. Papers*, pp. 170–171.

[15] de Graaf C., Young P. and Hulsbos D. (1995) "Feasibility of multilevel storage in Flash EEPROM cells", *Proc. ESSDERC '95*, pp. 213–216.

[16] Di Cataldo G. and Palumbo G. (1993) "Double and triple charge pump for power IC: dynamic models which take parasitic effects into account", *IEEE Trans. Circuits Syst.*, **40-I**, 2, pp. 92–101.

[17] Dickson J. (1976) "On-chip high-voltage generation in MNOS integrated circuits using an improved voltage multiplier technique", *IEEE J. Solid-State Circuits*, **11**, 3, pp. 374–378.

[18] Eitan B., Kazerounian R., Roy A., Crisenza G., Cappelletti P., Modelli A. (1996) "Multilevel Flash cells and their trade-offs", *IEDM 1996 Tech. Dig.*, pp. 169–172.

[19] Fazio A. and Bauer M. (1997) "Intel StrataFlash™ memory technology development and implementation", *Intel Technology Journal*, Q4.

[20] Fujita O. and Amemiya Y. (1993) "A floating-gate analog memory device for neural networks", *IEEE Trans. Electron Devices*, **40**, 11, pp. 2029–2035.

[21] Gastaldi R., Novosel D., Dallabora M. and Casagrande G. (1988) "A 1Mbit CMOS EPROM with enhanced verification", *IEEE J. Solid-State Circuits*, **23**, 5, pp. 1150–1156.

[22] Geppert L. (1998) "Solid State", in "Technology 1998 Analysis & Forecast", *IEEE Spectrum*, **35**, 1, pp. 23–28.

[23] Hemink G.J., Tanaka T., Endoh T., Aritome S. and Shirota R. (1995) "Fast and accurate programming method for multi-level NAND EEPROMs", *1995 Symp. VLSI Technology Dig. Tech. Papers*, pp. 129–130.

[24] Horiguchi M., Aoki M., Nakagome Y., Ikenaga S. and Shimohigashi K. (1988) "An experimental large-capacity semiconductor file memory using 16-levels/cell storage", *IEEE J. Solid-State Circuits*, **23**, 1, pp. 27–33.

APPLE INC.
EXHIBIT 1008 - PAGE 0402

[25] Hu C.Y., Kencke D.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 Dev. Letters*, **16**, 11, pp. 500–502.

[26] Jung T.-S., Choi 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 3.3V 128Mb multilevel NAND Flash memory for mass storage applications", *1996 IEEE ISSCC Dig. Tech. Papers*, pp. 32–33, 412.

[27] Jung T.-S., Choi 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 117-mm^2 3.3V only 128-Mb multilevel NAND Flash memory for mass storage applications", *IEEE J. Solid-State Circuits*, **31**, 11, pp. 1575–1583.

[28] Kencke D.L., Richart R., Garg S. and Banerjee S.K. (1996) "A sixteen level scheme enabling 64Mbit Flash memory using 16Mbit technology", *IEDM 1996 Tech. Dig.*, pp. 937–939.

[29] Kim K.-H. and Lee K. (1998) "An 8b resolution 360μs write time nonvolatile analog memory based on differentially balanced constant-tunneling-current scheme (DBCS)", *1998 IEEE ISSCC Dig. Tech. Papers*, pp. 336–337, 459.

[30] Kosaka H., Shibata T., Ishii H. and Ohmi T. (1995) "An excellent weight-updating linearity EEPROM synapse memory cell for self-learning neuron-MOS neural networks", *IEEE Trans. Electron Devices*, **42**, 1, pp. 135–143.

[31] Kramer A., Fabbrizio V., Mariaud X. and Raynal F. (1998) "1.5TXPS convolver using 5b analog Flash for real-time large-kernel image filtering", *1998 IEEE ISSCC Dig. Tech. Papers*, pp. 196–197, 437.

[32] Kynett V.N., Fandrich M.L., Anderson J., Dix P., Jungroth O., Kreifels J.A., Lodenquai R.A., Vajdic B., Wells S., Winston M.D. and Yang L. (1989) "A 90-ns one-million erase/program cycle 1-Mbit Flash memory", *IEEE J. Solid-State Circuits*, **24**, 5, pp. 1259–1264.

[33] Lanzoni M., Briozzo L. and Riccò B. (1994) "A novel approach to controlled programming of tunnel-based floating-gate MOSFET's", *IEEE J. Solid-State Circuits*, **29**, 2, pp. 147–150.

APPLE INC.
EXHIBIT 1008 - PAGE 0403

[34] Lanzoni M. and Riccò B. (1995) "Experimental characterization of circuits for controlled programming of floating gate MOSFET's", *IEEE J. Solid-State Circuits*, **30**, 6, pp. 706–709.

[35] Lanzoni M., Tondi G., Galbiati P., Riccò B. (1996) "Non-volatile EEPROM cells for analog circuit calibration", *Proc. ESSDERC '96*, pp. 135–138.

[36] Lanzoni M., Tondi G., Galbiati P. and Riccò B. (1998) "Automatic and continuous offset compensation of MOS operational amplifiers using floating-gate transistors", *IEEE J. Solid-State Circuits*, **33**, 2, pp. 287–290.

[37] Montanari D., Van Houdt J., Wellekens D., Hendrickx P., Groeseneken G. and Maes H.E. (1996) "Comparison of the suitability of various programming mechanisms used for multilevel non-volatile information storage", *Proc. ESSDERC '96*, pp. 139–142.

[38] Montanari D., Van Houdt J., Wellekens D., Groeseneken G. and Maes H.E. (1997) "Novel small-area read-out circuit for multi-level memories", *16-th IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, California, paper 6.2.

[39] Montanari D., Van Houdt J., Groeseneken G. and Maes H.E. (1998) "Novel level-identifying circuit for Flash multilevel memories", *IEEE J. Solid-State Circuits*, **33**, 7, pp. 1090–1095.

[40] Ohkawa M., Sugawara H., Sudo N., Tsukiji M., Nakagawa K., Kawata M., Oyama K., Takeshima T. and Ohya S. (1996) "A 98mm^2 3.3V 64Mb Flash memory with FN-NOR type 4-level cell", *1996 IEEE ISSCC Dig. Tech. Papers*, pp. 36–37, 413.

[41] Ohkawa M., Sugawara H., Sudo 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 J. Solid-State Circuits*, **31**, 11, pp. 1584–1589.

[42] Oyama K., Shirai H., Kodama N., Kanamori K., Saitoh K., Hisamune Y.S. and Okazawa T. (1992) "A novel erasing technology for 3.3V Flash memory with 64Mb capacity and beyond", *IEDM 1992 Tech. Dig.*, pp. 607–610.

[43] Pavan P., Bez R., Olivo P. and Zanoni E. (1997) "Flash memory cells - An overview", *Proc. IEEE*, **85**, 8, pp. 1248–1271.

APPLE INC.
EXHIBIT 1008 - PAGE 0404

[44] Razavi B. (1995) *Principles of Data Conversion System Design*, IEEE Press, Piscataway, NJ (Chapters 6,7).

[45] Rolandi P.L., Canegallo R., Chioffi E., Gerna D., Guaitini G., Issartel C., Lhermet F., Pasotti M. and Kramer A. (1998) "1M-cell 6b/cell analog Flash memory for digital storage", *1998 IEEE ISSCC Dig. Tech. Papers*, pp. 334–335, 459.

[46] Säckinger E. and Guggenbühl W. (1988) "An analog trimming circuit based on a floating-gate device", *IEEE J. Solid-State Circuits*, **23**, 6, pp. 1437–1440.

[47] Seevinck E. (1990) "Analog interface circuits for VLSI", in: *Analogue IC Design: the Current-Mode Approach* (Toumazou C., Lidgey F.J. and Haigh D.G., Eds.), Peter Peregrinus Ltd., London (UK) (Chapter 12).

[48] Shirota R., Hemink G.J., Takeuchi K., Nakamura H. and Aritome S. (1995) "A new programming method and cell architecture for multi-level NAND Flash memories", *14-th IEEE Non-Volatile Semiconductor Memory Workshop*, Monterey, California, paper 2.7.

[49] Shum D.P., Swift C.T., Higman J.M., Taylor W.J., Chang K.-T., Chang K.-M. and Yeargain J.R. (1994) "A novel band-to-band tunneling induced convergence mechanism for low current high density Flash EEPROM applications", *IEDM 1994 Tech. Dig.*, pp. 41–44.

[50] Sin C.K., Kramer A., Hu V., Chu R. and Ko P.K. (1992) "EEPROM as an analog memory device", *IEEE Trans. Electron. Devices*, **39**, pp. 1410–1419.

[51] Takeuchi K., Tanaka T. and Nakamura H. (1996) "A double-level-Vth select gate array architecture for multilevel NAND Flash memories", *IEEE J. Solid-State Circuits*, **31**, 4, pp. 602–609.

[52] Takeuchi K., Tanaka T. and Tanzawa T. (1997) "A multi-page cell architecture for high-speed programming multi-level NAND Flash memories", *1997 Symp. VLSI Circuits Dig. Tech. Papers*, pp. 67–68.

[53] Tanaka T., Tanaka Y., Nakamura H., Sakui K., Oodaira H., Shirota R., Ohuchi K., Masuoka F. and Hara H. (1994) "A quick intelligent page-programming architecture and a shielding bitline sensing method for 3V-only NAND Flash memory", *IEEE J. Solid-State Circuits*, **29**, 11, pp. 1366–1373.

[54] Tanaka T., Tanzawa T. and Takeuchi K. (1997) "A 3.4-Mbyte/sec programming 3-level NAND Flash memory saving 40% die size per bit", *1997 Symp. VLSI Circuits Dig. Tech. Papers*, pp. 65–66.

[55] Tanzawa T., Tanaka V., Takeuchi K., Shirota R., Aritome S., Watanabe H., Hemink G., Shimizu K., Sato S., Takeuchi Y. and Ohuchi K. (1997) "A compact on-chip ECC for low cost Flash memories", *IEEE J. Solid-State Circuits*, **32**, 5, pp. 662–669.

[56] Tanzawa T. and Tanaka T. (1997) "A dynamic analysis of the Dickson charge pump circuit", *IEEE J. Solid-State Circuits*, **32**, 8, pp. 1231–1240.

[57] Torelli G. and Lupi P. (1983) "An improved method for programming a word-erasable EEPROM", *Alta Frequenza*, **LII**, 6, pp. 487–494.

[58] Van Tran H., Blyth T., Sowards D., Engh L., Nataraj B.S., Dunne T., Wang H., Sarin V., Lam T., Nazarian H. and Hu G. (1996) "A 2.5V 256-level non-volatile analog storage device using EEPROM technology", *1996 IEEE ISSCC Dig. Tech. Papers*, pp. 270–271, 458.

[59] 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 1991 Tech. Dig.*, pp. 307–310.

[60] 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 1992 Tech. Dig.*, pp. 595–598.

[61] Yoshikawa K. (1996) "Impact of cell threshold voltage distribution in the array of Flash memories on scaled and multilevel Flash cell design", *1996 Symp. VLSI Technology Dig. Tech. Papers*, pp. 240–241.

APPLE INC.
EXHIBIT 1008 - PAGE 0406

7 FLASH MEMORY RELIABILITY

Paolo Cappelletti, Alberto Modelli

STMicroelectronics, Central R&D
Via Olivetti 2, 20041 Agrate Brianza (Milano), Italy
Paolo.Cappelletti@st.com, Alberto.Modelli@st.com

Abstract: With reference to the mainstream technology, the most relevant failure mechanisms which affect yield and reliability of Flash memory are reviewed, showing the primary role played by tunnel oxide defects. The effectiveness of a good test methodology combined with a proper product design for screening at wafer sort latent defects of tunnel oxide is highlighted as key factors for improving Flash memory reliability. The degradation of device performance induced by program/erase cycling is discussed, covering the behavior of a typical cell, the evolution of memory array distribution, and the single bit failure modes. Oxide traps are demonstrated to be responsible for the most critical failure mechanisms, like the erratic erase and the single bit data loss: the impact of stress-induced leakage current on data retention is shown to limit the scalability of tunnel oxide thickness. Finally, reliability implications of multilevel cell concept are briefly analyzed.

7.1 INTRODUCTION

Flash memory has become in the most recent years the star among non volatile memories because it offers the capability of being electrically erased and re-written, so far featured only by the expensive EEPROM's, at a cost comparable to the one of EPROM's.

Together with the desirable features which make it so attractive, Flash memory unfortunately combines also the yield and reliability issues of EPROM's and EEPROM's (Fig. 7.1), plus some additional ones which are specific of this technology. Writing operations involve both Fowler-Nordheim tunneling and hot

APPLE INC.
EXHIBIT 1008 - PAGE 0407

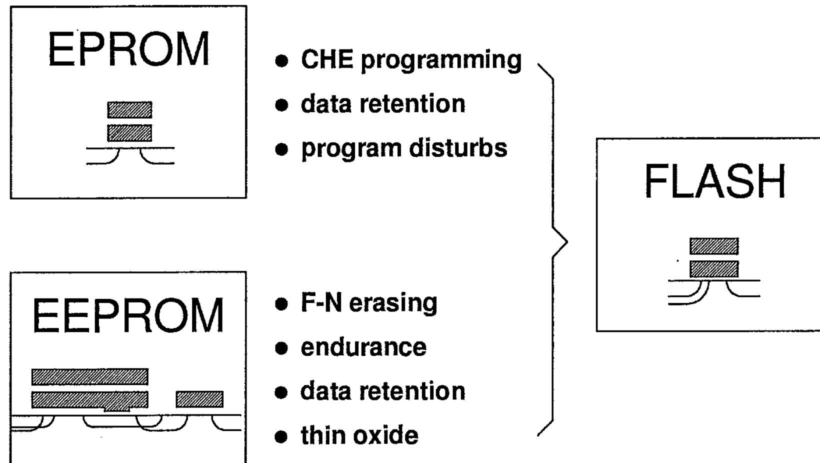


Figure 7.1 Flash memory combines yield and reliability issues of EPROM and EEPROM.

carriers; program/erase cycling endurance must be achieved without degrading data retention; the single transistor structure exposes memory cells to array disturbs and to over-erasure problems.

All these issues make the Flash technology one of the most difficult to be mastered, requiring a very accurate process optimization and a severe process control [1].

The single most important factor contributing to yield and reliability of Flash memory is the quality of tunnel oxide, both in terms of intrinsic properties and defect density.

Uniformity of tunnel current, which determines the width of the erased V_t distribution within a memory array, and oxide wear-out under current stress, which affects memory endurance, are much related to tunnel oxide process conditions. Point defects which cause a local increase of tunnel oxide conductivity are responsible for single bit over-erasure and single bit failures due to program disturb, the most relevant yield problems in Flash memories.

Major efforts have been devoted in the last few years to improve the way of evaluating and monitoring the quality of tunnel oxide, for what concerns both electrical test methodology and the related test structures [2–4].

Moreover, the high degree of testability of Flash memories allows to screen at wafer sorting latent defects which may cause single bit failures related to program disturbs, data retention and premature oxide breakdown.

With reference to the mainstream technology, this chapter will review the major issues impacting yield and reliability of Flash memory, it will discuss in

APPLE INC.
EXHIBIT 1008 - PAGE 0408

more details the failure mechanisms which limit memory endurance and it will introduce the problems related to multilevel storage.

The vastness and the complexity of the subject impose a choice on the way of organizing the presentation; instead of a tedious list of all failure mechanisms reported in the literature, this chapter will provide an insight, based on author's experience, of the key concepts and the most relevant issues, emphasizing the practical implications on process control and product design and testing. It will be like a guided tour in the labyrinth of Flash memory reliability: rather than showing everything, the purpose of the guide will be to show you the way out.

7.2 MEMORY ARRAY V_T DISTRIBUTIONS AND TUNNEL OXIDE "DEFECTS"

The role of tunnel oxide "defects" is so important that a discussion on Flash memory yield and reliability cannot start without a good understanding of what oxide "defects" does mean and how do they impact memory functionality and manufacturability.

In Chapter 2 the program and erasing mechanism of Flash cell has been described in great details, introducing the concept of over-erasing. This section will deal with the erasing of a memory array.

Fig. 7.2 shows typical distributions of cell threshold voltages in a memory array. The UV erased distribution is pretty narrow and symmetrical; a more accurate analysis would reveal a Gaussian distribution due to random variation of critical dimensions, thickness and doping which contribute to cause a dispersion of threshold voltages, either directly or through coupling ratios.

The programmed distribution is wider than the UV erased one but is still symmetrical; the enlargement is because most of the parameters which cause a V_t dispersion of UV erased cells are also impacting the threshold shift of programmed cells.

The distribution of threshold voltages after electrical erase is much wider and heavily asymmetrical. A more detailed analysis (Fig. 7.3) shows that the bulk of this distribution, including over 99% of cells, is again a Gaussian with a standard deviation larger than the one of programmed cells; from now on the bits owning to the Gaussian distribution will be referred to as "normal" bits. The left side of the distribution is made of an exponential tail of cells that erase faster than the average: these bits will be referred to as "tail" bits.

The dispersion of threshold voltages of normal bits is due to coupling ratio variations and it has been accurately modeled by Yoshikawa *et al.* [5]. That model is well confirmed by the comparison of erased V_t distributions obtained on the same array by two different erasing methods (Fig. 7.4); the distribution

APPLE INC.
EXHIBIT 1008 - PAGE 0409

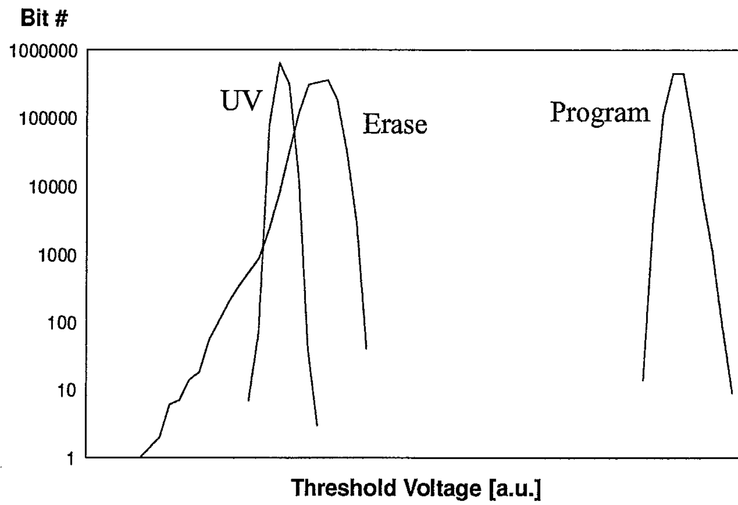


Figure 7.2 Threshold voltage distribution of a 1Mbit Flash array after UV erasure, after programming, and after electrical erasure.

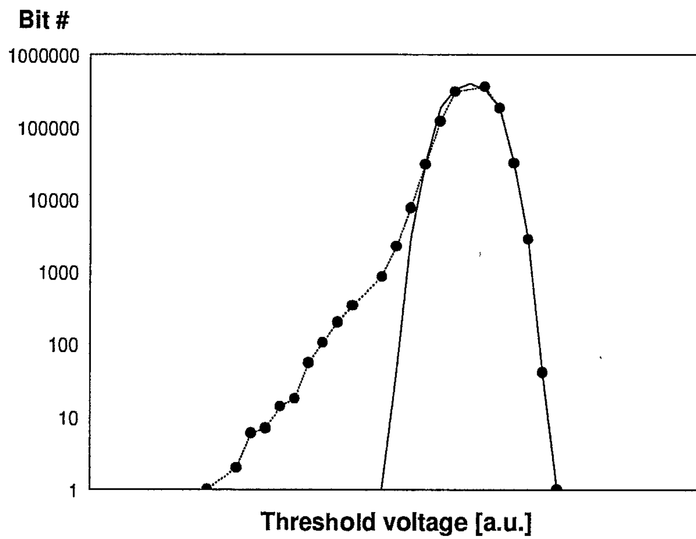


Figure 7.3 Experimental threshold distribution of a 1Mbit Flash array after electrical erasure (dots) and Gaussian fitting of the bulk of the distribution (solid line).

obtained by negative gate erasing scheme is wider than the one obtained with the source erasing scheme. Indeed the erasing speed is equally influenced by

APPLE INC.
EXHIBIT 1008 - PAGE 0410

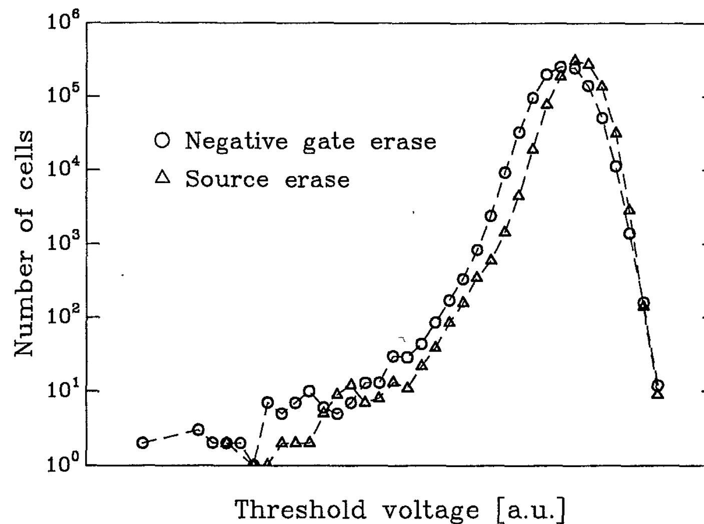


Figure 7.4 Threshold distributions of the same array obtained by two different erasing schemes; negative gate erase gives a wider distribution than source erase.

source and gate coupling ratios in the negative gate erasing scheme while the gate coupling ratio has a negligible impact on the source erasing speed.

Understanding the nature of tail bits is of key importance. As these bits erase faster than normal bits with the same applied voltage, one should assume that they are somehow "defective". However they are just too many for being associated to extrinsic defects.

Two intrinsic structural imperfections have been invoked to explain the nature of tail bits.

Muramatsu and coworkers [6] attributed the presence of a tail in the erased V_t distribution to the polycrystalline structure of the injecting electrode; barrier height variation at grain boundaries and/or dopant inclusion into the underlying oxide would cause a local enhancement of tunneling current.

An alternative explanation has been given by Dunn and coworkers [7] who have modeled the erasing tail as purely due to randomly distributed positive charges in the tunnel oxide. This model is solidly based on the well known existence of donor-like bulk oxide traps and on simulations which show the huge increase on tunnel current density caused by the presence of an elementary positive charge closed to injecting electrode (see Section 7.5.3.1).

The two models, schematically shown in Fig. 7.5, are both probably valid; the impact of polysilicon grain size on erased V_t distribution has been clearly demonstrated but oxide charges play for sure a role as well: Fig. 7.6 shows the V_t distributions obtained with a uniform channel FN/FN writing scheme, i.e.,

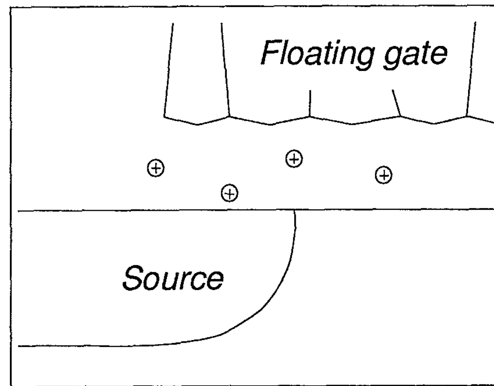


Figure 7.5 Schematic model of intrinsic defects responsible for the tail of the erased V_t distribution.

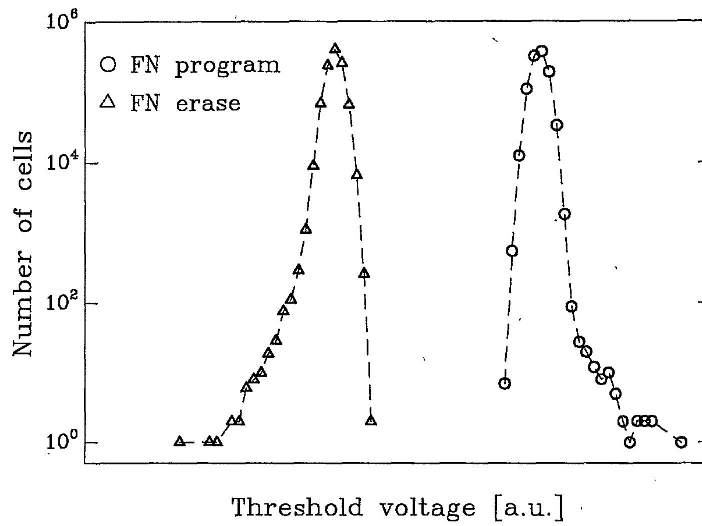


Figure 7.6 Threshold distributions of a Flash array obtained programming and erasing by Fowler-Nordheim tunnelling over the whole channel: both distributions show a tail.

APPLE INC.
EXHIBIT 1008 - PAGE 0412

applying positive voltage to the gate for programming and negative voltage to the gate for erasing: when the electrons are injected from the substrate, the cathode is monocrystalline but the tail is present as well!

As the exponential tail of the erased V_t distribution is mostly related to structural imperfections, i.e., intrinsic defects, it can be minimized by process optimization (for example working on tunnel oxidation, floating polysilicon and interpoly dielectric processes) but it cannot be eliminated; products must be designed taking into account the existence of such a tail.

The reliability of tail bits will be discussed, when relevant, in the following sections; it can be anticipated that, if taken as individuals, i.e., not considering the implications of being included in an array, these bits are as reliable as normal bits.

Normal bits and tail bits do not cover the totality of memory cells. Some bits, in the order of a few per million or even less in a very "clean" process, erase much faster than normal bits and they stand out also from the tail of the distribution; in an array which require a common mode erase, these bits would be normally over-erased rising the problems that will be discussed later on.

The physical defects responsible for these "fast-erasing" bits are likely different from the ones of tail bits; as they clearly do not belong to the statistical distribution of tail bits and they are very few, the "fast-erasing" bits are probably due to extrinsic defects, like tunnel oxide thinning caused by contamination.

The distinction between tail bits and fast-erasing bits may appear arbitrary, and in some extend it is because a sharp border cannot be defined (see for example the discussion of erratic bits in Section 7.5.3.1); however it is very important for understanding the implications on yield and reliability. For example, it has been proven that ppm level contamination of heavy metals or organic compounds in D.I. water or chemicals used for pre-oxidation cleaning can increase the number of fast-erasing bits without impacting the exponential tail of the erased V_t distribution. Moreover, while tail bits exhibit a good data retention, fast-erasing bits often show an enhanced charge loss in data retention tests.

Monitoring the density of extrinsic defects in tunnel oxide is instrumental for securing yield and reliability of a Flash memory production. Huge efforts have been spent to develop a test methodology to measure oxide defects and to effectively implement it in a process control strategy.

The Exponentially Ramped Current Stress (ERCS) [3] has been proven to be a very efficient test to detect defects in thin oxide. As it offers the optimum trade-off between sensitivity and testing time, it is best suited for intensive statistical process control. The method can be applied to simple capacitor test structures and therefore it can be effectively implemented for short loop monitoring of process equipment.

APPLE INC.
EXHIBIT 1008 - PAGE 0413

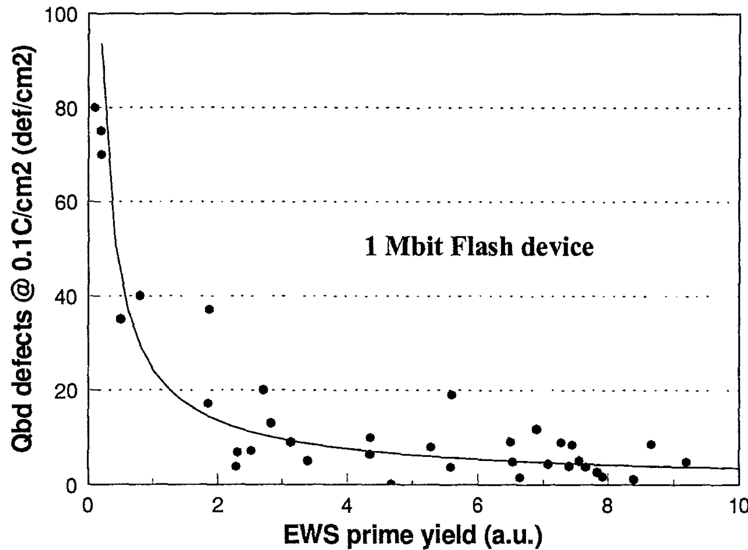


Figure 7.7 Correlation between tunnel oxide defect density measured on capacitors with a Qbd test and yield of 1Mbit Flash memory; each point represents the average value of one diffusion lot.

Moreover it has been demonstrated that the defect density obtained by Qbd measurements correlates with the yield of Flash memory: Fig. 7.7 shows that the yield of 1Mbit memory drops to zero when the Qbd defect density increases. As the yield loss in the lots with high defect density is mostly due to over-erased bits, it means that the physical defects responsible for premature oxide breakdown also cause individual bits to behave as fast-erasing and, vice-versa, at least some of fast-erasing bits have latent defects that would limit their endurance.

The same figure shows that for lowest Qbd defect density values, there is no longer correlation with memory yield data, which means that, while all defects detected by the Qbd test cause also a memory failure, not all the defects responsible for over-erasing can be detected by the Qbd test. This partial correlation can be easily explained considering that a memory cell is a very sensitive electrometer and it can reveal more subtle oxide defects than a Qbd test.

In order to improve the detectability of oxide defects, a test methodology has been specifically developed for Flash memory, which positively exploits the intrinsic sensitivity of memory cells and the steepness of their sub-threshold I–V characteristics.

The method is called Cell Array Stress Test (CAST) [4] and the structure utilized is an array of memory cells connected in parallel so that they can all be

APPLE INC.
EXHIBIT 1008 - PAGE 0414

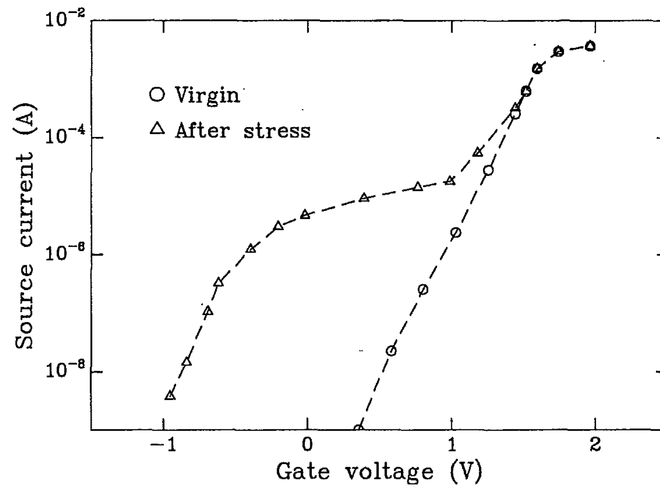


Figure 7.8 Transfer characteristics of a CAST structure before and after a negative gate electrical stress; the presence of a defective bit is clearly detected.

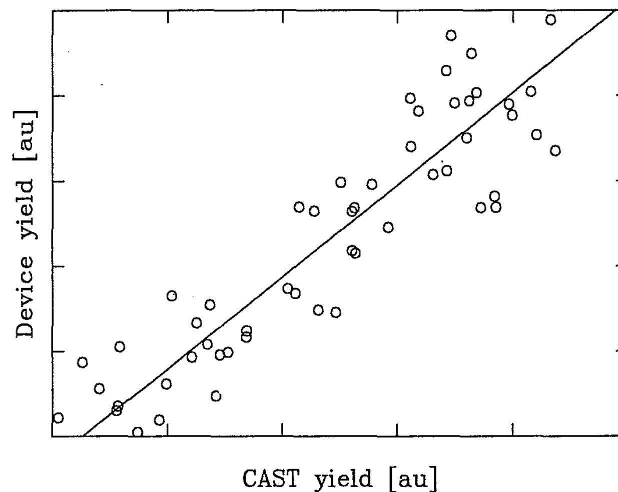


Figure 7.9 Correlation between CAST yield and the yield of 1Mbit Flash memory; each point represents the average value of one diffusion lot.

stressed and tested at the same time. Such a test structure perfectly resemble the array of a real memory and can be erased in the same way; the presence of fast erasing bits can be revealed by the analysis of the sub-threshold region of transfer I–V curve of the CAST structure (Fig. 7.8). The yield of CAST has been proven to correlate very well with the yield of actual memory (Fig. 7.9).

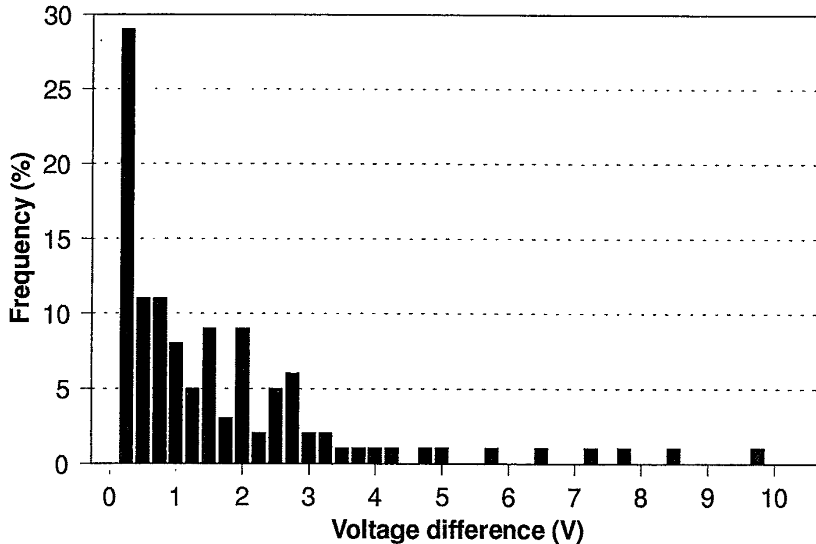


Figure 7.10 Histogram of CAST threshold shift measured on a fairly good yielding wafer; some structures show a very high shift.

The CAST structure provides a good vehicle with reasonably short cycle time for process development; moreover it does not require huge design effort and expensive testing equipment like a real memory.

For the subject of this section, CAST provides an additional piece of information which cannot be easily obtained by a real memory. As the most peculiar and easily detectable property of fast-erasing bits is that they stand-out from the exponential tail of the erased V_t distribution, one could be interested to know how far from the bulk of the distribution defective bits can still be found. That information cannot be normally provided by real memory because of the impossibility of biasing the gate to negative values in read mode and because of the saturation of sense amplifiers.

Such an information can be effectively obtained by CAST; Fig. 7.10 show for a pretty good yielding wafer the distribution of the V_t shifts of defective structures: there are structures with individual cells which stand out from the bulk of the distribution as much as 10V! The same results can be also read in a different way: while the electric field to be applied to the tunnel oxide for producing a detectable V_t shift on normal bits is in the range of 9–10MV/cm, most defective bits show a V_t shift at an applied electric field as low as 2MV/cm! CAST can provide a distribution of oxide defect conductivity thresholds, which is a much different and much more valuable information than the one obtained from measurements of electric-field-to-breakdown (Ebd) on capacitors.

APPLE INC.
EXHIBIT 1008 - PAGE 0416

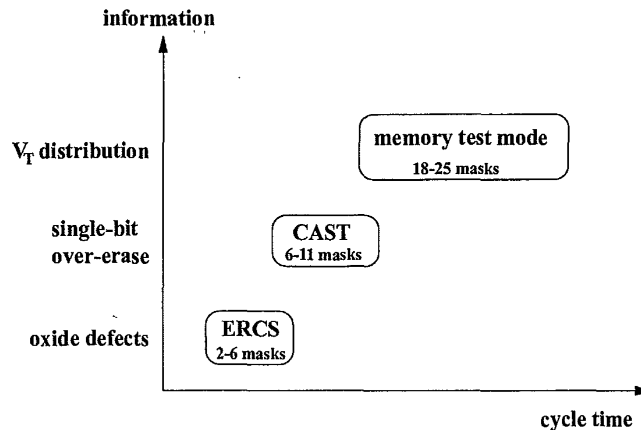


Figure 7.11 Different test methodologies give more detailed information on oxide quality at the cost of a longer cycle time.

Back to the question whether all fast-erasing bits belong to the tail of the erased V_t distribution, the probability of finding a representative of the exponential tail 10V apart from the bulk of the distribution would be less than 10^{-30}!

As a conclusion of this section on oxide defects, it is worth to enforce again the importance of a constant control and improvement of tunnel oxide quality for Flash memory yield and reliability. The described test methodologies, together with the information provided by the product itself, are all useful to develop and to fine-tune the process for high manufacturability: as schematically shown in Fig. 7.11, they provide different level of information at the cost of different processing cycle time.

7.3 MAIN YIELD AND RELIABILITY ISSUES

This section will illustrate the most relevant issues concerning yield and reliability of Flash memory.

7.3.1 Over-Erasing

Unlike EPROM's, Flash memories can be over-erased because their erasing mechanism is not self-stopping; as long as the erasing voltage is applied to the memory cell, electrons are continuously removed from the floating gate and the threshold voltage can eventually assume negative values if the erasing pulse is not stopped at the right time. Over-erasing is a potential cause of failure because a NOR-type memory array cannot be correctly read if it contains

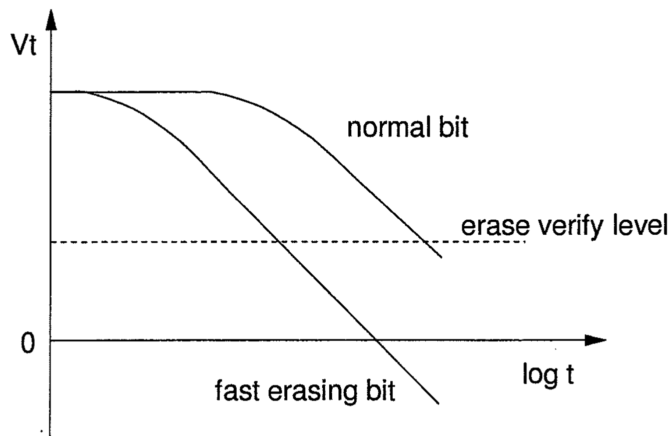


Figure 7.12 Threshold voltage as a function of time during erase for a normal bit and a fast erasing one.

depleted cells: all the cells connected to the same bit line of a depleted cell would be read as "1" irrespective of their actual content.

To stop at the right time the erasing operation, a quite troublesome erasing algorithm is utilized which is a sequence of short pulses followed by a verification of memory state after each pulse: the sequence is concluded as soon as all cells are successfully read as "1".

As all the cells in an array are erased simultaneously and the erasing voltage is applied to all of them, the time required to erase the slowest bit may be long enough to over-erase the fastest bits. The erase algorithms can only ensure that erasing is stopped as soon as all cells are written but it cannot prevent fast-erasing bits from being over-erased (Fig. 7.12).

Only an adequate process tuning combined with a proper circuit design can make the erasing mechanism reliably working. The erased V_t distribution must be taken into account in designing memory circuit and process must be optimized for narrowing it and keeping it under control. Redundancy must be used for repairing fast-erasing bits due to extrinsic defects.

Single bit over-erasing is typically the most relevant yield issue for Flash memory. Different soft-programming and self-convergence techniques have been developed to solve the problem by recovering over-erased or even by tailoring the V_t distribution [8, 9], but all the proposed methods are time and power consuming and they are not free from reliability concerns: an intensive utilization of these techniques to systematically narrow the erased V_t distribution is still to be proven. In any case would not be wise to utilize these techniques to

APPLE INC.
EXHIBIT 1008 - PAGE 0418

recover also heavily depleted bits; as they are associated to extrinsic defects it is more reliable to detect and repair such bits.

7.3.2 Program Disturbs

The failure mechanisms referred to as "program disturbs" concern data corruption of written cells caused by the electrical stress applied to these cells while programming other cells in the memory array. Flash memories are more sensitive to program disturbs than EPROM's because of the thinner gate oxide.

Two types of program disturbs must be taken into account: row disturbs and column disturbs.

Row disturbs are due to gate stress applied to a cell while programming other cells on the same word line. Let us refer to Fig. 7.13 which shows the schematics of a portion of the array and let us assume that the cell identified by the circle has being programmed. High voltage is applied to the selected row and all the cells of that row must withstand the gate stress without loosing their data. Depending on the data stored in the cells, data loss can be either caused by a leakage in the gate oxide (charge gain due to gate stress on "1") or by a leakage in the interpoly dielectric (charge loss due to gate stress on "0"). Fig. 7.14 shows the effect of gate stress on an erased normal cell; charge gain is due to Fowler-Nordheim tunneling from the channel to the floating gate.

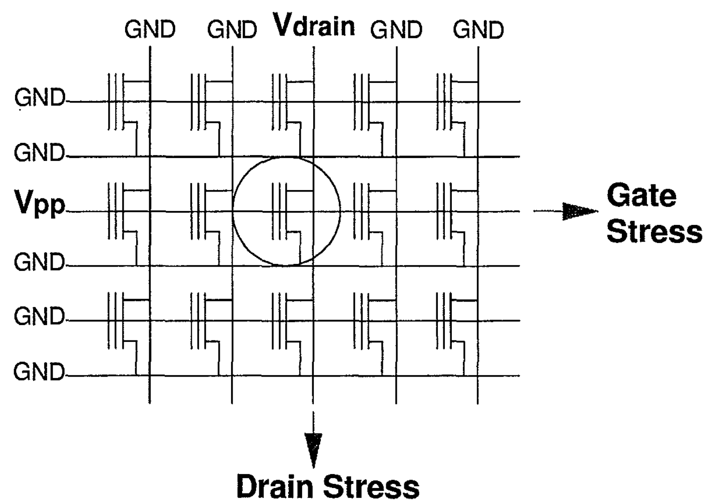


Figure 7.13 Schematic of a Flash array, showing row and column disturbs occurring when the circled cell is programmed.

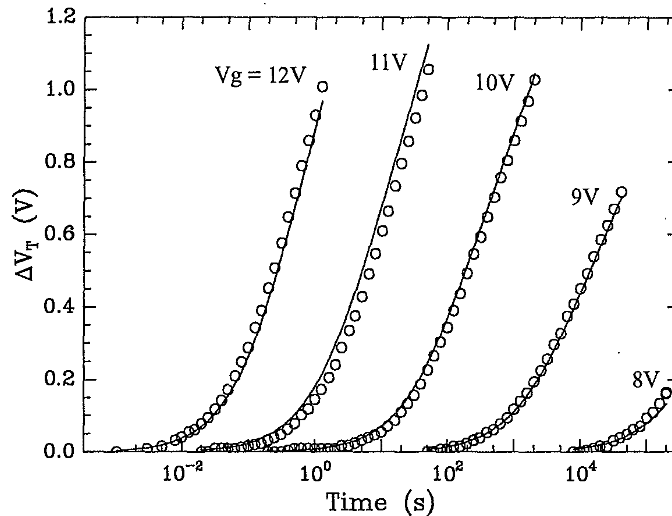


Figure 7.14 Threshold shift as a function of time for different gate stress voltages. The lines represent the best fit to the data using Fowler-Nordheim tunneling equation.

Worst case stress time depends on array organization and product specifications. Let us consider a byte programming with a maximum programming time of 50μs and let us assume an array organization with 128 bytes per row; the worst case stress time is

$$t_{\text{stress}} = 128 \times 50\mu\text{s} = 6.4\text{ms} .$$

Column disturbs are due to drain stress applied to a cell while programming other cells on the same bit line. Referring again to Fig. 7.13, all the cells on the selected column, except the one being programmed, have the gate grounded and a fairly high voltage applied to the drain. Under this condition, programmed cells can loose charge by Fowler-Nordheim tunneling from the floating gate to the drain (soft erasing); Fig. 7.15 shows the effect of drain stress on a programmed normal cell. Indeed the drain stress is quite similar to the normal condition for source erasing, just with lower applied voltage.

Worst case stress time is again related to memory organization and specifications. Assuming an organization with 1K cells per column and again 50μs maximum programming time, the worst case stress time is

$$t_{\text{stress}} = 1024 \times 50\mu\text{s} = 51\text{ms} .$$

It is worth to note that stress times have been calculated considering a single array, like for a bulk memory. The situation is much more critical for sectorized

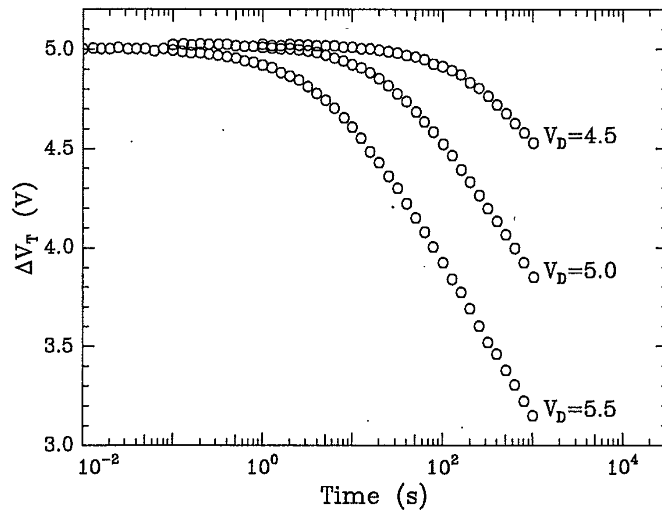


Figure 7.15 Threshold shift as a function of time for different drain stress voltages.

memories where, depending on the organization, the number of program/erase cycles may enter in the calculation of worst case stress time increasing it by orders of magnitude.

Let us consider, for example, a memory with sectors organized by columns and word lines shared by different sectors. Computing the worst case gate stress time, one must consider the case that one sector has to keep his data while the other sector are written as many time as for the specified endurance. Assuming that a word line is common to 8 sectors, that within each sector there are 32 bytes per row, that the maximum programming time per byte is again 50μs and that the number of program/erase cycle is 10^5, the worst case stress time is

$$t_{\text{stress}} = (32 \times 50\mu\text{s}) \times (1 + 7 \times 10^5) = 1.6\text{ms} + 1120\text{s}!\ .$$

It is clear that program disturb propagation among different sectors may have a dramatic impact on product reliability and it must be carefully considered by designers. The most effective way to prevent disturb propagation is to use block select transistors in a divided bit line and divided word line organization to completely isolate each sector from other sectors; but that is an expensive solution both in terms of array efficiency and process complexity. An alternative approach is to use substrate or source bias to reduce the stress on unselected sectors.

Program disturb is really a critical issue in Flash memory; cell and circuit must be designed with safety margins not only versus the stress sensitivity of normal bits but taking into account the tail of the distribution (Fig. 7.16). Still,

APPLE INC.
EXHIBIT 1008 - PAGE 0421

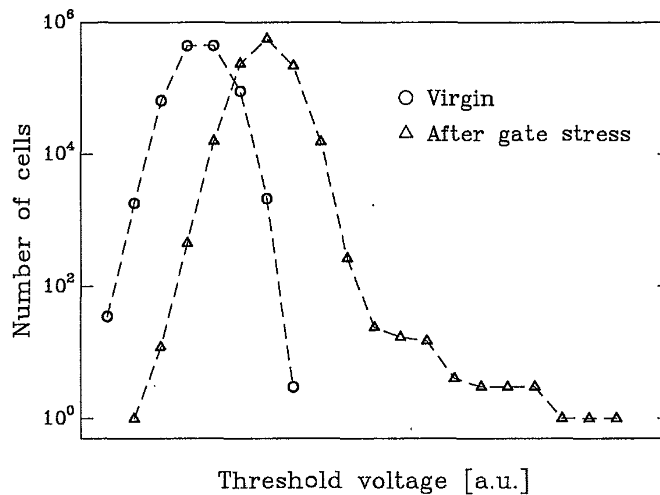


Figure 7.16 Threshold distribution of a 1Mbit array before and after a gate stress.

even when product and process are carefully designed, defects in gate oxide or interpoly dielectric can cause single bit failures due to program disturbs.

7.3.3 Read Disturb

When a cell is read, a gate voltage and a drain voltage are applied to the selected row and column, respectively. This condition can cause two kind of read disturb on a cell in the erased state. If we considered the cell that is read, an unwanted programming due to CHE injection can take place, if the drain voltage is not low enough. Moreover, all the cells in the erased state on the selected row are subjected to a low-voltage gate stress which can induce a tunneling current from the channel to the floating gate, resulting again in an unwanted programming. In both cases the oxide current is very low, but the worst case read-disturb time is of the same order of magnitude of the device lifetime.

Here again, safety margins can be achieved by a proper cell and circuit design, mainly in terms of oxide thickness and applied read voltages. A major concern arises when we consider the read disturb characteristics of a cell after extended program/erase cycling. Tunnel oxide degradation due to cycling may impact adversely on read disturb immunity, especially for very thin oxide, as will be discussed in Section 7.5.3.

APPLE INC.
EXHIBIT 1008 - PAGE 0422

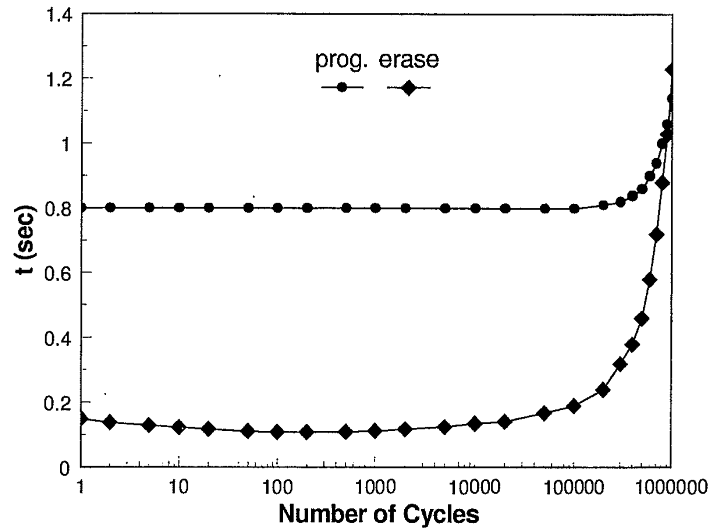


Figure 7.17 Programming and erasing time as a function of the number of program/erase cycles. Data refer to a 64kByte sector.

7.3.4 Program/Erase Endurance

Endurance of Flash memory can be limited by two failure modes: 1) the reduction of program/erase efficiency due to oxide aging, which brings to a parametric failure; 2) single bit failures due to tunnel oxide defects.

The first failure mode is related to a quite uniform and reproducible wear-out of memory cell performance; the mechanisms responsible for this degradation will be discussed in detail in Section 7.5.1. Over 10^3–10^4 program/erase cycles, both programming time and erasing time increase with the number of cycles (Fig. 7.17) and eventually exceed the specification limits. Cycling wear-out can be reduced by a proper device engineering and by the optimization of tunnel oxide process. However, once process and product are qualified for a given endurance specification, no major problem should come from lot to lot variation.

Actually, endurance problems are mostly related to single bit failures.

Section 7.5 will extensively cover the failure modes induced by program/erase cycling and Section 7.5.3 will specifically deal with single bit failures.

For the purpose of the present section it is sufficient to stress that, before any more sophisticated failure mechanisms, like the ones discussed in Section 7.5, extrinsic defects in tunnel oxide can likely cause single bit endurance failures due to premature oxide breakdown or degradation of data retention. These failure modes have been historically the major reliability issues for EEPROM's

APPLE INC.
EXHIBIT 1008 - PAGE 0423

which forced the introduction of on-chip error correction for high density memory, at the level of 1Mbit or even lower. As a matter of fact Flash memories are in production at 8Mbit and 16Mbit complexity level without on-chip error correction; that happens not because Flash memories are insensitive to defects in tunnel oxide but for the opposite reason, as will be clarified in Section 7.5.

7.3.5 Data Retention

As any non-volatile memory, Flash memories are specified to retain data for over 10 years. Possible causes of charge loss are:

- defects in tunnel oxide,
- defects in interpoly dielectric,
- mobile ion contamination.

Scaling technology and cell size, the demand for high quality processes is becoming more and more severe because of the reduction of cell capacitance and of stored charge. Fig. 7.18 shows a projection to the Gbit generation of cell capacitance and of the corresponding maximum acceptable concentration of mobile ions in interlevel dielectric. Similarly, also the maximum leakage through tunnel oxide and interpoly dielectric decreases from generation to generation and its value can be measured in few electrons per day.

A detailed discussion of data loss mechanisms and process solutions is out of the scope of this presentation; over 20 years of work on EPROM have generated

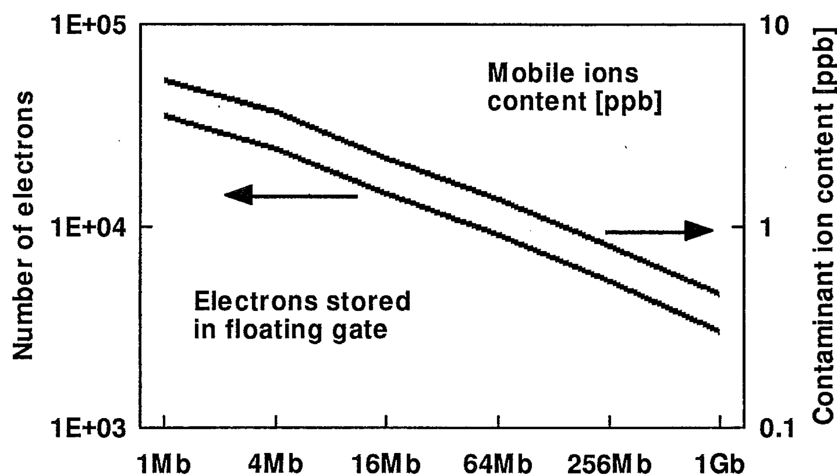


Figure 7.18 Stored charge and maximum content of mobile ions in the interlevel dielectric as a function of Flash generation.

APPLE INC.
EXHIBIT 1008 - PAGE 0424

a large literature on the subject and many good reviews are available (see for example [10]).

What is specific of Flash memory with respect to EPROM is the thinner gate oxide which makes the impact of oxide defects on data retention more critical. In addition, tunnel oxide degradation induced by program/erase cycling may eventually affect data retention, as will be shown in Section 7.5.3.

7.4 TESTING FOR RELIABILITY

As shown in the previous section, reliability issues in Flash memory are mostly related to oxide defects which cause single bit failures. We have also seen that the way oxide defects affect the behavior of a memory cell is by making it more sensitive to electrical stress than a normal cell. Therefore the most effective way to reveal oxide weakness is to apply accelerating tests, that can be performed at wafer sort utilizing specifically designed test modes (see Chapter 8) which are aimed to detect defective bits. In other words, the inherent sensitivity of Flash memory to oxide defects, properly enhanced by test modes, can be positively employed to identify and repair cells with latent defects.

A typical test flow for electrical wafer sorting is schematically shown in Fig. 7.19.

After usual continuity and parametric tests and the blank verification, a gate stress is applied to the whole array with an higher voltage than the one utilized in normal operations; stress voltage and time are set to cover the worst case programming stress. The stress is followed again by a blank check to detect any cell with anomalous charge gain. This test is aimed to detect bits that could fail because of row programming disturb in their erased state.

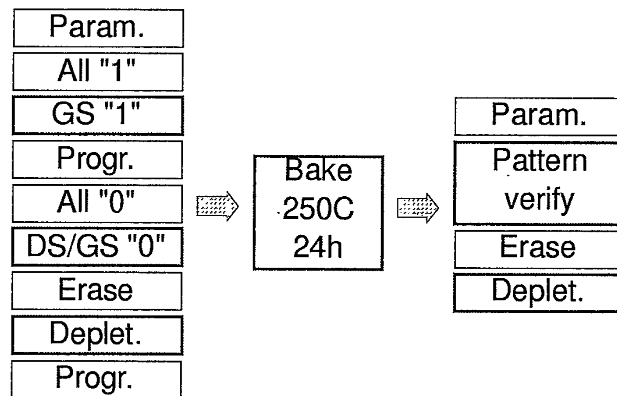


Figure 7.19 Typical test flow for electrical wafer sorting of a Flash memory.

APPLE INC.
EXHIBIT 1008 - PAGE 0425

Then, after an array programming and verification, a gate stress is repeated to detect bits that could fail because of row programming disturb in their programmed state and a drain stress is also applied, with similar criteria, to identify cells that could fail because of column programming disturb. Also in this case, an anomalous charge loss is detected by comparing the status of the array before and after stress.

The memory is then electrically erased and specific test searches for depleted bits. This is actually a test for functionality which by itself reveals defective cells that could affect product reliability.

The first part of sorting flow is competed by programming the whole array.

An accelerated data retention test is performed at high temperature, followed by a charge loss verification. Bake temperature and time are set to cover 10 year data retention for a given minimum activation energy; considering also the sensitivity of charge loss verification test, a typical bake of 24hrs at 250°C covers data retention failure mechanisms with activation energy higher than 0.6eV.

The second part of wafer sort flow is completed by a second erasing followed again by a depletion test.

Such a test flow, supported by properly designed test modes, can effectively detect and repair defective bits, significantly improving product reliability. Moreover it provides a very powerful wafer level monitor for statistical process control: yield results and engineering data coming from product testing are extremely valuable inputs for continuous process improvement.

With a good test methodology, potential reliability issues can hence be translated into yield issues. Then, the problem became how to reach and maintain high yields; to this aim the oxide evaluation methodologies described in Section 7.2 greatly help.

7.5 FAILURE MODES INDUCED BY PROGRAM/ERASE CYCLING

In the previous section we have seen the effectiveness of a proper wafer level testing in eliminating the impact of latent defects, greatly improving product reliability.

Remaining reliability concerns are related to failures induced by program/erase cycling because they cannot be adequately covered at wafer sorting; this sections will be dedicated to the failure mechanisms which limit Flash memory endurance.

7.5.1 Memory Cell Intrinsic Endurance

The wear-out of cell performances induced by program/erase cycling is due to gate oxide degradation. A typical result of an endurance test on a single han-

APPLE INC.
EXHIBIT 1008 - PAGE 0426

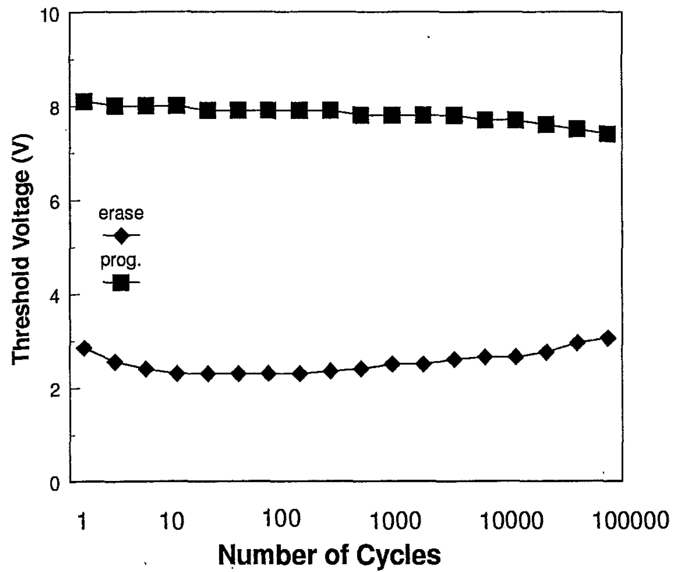


Figure 7.20 Programmed and erased threshold voltage as a function of the number of cycles. The experiment was performed on a cell using fixed program and erase conditions.

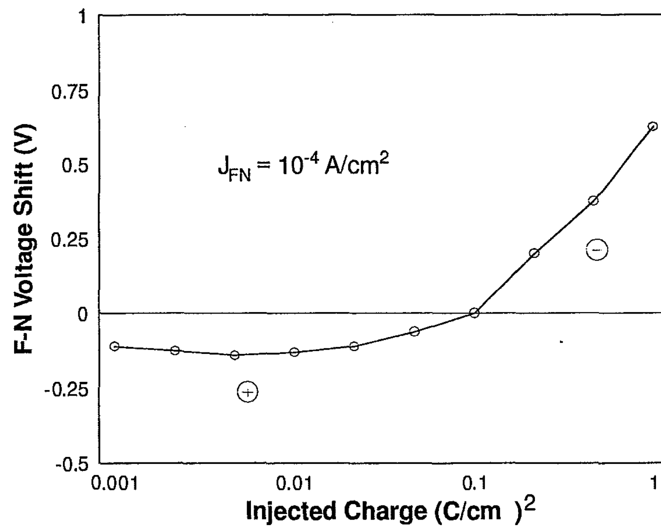


Figure 7.21 Variation of the voltage required to inject a constant current through the oxide of an MOS capacitor as a function of the injected charge. A negative variation corresponds to a positive charge build-up in the oxide and vice versa.

APPLE INC.
EXHIBIT 1008 - PAGE 0427

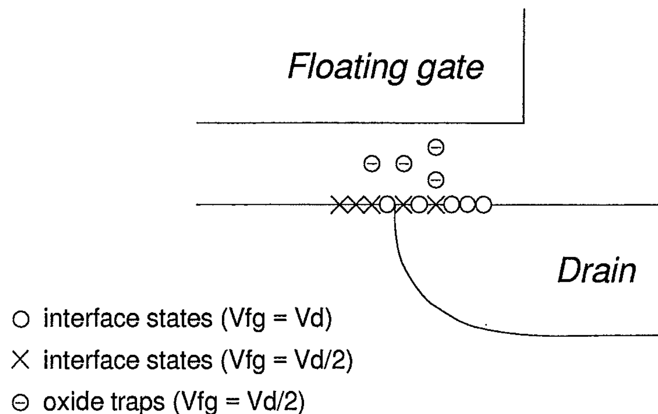


Figure 7.22 Schematic representation of the location of oxide charge and interface states generated by CHE injection under various voltage conditions.

dled cell is shown in Fig. 7.20 [11]; as the experiment was performed applying constant pulses, the variations of programmed and erased threshold levels give a measure of oxide aging.

The evolution of erased threshold voltage (V_{te}) is similar to the one typically observed in EEPROM's cell; it reflects the well known dynamics of net fixed charge in the tunnel oxide as a function of the injected charge [12], as reported in Fig. 7.21. Therefore the initial lowering of V_{te} is due to a pile-up of positive charge which enhances tunneling efficiency, while the long term increase of V_{te} is due to the generation of negative traps.

The reduction of programmed threshold voltage at high cycling numbers has been well explained by Yamada *et al.* [13]; it is attributed to oxide traps and interface state generation at the drain side of the memory cell (Fig. 7.22), a degradation mechanism which is inherent to CHE programming. According to the model of Yamada and his co-authors, a stress condition with $V_{fg} \sim V_d$, which corresponds to the beginning of programming when the floating gate is almost neutral and the cell operates in triode mode, is responsible for the generation of interface states over the drain region. At the end of programming, when the floating gate is negatively charged, the cell operates in penthode mode ($V_{fg} \sim V_d/2$) and interface states and bulk oxide charges are generated over the channel. This latest condition is the most relevant for the degradation of cell performance: interface states reduce electron mean free path, impacting on the hot carrier generation mechanism, and electrons trapped in the tunnel oxide modify the electric field at the injection point, reducing programming efficiency. Fig. 7.23 clearly shows the degradation of programming curves caused by program/erase cycling.

APPLE INC.
EXHIBIT 1008 - PAGE 0428

Nevertheless, a properly designed memory cell can reach 10^5–10^6 program/erase cycles with acceptable writing performance degradation and no detectable reduction of its transconductance (Fig. 7.24).

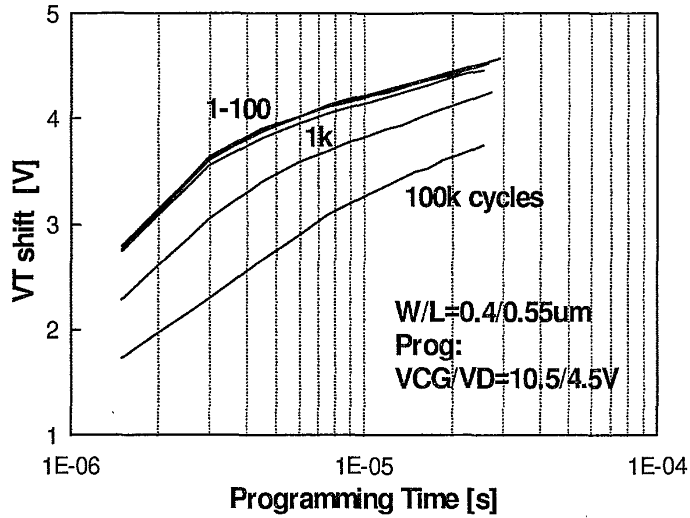


Figure 7.23 Programming characteristics of a Flash cell measured at various steps of an endurance test.

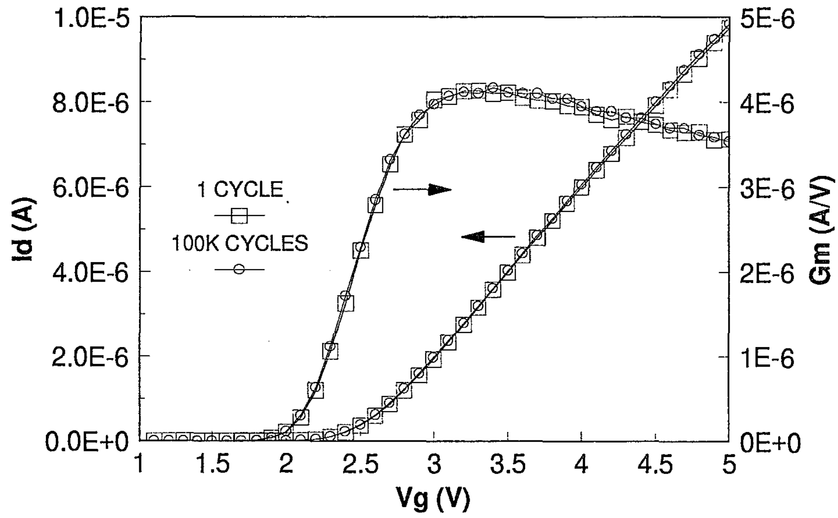


Figure 7.24 Drain current and transconductance of a Flash cell before and after an endurance test.

APPLE INC.
EXHIBIT 1008 - PAGE 0429

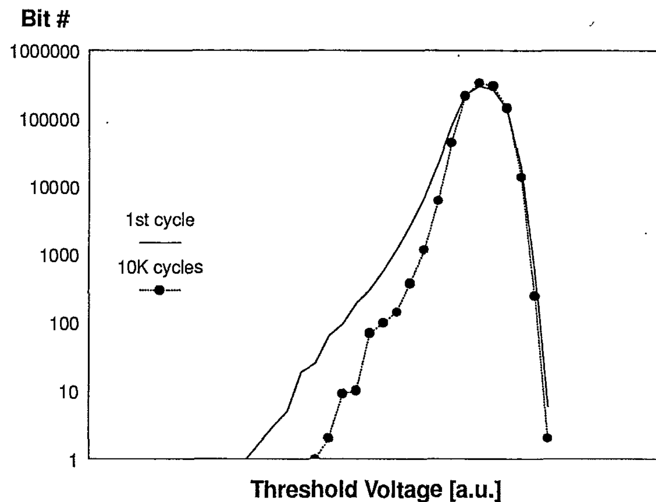


Figure 7.25 Erased threshold distribution of a 1Mbit array at the first cycle and after 10^4 cycles.

7.5.2 The Behavior of Tail Bits

After having discussed the endurance of a single cell, let us now consider a memory array.

The behavior of normal bits reflects the one of a single handled cell, as described in the previous section; concerns could come from tail bits.

Typically, program/erase cycling do not cause a broadening of the erased V_t distribution; on the contrary, the tail gets a little narrower increasing the number of cycles (Fig. 7.25). The enhanced wear-out of tail bits is consistent with the model that these bits erase faster because of a localized higher current injection caused by a peak of electric field. Indeed a larger current density corresponds to an higher generation rate of negative traps, i.e., a faster aging. Hence, the generated negative charge partially smoothes the peaks of electric field, making current injection more uniform and erasing speed of tail bits more similar to the one of normal bits.

7.5.3 Single Bit Failure Mechanisms

So far our discussion on Flash memory endurance has given a quite comfortable picture: extrinsic oxide defects are efficiently detected at electrical wafer sorting, performance wear-out is reproducible and compatible with product specification, erased V_t distribution does not worsen with cycles, on the con-

trary it gets narrower. We should hence expect that no endurance failures would ever happen in a properly designed Flash memory.

Actually, endurance failures are typically due to single bits and to quite sophisticated mechanisms; this section will present the most important ones.

7.5.3.1 The Erratic Erase Phenomenon. The most relevant mechanism of single bit failure in program/erase cycling reported so far is the erratic erase, which has been presented the first time by Ong *et al.* [14] and further investigated by Dunn *et al.* [7]. Erratic bits show an unstable and unpredictable behavior in erasing: their V_{te} changes randomly from cycle to cycle between two or more distinct values (Fig. 7.26), moving back and forth from the bulk of V_t distribution to the lowest part of the tail; erratic bits can cause over-erasing failures in the field.

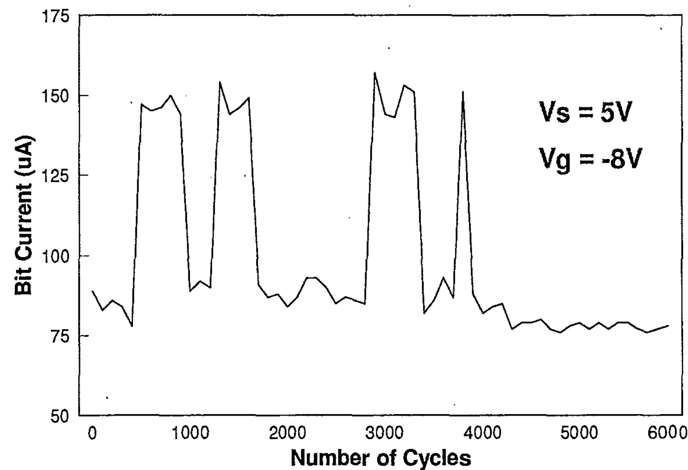


Figure 7.26 Drain current after erase as a function of the number of cycles for a cell exhibiting the erratic erase behavior.

Such a behavior has been attributed to hole trapping in the tunnel oxide: the statistical distribution of hole traps gives an extremely low but finite probability of having clusters of three or more positive charges (Fig. 7.27), whose electric fields overlap each other to produce a huge local increase of the tunnel current. In this condition, trapping/detrapping of an individual positive charge cause a detectable change in erasing speed and threshold level. This model has been quantitatively confirmed by WKB calculations [7, 14]; Fig. 7.28 shows the current density increase caused by single, double and triple elementary positive charges as a function of their distance from the injecting electrode: the increase of current density can be of 4 or 5 orders of magnitudes!

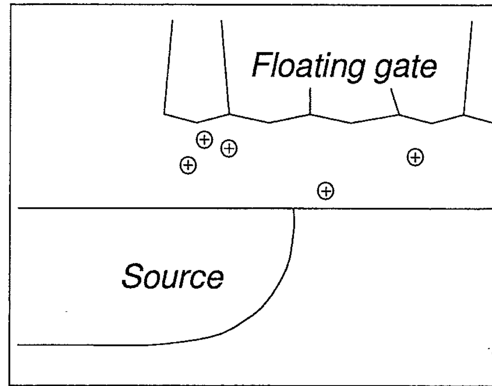


Figure 7.27 Clusters of positive charges may enhance the tunneling current during erase.

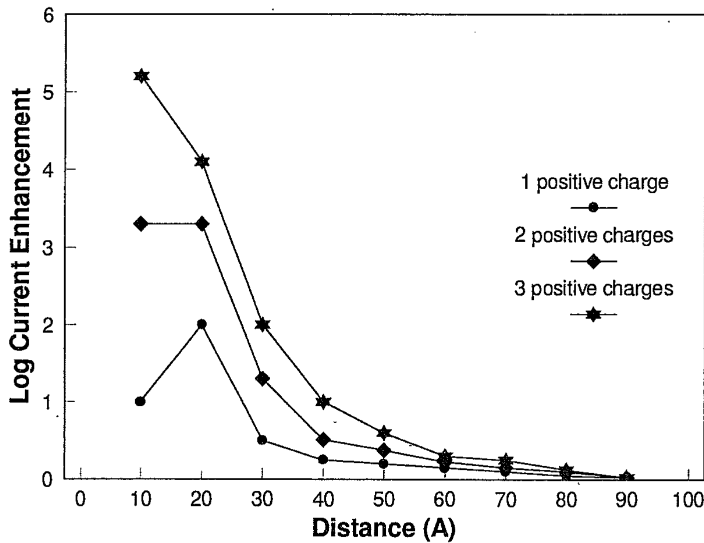


Figure 7.28 Tunneling current enhancement in a unit dielectric cell, 10nm per side, as a function of the distance from the injecting electrode of a positive charge cluster.

An experimental confirmation of the described model is reported in Fig. 7.29; the figure shows the erasing curves, i.e., the cell current as a function of erasing time, of an erratic bit measured in three consecutive cycles. In the first cycle the cell behaves as a normal bit and at the end of erasing the current reach the typical value of a cell in the bulk of the erased V_t distribution. In the following cycle, the cell starts erasing as in the first cycle; then the current suddenly increases at a much higher rate with an evident discontinuity in the

APPLE INC.
EXHIBIT 1008 - PAGE 0432

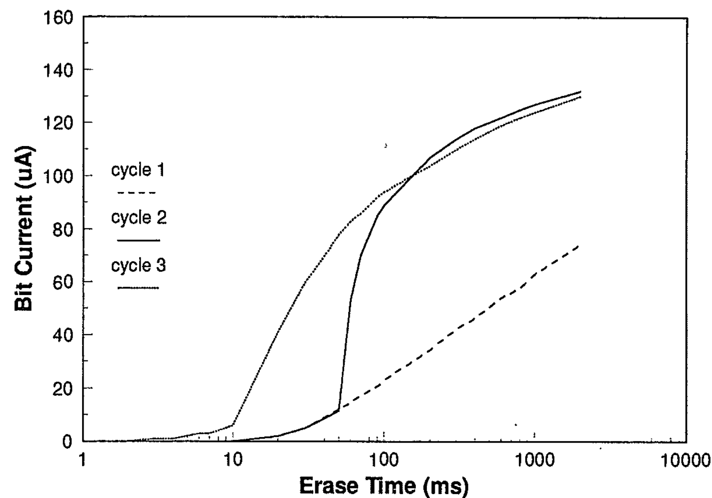


Figure 7.29 Cell current of an erratic bit as a function of erasing time during three consecutive cycles.

derivative, i.e., in the tunneling current; at the end of the second cycle the cell is over-erased and its current reach a value which corresponds to a depleted cell. In the third cycle, the cell behaves as a fast-erasing bit from the beginning and at the end the cell current is the same as in the second cycle. The curves shown in Fig. 7.29 can be clearly explained by the presented model: during the second cycle a hole is captured by one trap of the cluster responsible for the erratic behavior of the bit, while the other traps of the cluster are already positively charged; that determines the sudden increase of tunneling current in the second cycle. It is worth to notice that Fig. 7.29 somehow represents the picture of the capture event of a single elementary charge detected by an industrial product!

For better understanding the mechanism and for addressing the solution of the erratic bit problem, two relevant questions should be answered:

1. What is the "generation rate" of erratic bits versus the number of cycles?
2. Would a negative gate erase scheme eliminate the erratic bit phenomenon?

The first question is clearly related to the nature of the involved traps, their generation rate over cycling and their hole capture cross-section. Fig. 7.30 shows the histogram of the occurrence rate of "new" erratic bits in the first 10^4 cycles; it is obtained by a large sample of 1Mbit Flash memories failed in cycling tests for erratic bits and counting how many of them started showing an erratic behavior in the first thousand cycles, how many in the second thousand,

APPLE INC.
EXHIBIT 1008 - PAGE 0433

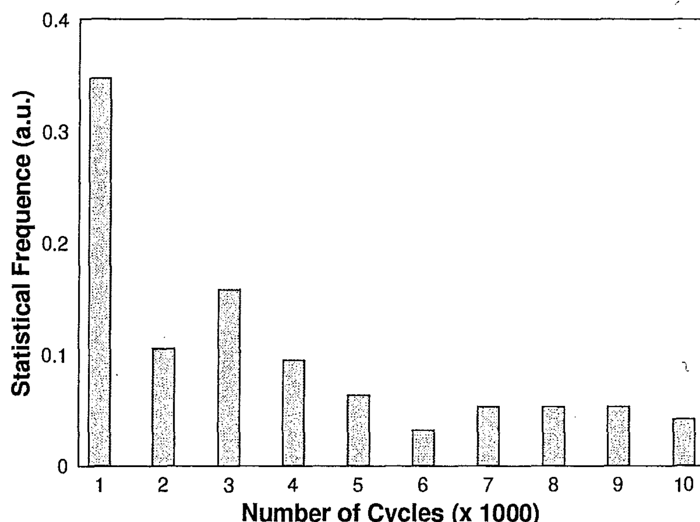


Figure 7.30 Frequency of occurrence of new erratic bits as a function of the number of program/erase cycles.

and so on. Increasing the number of cycles, the erratic bit occurrence rate decreases but it does not goes to zero even after 10^4 cycles. Such a behavior is quite consistent with the known evolution of positive charge in stressed tunnel oxide: the positive charge is most rapidly piled-up in the initial stage of current stress but it does not saturate. For practical purposes, Fig. 7.30 demonstrate that even a 10^3 cycling test would not effectively screen erratic bits.

The second question is related to the origin of trapped holes. If they are hot holes generated by avalanche multiplication in the depletion region of source junction, the negative gate erase scheme would significantly reduce the hot hole generation rate practically eliminating the erratic bit problem. That is not true: the data of Fig. 7.26 come from a memory that utilizes a negative gate erase scheme ($V_g = -8$V, $V_s = 5$V).

In summary, as the erratic behavior is due to statistical fluctuation of intrinsic oxide defects, the occurrence of erratic bits can be reduced by process optimization but cannot be completely eliminated; design solutions have been developed to take care of the problem at circuit level.

7.5.3.2 Single Bit Data Loss after Program/Erase Cycling. In Section 7.5.1 we have discussed the effect of oxide charges, generated by program/erase cycling, on cell performance, concentrating our attention on writing time. However the most critical reliability issues with the oxide degradation are related to data retention.

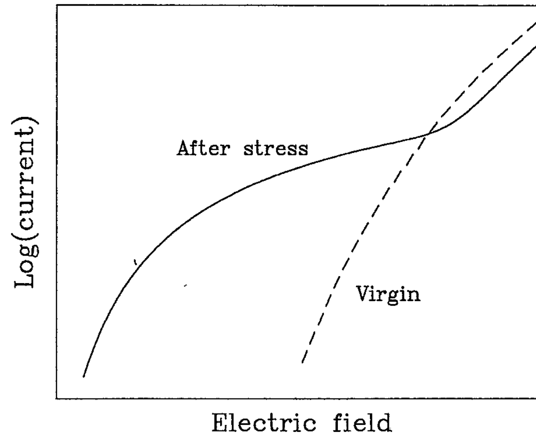


Figure 7.31 Qualitative behavior of the tunnelling current for a fresh sample and after a high field stress.

A high-field stress on thin oxide is known to increase the current density at low electric field; the excess current component which causes a significant deviation of I-V curves from the theoretical Fowler-Nordheim characteristics at low field is known as Stress-Induced-Leakage-Current (SILC). SILC is clearly related to stress induced oxide defects and, as far as conduction mechanism, it is attributed to a trap assisted tunneling (see Chapter 4 for a detailed discussion). The main parameters controlling SILC are the stress field, the amount of charge injected during the stress, and the oxide thickness [15]. For fixed stress conditions the leakage current increases strongly with decreasing oxide thickness below 10nm.

The qualitative behavior of the oxide current as a function of the electric field for a heavily stressed thin oxide is compared in Fig. 7.31 with that of a virgin sample. At high field the current decreases, the well-known phenomenon leading to the erase time increase after extended cycling. At low field the opposite situation occurs: with increasing stress time the current increases well above that of the fresh oxide.

This effect can be observed in a Flash cell as an enhanced sensitivity to low voltage gate stress after cycling. Fig. 7.32 shows the sensitivity to gate stress of a single-handled cell with a thin tunnel oxide as a function of program/erase cycles: increasing the number of cycles worsens the effect of gate stress. The figure reports also the result for a cell cycled with a double Fowler-Nordheim writing scheme, showing the enhanced degradation due to the higher field and to the bi-directional stress inherent to this writing scheme.

APPLE INC.
EXHIBIT 1008 - PAGE 0435

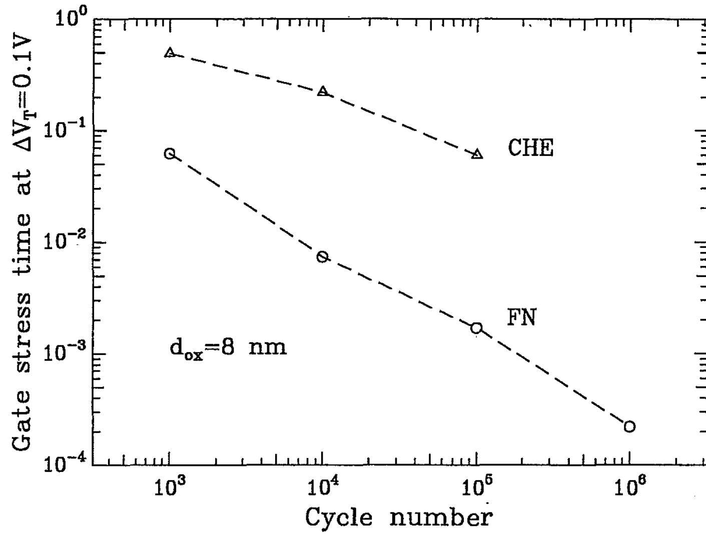


Figure 7.32 Gate stress time required for a 0.1V threshold shift as a function of number of cycles done before the gate stress. Stress voltage is 8V. Program/erase times are 0.005/100ms for CHE and 0.1/100ms for FN.

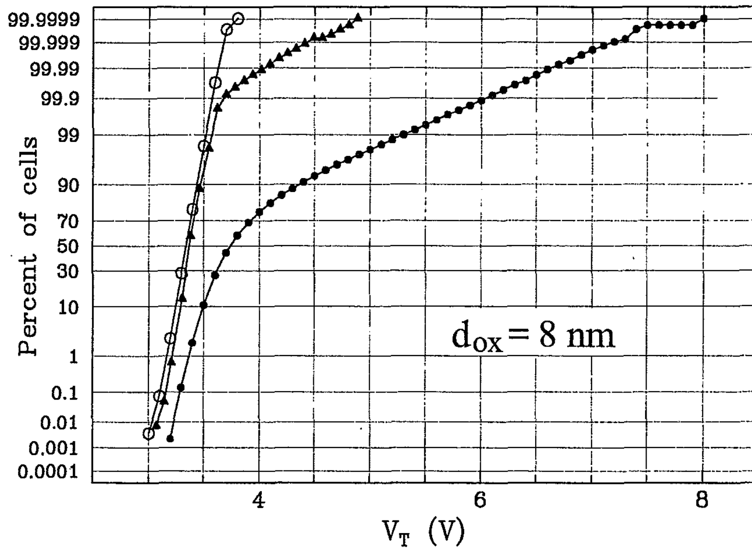


Figure 7.33 Threshold distribution for a 1Mbit array after gate stress on the virgin sample and on the sample cycled 10^5 times (closed circles). Stress time and voltage were 63h and 8V, respectively, in both cases. The open circles represent the distribution after UV erasure.

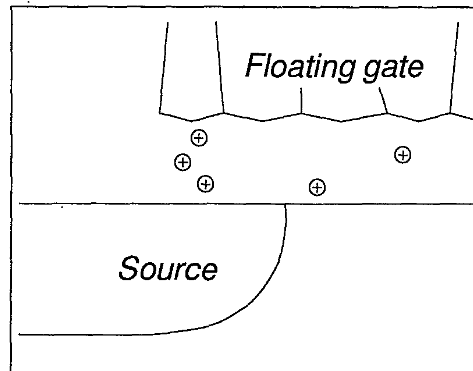


Figure 7.34 Schematic representation of a possible configuration of trapped positive charges giving rise to a leakage current enhancement.

Once again, the major concerns do not come from a typical cell, but from the tail of the distribution, as is shown in Fig. 7.33, where the results of a low-voltage gate stress experiment on a 1Mb array before and after cycling are compared. The different magnitude of the V_t shift for different cells can be explained again by the random spatial distribution of oxide traps responsible of the tunneling current enhancement, like for erratic bits; Fig. 7.34 gives a schematic representation of possible coordination of positive charges that would locally enhance the leakage current. The analogy with the erratic bit phenomenon is enforced by the erratic behavior of the anomalous SILC current reported by Yamada *et al.* [16].

A first impact of SILC on Flash reliability could be expected as an increased program disturbs sensitivity. Actually, the electric field in the case of program disturbs is high enough to fall in the region near the current crossing point in Fig. 7.31, so that no big difference is found between the program disturb characteristics before and after cycling.

The impact of SILC on read disturb and data retention is strongly dependent on tunnel oxide thickness.

For very thin tunnel oxide (below 8nm), SILC is not negligible even at electric field values as low as the ones typical of reading or data storage conditions. The electric field in the tunnel oxide of an erased cell in read mode is in the range of 2–3MV/cm, while it is in the 1–2MV/cm range for a programmed cell not biased; cycled memory cells must retain their data for years in such conditions. Many papers have been published on read disturbs after cycling in Flash memory based on FN programming with very thin tunnel oxide [17].

From SILC data obtained on capacitors, Runnion *et al.* [18] have shown that the leakage of tunnel oxides in the 5–7nm range after a stress equivalent

APPLE INC.
EXHIBIT 1008 - PAGE 0437

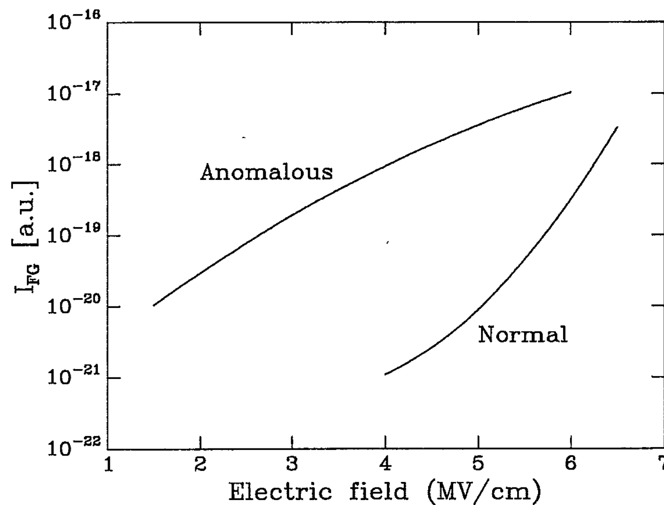


Figure 7.35 Leakage current as a function of the electric field in the tunnel oxide for a typical and an anomalous cell after 10^5 cycles. Data from ref. [16].

to 10^5 cycles largely exceeds the requirements for 10 year data retention or for read disturb immunity.

If that is the limit to tunnel oxide scaling one could estimate from capacitors, which give an average SILC value, Yamada *et al.* [16] have shown that the situation on a real memory can be worse because of single cells that exhibit a much higher SILC (Fig. 7.35).

Again, for setting the minimum reliable thickness of tunnel oxide, we cannot refer to the typical cell but we must give margin for the tail of the distribution; Fig. 7.36 shows the results of a room temperature retention test on a 1Mbit array with 8nm tunnel oxide thickness after 10^5 cycles: while almost the totality of the array does not present any detectable threshold shift, there is a tail of cell that loose charge.

Data retention after cycling is the issue that definitely limits the tunnel oxide thickness scaling. For very thin oxide the number of leaky cells becomes so large that even an error correction technique cannot fix the problem.

7.5.3.3 Gain Degradation. We have seen in Section 7.5.1 that a properly designed memory cell does not show any transconductance degradation after 10^5 cycles. Nevertheless, single bits may fail in endurance tests because of gain degradation.

Fig. 7.37 show the I–V characteristic of two degraded bits together with the normal characteristics of other bits of the same array not modified by cycling.

APPLE INC.
EXHIBIT 1008 - PAGE 0438

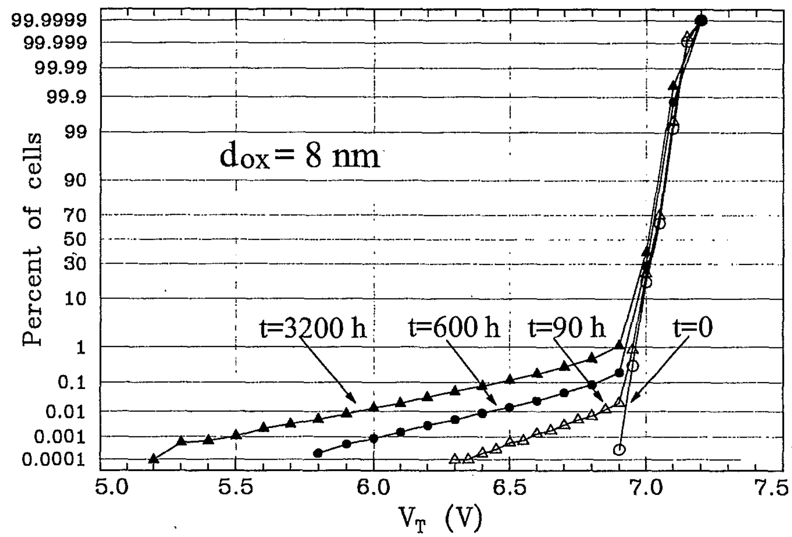


Figure 7.36 Threshold voltage distribution of a 1Mbit array cycled 10^5 times after program and at different time steps during storage at room temperature.

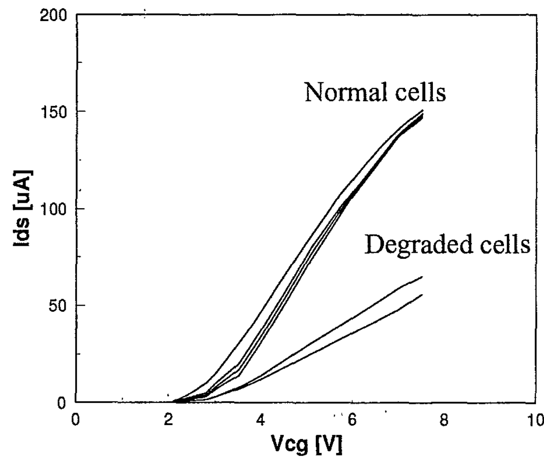


Figure 7.37 Transfer characteristics of two cells showing cycling-induced gain degradation, compared with those of normal cells in the same array.

The failure modes induced by such degraded bits can be different. They can cause an erase failure, because they force the whole array to be over-erased. They can cause a program failure, because their programming performance is degraded as well. They can fail in high temperature retention test, because the bake partially recover their original characteristics and the bit is no longer read as "0" after the bake.

The transconductance degradation is due to interface states generated by hot holes. There are different mechanisms which can cause an exceptionally high rate of hot hole injection and, consequently, a transconductance degradation in single bits.

The most likely source of hot holes is the erasing itself. As discussed in detail in Chapter 4, during the erasing hole-electron pairs are generated by band-to-band tunneling in the source region; holes drift towards the substrate and they gain energy from the electric field in the depletion region of the reverse biased source junction; the electric field in tunnel oxide over the channel is then favorable to the injection of hot holes. Source junction engineering is aimed to minimize the electric field and to prevent avalanche multiplication that would accelerate oxide degradation. Also circuit design helps in minimizing the risk of avalanche multiplication by limiting the current supplied to the source in erasing. Therefore, in a carefully designed memory, hot hole generation is minimized and cycling does not degrade the transconductance of memory cells. However, silicon defects or contamination can locally reduce significantly the source junction breakdown voltage; a defective cell may hence present a much higher hot hole generation in erasing than average because of locally enhanced junction electric field. The source current control is not effective for preventing avalanche multiplication in the defective cell, because it limits the total current of the array, but it cannot avoid that a single cell sinks orders of magnitude higher current than average. This model has been experimentally proved by spot light emission from the degraded bits during the erase operation.

Process and doping profile optimization can effectively minimize the endurance failure rate related to the above described mechanism. Moreover, a negative gate erase scheme greatly helps to solve the problem by significantly reducing the source junction reverse bias.

A second mechanism that can cause an enhanced hot hole injection in single cells is associated to the program operation and it is a quite complex combination of program disturb and over-erasing.

It may affect the cells with the lowest V_t after electrical erase (i.e., tail bits) and it is due to the drain stress these cells experience while other cells on the same bitline are programmed (Fig. 7.38).

A bit at the lowest edge of the erased V_t distribution has the floating gate positively charged. In drain stress condition, the combination of the positive

APPLE INC.
EXHIBIT 1008 - PAGE 0440

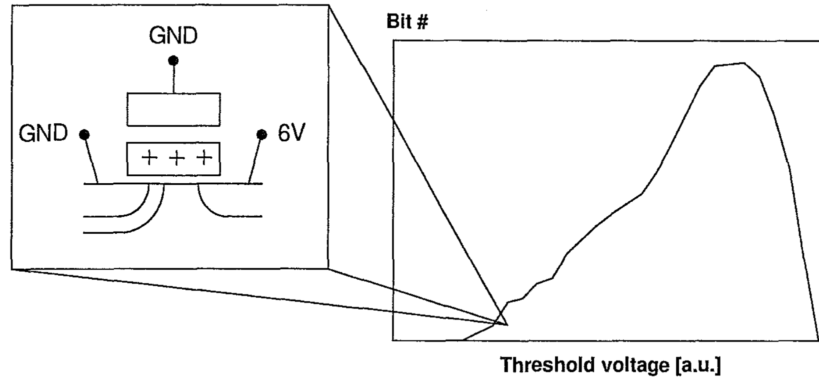


Figure 7.38 A cell in the erase tail can be slightly turned on under column program disturb condition. Drain-stress-induced gain degradation may occur as a consequence.

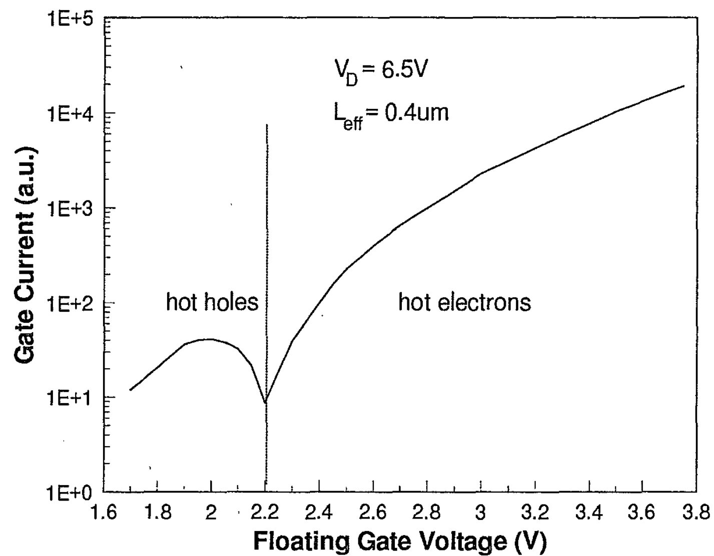


Figure 7.39 Floating gate current as a function of voltage, showing the sign reversal of the current at low voltage due to hot hole injection.

charge in the floating gate and the drain coupling effect can turn this bit slightly on, even with the gate grounded. High drain voltage and low floating gate voltage are the most favorable conditions for hot hole injection (Fig. 7.39). The cumulative effect of the interface damage produced at each cycle by the drain-stress-induced hot hole injection eventually causes the reduction of cell transconductance.

APPLE INC.
EXHIBIT 1008 - PAGE 0441

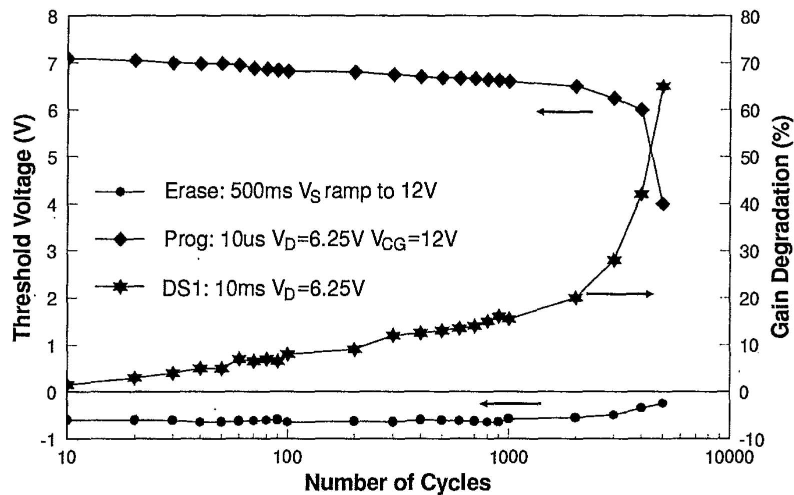


Figure 7.40 Threshold window and gain degradation as a function of the number of program/erase/drain-stress cycles.

Fig. 7.40 shows the dramatic effect of introducing a drain stress ($V_d = 6.25$V, $t = 10$ms) at every cycle in an endurance test on a cell which is intentionally over-erased. The degradation has been intensively studied on contacted-floating-gate devices stressed in different bias condition. The degradation rate (Fig. 7.41) is fastest for floating gate voltages equal or little above the value at which the hole and the electron components of gate current compensate each other so that the net current in Fig. 7.39 goes to zero.

The stress conditions reproduced in the single cell cycling experiment correspond to the extreme case of a bit at the lowest edge of the erased V_t distribution tail. They may be outside the normal operating condition, depending on the minimum accepted V_t for depleted bits and to drain voltage limit. On the contrary, those conditions very likely correspond to the typical case of some self-convergence schemes proposed for recovering depleted bits, whose reliability is therefore very questionable.

Device dimension (L_{eff}) and operating conditions (V_d, V_{fg}) have a very critical impact on the degradation rate. Cell with shorter L_{eff} are sooner, in terms of stress time and then in terms of cycles, influenced by this degradation (Fig. 7.42). For scaling-down cell size, this mechanism must be carefully taken into account: memory cell architecture, e.g., junction engineering or oxide robustness to hot carrier stress, must be optimized to reduce the impact of drain stress degradation.

APPLE INC.
EXHIBIT 1008 - PAGE 0442

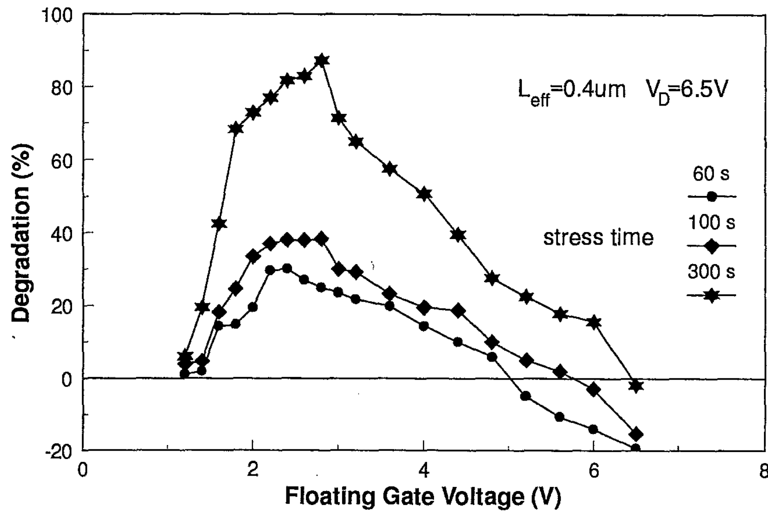


Figure 7.41 Gain degradation as a function of floating gate voltage during drain stress for different stress time.

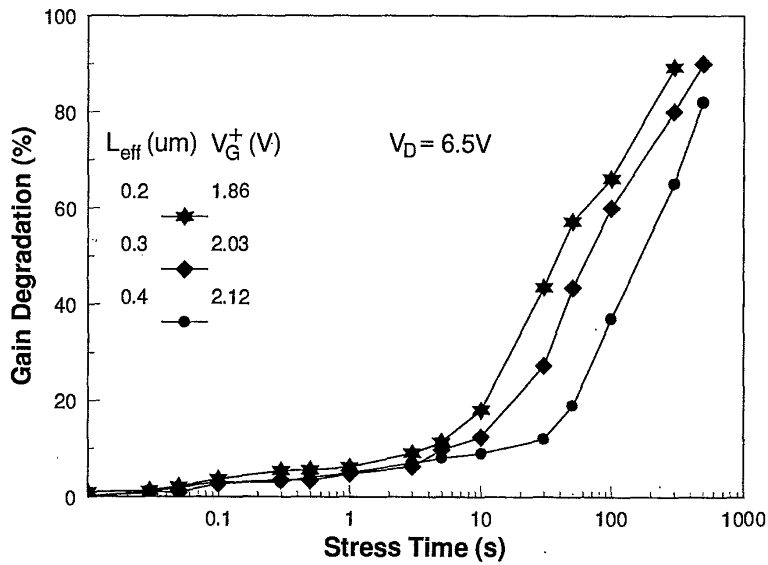


Figure 7.42 Gain degradation as a function of time during drain stress for different effective channel length.

APPLE INC.
EXHIBIT 1008 - PAGE 0443

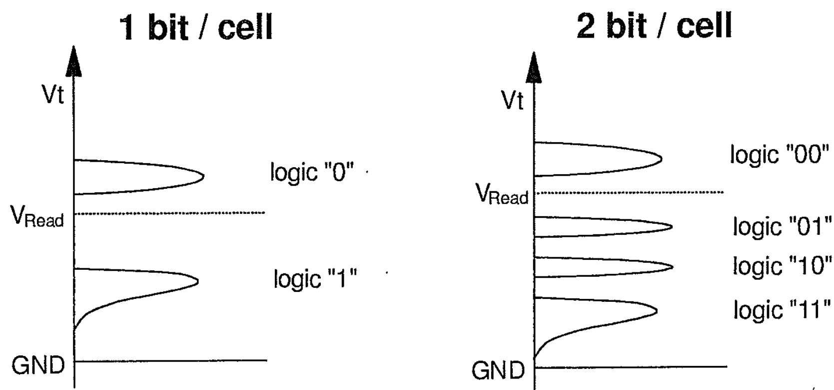


Figure 7.43 Threshold distributions for a conventional and a 2-bit/cell Flash memory.

7.6 MULTILEVEL STORAGE RELIABILITY

The concept of storing more than one bit per cell is based on the proven possibility of analog programming non-volatile memory cells.

In standard digital Flash memory, every cell stores one bit and two V_t ranges correspond to the two logical states "1" (erased cell) and "0" (programmed cell). Conceptually, nothing prevents from defining more V_t levels as different cell states in order to store more bits per cell: $4V_t$ levels would allow to store 2 bit/cell, $8V_t$ levels would allow to store 3 bit/cell and so on. The general rule is that 2^n levels are required to store n bit/cell.

In Fig. 7.43 the V_t distributions of a standard and 2bit/cell Flash memory are compared. The V_t range where to allocate the levels is limited on the lower side by the requirement that all thresholds must be above a minimum positive value to avoid bit-line leakage during read or program. On the upper side the V_t range is limited by the condition that the second highest level must be placed 0.5–1V below the gate voltage during read to be properly discriminated from the highest one. Read voltages much higher than 6V are not practical, both because reliability issues (read disturb) and, especially in low-voltage devices, because of drawbacks of voltage boosting well above V_{cc}. Thus defining more levels requires tighter V_t distributions and less margins between levels. The width of programmed levels can be strongly reduced by means of suitable programming algorithms.

The most accurate method is the staircase gate voltage ramp, consisting in a sequence of program pulses with a constant drain voltage and a gate voltage increased at each pulse by a constant amount. Each pulse is followed by a program verify operation. After an initial phase at low gate voltages, the V_t shift after each pulse is equal to the gate voltage step, thus allowing to

APPLE INC.
EXHIBIT 1008 - PAGE 0444

obtain a V_t distribution of the same order [19]. Very small distributions can be achieved, but at the cost of programming speed, because a large number of program/verify steps is required. Similar considerations apply to noise margins, that can be reduced by using high precision circuitry, but at the cost of access time. The best trade-off between accuracy and speed must be chosen and this leads to small reliability margins (of the order of few hundreds mV) for data retention and disturbs.

At the present state of knowledge, no specific failure mechanism of multilevel Flash is foreseen, but the impact of every failure mode previously presented must be carefully considered.

As far as performance degradation induced by program/erase cycling is concerned, no major problem should arise when dealing with multilevel storage. The same erase time increase with cycle number as in conventional memory is expected, apart from the small effect of the slightly larger amount of charge flowing through the tunnel oxide due to the increased V_t window. Similar consideration holds for programming, with the additional concern that accurate programming requires low V_g overdrive, which is supposed to be a critical condition for the injection efficiency degradation.

Gain degradation and erratic erase issues are the same as in conventional memory, at least in the case of CHE programming memory. The erratic behavior of the tunneling current may be a problem for those memory using FN programming, because a sudden increase of current during programming may lead to over-programmed bits.

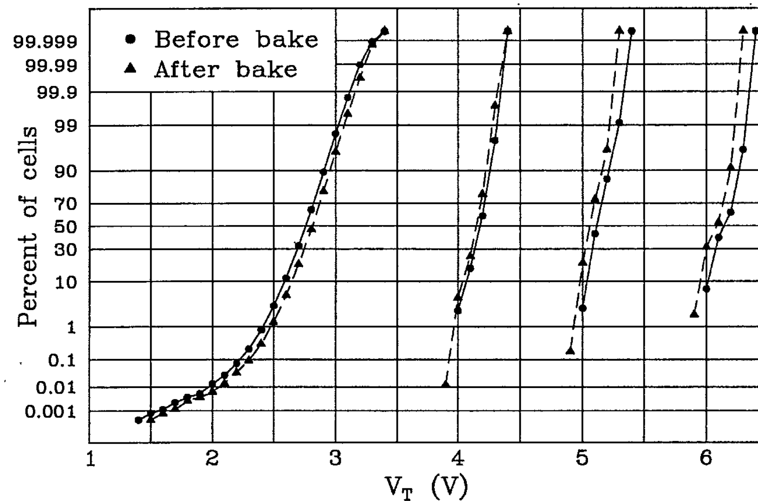


Figure 7.44 Four-level threshold distribution of a 1Mbit array before and after a bake performed at 250°C for 500h.

Data retention is obviously a key issue for multilevel storage. Intrinsic behavior is compatible with the requirements for 3 or even 4 bit/cell, as is shown in Fig. 7.44 where the result of an accelerated retention test is reported. The test condition (500hrs at 250°C) corresponds to more than 10 years at 100°C for a leakage mechanism having an activation energy higher than 0.6eV. It is worth to note that the V_t shift is roughly linear with the programmed V_t, so that the spacing between level is reduced by only a fraction of the maximum shift. This fact can be effectively exploited performing the reading operation by comparison with programmed reference cells [20], which experience the same charge loss of the array cell. A major issue is single bit data loss after program/erase cycling, discussed in Section 7.5.3.2. Here again, the problem is the margin reduction and the higher electric field in the tunnel oxide during storage, associated with the increased V_t window. While in a conventional Flash the electric field just before failure can be very small, in a multilevel Flash the electric field for a cell programmed to the uppermost level is necessarily much higher.

An enhanced sensitivity to program disturbs is expected for three reasons: the reduction of margins, the largest V_t window, and the increasing program time. The worst case is that of column disturb on a cell programmed to the highest level because of the large V_t shift with respect to the neutral state. Moreover, the programming drain voltage is the same as in a conventional Flash, but the program time is substantially increased. On the contrary, row disturb should not be affected by the program time increase, at least for a staircase programming algorithm, since during most of the time the gate voltage is rather low and only the last few pulses will contribute to the disturb.

A major reliability concern is that of read disturb, which affects cells in the lowest V_t state. Besides margin reduction, the higher read voltage with respect to a conventional Flash has a strong impact on read disturb immunity, especially for very thin tunnel oxide. In the case of a virgin cell with 8nm tunnel oxide, the read voltage should be below 7V to guarantee 10 years lifetime, but the worst situation is clearly that of a cycled cell, where SILC poses much more severe constraints both on oxide thickness and read voltage. The data of reference [16] suggest that even with 10nm oxide thickness a read voltage of 6.5V is able to induce on some anomalous cells after 10^5 program/erase cycles a V_t shift of several hundreds mV in a time scale of few days. Even if this kind of anomalous SILC occurs with very low probability (at ppm level or less), in large density multilevel Flash it may have a strong impact on reliability.

From the above considerations we conclude that, in spite of a good intrinsic behavior, the practical exploitation of multilevel memory will be limited by the impact of single bit failures on product reliability. Most likely, multilevel concept will be initially implemented for fault tolerant applications (audio, video) or in combination with error correction techniques.

APPLE INC.
EXHIBIT 1008 - PAGE 0446

7.7 CONCLUSION

The success of Flash memories in the semiconductor market has grown together with the understanding of their reliability issues. In the early stage of their introduction, there was a diffused concern among the system manufacturers in moving from the well known EPROM's to the attractive but just born Flash memories and that concern was mainly about reliability and manufacturability. The following years somehow confirmed that mastering Flash memory production is not an easy task and twice in the first half of the nineties Flash memory market went into supply shortage. Now, Flash memory is widely accepted as an established and reliable technology.

Realizing the key importance of tunnel oxide quality and learning how to improve and monitor it has been instrumental for achieving industrial manufacturing standards as well as reliability standards.

The extensive investigation of failure mechanisms has built a solid knowledge base for improving memory cell and product design.

Intrinsic degradation mechanisms, responsible for the wear-out of device performances in program/erase cycling, are fairly well dominated; through a proper optimization of cell architecture, their impact can be minimized to push the endurance limit into the 10^5–10^6 range.

A quite tricky failure mechanism such as the erratic erase has been identified and deeply studied: solutions have been addressed both at process and at circuit level.

The impact of Stress Induced Leakage Current on data retention will limit the scalability of tunnel oxide thickness, unless error correction techniques or in-system data refresh are utilized to cope with the problem.

Multilevel storage feasibility is demonstrated and assessing the reliability of Flash memory storing more than one bit per cell will be the challenge of next years.

References

[1] Cappelletti P., Panchieri A. and Ravazzi L. (1993) "Mastering key factors which affect Flash memory reliability". Proc. *ESREF*, Bordeaux (France).

[2] Cappelletti P. and Ghidini G. (1995) "Evaluation methods on thin oxide for Flash memories". Proc. 7th *Workshop on Dielectrics in Microelectronics*, Crete (Greece).

[3] Cappelletti P., Ghezzi P., Pio F. and Riva C. (1991) "Accelerated current test for fast tunnel oxide evaluation". Proc. *ICMTS*, 4, p. 81.

APPLE INC.
EXHIBIT 1008 - PAGE 0447

[4] Cappelletti P., Bez R., Cantarelli D. and Ravazzi L. (1994) "CAST: an electrical stress test to monitor single bit failures in Flash-EEPROM structures". Proc. 12th *Workshop on Non Volatile Semiconductor Memory*, Monterey, California (USA).

[5] 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.*, p. 595.

[6] Muramatsu 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 scaling principles for Flash memory". *IEDM Tech. Dig.*, p. 847.

[7] Dunn C., Kay C., Lewis T., Strauss T., Schreck J., Hefley P., Middendorf M. and San T. (1994) "Flash EPROM disturb mechanisms". Proc. *Int. Rel. Phys. Symp.*, p. 299.

[8] 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 Tech. Dig.*, p. 307.

[9] Hu C.-Y., Kencke D.L., Banerjee S.K., Richart R., Bandyopadhyay B., Moore B., Ibok E. and Garg S. (1995) "A convergence scheme for over-erased Flash EEPROM using substrate-bias enhanced hot electron injection". *IEEE El. Dev. Lett.*, **16**, p. 500.

[10] Crisenza G., Clementi C., Ghidini G. and Tosi M. (1992) "Floating gate memories reliability". *Quality and Reliability International*, **8**, p. 177.

[11] Cappelletti P., Bez R., Cantarelli D. and Fratin L. (1994) "Failure mechanisms of Flash cell in program/erase cycling". *IEDM Tech. Dig.*, p. 291.

[12] Olivo P., Riccò B. and Sangiorgi E. (1983) "Electron trapping/detrapping within thin SiO_2 films in the high field tunneling regime". *J. Appl. Phys.*, **54**, p. 5267.

[13] Yamada S., Hiura Y., Yamane T., Amemiya K., Ohshima Y. and Yoshikawa K. (1993) "Degradation mechanism of Flash EEPROM programming after program/erase cycles". *IEDM Tech. Dig.*, p. 23.

[14] 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.*, 7A-2, p. 83.

APPLE INC.
EXHIBIT 1008 - PAGE 0448

[15] Moazzami R. and Hu C. (1992) "Stress-induced current in thin silicon dioxide films". *IEDM Tech. Dig.*, p. 139.

[16] Yamada S., Amemiya K., Yamane T., Hazama H. and Hashimoto K. (1996) "Non-uniform current flow through thin oxide after Fowler-Nordheim current stress". Proc. *Int. Rel. Phys. Symp.*, p. 108.

[17] Kato M., Miyamoto N., Kume H., Satoh A., Adachi T., Ushiyama M. and Kimura K. (1994) "Read-disturb degradation mechanism due to electron trapping in the tunnel oxide for low-voltage Flash memories". *IEDM Tech. Dig.*, p. 45.

[18] Runnion E.F., Gladstone S.M., Scott R.S., Dumin D.J., Lie L. and Mitros J. (1996) "Limitations on oxide thickness in Flash EEPROM applications". Proc. *Int. Rel. Phys. Symp.*, p. 93.

[19] Calligaro C., Manstretta A., Modelli A. and Torelli G. (1996) "Technological and design constraints for multilevel Flash memories". Proc. *ICECS*, p. 1003.

[20] Bauer M. *et al.* (1995) "A multilevel-cell 32Mb Flash memory". *ISSCC Digest of Technical Papers*, p. 132.

APPLE INC.
EXHIBIT 1008 - PAGE 0449

8 FLASH MEMORY TESTING

Giulio Casagrande

STMicroelectronics
Stradale Primosole 50, 95121 Catania, Italy
Giulio.Casagrande@st.com

Abstract: This chapter is not aimed at providing a complete testing theory about Flash; its objective is to present and analyze the most critical aspects related to Flash testing, the tools and methods to improve their testability; to give an idea of the test flow, and of its relation with the excellent quality and reliability that Flash have reached. Aspects related to test cost and productivity are also presented.

The subject is seen from the viewpoint of the Flash manufacturer and treated in very practical terms, with the intent to give an insight into these aspects to the non-expert reader.

Although most of the aspects may be valid for other Flash technologies, this chapter refers to the mainstream Flash technology: NOR architecture, erased by Fowler-Nordheim, programmed by Channel Hot Electrons.

The subjects of testing Known-Good-Die, Flash Cards or embedded Flash are not presented: each one would have required a dedicated chapter.

For the reader interested in a more theoretical and formal insight into Semiconductor Memory testing, excellent books exist (e.g. [1]); for the test engineer with the need to go more deeply in the practical details of Flash testing, exhaustive datasheets and application notes are published by Flash manufacturers.

8.1 INTRODUCTION

8.1.1 *Impact of Testing on Product Cost*

Before discussing the test subject in technical terms, it may be worth giving an idea of the economic aspects of it. The semiconductor industry is well known

APPLE INC.
EXHIBIT 1008 - PAGE 0450

for requiring huge amounts of investments: as a rule of thumb, to generate an increase of revenues of X million dollars, the same amount of money must be invested in process facilities and equipment. For Flash production, the percentage of testing related investments may be in the 10% range, significantly over shadowing the investments related to assembly.

Testing impact on the cost of the finished product may be in the 5 to 15% range, largely varying with volumes, product and process maturity, complexity, etc.

The product lead time is also impacted by testing: testing on wafer and after assembly may contribute with 1 to 2 weeks to the cycle time of the Flash packaged product.

8.1.2 Impact on Product Life Cycle

Testing is a pervasive activity which, in its different aspects, follows a Flash product during all its life (see Fig. 8.1). In the product development phase, design for testability aspects must be considered and product design and test development therefore proceed concurrently. Design debug and validation may be carried out with the support of specific test hardware and software and it is generally the result of a teamwork involving designers and test engineers.

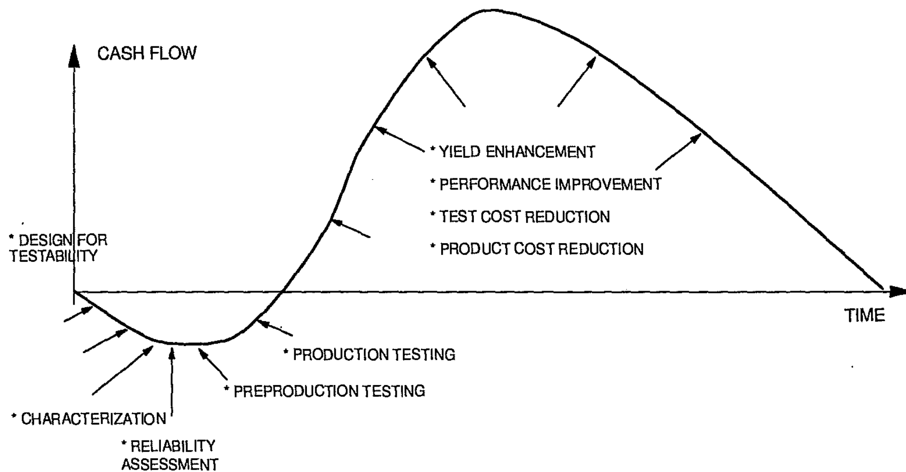


Figure 8.1 Testing versus product payback curve.

The products are qualified through an extensive characterization, a reliability assessment scheme and small production trials. The quality of the testing work (in its broad sense) done in the product development and qualification phase has

APPLE INC.
EXHIBIT 1008 - PAGE 0451

a large influence on the time-to-market, time-to-volume and on the economic success of the product.

In the volume production phase, the highly competitive nature of the Flash business still requires a continuous effort in the testing area addressed to yield enhancement, performance improvement, zero-defectivity, test cost reduction and product cost reduction in general. As common for all memory families, a Flash product may be shrunk or ported in a more aggressive technology: this requires a repetition of most aspects of the development-qualification-industrialization cycle.

8.1.3 Objectives of Production Testing

Production testing may be defined as a screening sequence that contributes to guarantee that the product shipped complies with a certain specification in terms of performance, quality and reliability. Production testing has also other collateral objectives like, for example, the collection of data that allows better control and fine-tuning of the whole manufacturing process and progress in the yield enhancement process. Production testing also serves to differentiate the same product by different classes of performance (speed, Vcc range, temperature range, erase/program performance ...). Similar to all memory families, most Flash products make use of redundancy to improve yield, and production testing has also the objective of defect identification, diagnosis and fault repairing, utilizing at best the spare resources available.

All the above, with very stringent limits in terms of cycle times impact overall test cost and investments required.

8.1.4 Testing Versus Quality and Reliability

Production testing is not the predominant factor to guarantee quality and reliability of a Flash product; it is only one of them. Other key factors include:

- design of the cell, process and product;
- debug and characterization of process and product;
- the manufacturing machine: equipment, resources, materials ...;
- the overall process control;
- the reliability assessment of process and product;
- the overall know-how of the manufacturer.

APPLE INC.
EXHIBIT 1008 - PAGE 0452

8.2 FLASH TESTING ASPECTS

8.2.1 Flash Functional Model

A functional model of a Flash is reported in Fig. 8.2; an exhaustive description of the Flash structure from the architecture and design viewpoint is reported in Chapter 5. Here the model intends only to show how, in spite of its rather simple usage from the user standpoint (ROM-like for reading and SRAM-like for the erase and programming commands) the internal structure has become quite complex in order to make the difficulties of the erase and programming operations "invisible" to the user; this implies a similar complexity for the testing.

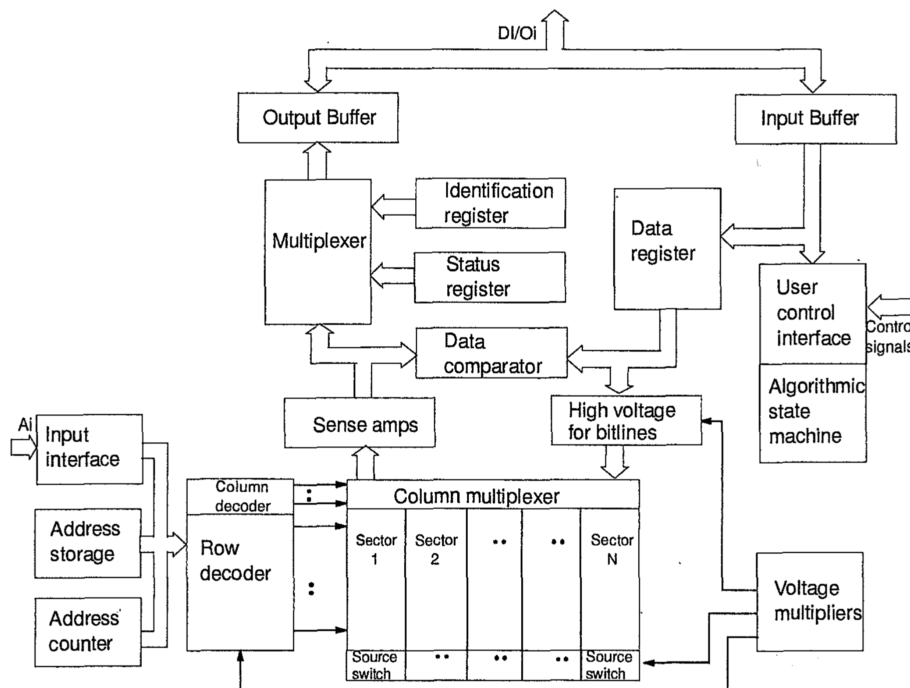


Figure 8.2 Flash functional model.

8.2.2 Oxide Stress in a Flash

Thin oxides are one of the most critical elements of the proper functionality and reliability of a VLSI product; for Non Volatile memories they are definitely the most critical.

Table 8.1 Breakdown of critical areas/stresses (Flash 4 megabit, 5V only).

	MEMORY ARRAY	PERIFERY (total)	PERIFERY (high voltage)
Oxide area	$1mm^2$ (tunnel) $2mm^2$ (interpoly)	$2mm^2$	$1mm^2$
Thickness	110Å(tunnel) 250Å eq (interpoly)	200Å	200Å
Operating field	10MV/cm (tunnel) 2.5MV/cm (tunnel,interpoly)	2.5MV/cm	5MV/cm
Charge passed	$0.1C/cm^2$ (tunnel)	–	–
Operating life	50hrs (tunnel at 10MV/cm) 20 years (tunnel, interpoly at 2.5MV/cm)	20 years	50hrs
Requirements	Retention $R > 10^{24}$Ohm	Good insulator $R > 10^9$Ohm	Good insulator $R > 10^9$Ohm

This concept is illustrated in Tab. 8.1 (referred to a 4 megabit example): the total area occupied by thin oxides in the array is comparable to that in the periphery, but the tunnel oxide in the array receives a much higher stress (10MV/cm compared to 5MV/cm for the part of the periphery working at high voltage); furthermore, while in the periphery the oxide must only guarantee a good insulation (tens of MΩ) to allow the MOS transistor to switch properly, the cells in the array must guarantee the retention of data for many years: in terms of resistivity there are about 15 orders of magnitude difference.

This explains why most of the testability tools and most of the attention in terms of screenings are addressed to the matrix.

8.2.3 Flash Testing Aspects

Testing of a Flash product must take into account the list of aspects reported in Tab. 8.2.

A peculiar aspect of Flash memory testing relies on the fact that the erase and write mechanisms are significantly slower than the read operation: erasing or writing a sector requires a time in the order of magnitude of the second (or fraction of it) while reading the same sector takes only milliseconds; a major difference with respect to DRAM and SRAM where write takes the

Table 8.2 Main test aspects for a Flash.

Main test aspects
DC TEST/PARAMETRIC
AC READ/COMMAND INTERFACE
ALGORITHMIC STATE MACHINE
ERASE PERFORMANCE
WRITE PERFORMANCE
ERASE/WRITE DISTURBS
REDUNDANCY: DIAGNOSIS AND REPAIR
RETENTION
ENDURANCE
LIFE TEST (READ)

same short time as read. This implies that for Flash it is rather impractical to conceive a test flow that checks analytically, in a linear sequence, the different aspects independently; instead, the test flow must be designed in order to try to minimize the number of erase/program cycles and the coverage of the various aspects must be distributed along the flow itself. Also the defect diagnosis and repair may be applied at different steps in the flow.

8.2.4 Conceptual Test Flow

In order to allow the reader to understand the following paragraph on Testability Tools and Fault Repairing, a conceptual production test flow for a Flash product is introduced here (Fig. 8.3): the subject will be treated more in details in Section 8.4.

A first wafer sort test includes a complete coverage of most of the aspects: all DC and parametric tests, read tests, functional tests of the command interface and of the algorithmic state machine, erase and program performance and related disturbs. AC conditions are generally relaxed at probe test.

The (almost) complete Wafer Probe testing also allows for the detection of defects during the various test routines, defines whether they are reparable and decides the repairing strategy. Most Flash producers utilize flash-based non volatile elements to permanently store the address of the spare rows or columns

APPLE INC.
EXHIBIT 1008 - PAGE 0455

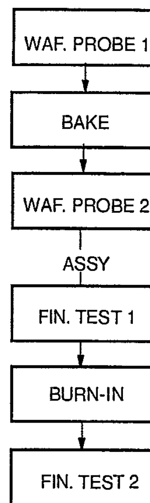


Figure 8.3 Flash conceptual test flow.

utilized for repairing; in this case the repairing may also be made during the first wafer sort test. Key parameters that may be collected and monitored to provide a short-loop feedback to the process may include: Icc current in active and stand-by conditions, cell current, erase and programming times, threshold voltage after programming, etc. The prime (non repaired dice) yield, the yield including repaired dice, the number of hard failures and the calls for repairing at the different test routines are recorded and the analysis of these data represents a major source of information for the yield enhancement activity.

A bake step at a temperature in the range of 250°C is typically performed on all programmable non volatile memories (typically 24 hours or more) to accelerate potential retention failures and a second wafer test step is used to screen them out.

After packaging, a typical test flow may be made of two steps at different temperatures (e.g. 25°C and 80°C for a commercial range product); all aspects are covered at the worst case corner conditions. AC performance is fully tested using worst-case patterns.

A burn-in step is often inserted between the two steps to accelerate single cell failures which are sensitive to the combination of voltage and temperature and may also be caused by the stress related to the assembly process; 125°C or slightly more (for several hours) is the temperature typically used (limited by the characteristics of the plastic package) and a supply voltage 30 to 50% higher than the specified Vcc.

APPLE INC.
EXHIBIT 1008 - PAGE 0456

8.3 FLASH TESTABILITY TOOLS

Since Flash was born, ad-hoc testability tools have been designed into the products in order to facilitate the technology learning curve and its reliability improvement; in the more recent Flash generations characterized by embedded erase/program algorithms, larger memory sizes subdivided in many sectors and with internally generated erase/program voltages, the need for designed-in testability tools is even increased.

Test tools may be addressed to different purposes:

- array characterization and screening;
- test productivity;
- design testability;
- redundancy;
- product differentiation.

Most test tools can be used in the production test flow, provided the related test routine is within acceptable test times (or $\ll$ seconds). Test modes are extensively used, with less stringent time constraints, during product debug, product/process characterization, for reliability assessment, failure analysis, etc. Test circuitry can impact the die size by 2–5%, the impact reducing with memory size and the progressing experience on Flash.

8.3.1 Focus on Cell and Technology

The most common test modes in this category are addressed to:

- directly access the cell characteristics (array, reference cells, ...);
- generate high voltage stress modes for fast defect screening;
- screen out depleted or low-Vt cells;
- set the threshold of reference cells;
- modify sense conditions or cell bias conditions during the different operations for debug, margining or characterization.

Some of the most commonly used test tools are described in the following paragraphs.

APPLE INC.
EXHIBIT 1008 - PAGE 0457

8.3.1.1 Direct Memory Access. A primary tool for failure analysis is DMA (Direct Memory Access), see Fig. 8.4: a pass transistor bypassing the sense amplifier allows the direct access from the I/O pad to the selected bit line. The sense amplifier circuitry is disabled and the output buffer tristated. Vpp may be applied as a supply voltage to the row decoder in order to vary the gate voltage of the addressed cell over a wide range. The scheme could be repeated for all the 8, 16 (or more) I/O pins. The characteristic of each individual cell inside the array can be observed on the I/O pins as illustrated in Fig. 8.4.

On top of the obvious use for failure analysis, the DMA tool can be used to monitor the typical cell's current for production control, or for process/product characterization. A tight distribution of the cell currents inside the array (after UV erase, or after electrical erase, or after programming) is one of the key issues for a Flash device and is commonly used to evaluate a new process, process changes or shrinks.

The limitation of the method is in the intrinsic slowness of the testers' parametric units: tens of milliseconds typically. To better exploit this capability some Flash testers feature fast parallel parametric units (one unit per I/O) that allow measuring the current of 16 cells in few milliseconds. Nevertheless a full cell current distribution for a large Flash memory like in Fig. 8.5 may require a few hours.

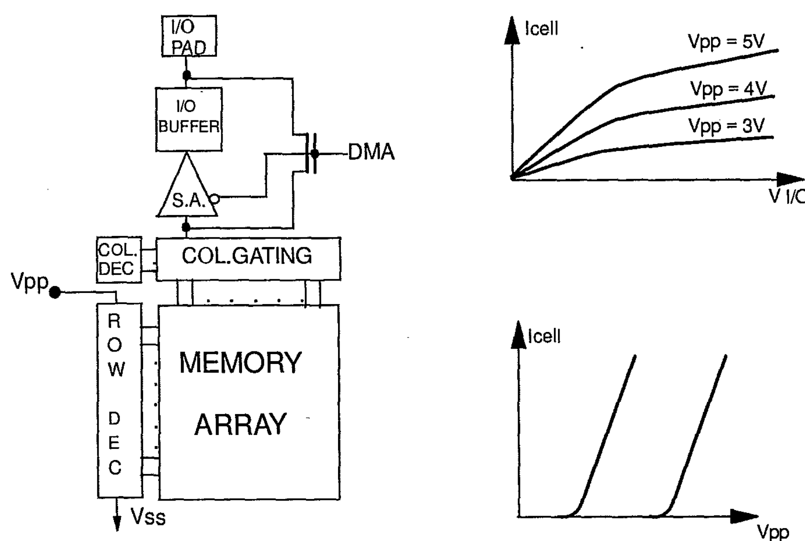


Figure 8.4 Direct Memory Access.

APPLE INC.
EXHIBIT 1008 - PAGE 0458

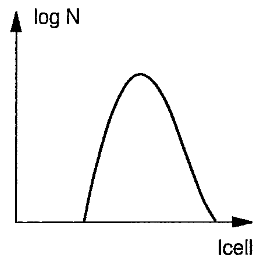


Figure 8.5 Distribution of cell currents inside a full chip obtained through DMA.

A faster (and more practical) analysis can be performed in a few minutes, by just exploring the tails of the distribution.

8.3.1.2 Vt Measurement. Most non volatile memories utilize a double-ended sensing scheme (see Chapter 5): the current of the addressed cell is applied to a load (transistor), the voltage output is feeding one input of a voltage comparator; the other side is fed with the output of a similar structure where the current of a reference cell is applied to a load which is in a given ratio (e.g. 2) with the load on the matrix side.

With little modification this sensing scheme can be altered to allow for the measurement of the threshold of the matrix cell: in the example illustrated in Fig. 8.6, a fixed reference current and a ratio 1 between the loads are used. By applying a varying voltage to the gate (e.g. supplying the row decoder with Vpp), the Vt (at a given current level) can be measured as the Vpp at which the I/O pad will toggle.

This method allows the Vt measurement to be performed in the same time as a read operation.

In practical terms the limiting factor is the error-counting time of the tester: a Vt go-no-go test on a large memory (or counting the cells with Vt below a given value, and produce a topological bit-map if desired ...) can be obtained in the range of minutes. To characterize the Vt of only the top or bottom of a distribution can be made in less than a second. Testers may provide techniques for fast error count by I/O to improve speed.

From the circuit point of view the limitation is imposed by the minimum operating voltage of the row-decoder (about 1 volt): no Vt below that point can be measured except by extrapolation; to help this operation the Iref can be made variable.

Vt measurement is largely used in production testing to guarantee margins after programming, retention bake, erase, etc. Examples of utilization of the tool for process/product characterization are illustrated in Fig. 8.7: the erase

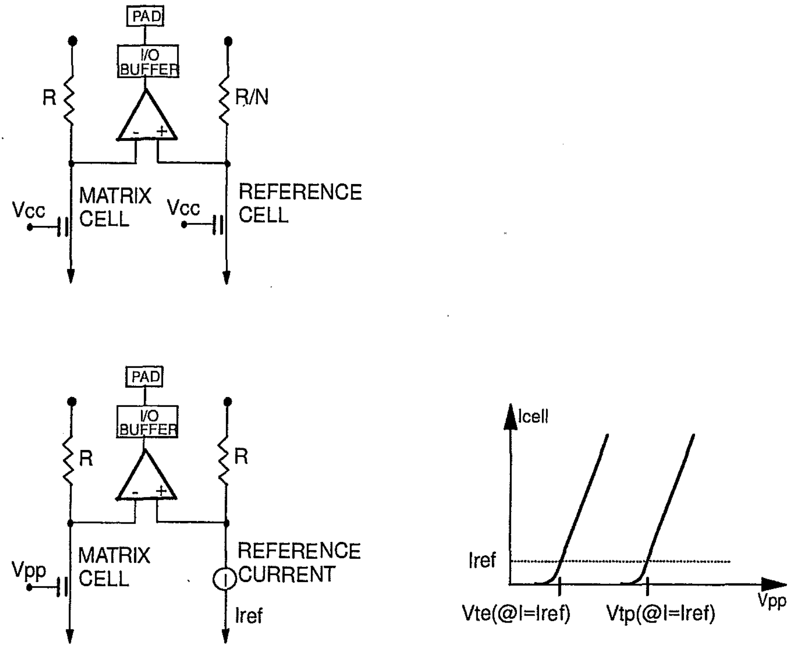


Figure 8.6 Sense scheme in read (top) modified (bottom left) to allow for on-chip cell Vt measurement (bottom right).

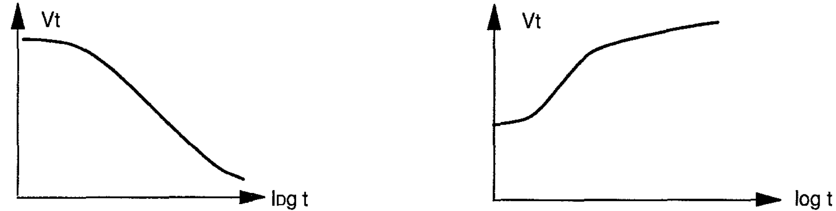


Figure 8.7 Erase (left) and programming (right) characteristics of single cells obtained on-chip by means of the Vt measurement tool.

and program characteristic of a single cell inside the product can be calculated and compared to the ones obtained on elementary structures during process development.

Another example of utilization of Vt measurement can be the one illustrated in Fig. 8.8: 3 characteristics of cells are reported in the right side (well erased, well written, and, in between, a marginal one). In the hypothesis that a double ended sensing with ratio N is used, the dashed line represents the reference cell's current divided by N. By marginally erasing or writing a cell, a relationship

APPLE INC.
EXHIBIT 1008 - PAGE 0460

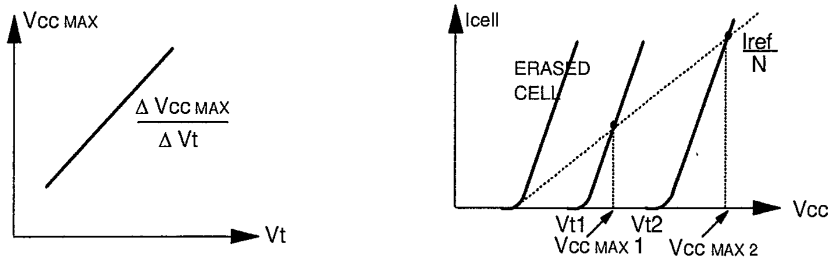


Figure 8.8 Characterization of the sense amplifier behavior obtained on-chip through the use of the Vt measurement tool.

can be determined between the Vcc_{max} (i.e. the maximum Vcc at which the cell is correctly sensed) and its threshold (on the left) in order to check whether the sensing performs as expected.

8.3.1.3 Stress Modes. Potential failure modes of Flash technology are charge gain or charge loss failures caused by anomalies in the tunnel or in the interpoly oxides, generally impacting single cells. These are emphasized by the so-called array disturbs (see Chapter 7) mostly associated with the high voltage stress applied to the word lines and bit lines while programming; the most typical can be:

- charge gain on an erased cell associated with Gate stress;
- charge loss on a programmed cell associated with Gate stress;
- charge loss on a programmed cell associated with Drain stress.

Flash chips typically incorporate test modes to allow fast activation of disturb-sensitive cells like:

- Gate stress: all word lines (together) at high voltage (Vpp), bit lines disconnected;
- Drain stress: all bit lines (together) at 5–6 Volt, word lines grounded.

Disturb-sensitive cells are screened by detecting the effect of the stress on the cell's threshold or on the Vcc_{min} or Vcc_{max}: as illustrated in Fig. 8.9 a charge gain on an erased cell increases the minimum Vcc at which the cell can be correctly read (left) and a charge loss on a programmed cell reduces the maximum Vcc at which the cell can be read (right).

APPLE INC.
EXHIBIT 1008 - PAGE 0461

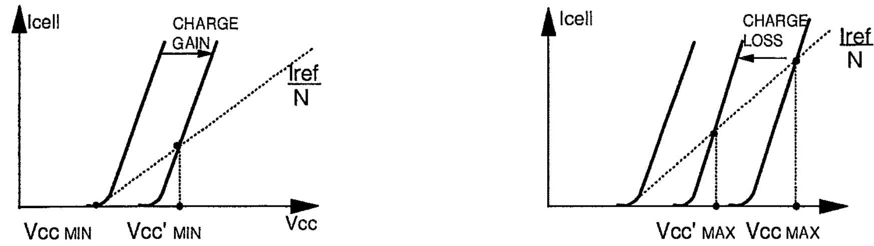


Figure 8.9 Impact of charge gain (left) and charge loss (right) on the minimum and maximum Vcc operating range.

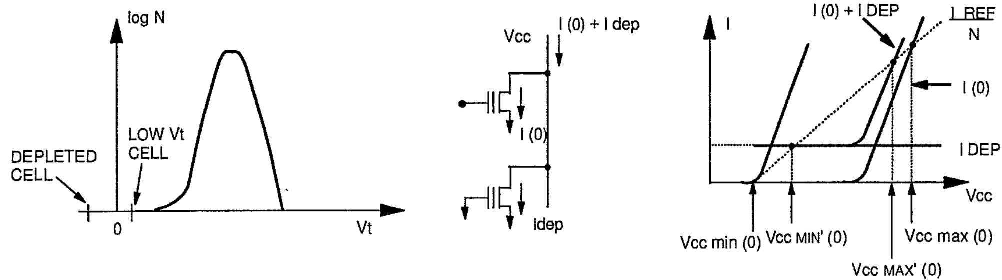


Figure 8.10 Example of cells' Vt distribution of a full array (left); impact of a depleted cell on the Vcc operating range (right).

8.3.1.4 Depletion/Low-Vt Test. Depleted cells after erase represent a major potential cause of failure for the NOR Flash technology; the current drawn by the depleted cell can disturb the read of all programmed cells within the same bit line (or sub-bit line in case of sectorization by rows); Fig. 8.10 (right) shows the effect of a depleted cell on the sensing: the minimum supply voltage at which the programmed cell can be correctly sensed is increased from $Vcc_{min}(0)$ to $Vcc_{min'}(0)$, but also the $Vcc_{max}(0)$ is reduced to $Vcc_{max'}(0)$.

Low threshold cells (cells with Vt few hundred millivolt higher than zero) could also increase the probability of hard or erratic depletion failures or of gain degradation.

Depleted or low Vt cells can be detected by specific test modes for characterization, screening, or to activate recovery algorithms, etc.; a simplified scheme is shown in Fig. 8.11.

Depletion Test. Row decoder disabled, row decoder "grounded" at 0 Volt → all word lines grounded; if DMA (direct memory access) is enabled a depleted

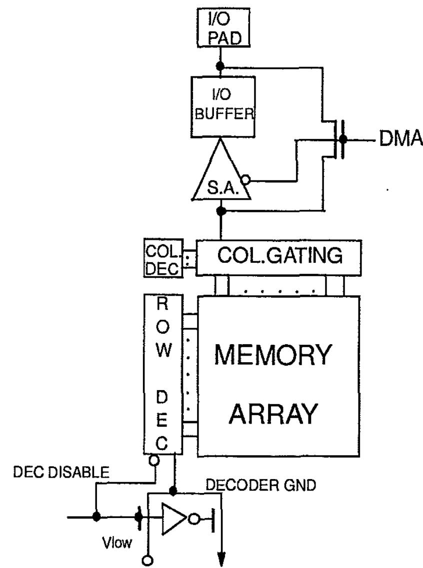


Figure 8.11 Simplified scheme of a testability tool for detection of depleted or low-Vt cells.

cell's current can be seen on the I/O; if the sense is set for normal operation, it should detect a zero for Vcc down to the minimum voltage at which the circuitry stops working; if a 1 starts to be sensed at a higher voltage, a depleted cell could be present on the bit line. Similarly, the depleted cell can be detected using the Vt measurement mode.

Low Vt Test. Row decoder disabled, row decoder "grounded" at a positive (V_{low}) voltage $\rightarrow$ all word lines at V_{low}; DMA allows for the detection of a cell with Vt lower than V_{low} (on the addressed bit line). The lower-side of the erase distribution can also be explored with this tool.

A depletion test can also be performed without the use of a specific depletion test mode: a single row is programmed, the rest of the array remains erased; addressing each cell in the programmed row at low Vcc (below 4 to 5V) with DMA enabled, a current on the bit line would likely indicate a depleted cell (varying the I/O voltage the depleted cell current would be flat ...).

Also with the sense in normal operation an abnormal $Vcc_{min}(0)$ could indicate a depleted cell.

The depletion tests described above can only locate the bit line affected (and more than a depleted cell's current can be cumulated ...). Scanning the bit line using DMA or Vt test can then allow to identify the exact cell location.

8.3.2 *Focus on Test Productivity*

Due to the intrinsic slowness of the Erase and Program mechanisms, test time has always represented a concern for the test engineer. For a first generation, bulk-erasable Flash, the chip erase time and programming times were both in the range of a second; consequently a test pass would take in the range of ten seconds (the numbers are related, as is the rest of the chapter, to a NOR Flash programmed by Channel Hot Electrons and erased by the Fowler-Nordheim mechanism). Second generation products, featuring sectorization and larger in size, would be much slower to test: for a 16 megabit with 32 sectors the erase time would be multiplied by 32 times and the programming time by 16 times resulting in the range of a hundred of seconds.

To avoid unacceptable costs, the possibility of parallelism has been exploited, both internal to the chip and external.

Two basic test features are often provided internally:

• parallel programming: few bytes or words can be programmed together.
The limitation of the technique is in the magnitude of current required to program a bit (hundreds of microamps for the mainstream Flash); by increasing the number of bits, internal resistive voltage drops are increased making the drain voltage lower and, consequently, making the program slower and misrelated with the normal operation. In particular, this issue becomes critical for single-supply, low-Vcc (3V or below ...) products, where the drain voltage needs to be provided by internal charge pumps.

For easier implementation of the test mode, the bytes are often programmed all with the same content; in some products parallel programming might be offered as a specified feature also to the user: a page register is provided to allow for storing the random content to be programmed in parallel.

• Parallel erase: several or all sectors erased in parallel.
Initially introduced as a test mode, this capability is now widely offered as a specified feature; the command to erase any combination of sectors simultaneously can be issued to the chip, and the erase algorithm takes care of the parallel erase. The limitation comes from the source current required by the sectors being erased (range of the milliampere ...); again, this has an impact on the size of the internal charge pumps on single-supply low-Vcc products.

8.3.3 *Focus on Design*

The complex logic structure of a Flash requires that specific testability features are addressed to the design, mainly because:

- the logic is mostly sequential;

APPLE INC.
EXHIBIT 1008 - PAGE 0464

- the logic machine does not generate directly digital outputs available on the pads; instead the logic machine generates outputs that activate analog inputs to the array which in turn produce effects in the array status. Only the array status can then be read on the I/O pads;
- an incorrect behavior of the logic machine or of bits in the array (caused by incorrect design or process or by manufacturing defects) would then be very difficult to understand and to screen.

In addition, testing the logic machine alone before the costly testing of the array could be a way to save test time. To support this, Flash testers are generally required to provide Vector Testing capability (or the possibility to test a logic structure by providing input vectors and comparing the outputs of the structure with output vectors provided by the tester); this was not usually required by previous Non Volatile Memories.

The testability features of more interest can be designed to:

- allow for testing the PLA(s) as combinatorial networks;
- perform independent test of registers/counters;
- control the matrix from the tester rather than from the internal state machine;
- measure or force significant logic or analog nodes;
- execute erase/programming algorithms with altered logic or analog conditions with respect to user modes for screening purposes.

All the above capabilities are useful for debug, characterization, production testing and failure analysis.

8.3.4 Flash Design Testability: an Example

An example of a testability structure designed into a Flash memory is here described; the purpose is to provide good observability and controllability of the logic/analog circuitry and of the array (Fig. 8.12).

The circuitry specifically added for test purposes consists of:

- an 8 wires bi-directional bus connected to I/O pads;
- 3 wires for analog signals connected to address pads;
- latches already available or specific test registers to store test signals;
- multiplexers to control the access to the bus or to the analog wires;

APPLE INC.
EXHIBIT 1008 - PAGE 0465

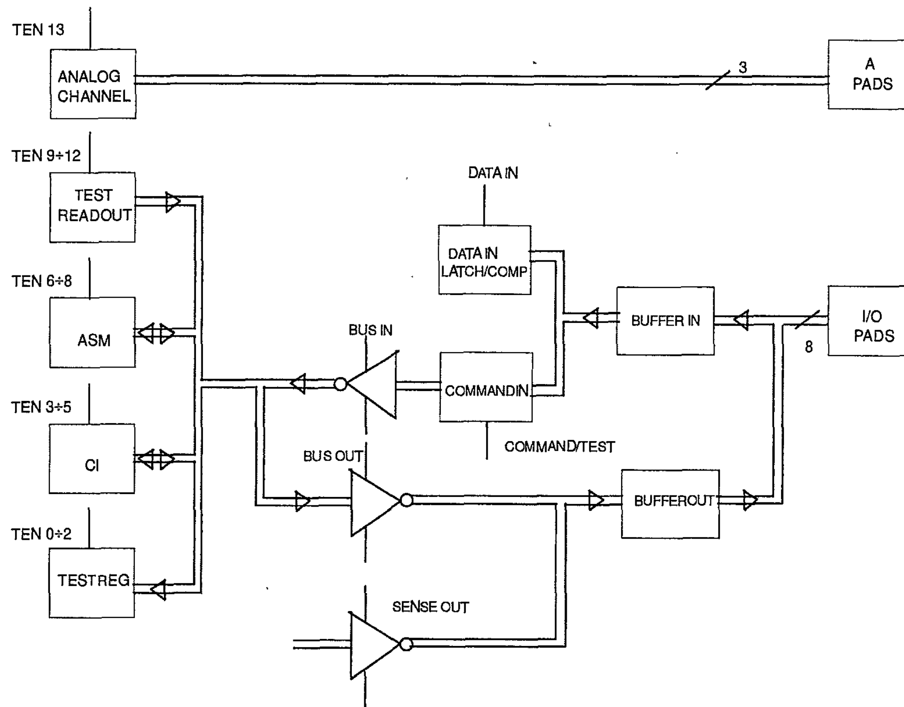


Figure 8.12 Example of a Flash design-for-testability structure.

- 4 address pins that are decoded to select the multiplexers;
- a simple D/A converter supplied by Vpp, controlled by test bits to generate variable analog voltages;
- two third level (12V) detectors to provide hardware protection against unwanted activation of test modes.

After providing a third level on the two input pins (Figs. 8.13 and 8.14), the code for the test mode is issued to the Command Interface as for user modes; the Output Enable pin selects the direction of data on the I/O. A DU (Don't Use) pin is utilized to switch between test and user mode. After the successful acceptance of the test mode start sequence, the Command Interface allows the information to be passed to the test latches. Each latch modifies the behavior of a circuit, such that it still may be used in other operations of the device (e.g. program, erase, read, etc.). Considering that a latch produces a function, these functions can then be combined to produce a complex test function, allowing a flexible approach to testmode development.

APPLE INC.
EXHIBIT 1008 - PAGE 0466

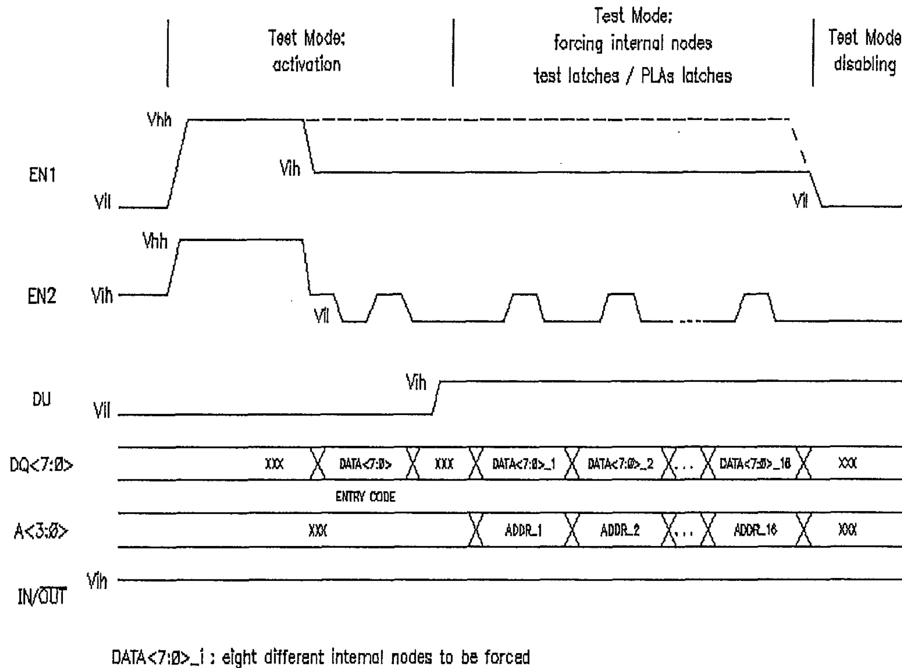


Figure 8.13 Example of test-mode activation: controlling internal nodes or loading test registers.

Fig. 8.13 illustrates an example of forcing test data, Fig. 8.14 an example of reading test information.

In this example 92 nodes could be observed via the different test modes and 128 nodes forced, among which the inputs and outputs of the algorithmic state machine, connected to the test bus as in Fig. 8.15.

8.4 FAULT REPAIRING

In all types of memories the number of transistors per square millimeter is much higher in the array than in the periphery due to the more regular structure of the array and, often, also due to the use of tighter layout rules; also, as seen in Section 8.2.2, the stresses and the requirements are higher for the cells compared to the periphery transistors. As a result, the defect density in the array is typically much higher than in periphery. For this reason most memories utilize fault repairing as a way to increase yield and reliability.

Redundancy is the most commonly used method for fault repairing in Flash; Error Correction could be considered as a way to increase yield, but it is not of relevant interest for standard Flash.

APPLE INC.
EXHIBIT 1008 - PAGE 0467

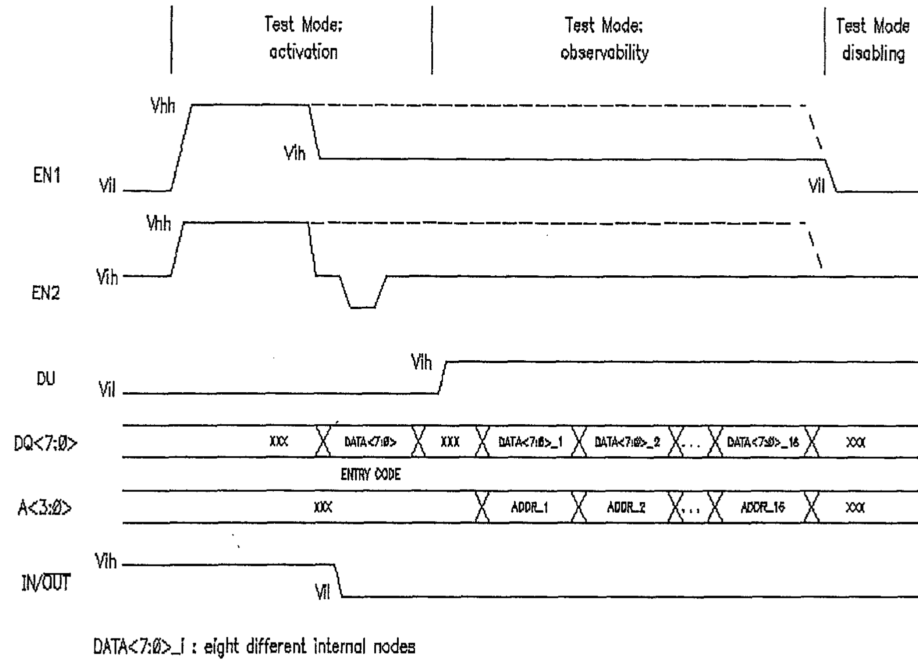


Figure 8.14 Example of test-mode activation: observing internal nodes or reading test registers.

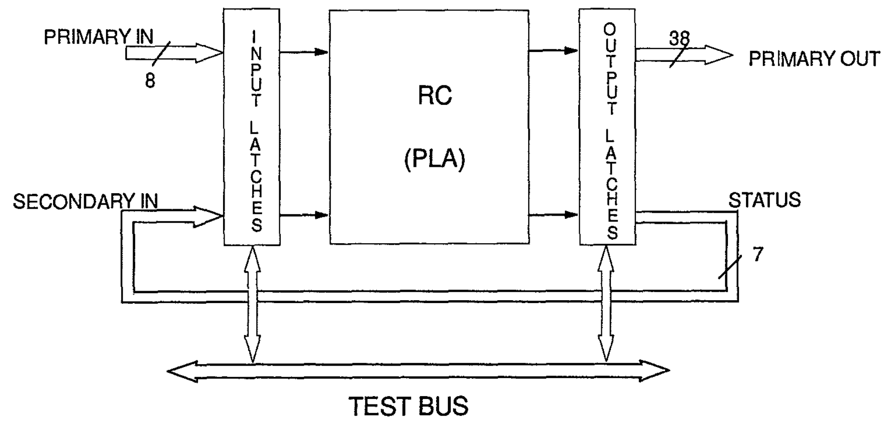


Figure 8.15 Access to the algorithmic state machine for direct testability.

APPLE INC.
EXHIBIT 1008 - PAGE 0468

8.4.1 Error Correction

Error correction is more used as a technique to improve reliability at board level when using large arrays of memories, but it could be considered also (at least in theory) as a way to increase yield inside the memory chip. A detailed analysis of error correction is reported in Chapter 5. As a practical example, a Flash organized X16 could use an Error Correction Code (ECC) scheme with the objective to correct up to a single bit error per word; the requirements would be 5 bits added to each word (30% overhead for the array) plus ECC circuitry in the data-out path to extract from the (potentially faulty) 21 bits the corrected 16 bit word and, in the data-in path, to generate for each word the additional 5 bits to be programmed (few percentages in die area plus a delay in access time); sense amplifiers and column selection need to be also increased by 30%.

The impact on yield depends on the level of defectivity and on the dominant types of defects; for example, a very high level of single bit random errors can be corrected almost completely; on the contrary, clusters of bits or a failing word line cannot be corrected.

In the example, in synthesis, the overhead required would be 20 to 30% plus a penalty in access time of several nanoseconds; the benefit in yield and reliability is widely variable according to level and type of defectivity.

The above considerations, on top of the excellent reliability demonstrated by Flash, explain why standard Flash generally does not use ECC; on the contrary, ECC can be more convenient when the architecture makes use of very long words (low impact of the correction bits in %) and/or serial access (access time penalty not so important); ECC might be also required for multilevel Flash technology.

8.4.2 Redundancy

Redundancy is widely used in all memory types as a way to increase yields. A functional scheme often used is illustrated in Fig. 8.16: few spare word lines and few spare bit lines are added to the matrix array; the access to each spare resource is controlled by a CAM (a group of Content Addressable Memory elements) whose content can be programmed permanently. If a word line in the regular array is found to fail, its address will be programmed into the CAM. After that, each time the "failing" word line is addressed, the CAM will produce a "hit" signal that will disable the regular decoder and will activate the spare word line. The same concept is valid for the bit line.

To store permanently the desired address in the CAMs, Non Volatile Memory elements are used. Most manufacturers utilize Flash cells which are very convenient from the process and testing standpoints.

APPLE INC.
EXHIBIT 1008 - PAGE 0469

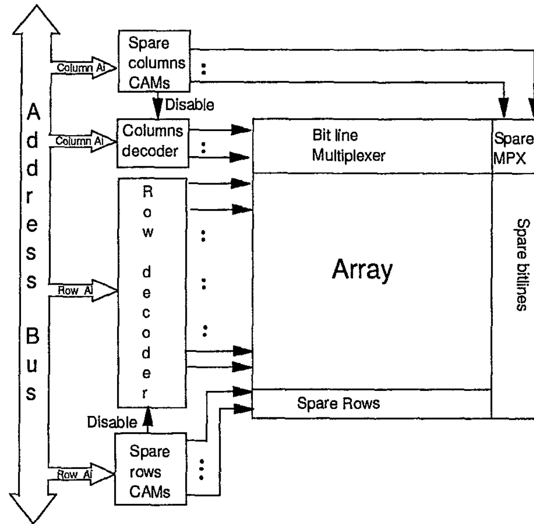


Figure 8.16 Simplified redundancy functional model.

Much less convenient in Flash are fuse elements to be laser blown (like in DRAM) which add manufacturing complexity, or polysilicon fuses to be electrically blown, also difficult to manufacture: both options are also inflexible as they are one-time-programmable.

In reality the scheme in Fig. 8.16 becomes much more complicated due to the data width X8 or X16 or more, due to the organization by sectors and to the need to avoid any penalty in performance (access time, current consumption). Sometimes entire sectors or groups of bit lines or word lines may be substituted together to account for large defects; or the repaired element can be a subbitline. To repair a single bit is impractical.

The cost in terms of die size is generally very moderate (few percent). The yield improvement obtainable can be extremely variable: in the order of a factor of ten at the early stage of the learning curve for a large memory in an advanced technology, while for a small memory in a mature technology the increase will be small (a few percent). It is common in memories to shrink the products as soon as the learning curve of a given technology progresses; this has the effect of increasing the killer defect density and hence to increase once again the positive impact of redundancy on yields.

For the above reasons redundancy is used in most Flash products to lower the die cost.

Redundancy has no direct impact on reliability (self-diagnosis-and-repairing is an appealing idea but is too complex to be used in the state of the art).

APPLE INC.
EXHIBIT 1008 - PAGE 0470

Nevertheless, the availability of several spare resources allows to apply very severe screenings on the array cells' distribution which contribute significantly to the excellent reliability of Flash.

The fact that a Flash product utilizes redundancy is totally transparent to the user; instead, it impacts significantly the testing from the manufacturer side in terms of test procedures, need for testability tools and requirements of the tester used.

8.4.2.1 Diagnosis and Repairing. Redundancy allows the repair of only (part of) the defects in the array and in the row decoder and column selectors; parametric failures, failures in the data path, I/O or address buffers, failures in the algorithmic state machine or in the Erase/Programming circuitry cannot be repaired.

Defect detection and repairing is mostly done during first Wafer Probing and it is mainly composed of three conceptual steps:

1. fault detection and real time memorization of fail locations;
2. analysis of errors logged to decide if and how DUT should be repaired;
3. repairing.

To support these steps the memory testers feature:

- Error memory;
- a small Fail RAM that can be used to store a list of failing address locations, or two vectors mapping rows and columns of the array respectively, where the content of each bit in the vector indicates if a failure has been detected in the correspondent row or column; the two vectors are then used to make a rapid diagnosis to decide if the part is reparable or not;
- an Error Catch Memory mapping topologically the DUT array on a bit-per-bit basis;
- scramble RAM mapping topologically rows and columns (and sometimes I/Os) in the DUT array;
- dedicated redundancy processor.

During each functional test step (read a pattern, program a pattern, erase a sector, depletion test ...) one or more failing bits can be detected and must be stored in real time by the tester in an error memory which maps topologically the Flash memory under test. The errors detected during several test steps can

APPLE INC.
EXHIBIT 1008 - PAGE 0471

be cumulated in the error memory. After a single (or a group of) test step, the error memory must be analyzed to diagnose if, given the spare resources available, the DUT is reparable and what is the best combination of spare resources that should be used.

Actual activation of spare resources is typically made only when necessary during the flow (because the failures detected would disturb the following tests, if not removed), otherwise it would be made only at the end, when the error map is complete and the repairing strategy can be "frozen": "failing" addresses are stored in the CAMs and the relevant spare rows or columns activated permanently. In the following test steps, the DUT will behave as fully functional and the fact that some repairing has taken place will not be visible to the user.

On the subject of fault detection, it must be noted that, while detecting a fail during a read is straightforward, this is not so easy for erasing failures: for example, a single bit that cannot be erased in a sector; if the standard "user mode" internal erase algorithm is used, the sector erase would fail for "exceeding time limit"; after that a read could easily detect the failing bit, but the other cells in the sector would probably have thresholds which have been pushed towards negative values and the full sector would probably fail the next programming step. The example makes clear why test-mode erase routines need to be used and careful monitoring of the erase distribution and of the anomalous bits behavior must be applied to detect properly and safely potential failures.

Also for programming, test modes should be used in order to allow a screening more severe than the specification.

Row redundancy is also an issue: a word line in the array that is substituted by a spare one will still be physically there for the life of the product; the design of the internal erase routine must still take care of it to avoid that it could cause depletion during the many erase operations applied to the same sector during the life. Also as a consequence, not all the defects that could be repaired by spare rows will guarantee a safe operation during the product life: the defect diagnosis routine must identify and test the behavior of defective rows to decide if repairing will be safe or the part need to be rejected.

8.4.2.2 Testability Tools for Redundancy. In order to guarantee test coverage and reliability of the redundancy circuitry (containing Non Volatile Memory elements), to maximize the impact on yield at a low cost in terms of test time, some test tools are typically provided by the design; the most common may allow to:

- functionally test the spare resources before activation, called shadow-made;

APPLE INC.
EXHIBIT 1008 - PAGE 0472

- fast programming of CAMs (erasing of CAMs);
- marginal testing of CAMs (erased or programmed);
- fast detection of repaired addresses;
- disable redundancy;
- DMA of CAM cells.

These last two are better suited for engineering analysis than for production.

In addition to the above, all the test modes available to better analyze the matrix can also be applicable to the cells inside the spare resources.

8.5 PRODUCTION TESTING

After describing how Flash are designed with a lot of attention to the testability of the product and technology, the production testing can now be treated more in detail.

Production testing at the manufacturer's side is a sequence of tests that is sufficient to screen and guarantee a "known" product, in other words a product which has been produced inside process control and parametric test limits, whose process and design have been developed with proper know-how and extensively characterized for performance and reliability.

The focus at Wafer Probe testing is more on:

- screen for DC and gross functional failures;
- diagnosis of faults for feedback to process;
- diagnosis of faults and repairing;
- parametric data collection for process control;
- screen for reliability;
- fast removal of unreparable failures.

Fig. 8.17 reports a schematic example of Wafer Probe flow: the redundancy detection, diagnosis and repair is distributed among the various steps; all steps are oriented to guarantee margins for quality and reliability, using the testability tools described in the previous sections, rather than to simply guarantee specifications.

At Final test the focus is on:

APPLE INC.
EXHIBIT 1008 - PAGE 0473

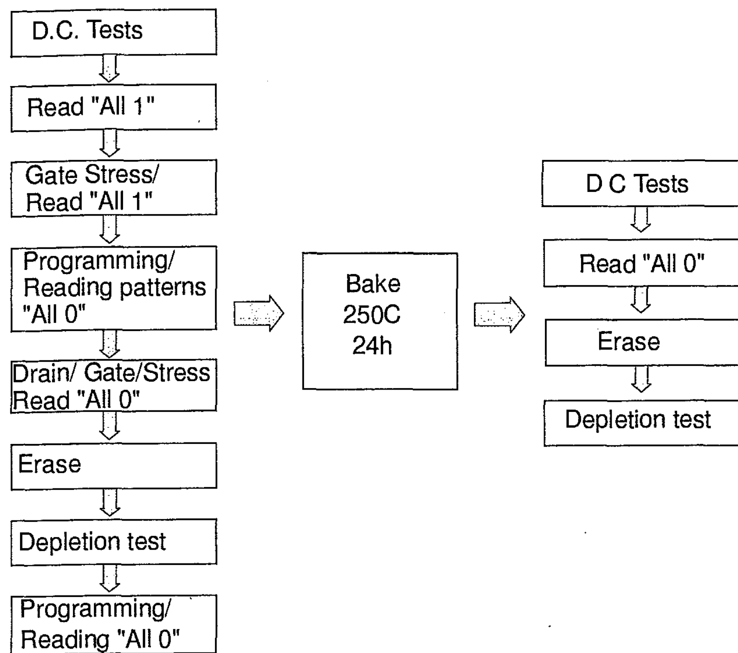


Figure 8.17 Schematic example of Wafer Probe testing.

- screen for DC and gross failures (after assembly; at different temperatures);
- guarantee (and classify) AC performance across temperature range;
- fully check adherence to specification;
- screen for reliability.

8.5.1 DC Tests

Icc currents (and Ipp for dual supply products) in the different specified conditions and output DC levels are tested at wafer probe and final test.

In particular, the availability (in the specification or as a test mode) of a Power Down test, where the current expected is below a microampere, not only allows the rapid screening of most of the gross failures at wafer probe, but also constitutes a kind of partial static IDDQ test for significant parts of the circuitry with benefits for reliability.

APPLE INC.
EXHIBIT 1008 - PAGE 0474

8.5.2 *Functional Testing*

Flash memories have significantly different characteristics if compared to DRAMs (the testing of which has been extensively studied and a vast literature exists). A few of the main ones are:

- erase and program are 3 to 4 orders of magnitude slower than read;
- erase is only by sector; after UV erase the (non defective) array starts from *all* 1;
- the cell provides to the sense amplifier a static current (tens of microamperes) rather then a small (tens of millivolts) voltage difference.

As a consequence, the concept of "order of pattern" does not apply immediately; instead, erase and write operations are to be minimized.

Among the different types of faults related to the array cells, the ones of "resistive" nature are by far the most likely (stuck-at faults, bridging faults ...): these are relatively easy to detect with simple patterns; more subtle types of failures in the array cells are the ones related to oxides defects that are discussed deeply in Chapter 7.

In the periphery circuitry some possibilities exist of faults of "capacitive" nature, particularly where the signals are not rail-to-rail (or digital), like in the sense-amplifiers and their inputs (bit-lines) and output signals (data-lines), in the timing chain or in the Input or I/O receivers when operated at TTL levels. The characterization must be designed (based also on the knowledge of the product architecture) to exclude intrinsic capacitive-coupling faults and the production testing must be able to screen the possible defect-related ones.

A non-exhaustive list of (gross) functional failures may be:

- fault in addressing, decoding, data path, I/O;
- fault in erase/programming circuitry, command register or algorithmic state machine;
- single or cluster of bits that do not read/erase/program;
- couple of bits stuck (sharing the same "open" contact, or with floating gates shorted together);
- bit-line open, short (or leaky) to ground, source line, word line or other bit line;
- same for a sub-bitline, including short to main bit line;

APPLE INC.
EXHIBIT 1008 - PAGE 0475

- word line open, short (or leaky) to ground, source line, other word line, bit line or sub-bitline.

The list is much longer, but most of the failures can be detected with simple patterns which have the advantage of requiring a few erase/programming steps and allow easy diagnosis of the failure type for feedback to the process and for repairing.

Some of the most common and effective may be:

- ALL ONES: detects some of the gross failures also if the DUT is only UV erased; detects erase failures if performed after electrical erase;

- DIAGONAL OF ZEROES or FRAME OF ZEROES, programmed over ALL ONES: programming a small percentage of bytes, allow to detect most of the failures related to bit line, word line or periphery circuitry;

- CHECKERBOARD (TOPOLOGICAL): CK-ODD, or (for a X8 data width) writing 00 in all the ODD addresses, or its complement CK-EVEN. Executing both patterns, using a MARCH sequence (for each location: read FF, write 00, read 00), is very effective in detecting stuck-at and bridging faults in the array, decoding, addressing, data path and programming circuitry.

In synthesis the gross functionality of a Flash could be tested with only one complete erase/program/erase cycle using ad-hoc patterns and intermediate read verifications.

Of course much more complex screenings are required to guarantee full-spec performance and reliability (see following sections).

The functionality of the algorithmic state machine is somewhat more complex to test exhaustively without excessive test time cost (see Section 8.3.3): if ad-hoc test modes are available, it can be tested independently as a logic block.

8.5.3 AC Read/Command Interface

The characterization must provide the worst case conditions (patterns, addressing sequences, supply voltages, temperature, erase and programming conditions of the patterns themselves) that are necessary and sufficient to test and classify in speed classes the DUTs in production. Complementary patterns may be used to test all bits in both states; all different specified types of read modes are tested (address access time, access time from chip enable, from output enable and combinations of them). Different Vcc are used and check at two temperatures is also usual.

APPLE INC.
EXHIBIT 1008 - PAGE 0476

Also Flash products often feature very-low-current power-down modes that require a longer "recovery time" to read the first data, instead of the usual read access time from chip enable; those need also to be characterized and tested.

Except for the read operation, for which the Flash is operated like a classical ROM or EPROM, all other operations are initiated and monitored through a Command Interface (see Chapter 5), through which commands and data are passed to and from the Flash through the I/O pads. The AC parameters specified for the Command Interface are also tested by simply using worst-case timing conditions during erase/programming at final test.

8.5.4 Erase/Program Performance; Endurance

Erase and programming times are tested both at wafer probe and at final test, taking care also of the margins necessary to guarantee that the DUT will remain within specification after the number of erase/programming cycles guaranteed for the product (usually 100K), knowing from the characterization/qualification of the product/technology the expected degradation of the erase and programming times.

The main emphasis, however, particularly at wafer probe, is on the screenings applied to the cells which are at the top and bottom of the distribution after erase or show anomalous programming or erase behaviors; these, in addition to the application of severe stress modes (see Section 8.3.1.3), guarantee the endurance with very low defectivity levels.

8.5.5 Reliability

As shown in Fig. 8.3, the Wafer Probe testing generally includes a high temperature bake between a first and second test steps: the first pass ends with the programming of an ALL 0 (or almost ALL 0) pattern. After the bake, the same pattern is tested. For a more efficient screening, the Vt margin of the worst programmed cells may be recorded before bake and the screening after bake may be based on a maximum allowed threshold shift, rather than on a go-no-go read test.

At final test a burn-in test may be used mainly to screen oxide defects which can only be activated by a combination of temperature and voltage stresses.

8.6 TEST PRODUCTIVITY

Some considerations on Flash test times are reported in Section 8.3.2: given the 1 second intrinsic erase and program time for a sector and in spite of the possible use of internal parallelism, the test time for a single unit remains in the order of tens of seconds. For this reason Flash manufacturers are using

APPLE INC.
EXHIBIT 1008 - PAGE 0477

parallel test systems for production testing, both at wafer and at packaged levels. The productivity of a parallel test system for Flash depends heavily on the tester structure and on a number of aspects that will be discussed in the next paragraphs.

8.6.1 Impact on Tester Structure

In very simple (hardware) terms, a tester is functionally made of few basic building blocks:

- System supervisor/controller;
- Power supplies;
- Parametric forcing/measurement units (PMU);
- Pattern generators;
- Pin electronics;
- Pattern memory (+ vector memory);
- Error memory;
- Redundancy processor.

A parallel tester can be made according to the scheme of Fig. 8.18, where all the resources are shared by the N devices (DUT) tested in parallel, except for the Pin Electronics that need to be replicated locally to force and/or sense the signals on the pins of each DUT. The drivers of the tester must be buffered from the main driver and from site to site in order to avoid a short on one DUT to cause failures on the other devices.

At the other extreme, a fully parallel test configuration replicates all the building blocks for all the N DUTs (Fig. 8.19). The first configuration would be typically of lower cost, while the second would allow a better productivity.

Many intermediate possibilities exist between the two extremes. For Flash testing, in particular at wafer level, the tendency is toward testers which have (almost) all the resources duplicated per DUT (tester-per-site).

The improvement in productivity that can be obtained using a parallel system versus single testing (given a certain test flow) is largely dependent on the tester configuration but also depends on many other aspects related to the product.

To simplify the matter we can consider separately the tester impact, making the hypothesis that there is no impact from product aspects (yield 100%,

APPLE INC.
EXHIBIT 1008 - PAGE 0478

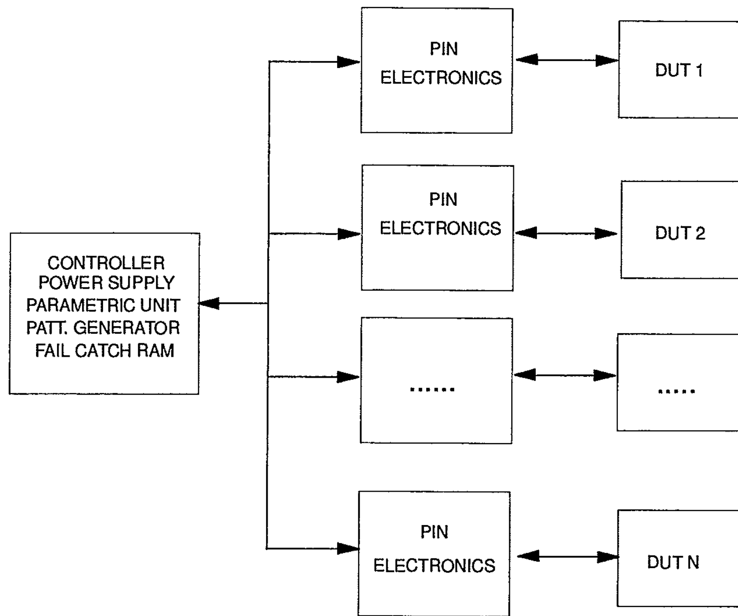


Figure 8.18 Simplified scheme of shared-resources tester.

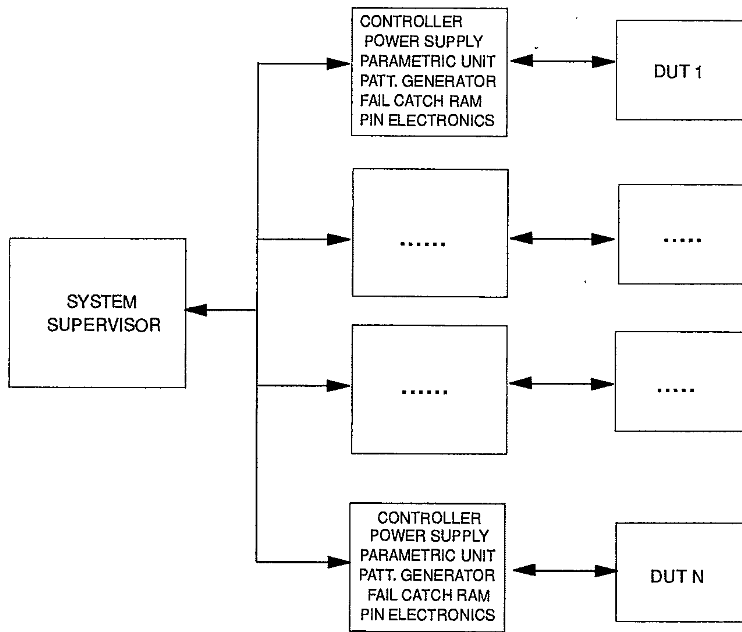


Figure 8.19 Simplified scheme of fully-parallel tester.

APPLE INC.
EXHIBIT 1008 - PAGE 0479

no dispersion on performance); we can then estimate the theoretical Parallel Productivity Factor (PPF) that a tester can obtain (with a given test program):

$$\mathrm{PPF(tester)} = \frac{N}{P + N(1 - P)}$$

where P is the fraction of the testing program that can run fully in parallel for the N DUTs, while $(1 - P)$ accounts for the routines that need to be run sequentially one DUT after the other.

As an example, DC measurements or setting of reference cells can be made in parallel if a PMU per site is available (and the relevant test time will be in the P fraction); if the tester features only one PMU the above tests will be repeated sequentially for the N DUTs.

Testers specifically designed for Flash (like Fig. 8.19) obtain a Parallel Productivity Factor very close to N which is the theoretical limit; in other words the tester can test N DUTs in the same time as a single one. Unfortunately product related factors introduce a significant reduction.

8.6.2 Parallel Testing Final Test

To test packaged parts, one or more handlers are connected to the parallel tester; very high level of parallelism (e.g. 32 or even 64) can be achieved conveniently at the state of the art.

The factors that reduce the productivity with respect to the one allowed theoretically by the ideal tester are few:

- yield and failure mode;
- Erase/Programming time differences among different DUTs;
- indexing time (or time to load/unload the batch of N DUTs).

They cannot be easily reduced to an analytical model but the results are not so far from the theoretical ones; see Fig. 8.20, where the average test time per DUT is reported for 3 different test systems: it can be seen that testers specifically designed for large parallelism (like B and C cases) show only a small deviation from the maximum achievable $\mathrm{PPF} = N$.

8.6.3 Parallel Testing at EWS

Compared to the Final Test case, parallel testing at EWS is more challenging and can be less productive for the following reasons:

- the probe card technology (considering that Flash have typically 40–60 pins);

APPLE INC.
EXHIBIT 1008 - PAGE 0480

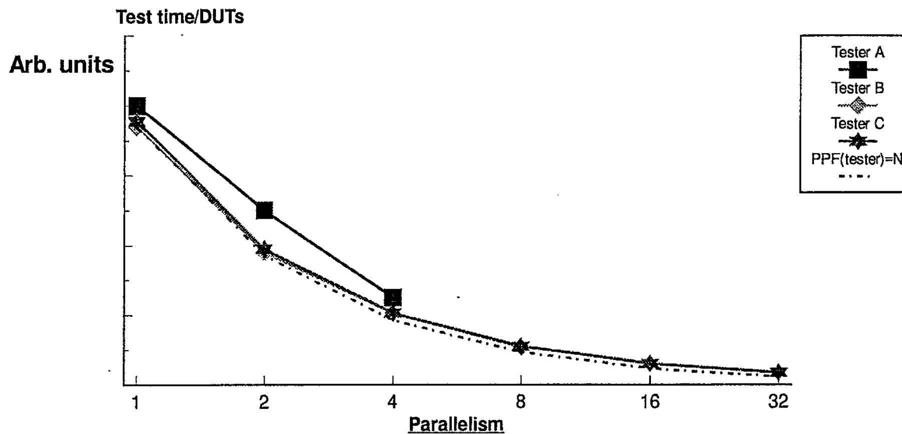


Figure 8.20 Curves of test time/DUT obtained with different parallel testers.

- yields at the early stages of a new technology may be low;
- inefficiency due to the edge of the wafers;
- testing based on distributions of erase time and program time, in which the tester checks the state after each pulse of each address: this is generally slower and less productive on parallel testers with respect to Final Test where the internal (user mode) Erase/Programming algorithms are typically used.

Due to the above reasons, the level of parallelism used in Flash at EWS is now mostly at the 4 or 8 level but probe technology for 16 or 32 is also being introduced.

At EWS the role played by the structure of the tester is even more important due to the extensive use of test-modes and of redundancy.

8.7 PRODUCT CHARACTERIZATION

The objectives of product characterization (from the manufacturer side) are several:

- assess (on few lots) that the product meets fully the specification with enough margins;
- assess, through test modes, that the performance of the array is correct with margin;

APPLE INC.
EXHIBIT 1008 - PAGE 0481

- check that the behavior of the parameters (not only the ones specified to the user, but also some critical nodes accessible for testing) is in line with design, process, physics' expectations;
- assess that the product can tolerate the expected process variations;
- extract information (worst-case patterns, worst-case corners, limits and guard-bands ...) for the production testing.

Of course, in case of problems detected, the characterization will serve as a feedback to design and process engineers to investigate and correct the issue.

All parameters are characterized versus at least Vcc and temperature; for erase and programming also Vpp needs to be considered, for a dual supply product; access time performance needs to be assessed through the use of several patterns, addressing sequences, read modes and parameters specified.

In SRAMs and DRAMs the write path is just as critical as the read path; this is because the write sequence is controlled by a fast and asynchronous timing chain and, as a result, it is more susceptible to variation from part to part, and then as critical as read in terms of speed screening and classification.

On the contrary, in Flash the write commands are latched and the internal operations are clocked.

The characterization of AC write cycle is basically the characterization of the Command Interface.

The complete program or erase operations are 3 to 4 orders of magnitude longer then the read one and, as a result, sufficient design margins are allowed to avoid speed issues in production.

The read cycle is however more critical as far as speed. In the characterization the Flash must be investigated as a ROM, taking into account data patterns in the array, architecture and design choices made. Typical data patterns used are electrical and topological checkerboards, diagonal patterns, random or with different percentages of programmed bits. The data patterns must also be used to identify I/O interactions resulting from cross talk or parasitic coupling among sense amplifiers or in the read path signals.

Addressing sequences must be used in combinations with the data patterns to highlight different types of problems. For example, address complement scan combined with a checkerboard pattern is classically a worst-case for input levels. Address sequence used for interaction between row access and column access is the Butterfly-like (transition from each individual cell to all other cells in the same row and column). ATD (Address Transition Detection, see design section) pulse generation may be checked, as a worst-case, reading a pattern following a Gray code address sequence. Also read access time after read-abort cycles need to be characterized: in other words, addresses or chip-enable changing at a cycle time lower that the guaranteed one, followed by a

APPLE INC.
EXHIBIT 1008 - PAGE 0482

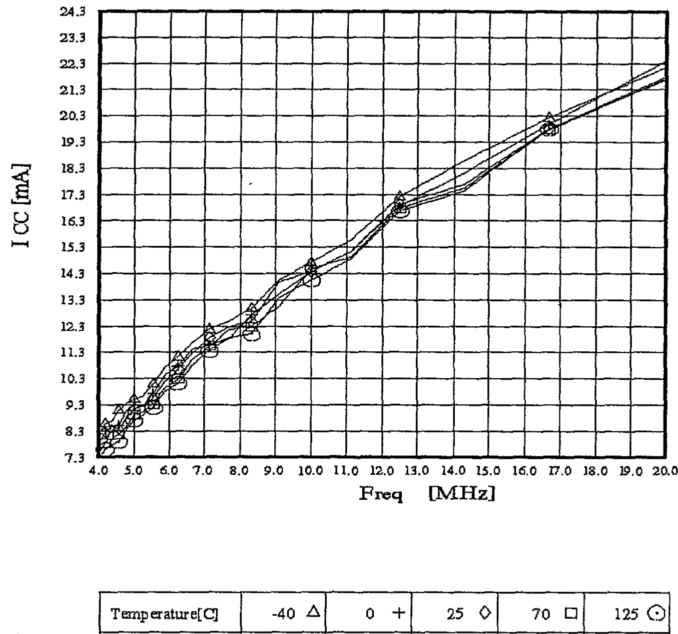


Figure 8.21 Icc active versus operating frequency at different temperatures.

normal read cycle. Other patterns reported in text books exist [1] and may be chosen based on the design.

To mimic applications, an address format with return to 1 and return to 0 may be used to reflect the pull-up or pull-down resistor used on the bus.

In Figs. 8.21 and 8.22 examples of data sheet's parameter characterization are illustrated: Icc versus operating frequency in read mode at 5.5V Vcc and Erase time versus Vcc.

A full characterization of a product is comprehensive of a detailed check of the evolution of the embedded algorithm handled by the Program/Erase Controller.

As an example it is reported in Fig. 8.23 the output of a testing program which performs a sector erase starting from an array whose content is FFFF: within the algorithm the array is pre-programmed to 0000 and then erased back to FFFF.

The testing program tracks the evolution of this operation performed (on a 4M Single 5V supply) by measuring the current drawn from the power supply (right y axis) and the status of the array content on a bit-per-bit basis (left y axis). The tester samples the status of bits every 50ms (the sampling steps on the x axis).

APPLE INC.
EXHIBIT 1008 - PAGE 0483

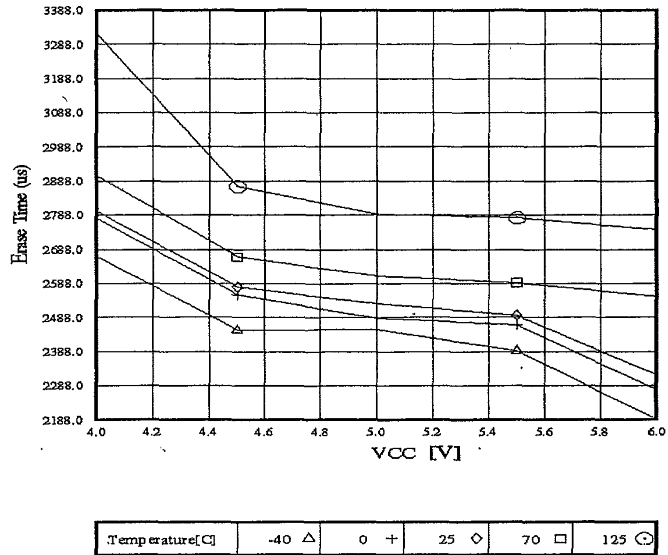


Figure 8.22 Erase time versus Vcc on a single supply product.

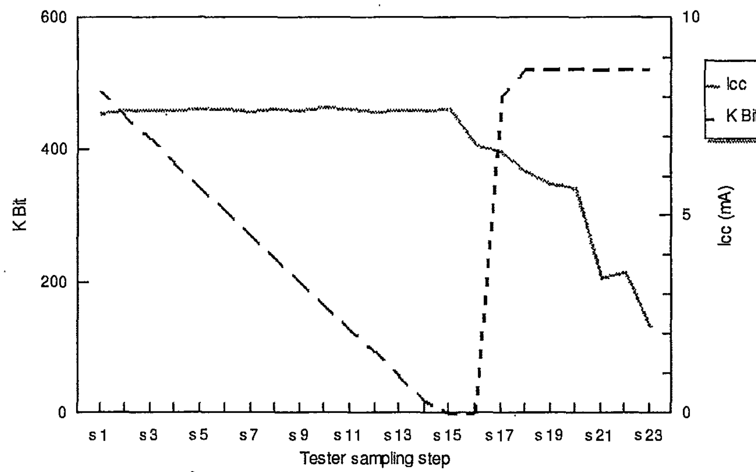


Figure 8.23 Detailed characterization of an erase cycle performed by the internal Program/Erase Controller (4M Single Supply). On the left y axis is the number of bits which are found to be erased at a certain tester sampling step (x axis): on the right y axis is the Icc current.

It can be appreciated the linearity in time of the Byte program operation with an associated constant current consumption.

During the erase operation the array is rapidly brought (couple of normalized steps) to a condition where the content is read as FFFF, but the internal algorithm requires more time (six normalized steps) to guarantee the correct margining of the erase operation.

A program like this gives at a glance a clear indication of the correct partitioning of the timings settled by the internal state machine and also gives an indication of the real power consumption during the various phases (program pulse, erase pulse, over-erase correction, ...).

Characterization of a lot may be made exhaustively on a small sample (maybe 10–20pcs) and integrated by a more statistically valid sample (50–500) on a reduced version of the characterization package; parameter data collected during production testing of the preproduction lots are also compared with characterization data and used for correlation and performance assessment.

For a better estimate of the variations in yield and performance that the process will cause due to its variations inside the process window, corner lots may be launched where the most important lithography and process variations are tried on purpose on few wafers (e.g. active area variation, polysilicon size variation, thresholds variations, etc.).

8.8 CONCLUSIONS

After several generations and a more than a decade's learning curve, Flash has become a very popular and easy to use memory technology, featuring excellent performances, quality and reliability.

This chapter shows how this is guaranteed on each device shipped through an effort that goes from memory cell, process robustness conception, design for testability and reliability, state-of-the-art manufacturing and process control, exhaustive characterization and reliability assessment and, finally, severe production test screening.

The chapter describes the testing aspects from the manufacturer's point of view with the intent of helping the final user in achieving a better knowledge of the product.

The message here communicated to the reader-user is that a trustful manufacturer tests each device (in broad sense: from design to final testing) with a sequence that is incomparably more severe and thorough than any acceptance test or characterization than a user can devise: to take the best profit of that I suggest that the user correlates his test results with the manufacturer's ones (not only in the go-no-go approach but in full detail) for mutual benefits.

APPLE INC.
EXHIBIT 1008 - PAGE 0485

While the Flash manufacturer has succeeded in making, through smart design, the Flash technology easy to use, the applications have become significantly more demanding and diversified and, as a consequence, new performances have been added to the Flash, again increasing the potential complexity for the user.

Here again, a strict relation between component and system manufacturer, from system conception to volume production, becomes vital to optimize cost, performance and time to market for both players.

Acknowledgments

I am indebted to many colleagues for contributing the text and figures and for feedback to the original draft: V. Malhi, P. Cappelletti, M. Dallabora, G. Bacis, B. Beverina, S. Ghezzi, C. Golla, A. Ferris.

Thanks to R. Capizzi and C. Pennati for drawing the figures and correcting the manuscript and for their patience.

Finally, I am grateful to Tiziana, Lorenzo and Marco for loving me even when I spend my time at home working instead of enjoying the life and playing with them.

References

[1] Van de Goor A.J. (1991) *Testing semiconductor memories.* John Wiley.

APPLE INC.
EXHIBIT 1008 - PAGE 0486

9 FLASH MEMORIES: MARKET, MARKETING AND ECONOMIC CHALLENGES

Bruno Beverina[1] and Philippe Bergé[1]
with contributions by
C. Kunkel[1], G. Moy[2], A. Damiano[3], R. Ferrara[3], A. Re[3]

[1]STMicroelectronics,
[2]IBM, [3] Magneti Marelli

Abstract: We describe in this chapter the dynamics of the Flash memory market, its segmentation and its expected evolution.

The typical characteristic of Flash memory technology, its flexibility, is seen as the main factor which explains the strong evolution of the demand, generating continuous new field application with the typical pervasiveness of the innovative semiconductors.

But the flexibility also determines the peculiar position of this product in the market. The Flash memory does not correspond to our definition of "commodity", but follows the same price pressure as the big commodity market, represented by DRAM, and dictated by the technology learning curve and by the classical macroeconomic supply/offer models. Flash memories are not a

APPLE INC.
EXHIBIT 1008 - PAGE 0487

dedicated product, but can, accordingly to the environment, appear sometimes as a standard part, or as an application specific circuit. For all applications, however, Flash always plays a strategic role.

A survey of the many applications presented, with an inside focus from the industry and the point of view of the system designer, is somewhat privileged to give a idea of the many great innovations Flash is making possible.

9.1 INTRODUCTION

Similarly to what DNA represents for living beings, Flash memory has the fundamental property of providing life to the systems in which it is used. It allows them to remember what they are and what they should do. Moreover, it gives them life because the system's code or behavior can be changed by the system itself, thus allowing life propagation and reproduction as in living organisms. The system can keep track of its own experience, make comparisons leading to decisions and reuse learnt information. Maybe someday it will even develop original concepts. The fundamental reason why Flash memory is behind these new possibilities is because data is not lost when the system is temporarily switched off. Flash memory is the first memory that is non-volatile *and* re-writable, offering at the same time a high level of performance and, most important, at low cost.

Since the first Flash memories were developed at the end of the '80s, initially emulating EPROMs with re-write capability, the inherent flexibility and power of the Flash memory have come a long way. Power supplies have been reduced from dual 5/12V to single 5V, and then reduced again to single 3V, with the search continuing to obtain 1V memories. The endurance cycling capabilities, the number of erase/write cycles each cell of the memory can withstand, has increased from the original 100 or so, to over 100,000 with a next target of over 1 million. Access times have been pushed to the limit and now are equal to those of DRAMs. Moreover, sophisticated chip architectures have been developed to deal with important issues such as reliability and test coverage, time and cost, and to improve the application flexibility in operation. Single transistor cell sizes have been shrinking, year by year, with cells of less than $1\mu m^2$ targeted for the 256Mb to 1Gb era.

The applications of Flash memories have also moved away from simple replacement of EPROMs, and now have invaded the fields covered by EEPROM and SRAM; in the future, a part of the DRAM market could be challenged. Designers also dream of replacing magnetic media with Flash memories, a transition which is at the moment possible only for very specific applications. Flash memory represents the state of the art of the semiconductor industry in terms of technology, process and design and is, at the same time, a challenge and opportunity for system designers. In fact, Flash is positioned in an area that

APPLE INC.
EXHIBIT 1008 - PAGE 0488

crosses over the application areas of different types of memories, thus making it an interesting tool for new approaches in marketing. Out of the plethora of many early technical solutions to Flash memory seen so far, the NOR architecture is the winning solution today. But with the ever widening range of applications, other architectures will achieve a significant share of the market by the end of the century.

Flash memory, then, is an enabling technology inasmuch its limits are far beyond the boundaries of the markets we can envisage today.

If you look at the low magnification optical micrograph of a relatively large sized memory dies produced by the semiconductor industry today (DRAM, EPROM or ROM, 16Mb density for example) you can immediately recognize that at least 70% of the surface area is taken up by the memory matrix itself. The command/control and I/O occupy only 30% of the area and originate, as a consequence, 30% of the cost and 30% of the manufacturing difficulties. But it is not like that for Flash memory. Even a relatively large Flash memory has only 40% of its area devoted to the memory matrix, while the rest has as much as the same complexity and nature as microcontroller and analog circuits. It is no surprise therefore that, to improve system performances, density, costs and reliability, there will be a further push for a widespread integration of Flash memory, thus enhancing the responsiveness and interactivity of any electronic appliance.

The evaluation of the Flash memory market is dependent on the way statistics are built. Circuits devoted to specific applications often are not counted in the Flash memory share, simply because of their function, even if they allocate 80% of the die area to Flash memory and 20% to DSP. Later in this chapter we will discuss the evolution of Flash applications and of the Flash market and we will see that in the near future the size and impact of Flash memory could be enormously bigger of what we can see today.

9.2 MARKET SEGMENTATIONS

In less than a decade, Flash memory has become the first non volatile memory market, ahead of masked ROM and EPROM. In addition Flash is now ranked third in the total memory market, after DRAM and SRAM (see Figs. 9.1 and 9.2).

One of the main reason of this success is the versatility of the Flash memory which brings benefits in all traditional market segments, Telecom, Computer, Consumer, Automotive and Industrial. This versatility leads to a combination of many different applications and many different functions from which different possible segmentations can result.

APPLE INC.
EXHIBIT 1008 - PAGE 0489

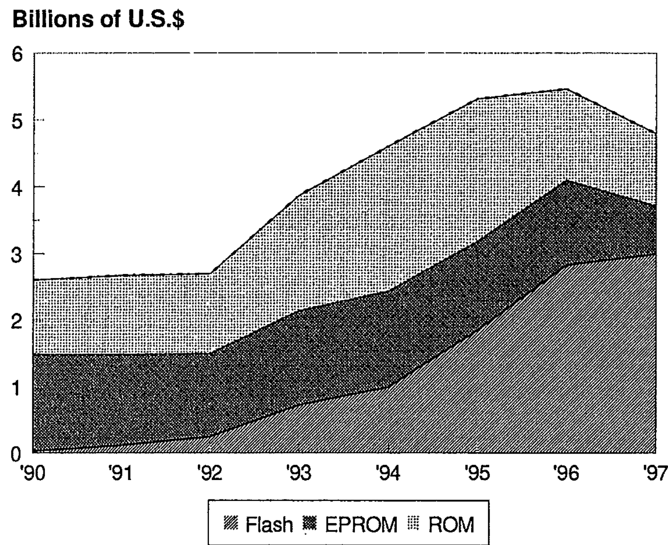


Figure 9.1 Flash market history.

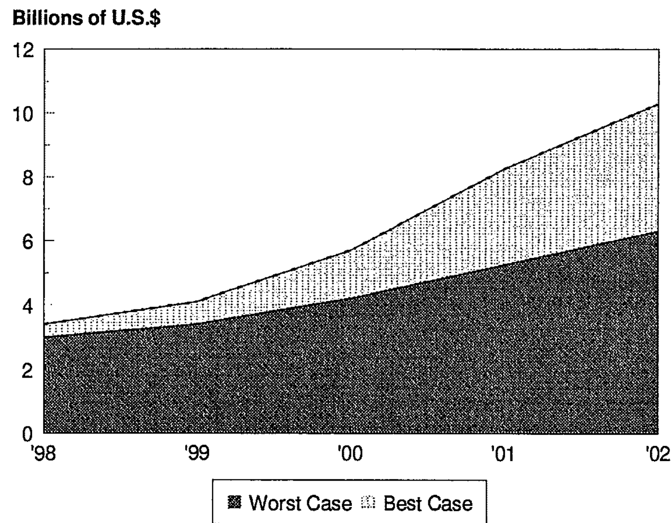
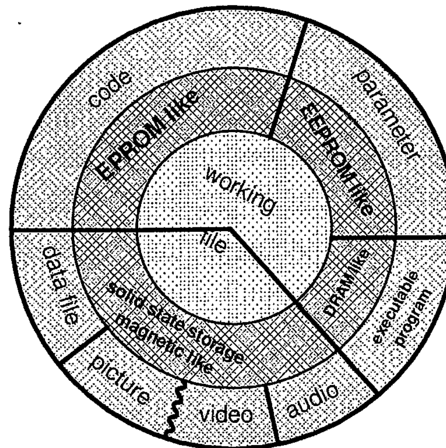


Figure 9.2 Flash market projections.

APPLE INC.
EXHIBIT 1008 - PAGE 0490



Note: Dimensions do not reflect relative market sizes

Figure 9.3 Flash segmentation.

9.2.1 Application Segments and Subsegments

In the following, we will attribute Flash memory to different application segments, depicted in Fig. 9.3. The first question to be answered is:

Is your memory "working" or not? A memory which contains the information which is mandatory for the execution of the application in which it is embedded is called a "Working" memory: it can be either volatile or non-volatile. A "Working" Memory can be used to store the code of a micro-controller, configuration parameters (such as set-up parameters) of the application, as well as higher level executable programs of a computer (word processor documents, worksheets ...) and also temporary computing information as in RAM.

On the contrary, a "File or Mass Storage" memory has the mission to store information which is not necessarily executed in an embedded application itself: this is typically the storage of voice (Digital Answering Machines) or pictures (Digital Still Cameras) or any data files.

The further question is:

What is your memory "like"? The versatility of Flash allows it to emulate many previously available technologies. Flash are "EPROM like" in the sense that they can replace any EPROM in any application, where they are used essentially for code storage.

APPLE INC.
EXHIBIT 1008 - PAGE 0491

When first introduced, Flash were used only as EPROMs: users had to pay extra money over the EPROM cost essentially for the security or comfort offered by the on line reprogrammability, which was impossible with OTP EPROM or tough and expensive to execute with UV-EPROM.

Today, after some years of manufacturers and users experience, the cost premium of Flash is collapsing and Flash are used more and more as real Flash with their in-system reprogrammability feature: the BIOS of your PC was initially stored in an EPROM, then some manufacturers moved it to Flash, which eased the new product launch by allowing quick reprogramming of the Flash on the boards immediately before shipping, in case an early bug was found. Now, all PC manufacturers offer the possibility to upgrade their PC by reprogramming the Flash BIOS after purchase. "EPROM like" is the pioneer subsegment and still the major one. However, its share is expected to reduce in the future as Flash, with its advanced features, will play the role of an enabling technology for many new applications, allowing them to emerge.

To some extent Flash are also "EEPROM like" and then can, in some cases, replace EEPROM for parameter storage. This is why some blocks of uneven sectored Flash are called Parameter Blocks. Flash technology allows electrical erasing like EEPROM, however EEPROMs retain the advantage of their granularity. Flash can be erased only by sectors, from Megabits down to a few bytes, while EEPROMs can be erased byte-by-byte. So the Flash cannot directly replace EEPROM when the granularity required by the applications is high. In some space constrained applications requiring both Flash and EEPROM functionality, Flash are used to "emulate" the EEPROM granularity. The software stores parameters as linked lists in a sector of the Flash, keeping a second sector ready to take over when the first is full, then erasing the first one, and so on. EEPROM emulation in the Flash requires to find an efficient way to allow the reading of the software code while updating the parameters, which corresponds to the read-while-write operation with a Flash. This can also be handled by software, at the cost of a few kilobytes of code stored in SRAM in the host computer, provided that the Flash has the capability to suspend/resume the program and erase function at any moment with a latency time which will make it transparent to the host.

The linked list principle can also be used with a so called dual bank Flash, which features two physically independent memory banks, offering by hardware simultaneous read/program of different sectors.

However, the weakness of the above mentioned linked list software emulation is that the access time worsens dramatically after each parameter update. The optimal solution from performance stand point is probably the Flash plus (tm) that actually merges on the same chip Flash and EEPROM technologies, for a true integration of the EEPROM in the Flash.

APPLE INC.
EXHIBIT 1008 - PAGE 0492

In the early '90s Flash was said to be potentially "DRAM like". Many memory manufacturers started to dream (daydream or nightmare?) that Flash memories would in the next 5 years replace all the DRAMs in the market place. In reality Flash will never replace most of the DRAMs when the application requires fast programming and erasing time, as in Video Frame memory or PC's main computing memory. However, instead of downloading your wordprocessor from your HDD to your DRAM for execution, you might execute it directly from a non-volatile Flash. Flash would then replace a part of the DRAMs in your systems.

The major excitement in the Flash market came when people started to dream that Flash could be "Magnetic like" and would replace all HDD, opening a new Golden Age for semiconductor memories. Very quickly it was understood that this will not happen in a foreseeable future.

The Mbyte growth requested by the standard HDD applications (Desk Top and Note Book PC) was so high (near to 50 times/decade) and the recording density per unit area of HDD soaring at such a pace (100 times in a decade since the early '90s) that Flash could not close the $/Mbyte cost gap with the HDD. Nevertheless Flash will grow as a "Solid State Storage" device in new consumer applications which will take full benefit of its advantage in term of ruggedness, weight and power consumption, such as PDAs, Digital Still Cameras and Voice recorders.

In these cases the success of the Flash is directly linked to the commercial success of the applications and, vice-versa, the market boom for these applications will be linked to the Flash capability to meet the consumer market requirements in term of $/Mbyte, and user friendliness.

Many efforts have being deployed by the Flash industry in this direction, first with the credit card size PC cards and more recently with the "stamp" size Flash memory cards, for which three potential standards are fighting on the market place:

1. "Miniature Cards" allow the maximum memory size as they can contain few Flash devices with no extra silicon, while the standard S/W (Flash File System) is stored in the end application;

2. "Compact Cards" provide a full emulation of the HDD standard ATA interface at the cost of embedding an ASIC on each card. ATA is the acronym for Advanced Technology bus Attachment, which is a high level interface for Hard Disk Drivers standardized since 1989;

3. "SSFDC" (Solid State Flash Disk Card) can be seen as a package for a single serial access Flash.

APPLE INC.
EXHIBIT 1008 - PAGE 0493

9.2.2 Technology, Performances and Applications

The proliferating applications that are driving the Flash market growth are so diverse that they require, or at least desire, very different features; or that they prioritize very differently the different features of the Flash memory such as the power supplies, access time, package ... This results in a proliferation of Flash solutions which originates from the technology itself and in different cell organizations such as NOR (including its DINOR derivative), NAND or AND, along with Multibit/cell, leading to a variety of features which best fits the requirements of the different market application. This variety of technological developments and solutions does not help the electronic industry to simplify the matter and to ease the choice of the best Flash memory, but the associated innovations will boost forward the electronic industry itself. Main features and applications can now be gathered in the three following streams (see Figs. 9.4–9.7):

1. Standard Market, mostly an "EPROM like" market, tending toward very fast access time and covering a wide spectrum of densities;
2. Mobile Applications, requiring lower and lower energy consumption and thus very low supply voltage;
3. Mass Storage, requiring very high densities as well as a very low cost per bit.

	Key Applications	Density	Key Features	Comments
STANDARD	Auto, HDD, PC Bios, Networking, Modems, Set Top Box...	Wide Range From 1 Meg to 16 Meg	5V=>3.0 V Access Time	"EPROM Like" NOR type
MOBILE	Cell Phones Pagers	Medium 4, 8, up to 16 Meg	Low Energy: 3.0V=>1.8V Chip Size Package	Market Driver NOR type
MASS STORAGE	Dig.Cameras Voice Rec. PDA Flash arrays	High 16 Meg +	Very low cost/bit Random Access	Still Uncertain Forecast

Figure 9.4 Flash market streams.

APPLE INC.
EXHIBIT 1008 - PAGE 0494

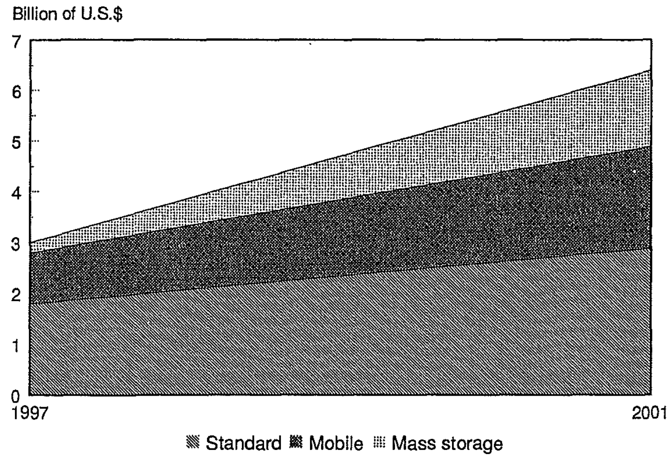


Figure 9.5 Flash market segment value.

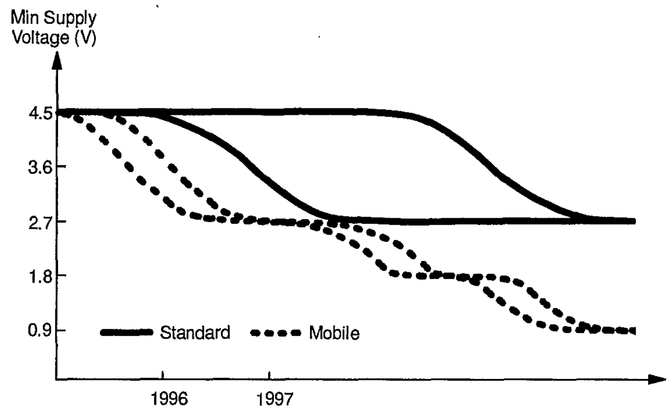


Figure 9.6 Market streams: supply voltage.

Another trend linked to technology and performances is the one toward "integration". Integration is a strong and historical trend in semiconductors which continues to accelerate, boosting system performances and reducing board space. However, embedding Flash in the CPU is not *the* single best solution. Repartitioning of Flash-based systems offers many different integration solutions having their own best factor of merit (see Figs. 9.8 and 9.9).

Packaging a stand alone memory in the tiny Chip Size Package, makes it the most cost effective, most flexible solution and gives the fastest time to volume,

APPLE INC.
EXHIBIT 1008 - PAGE 0495

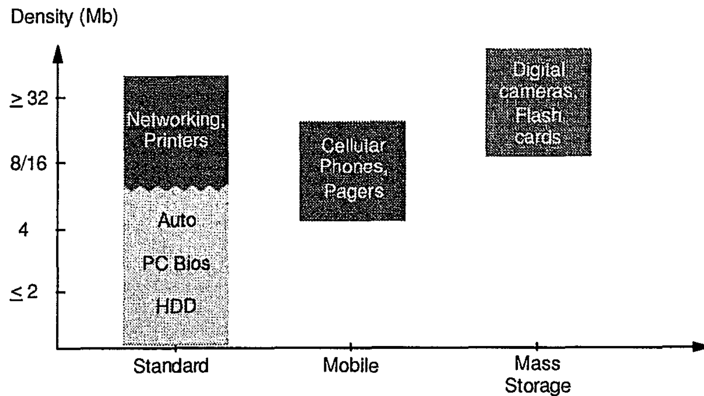


Figure 9.7 Market streams: density.

	Cost	Perf	PCB Area	Max Mem Size	Time To Vol	Mem Size Flexibility
Stand Alone (C.S.P)	**5**	3	3	**5**	**5**	**5**
Embedded (ST9,10,20...)	3-4	**5**	**5**	3	3	2
Mixed Embedded + Ext Flash	4	4	4	4	3	**5**
Embedding Logic	4	4	4	4	3	2
M.C.M (bare dice)	2	4	4	**5**	4	4
Mixed Mem Techno	3-4	3	4	4	3	2

Figure 9.8 Factors of merit for the various Flash memory implementations. A high value corresponds to a better performance.

while embedding the Flash only with glue logic also turns out to be a good compromise.

As mentioned before, Flash memories are used in all market segments (see Fig. 9.10). The business size and highest growth is currently linked to the Telecommunication segment, first because of the digital cellular phone, which has such an importance that one might wonder if GSM stands for "Growing Size Memory" or "Growing Software Monster" ... Moreover the need to exchange information, linking all computers together with standard protocols, has propelled the demand in Networking routers and of course modems, both of which benefited from the Internet growth wave.

APPLE INC.
EXHIBIT 1008 - PAGE 0496

	Cost Increase Factor	Cost Reduction Factor
Wafer Cost	Mask # + 50% vs. stand alone Micro + 20% vs stand alone Flash	
Die Size	+ : no shrink path for Flash	Flash logic partially replaced by CPU
Yield	Increase of die size & # masks, Yield multiplication effect, Flash speed distribution	
Test Cost	Flash side tested with sophisticated test equipt of Micro	
Package		One Package only

Figure 9.9 Cost impact of integration.

2001	Flash TAM (*) %	End. Equip. (Mu)	Density	Key Features	Technology
Cellular + Pagers	40	>200	Medium 8-16 Mb	very low energy	NOR
Digital Set Top Boxes	9	>30	Medium 16-32Mb		NOR
Auto (EMS, GPS)	9	>50	Medium 4-8Mb	Speed	NOR
HDD	5	>250	Low 2Mb	Speed	NOR
Networking (Routers)	9	>10	High 192Mb		NOR
PC Bios	9	>150	Low 4Mb		NOR
Digital Cameras	8	>10	High 64-128Mb	$/bit card	NAND, AND, Multilevel
Flash Arrays / Pda	3	>10	High 8-64 Mb	Random access	NOR

(*) TAM = Total Available Market

Figure 9.10 Major applications.

APPLE INC.
EXHIBIT 1008 - PAGE 0497

In computers the key applications are currently the PC BIOS and the Hard Disk Drive Operating System, although the latter is likely to lose some of its importance in the next five years because of integration. In the future, the computer segment will grow in importance with the take-off of Personal Digital Assistants and their need for embedded Flash arrays and memory cards.

The development of Digital TV, via satellite or cable, is allowing users to see programs all over the world, making the Set Top Box among the top Flash potential users. In consumer, another potential major growth area is the Digital Still Camera, also linked essentially to the Flash memory Card. This market will undoubtedly develop as a computer peripheral, initially in the special interest field, but we cannot assume that it could really become a consumer-like product which will substitute the current films, especially if one consider that a brand new film standard (APS or Advanced Photo System) is just being promoted by the leading film and camera makers.

Automotive has been among the pioneer Flash applications primarily in Engine Management Systems and in Automatic Gear Boxes and is reinforcing its position with the growing demand for GPS or Global Positioning Systems, which from high end cars will expend to lower grades.

One might even believe that high class cars will eventually be the convergence point of all advanced electronic, gathering video, telephone, fax and computer function on top of the standard powertrain management and navigation systems.

There seems to be no emerging application in the industrial segment which has a minor weight in the global Flash usage. Although we should not forget to mention robotics, which is probably currently the top user of PC card Flash memory cards.

9.2.3 Segment Dynamics

For all current manufacturers, the Flash market was initially an evolution of the EPROM market, however it evolved in a very different way due to the dynamics of the driving applications, giving more and more importance to "windowed markets" (see Fig. 9.11). We call a market a "windowed market" when it requires specific features in a given timeframe.

The oldest example is the computer market with respect to DRAM density. Each new generation using a new and higher DRAM density, leaving no market for the old and lower densities.

This is now the case for the cellular phone market with the power supply of Flash Memories. Windowed markets are tough to enter for the followers: if you miss the window, you can throw away all your development efforts!

APPLE INC.
EXHIBIT 1008 - PAGE 0498

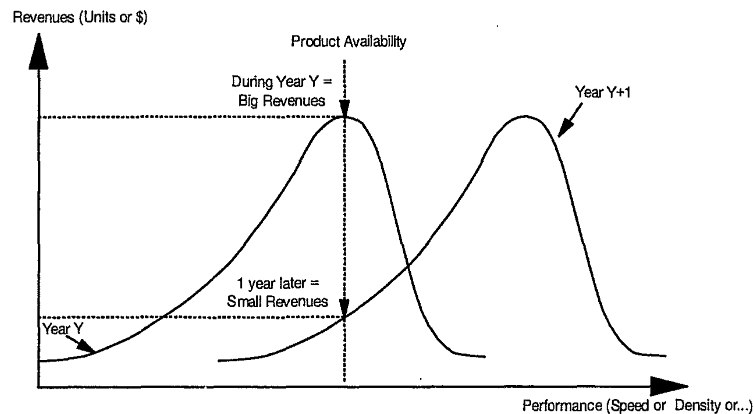


Figure 9.11 Windowed market curves.

Another important factor is the lifetime of the application. For most applications the trend today is to have shorter and shorter lifetimes, down to 6 months design and 6 months for volume production for many consumer-related high volume applications such as the PC.

When this is combined to windowed market applications, it becomes an extremely tough business: not only you cannot miss the design-in window, but also you will have little time to get the pay-back of your investments.

Conversely, when combined to a non-windowed market, a follower who missed one design-in window can expect to be in the next one which will come out shortly ...

9.2.4 Commodity or Non-Commodity?

Commodities are standard, interchangeable products which have very little differentiation (like soap!). They are very easily interchangeable and price and availability tend to become then the major selection criteria. Because of the history of DRAMs and even EPROMs, customers tend to expect Flash memories to be mere commodities. Actually the situation is much more complex, and is driven by the following factors.

a) Flash Memory Product Maturity. Only a few Flash memories can meet the criterion of full interchangeability with many sources, these are the dual Voltage 1 and 2 Meg Bulk Erase. From 4 Meg onwards and for most low density Single Voltage devices, it is very difficult to find more than one fully compatible second source. Both Flash manufacturers and customers are responsible for this situation.

Customers expressed both their very specific technical wishes (dreams?) and actual technical needs, forgetting that alternate sourcing was also a need and that a demand for a too wide variety of possible features would very likely lead to incompatible solutions.

The Flash manufacturers tried their best to differentiate themselves from their competitors to gain niches.

"How to build a position on the market?" "How to provide decisive competitive advantage to the customers and beat competition?" These are typical questions that must be answered by the manufacturers of these pseudo-standard products and which leads to product differentiation.

b) Segment Dynamics. Obviously, the windowed markets requiring advanced features, such as Digital Cellular Phones, cannot or at least should not consider Flash memories as commodities, since only the few market leaders will be ready in the window. Besides, most of these customers usually deal with non-commodity products and are also considering very seriously the integration of the Flash. Solutions embedding Flash in the base band (for instance, one chip including the whole or part of Microcontroller, DSP, SRAM, Flash) can not be considered a commodity anymore. Moreover, since fully compatible Flash are tough to find, some customers do not hesitate to include some hardware and software tricks allowing Flash interchangeability in their application.

c) Application Maturity. Flash memories are using more and more all their in-system reprogramming capabilities, which means that each application contains software linked to the Flash to manage these program/erase features. As a result, in many cases the choice of possible suppliers is finalized at the design-in stage when the software is developed. Any new "compatible" supplier will not necessarily be compatible with the above mentioned software, sometimes just because of the electronic ID of the Flash. So at design-in stage, Flash memories must be considered as non-commodity, which means that the Time to Market or, better, the Time to Volume (because high yield cost effective production capacity has been one of the critical problem of Flash history) is the *key* success factor for the Flash manufacturer. A supplier arriving after the elaboration of the application software will have to struggle to be accepted in the application.

Customers also require a high level technical support which will facilitate the availability of the application software.

Despite all the distinctions reported above, customers tend to associate memories to mere commodities, and expect the Flash to be no exception. The main reasons for this are two. First, the obsession with price reductions. Most of the leading Flash applications are indeed extremely cost sensitive and many

APPLE INC.
EXHIBIT 1008 - PAGE 0500

customers believe that multiple sourcing is necessary to negotiate lower prices, while they forget that their competitors can then do the same and that a proper technical relationship with the suppliers can yield much more effective cost reduction schemes.

Second, the trauma of volume shortage. Despite a very short history of the Flash market, many customers have already experienced two major shortages that left some trauma in their mind and one must admit that the fear is justified since the size of the Flash market is such (around 3B$ forecasted for '97), that the fact of anticipating or postponing a single new fab (1B$ sales) is sufficient to create a major unbalance of the market.

The first conclusion is that Flash memories have to be considered as Application Oriented Standardized products. The next is that another, less obvious, key success factor for the Flash manufacturer is to understand and adapt himself to this contradictory situation of supporting design-in as non-commodities, helping customers to differentiate their Wishes from their real Needs and ensuring that when in production the product would at least behave like a commodity. A commodity with whenever possible an alternate source, enough independent and dedicated production capacity to support the selected customers and proactive and regular lowering of the price along with cost improvements ... in other words this needs building a special relationship between the vendor and his customers ...

Vice-versa, the key success factor for Flash customers is teaming-up with the Flash manufacturers that meet the above criteria.

9.3 CUSTOMER/SUPPLIER RELATIONSHIP

As mentioned before, Flash memories are *not* commodities, but the customers must have the security of supply and the cost reduction path of commodities, while the Flash manufacturers must succeed in getting an acceptable payback from the higher investments linked to Application Oriented Standardized products. What is the secret? Does it lie in Partnership? This word has been overused in recent years and has possibly lost its meaning. Engagement, in the sense of betrothal, is probably the right word as it contains the concepts of mutual choice, trust, but limited commitment as well.

The Flash manufacturer will target or choose its customers in order to reach a proper balance between several criteria.

Every one looks forward to having the fastest volume growth but this volume has to be consistent with the planned capacity increase. Obviously all suppliers strive to secure the filling up of their fabs. However, targeting oversized customers would be at best a waste of time or worse can lead to a disaster if the supplier becomes the limiting factor of the growth of its customer.

APPLE INC.
EXHIBIT 1008 - PAGE 0501

The fastest growing businesses are often the more risky and unstable ones, so a minimum of business stability will have to be sought either by having a high enough mix of steady business or by ensuring that the fast growing markets respond to out-of-phase market cycles.

Innovation seldom comes from nowhere but most of the time from an encounter. Get the right engineers of a leading Flash manufacturer to exchange views with a customer which is the leader in its field ... new ideas will spring out and boost both industries.

In business trust can come from experience (but it is then a very long process), from reputation (which is long as well and not so reliable) or (which is the best and faster way) from transparency. This transparency concerns for instance the development roadmaps, the application impact of the process changes. In the last decade some customers were expecting this transparency to apply to the cost structure of their supplier in order to contribute to the cost reduction by the effect of margin control. The truth is that cost reduction cannot come in a durable way only from price pressure but rather from an intense collaboration allowing to eliminate the useless costs. In other words, the duty of the supplier is to help its customer to avoid paying for what it does *not* need or to make the difference between what he wishes and what he really needs.

This "engagement" stage cannot be maintained without a wide surface of contact between the two companies: as well as the sales force to purchasing department relationship it has to include top management and a wide spectrum of the product operating organization (design, planning, marketing ...).

9.4 THE DEVELOPMENT OF THE FLASH MARKET

Flash memory market has exploded mostly through the "EPROM like" applications where Flash rational advantages could barely justify the negative impact of high cost and tight supply. Flash memories turned out to be just like your car's power window, you could really perfectly well do without it but once you tried ...

How can we explain this unexpected growth from different points of view, including that of the theories of consumer marketing? These theories are more sophisticated and subtle than those normally used in industrial marketing, particularly in the field of consumer behavior and can apply easily to industrial marketing, if one assumes that all decision makers are mere consumers and that the companies themselves have a soul like any consumer.

The Flash market was born from a shift from EPROM, but it did not spread out homogeneously in all segments as they were not equally impacted by the different accelerating and limiting factors. If we look at it from the purely

APPLE INC.
EXHIBIT 1008 - PAGE 0502

rational side the limiting factors have been the price premium over EPROM and the tight supply. Indeed, the price of the Flash long exceeded twice that of EPROM and it needed the customers to be capable to envision economies at least at the level of the total cost of ownership or even at the level of global manufacturing cost. For example, on board programming allows one to produce the board and to stuff and test it by using the test program contained in the Flash itself. The code which corresponds to the end customer or to the specific application is downloaded just prior to shipping.

Supply shortage has been really the major factor limiting Flash expansion. It is a matter of fact that this young market has experienced already two major shortages, in '93 and '95, which impacted mostly the smallest customers and particularly the ones served through distribution. They had to drop projects at the end of their development phase or at best had to reduce their production ambitions. No wonder if industrial and distribution customers are the laziest among the Flash adopters classes.

Let us try now to picture the other classes. The innovators or pioneers were coming from leading and innovative US computer and automotive manufacturers. Europe brought the generation of early adopters with the first French set top boxes and then with the Nordic digital cellular phones. Then the adoption of Flash memories started to spread out in South East Asia with the motherboard manufacturers and some Hard Disk Drive producers. Japan remained marginal for a long time and can be considered as the core of the late majority, but the market really took off there with the growth of the digital cellular phones, which started after the recent earthquake of Osaka.

Obviously the decision process in favor of Flash of the first adopters was not motivated by pure utilitarian benefits. What happened is that many design engineers were actually driven by the hedonic benefits of Flash memories, such as the fun to be capable to envisage potential advanced functionalities, such as in-system re-programming, but they were justifying their choice by pure rational and utilitarian aspects such as the flexibility or the procurement cost.

Moreover, someone's buying behavior has to be consistent with his self-image. The standard non-volatile memories were becoming somewhat old fashioned and "low tech" as a pure commodity. Flash came as the high tech and non-commodity alternative which cheered up the self-image of the technical decision makers who are the designers and the qualification people.

One would conclude too easily that the need for Flash memories has been created by the marketing action itself. The truth is that needs are never created but that it is a matter of activating and satisfying underlying basic needs such as security, ownership, self-image and accomplishment. Security and the freedom of making mistakes without serious consequences were obviously the first concerns of the non-volatile memory users; this explains why OTP EPROM

APPLE INC.
EXHIBIT 1008 - PAGE 0503

never substituted UV-EPROM. Flash oversatisfied this need, giving the opportunity to correct errors even when the end product is in the field. As mentioned before, self-image had also an important role.

Although its effectiveness is often denied by the experts, we should not forget, in the same spirit, to deal with subliminal persuasion. The Flash name itself and the logo originally adopted by the first leader of the market (a stylized "Flash of lightning" symbol) contribute to promote the idea that Flash memories can satisfy the above mentioned needs. They suggested power, as the one that Zeus holds in his hands, as well as association with speed-of-light performances (and thus not only fast cancellation, but also high performances and even quick design time).

Communication had a great impact on the evolution of the Flash market. Press releases containing over simplified messages about price differential and announcing the disappearance of the current alternate solutions lured many customers and turned out to become true much later. Product specific communication did not expand so much due to the non-commodity aspect of this market. A direct communication focused on the key targeted applications and on customers turned out to be more efficient. Word-of-Mouth communication has been equally important but on the other side, since it is particularly effective with negative information, it hurt considerably the growth of the Distribution and mass market after the first major shortage of '93.

Knowledge is another key issue for the expansion of a new product, essentially in the field of usage and price. Usage problems have been quickly understood by the leading Flash manufacturers, since Flash are much more application sensitive that the previous generations of non-volatile memories; they developed very detailed application notes and user-friendly software routines that made the designers life much easier and development time much shorter. Regarding price, it looks as if the industry used the concept so that non users would get a wrong idea about it. The press was relating messages of low prices that struck the designers while the purchasers were actually paying a high premium over EPROM ...

One could not end a chapter on Flash marketing without mentioning the effects of innovation. We have already mentioned that Flash became an enabling technology boosting brand new applications. Moreover the kind of innovation required by Flash memories is continuous and smooth, so it has the least disrupting influence on the existing production processes thus allowing a smooth transition from the former non-volatile technologies.

APPLE INC.
EXHIBIT 1008 - PAGE 0504

9.5 FLASH MEMORY AND THE "ECONOMY"

The overall evolution (or is it a revolution?) of semiconductors throughout its 50 years of history has been a continuous, relentless demonstration of the Shumpeter principle of "creative destruction". What is more, in the field of memories, Flash is the perfect paradigm of this principle. The social psychologist, Amabile, says that a product is creative if it is novel and appropriate, useful, correct or a valuable response to needs and that the task it fulfils is heuristic rather than algorithmic.

In system development today and tomorrow it is clear that the final customer, who can also be the consumer himself, is less and less the classical stereotype of "only the rich buy the high priced, high performance" or "the masses buy only low cost, low performance". Today the "value customer" has emerged who rightly wants to buy high value, but at low cost. Electronics has managed to satisfy this demand, thanks to semiconductors, giving the possibility to everyone to buy equipments that 40 years ago could only have been bought by big companies. Moreover the performance of this new, low priced product is today orders of magnitude higher than the products offered at the origin of this "value" revolution. The Flash memory, with its inherent flexibility, is today and will continue to be at the forefront of enablers for this revolution.

The previous chapters in this book have tried to explain why Flash memories are at the leading edge of semiconductor technology and design. They are also different from many other logic ICs and even from DRAM. Flash memory is not only a "power business". Even if the technology and the money are available, a fab with a well controlled manufacturing machine running like a clockwork and pouring out cheap ICs is not enough, since Flash still needs a deeper knowledge of process physics, of device know-how and system know-how with respect to other technologies. The winners of tomorrow's Flash markets will be those able to meet the challenge of the system on a chip, bringing together more and more sophisticated technology with innovation and system know how. Flash technology today is an open door to future systems on a chip.

For many years, a new concept of marketing has been sweeping the world, known as "relationship marketing" as opposed to the traditional "transactional marketing". Transactional marketing is typical of commodities, dominated by the traditional marketing mix approach (with special emphasis on price), while relationship marketing shifts the emphasis to the management of the important strategic relationships between companies. Cost moves to value. The product gives a competitive advantage to the customer. Buyers and sellers become partners. Design becomes a symbiotic affair among equipment designers and component design. The virtual company becomes a reality. We are seeing this

APPLE INC.
EXHIBIT 1008 - PAGE 0505

in the field of memories today, and Flash is just crossing this point of marketing concept movement.

One very special characteristic of semiconductors, and memories in particular, is the push for higher and higher density. This leads to more and more expensive production systems. A fab today has a value of some billions of dollars, orders of magnitude higher than the fabs of only a decade ago. When the expensive, complex technology development and fab acquisition is added to the design complexity of Flash memory, then the overall development costs become exponentially bigger.

Customers however would like Flash memory to be tailored to their needs, an Application Specific Standard Flash. This will force the competing memory products manufacturers to find some new dimension of competition and collaboration. The high complexity of the combined product and technology, and the cost of the fabs tomorrow, lead to the conclusion that not everybody, in fact not many players, will be able to invest, compete, make profit and grow. Moreover, the evolution of system performance will shorten the product life cycle, leading to the fact that there cannot be many suppliers for a single product supplied to a single customer. Even if the customer desires to have more than one supplier.

One can even speculate if pseudo-standard products like Flash memories could, in a highly dynamic market, originate new macroeconomics environments. The "oligopolistic free competition" non-volatile memory market could be an example of the macroeconomics environments of tomorrow.

Flash is the new non-volatile memory. Flash is the product that allows integration of life giving renewal to a system. Flash is the enabling technology for system evolution. So will it be also the catalyst for a macroeconomics revolution?

9.6 APPLICATIONS MORE IN DETAIL

9.6.1 Survey (by P. Bergé - STMicroelectronics, Memory Product Group)

The traditional and main applications for Flash memory are in the storage of the software code and operating systems for microprocessors. In the PC, which is an application with which most people are familiar, the BIOS program is the firmware which links the hardware to the operating system (for example, Windows 95™). It is fundamental to the PC, as it contains the boot-up sequence of instructions that allows the processor to find and execute its first instructions when the equipment is turned-on. Most PCs today contain 1, 2 or 4Mbit of Flash memory. This allows an easy, in-system, upgrade of the BIOS. For example, to upgrade the operating system BIOS, while maintaining system integrity, the boot block of the Flash memory, which is normally write and erase protected, maybe re-programmed. As PCs move down in size to the laptop and

APPLE INC.
EXHIBIT 1008 - PAGE 0506

notebook, they require larger amounts of non-volatile memory. Notebook PCs have to integrate additional functions compared to their desktop counterparts, such as energy management. Personal Digital Assistants (PDAs) have no hard disk drive storage and store the whole of their operating system in non-volatile memory. Many of them are today using socketed masked ROM, but there is no doubt that they will soon realize the savings in terms of total cost of ownership and extra functionality that can be brought about by the use of Flash memories.

The computer industry is being converted to Flash memory, way beyond the PC mother board. Almost all peripherals, such as disk and tape drives, CDROMs, DVD players and most add-on boards like video and sound cards, require upgradeable non-volatile memory of 1 to 4Mbit for the storage of program code. Page and laser printers are even larger potential users of Flash memory, to store more complex page description languages and image rasterisation programs. Their market is growing particularly fast as a side effect of the Internet as a way of exchanging more information, especially images.

The Internet, and networking in general, are among the big drivers of the Flash industry. They are the best theoretical demonstration of the versatility of Flash memories as a code storage medium. All network equipment is linked together and remote program updates are relatively easy and respond to the need for frequent software upgrades resulting from the fast progress and change in the complex communication protocols. This applies to both small equipment, with small Flash memory of a few Megabit, like modems or network interface cards, to larger equipment containing Megabytes of Flash memory such as network routers and PBX.

The end of this century will be strongly influenced by the explosion of nomadic, portable communication tools. So will the Flash market. Cellular phones are a key Flash market driver, accounting for about 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 have quadrupled from 4 to 16Mbit. The communication protocols are extremely memory intensive, particularly for the European GSM and the Japanese PDC, not to mention the emerging multi-standard systems. The user interface software is becoming more and more sophisticated to make the phones more user friendly and allow manufacturers to differentiate their products from competitors, and customers really do use all the features! Here the Flash memory has become the memory of choice, especially after some manufacturers experienced the high costs of recall of phones from the field and upgrading due to the discovery of software bugs.

One of the earliest market segments to realize the value of using Flash memory has been automotive. In the late '80s the automotive industry had already understood the potential advantages of using Flash memory and started to ap-

APPLE INC.
EXHIBIT 1008 - PAGE 0507

ply them in the engine management and automatic gear change systems. One driving issue was the need to re-program the memory in the application, as the electronic modules were encased in resin protective blocks which prevented any physical component replacement. The use of Flash memory allows the board to be manufactured with one memory content tailored for testing, and then to be re-programmed according to the car model in which the module will be installed. The use of Flash memories also allows field upgrades by local garage personnel and dynamic adaptation of the equipment to control the engine according to the wear through its lifetime.

With the strong trend of electronic pervasion in the car, there is no doubt that applications will arise for which the Flash memory will find new uses in the body electronics area. This has started with the introduction of car navigation systems carrying the software code for interaction with the extensive Global Positioning System, and it will expand both through the wider spread of such systems and through their expansion into a broader concept of car multimedia integrating also games and entertainment capabilities.

In the field of entertainment, digital set-top boxes are absorbing a growing part of the Flash memory production. Here again is an emerging market in the take-off phase, also driven by increasing functionality and the use of much bigger software code. Set-top boxes are now running with Megabyte sizes of computer-like operating systems that the operators want to be able to upgrade "over the air", and there is no alternative to Flash memory. The availability of bigger, cheaper and more flexible Flash memories will fuel the evolution of the basic pay-TV functions toward interactivity, adding features such as Internet browsing and home shopping.

The historical trend to use Flash for code storage will ensure a healthy growth to the Flash industry. But the dramatic reduction in the cost per bit of Flash is now giving birth to file storage applications which are very likely to become a significant driving factor in the next few years. Unlike the code storage area, Flash has not yet won the battle against the alternative magnetic storage solutions. What we can expect however is that Flash will grow in its own new, innovative markets.

The take off of the PDA market will generate the long predicted business for data file storage, here the Flash is carried mostly in the format of PC Cards. Another driver is in image storage and the use of Flash in digital cameras. As digital camera image quality increases and prices fall, this will be complemented by increased Flash memory storage in small form factor cards. The camera image capacity will be competitive with standard films, with acceptable quality and the digital cameras have the big advantage that their pictures can be transmitted over the Internet.

APPLE INC.
EXHIBIT 1008 - PAGE 0508

Finally, voice storage will also benefit from Flash solid state media. Initially the use of Flash was introduced in answering machines, but the trend is to extend its use to voice recording functions in embedded portable systems such as cellular phones, PDA, pagers, etc. Once again the Flash memory is closely associated with the new, nomadic world that is fast becoming a reality.

9.6.2 Flash in Mobile Phones and Terminals (C. Kunkel - STMicroelectronics, Europe Telecom Business Unit)

Semiconductor memory is an important part of mobile phones of today. Due to the demanding needs of these mobile phones (small size, long talk and standby times, low cost) and the fact that mobile phones have entered into the consumer marketplace, the high volumes of production justify memory devices adapted to the needs of mobile phone manufacturers.

The following text will firstly explain the types of memory devices used in mobile phones today, and then concentrate more closely on the Flash device, in particular, why it is needed, how it is used, and finally Flash needs for the future.

Memory Usage Needs in Mobile Today

Presently three types of memory devices are used in today's mobile phones. These different types are needed due to the differing modes of operation required.

Flash memory is used to store the code for the Protocol Layer Stack (Layers 1, 2 and 3). This code implements the telecommunication functionality. The other major use of the Flash is to store the Man-Machine-Interface, or MMI. The MMI is the user interface between the keyboard/display and the Protocol Stack Layer 3. The MMI memory size can vary greatly due to number of languages supported, graphical display support, etc. In today's mobiles, 300 to 400 Kbytes are dedicated to Protocol Layer Stack storage and the rest to MMI. The GSM Protocol Layer Stack, a subset of the ISO/OSI 7 layer model, is the communication control protocol defined by the GSM specifications and must be implemented in all GSM terminals.

EEPROM memory is mainly used to store hardware related parameters. These parameters are computed, mobile per mobile, on the production line and stored in the EEPROM during the mobile production process (e.g., TX ramp up/down correction parameters). The EEPROM is also used to store user set-up parameters, e.g., language, ringing tone, etc. As read speed is not a critical issue, but package size is, the preferred EEPROM interface is serial (e.g., I2C or SPI). I2C stands for Inter-Integrated Circuit and is a 2 wire serial

APPLE INC.
EXHIBIT 1008 - PAGE 0509

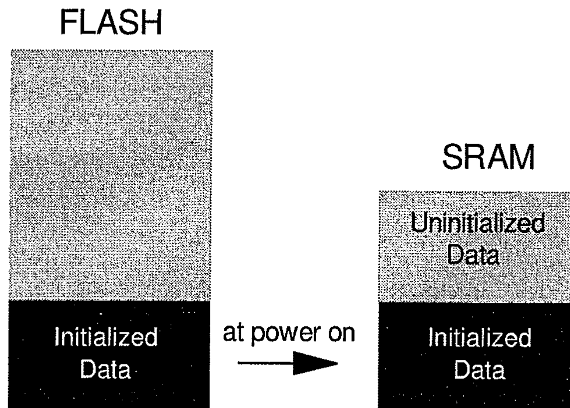


Figure 9.12 Handset power-on sequence.

bus protocol; SPI stands for Serial Peripheral Interface and is a 3 wire serial bus protocol.

SRAM memory is used to manage dynamic data. Initialized data is copied from the Flash to the SRAM at start-up (see diagram below), this data is read and write as long as the mobile is ON, the data is lost at power OFF. Also in the SRAM, we normally find a copy of all or part of the SIM card contents. As the SIM card is accessed in a serial manner, the access time can be long and hence the user does not want to wait a few seconds when accessing a phone number in the SIM card (see Fig. 9.12).

Flash Operation and Usage in Mobile Today

In normal mobile phone use, the Flash is used as a read-only memory. A typical configuration found in production today is 4/8Mbit organized as either x8 or x16 (depending on byte or word read access). The software is directly executed from the Flash by a microcontroller unit (typically 16-bit) using a frequency derived from the 13MHz GSM reference clock. Typical read access times required are between 80ns and 120ns. The supply voltage today is between 2.7V and 3.0V.

As the Flash is basically used as read-only-memory, the question needs to be asked as to why Flash is used instead of an OTP. The reasons are simple: code stability and time to market. GSM is a very complex standard. Before a mobile can be sold it must pass certain tests. This is called FTA or Formal Type Approval. These tests are the minimum that must be satisfied. Additionally, field tests are performed throughout the world on many different GSM networks, as each network does not behave exactly the same. It is not uncommon that the

APPLE INC.
EXHIBIT 1008 - PAGE 0510

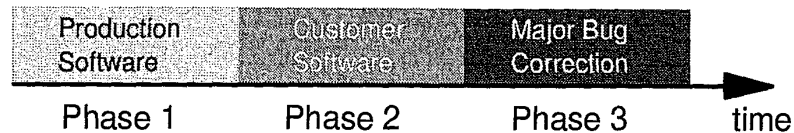


Figure 9.13 Flash programming phases.

Protocol Layer Stack code and MMI are frozen day(s) before the production ramp-up begins. Hence the attractiveness of Flash as a ROM memory to get to market quickly. Even if a manufacturer could live with OTP cycletimes, another problem arises. Field trials continue even after the mobile is launched in production and several more updates of the code will be made during the mobile's lifetime. Again, another advantage of Flash over OTP.

Flash Programming

The Flash is normally only used in the programming mode at time of production. The diagram below illustrates the three major phases when the Flash is programmed (see Fig. 9.13).

Phase 1 - Production Software. During this phase, the Flash is programmed with a software called "production software". This software contains, among the standard Protocol Stack and MMI software, special test routines that are needed in order to ensure that the mobile terminal is tested to FTA standards and that the production out-going quality is met. The Flash may be preprogrammed and then soldered to the PCB, or it may be soldered to the board unprogrammed and then programmed during the production flow. On-board programming can be a production bottleneck depending on Flash program speed. At the end of Phase 1, the mobile phone has been tested successfully and is ready for Phase 2.

Phase 2 - Customer Software. The mobile phone is now ready to be personalized based on the customer/operator in question. At this point in time the Flash is completely re-programmed with some or all of the following software:

- generic L1/2/3 + MMI software;
- extra features or new services (depending on operator);
- other operator security features (e.g., SIMLOCK, that is a method to restrict a GSM terminal to operate only with a chosen SIM card, and thus a chosen operator. Operators use this technique to keep their subscriber base, as the GSM terminal is subsidized in the user subscription price);

APPLE INC.
EXHIBIT 1008 - PAGE 0511

- new language (language not part of standard MMI).

At the end of this phase, the mobile phone is minimally tested again (to be sure programmed software has been correctly stored in Flash) and the mobile phone is ready to be shipped to the customer/operator.

Phase 3 - Major Bug Corrections. Once the mobile phone enters into volume shipments and deliveries, many more people are using or, otherwise stated, "testing" the software. It is not uncommon for some major and/or minor bugs to be discovered in this phase. Depending on the severity, a major bug correction may be planned. This implies that the Flash must be updated via field service personnel without physically removing the Flash (it must be reprogrammed in the system). These major bug correction changes are obviously kept to a minimum due to the cost and inconvenience involved.

As previously explained, it is obvious that, during Phase 2/3 when the Flash is already mounted on the PCB, a key Flash parameter is the programming time.

It is correct here to underline that the programming voltage is the premium feature and that the strive is to minimize it. Today it is advisable not to quote the 12V, in fact:

- this is the picture of the old design;
- at present 5V products single power supply do not pay any performance slow-down with respect to dual voltage products;
- in the case of the new (3V), the next (1.8V) and the future (0.9V) board design, memory manufacturers will have to adapt the program voltage of the acceleration pin to be used in factory (from 5V first to 3V then).

It won't be possible to apply 12V to products designed in 0.25–0.18μm technology and targeted to voltages below 3V (due to reliability problems such as oxide breakdown, hot carrier degradation, etc.).

In synthesis, programming time is the challenge, but:

- no penalty on 5V/3V single power supply;
- acceleration pin will be on 1.8V/0.9V in future products but with a voltage not higher than 5V.

Future Challenges for Flash

To understand the future requirements for Flash, it is helpful first to understand from the terminal manufacturer point of view their Careabouts.

APPLE INC.
EXHIBIT 1008 - PAGE 0512

Terminal Manufacturer Careabouts. From a user's point of view, users are looking for the following:

- long talk time;
- long standby time;
- smallest physical size possible.

From a cost point of view, manufacturers are interested in the following:

- reduced component count;
- lowest cost per functionality implemented.

These Careabouts translate to the following regarding Flash.

Terminal Manufacturer Careabouts for Flash.

- high density to support MMIs for high-end/low-end terminals;
- lowest read voltage possible to maintain 80–120ns read access time;
- single voltage operation for read/pgm/erase to reduce number of separate power supplies and therefore number of DC/DC converters and LDOs (Low DropOut voltage regulator) required;
- access time performance will have to be boosted by different device organization (page mode/burst mode) to achieve read sequences from 30ns to 15ns.
  Flash memory access time it is already the bottleneck for DSP (running at 100MHz and higher) and for MCU (running at 50MHz and higher): any time saved in accessing the Flash it's a direct improvement in the system performances;
- security features to protect memory content will be of interest;
- integration: by the next year combined memory and logic functions will be available for the cellular chip set;
- chip scale package will dominate the development even if it is not for granted that the μBGA™ (micro Ball Grid Array) will be the dominant one. μBGA is a trademark of TESSERA and corresponds to a technology where the package consists essentially in the die itself.
  μBGA has the advantages of being the most aggressive chip-scale package in terms of dimensions (x, y, thickness), but has the big drawback of not

APPLE INC.
EXHIBIT 1008 - PAGE 0513

being transparent to die shrinkage and has as of today an higher cost. FBGA (Fine pitch Ball Grid Array) is a more mature technology, which has the major advantage not only of being transparent to die shrink, but of being compatible of a multi-chip packaging which could be the name of the game in the evolution toward the single chip solution (possibility to start with the dual chip/single package in advance) or a real alternative to that aimed to optimize cost and complexity at the silicon technology level. FBGA uses the traditional plastic molding approach of the Ball Grid Array (BGA) and the ball pitch is $\leq$ 1mm.

Thus, the challenges facing Flash for the future are numerous:

- higher densities to support ever elaborate MMIs;
- trend to emulate EEPROM in Flash. This is already being implemented in M39432 Flash device;
- trend to integrate Flash into Digital Baseband device to reduce parts count. This requires a compatible cost-effective process;
- trend to lower the read voltage to that of the rest of the digital section. 1998 will see 1.8V and interest is to go to sub 1V by 2000;
- increasing use of data applications require the MCU to operate at speeds greater than 13MHz. Therefore reduced read access times are needed to operate MCU at max speed with no wait-states (sub 80ns);
- a program access time adequate to allow recording of voice samples directly (voice-memo function).

9.6.3 Flash in the BIOS (Gan Moy - IBM)

Introduction

In the computer industry, the key application of Flash memory is for program storage. This program storage is required to initialize the hardware and activate the computer. An example of the program storage is BIOS (Basic Input Output System). The function of the BIOS is to help the application program and operating system perform tasks on hardware. Another application for program storage is to store the boot up code for the operating system. This minimum program code is required to load the operating system and application code stored in mass media such as disk and tape drive into main memory (DRAM based) to bring the intelligence into the computer. Another application for program and control code storage in computer is network computing. The

APPLE INC.
EXHIBIT 1008 - PAGE 0514

Network Router uses Flash memory to store the Router table, control code and network operating system to control the hardware and routing of information between networks. The input/output (I/O) device such as the keyboard, disk drive, tape drive, and printer also use Flash memory to store the code to interact between the hardware and the operating system. This chapter will briefly describe the various Flash memory application usage on the IBM computing system. This includes the personal computer (IBM PS/2), IBM midrange computer (AS/400, RS/6000), IBM mainframe (S/390), Networking hardware like Router, 5494 remote control units and various I/O devices attached to the computer system such as disk drive, tape drive and printer.

The Personal Computer

The Personal computer uses Flash memory to boot up the system and starts to load the operating system, application program and diagnostic program from the hard drive into main memory. The final code loaded into main memory can be 8 Mbytes to 24 Mbytes depending on the operating system and application being loaded. There are also small density Flash memory used on various adapters (BIOS). Examples are the display adapter (monitor attachment), printer adapter (printer attachment), disk drive adapter (floppy disk drive and hard disk drive attachment) and the communication adapter (Local Area Network attachment and telephone interface attachment). They are used to store the control code for the controller on the various adapter cards.

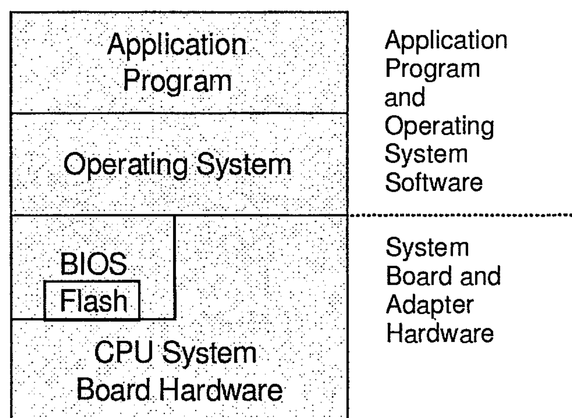


Figure 9.14 Hardware and software layer of PC.

RS/6000

The IBM RS/6000 system is utilized in two main application areas. One application is in the technical and scientific community that have high demand on number crunching and powerful graphic requirements. The second application is in the commercial application such as banking, manufacturing, and retail. This commercial application is for database manipulation, fast transaction processing and fast communication with other computers.

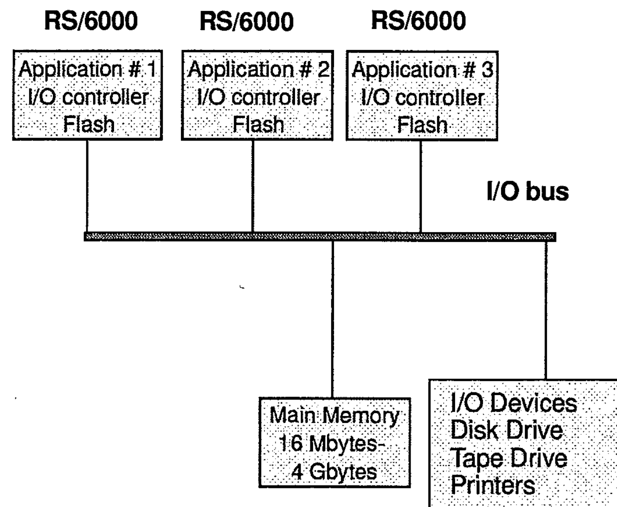


Figure 9.15 Three applications symmetric RS/6000 multiprocessor system.

The IBM RS/6000 uses the AIX (Advanced Interactive eXecutive) operating system. AIX is IBM's enhanced version of the UNIX operating system. Inside the RS/6000 system, the core processor is a RISC (Reduce Instruction Set Computing) processor. IBM chose to use the RISC vs. CISC (Complex Instruction Set Computing) processor is for better system performance. The advantage of the RISC processor is that it is customized to run efficiently for the application it is targeted for. It also has a very simple instruction set which can be executed in a single clock cycle. The CISC is more popular with the Personal Computer because it can run more than one operating system. However, the instruction set requires more than one clock cycle to execute. RS/6000 can be configured into the multiprocess or system. It adopts the concept called SMP (Symmetric Multiprocessor System). This concept involves using multiprocessors and sharing a common set of resources such as memory, disk, tape drive and printer. Each processor can run different application programs simultaneously.

APPLE INC.
EXHIBIT 1008 - PAGE 0516

The Flash memory application in the RS/6000 is used to boot up the computer, the system configuration for the I/O controller and the diagnostic code for the system after powering up and monitoring the status of the hardware.

AS/400

The AS/400 is IBM's mid-range business computer. The AS/400 is the follow on to the IBM system /3X series. The AS/400 is a multi-user computer system (a single computer can interact with more than one user at a time). In a single AS/400 processor system, the system processor is the core of the computer system. It has the hardware to execute the application software and the I/O bus to communicate with the I/O processor. Each I/O processor has its own unique responsibility to control and work with the I/O device attached to it and communicates back to the system processor. There are several I/O processors tied to this bus. An example of the I/O processor is the workstation I/O processor which is used to connect different types of terminals, workstations, and printers. They all need to work together with the AS/400 system. The Storage I/O processor is used to connect the disk drive, tape drive and diskette drive as a mass storage media for the AS/400 system. The communication I/O processor is used to communicate the remote digital network via the telephone line or IBM 5494 remote control unit which connects the remote workstation to the AS/400 system.

In addition to the system processor, there is a service processor attached to the system processor and the I/O buses. The function of the service processor is used to boot up the system (Flash memory), monitor the AS/400 system's

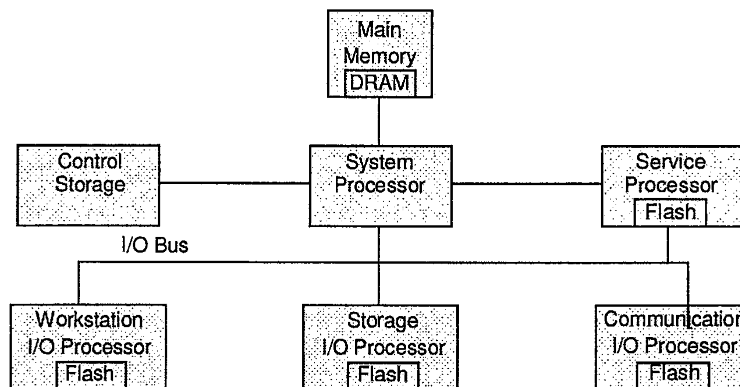


Figure 9.16 AS/400 single system processor.

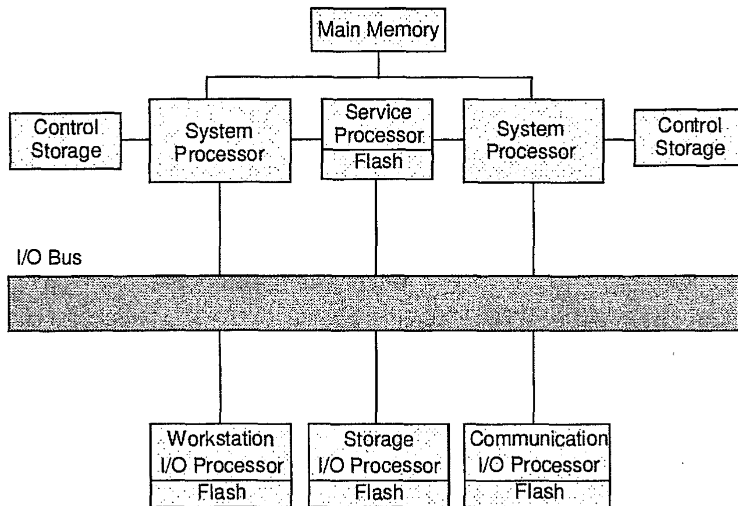


Figure 9.17 Two-way multiprocessor AS/400 system.

status and reports errors back to the system operator via the operator control panel and also stores the error condition of the Flash memory.

To further improve the processing power of the AS/400 system, IBM also offers a multiple processor system called N-Way multiprocessor system. In the N-Way multiprocessor system, each processor cooperatively executes a single application via the OS/400 operating system.

S/390

The S/390 is IBM's mainframe system. It is the follow on to the S/370 and the E S/9000 system. Like the mid-range system, the S/390 is a multi-user and multiprogramming computer system. In a multi-user computer system, it is capable of supporting many users in a single computer system. An example is the airline reservation and banking system used by the customer representative to process or inquire about a customer's account. In a multi-programming computer system, it can process multiple application programs in a predetermined time slice. A multiple application program can start simultaneously rather than wait for some long application program to finish before a new application can start.

Due to the complexity of the S/390, it is not sufficient to use the Flash memory to boot up the system. It requires a server (PC types) to boot up the system in sequence, and to load the operating system and application software. The server also performs control duty as well as monitor the systems well being

APPLE INC.
EXHIBIT 1008 - PAGE 0518

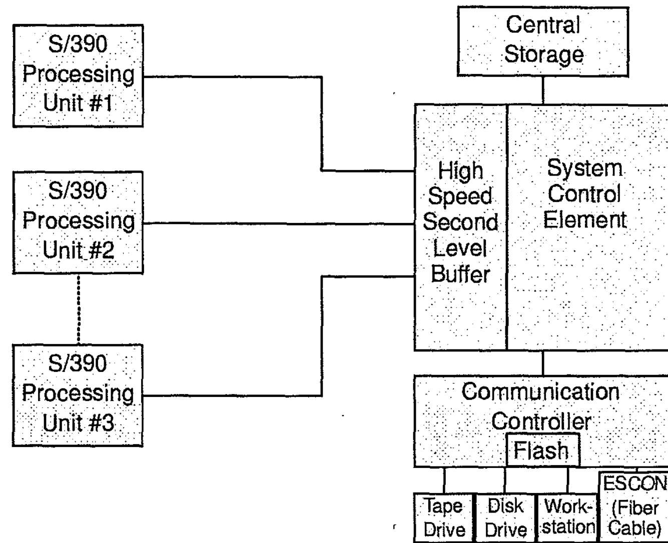


Figure 9.18 System 390 N-Way processing unit.

and it reports to the system programmer any problems associated with the system. The S/390 employs a modular concept, whereby it grows with the customer as its application requirements change.

The customer can start with a single processor unit. As its processing or storage requirement grows, it can add additional processor units or disk drives to enhance the current system.

Network Computer (NC)

The IBM network computer falls between a terminal and a personal computer. The Network computer is intended for a low cost terminal replacement, yet it is powerful enough to run on the windows environment and emulate many types of graphic terminals. IBM currently offers the series 100, and 300 system (series 300 is twice the processing speed of series 100 system).

The whole concept of the network computer is to have an operator push a power on button, boot up the system from a network server (using low density Flash memory to start the boot up process) and load in the operating system into the main memory of the NC and execute the client's request in a server environment. Once the request result is ready, the server sends the result back to the client. The future NC may include a PCMCIA Flash memory to store the operating system.

APPLE INC.
EXHIBIT 1008 - PAGE 0519

The advantage of having a NC is that there is no need to have multiple copies of the application program on multiple client's desktop (all clients are running the same version stored on the server). It is also easier to update the latest version of an application program (one copy server update versus multiple copy update on the client's desktop). The NC is also less likely to contaminated with a virus because there is no hard disk to write to.

Networking Hardware

Without the Networking Hardware, the Computer Network will not exist today. Midrange and High-end computer systems would sit lonely in a corner of an office building and the PC will be sitting alone in someone's desktop working independently. The computer will not be able work together and share the information with each other locally or in a global environment. The LAN (Local Area Network), WAN (Wide Area Network), Intranet (campus environment) and Internet (global environment) are all made possible by Switch, Bridge, Router and its software.

Computer Network communication is very similar to our telephone network communication. There are several factors to consider: cost (cabling and data communication equipment), distance (how far it goes before a repeater is needed), reliability, capacity, delay (guarantee of service), fairness and protocol compatibility. In order for a computer to communicate to another computer in a network, we must have a compatible path between the two points. If the network has 100 computers, it needs 99 connection points from that computer to the other computer on the network before it can talk to each other directly. The cost of cabling this network is too expensive. The more efficient and less costly method is to bring each computer communication to a centralized Switch and by using Bridge and Router, we can reduce the cost and solve the communication problem mentioned above. In IBM, low density Flash memory is used to store the control code for the Ethernet, Token Ring and FDDI (Fiber Distributed Data Interface) adapter. For Switch, Bridge and Router, the higher density Flash Memory is used because it needs to support multiple protocol, store the network topology and network operating system software.

Input/Output Devices

Input and output devices such as keyboard, disk drive, tape drive and printer, Flash memory are used in this area mainly for control code storage, and for the configuration of the I/O devices and optional feature for the customer. During manufacturing, IBM also uses the Flash memory to load in the various test programs to perform the manufacturing test for its I/O devices.

APPLE INC.
EXHIBIT 1008 - PAGE 0520

Summary

The Flash Memory used in the computer industry is mainly for program storage. The biggest challenge for the Flash Memory technology for the computer industry is to supply the industry with Fast Flash memory to catch up with the Microprocessor and Microcontroller speed (greater than 200MHz for reading and writing). Computer applications required a Common Flash Interface and a standard programming command among different Flash technologies. The network controller speed and bandwidth is increasing as well. Network computing requires a higher density Flash memory, yet smaller sector size to accommodate the smaller write time. All these requirements are in addition to making a reliable and quality Flash Memory device for the computer industry.

9.6.4 *Flash in Automotive (Antonino Damiano, Riccardo Ferrara, Antonio Re - Magneti Marelli, Electronic Systems Division)*

Introduction

The need for low cost, re-programmable non-volatile memory has been fundamental from the start of electronic applications in the automotive market. This need arises in both the development and the pre-production phases.

During development, trials and adjustments of the product require continual software updates. During and after starting production there is also a time when the latest versions of software must be installed before the programs are finally frozen and masked into ROM, either stand-alone or embedded into the microcontroller. The eventual use of ROM is essential to reach low costs for high volume production.

The answer from the semiconductor makers to this market requirement has been for many years the EPROM or OTP, which even if they are adequate for development, are completely unable to meet the requirements as a production solution because of high costs and the difficulty of re-programming.

The arrival and development of Flash technology has made this new type of memory very interesting for automotive applications. The continual reductions in cost of these memories and the possibility to integrate them in single chip microcontrollers operating over an extended temperature range (−40 to +125°C) make their use interesting also for volume production (over 1 million per year).

This contribution discusses the use of Flash memory in automotive applications and the advantages that they bring, above all in terms of development cost reduction and product cost.

In the first part, the problems associated with the downloading of software and the on-board programming of the Flash are discussed; in the second part

APPLE INC.
EXHIBIT 1008 - PAGE 0521

the use of Flash and the part it plays in cost reduction are presented (for example the elimination of the EEPROM through emulation in Flash memory); and finally the current state-of-the-art of the use of Flash in automotive is presented with views about future evolution.

The Engine Control Unit (ECU) is taken as an example throughout the article.

Downloading of Software

In the first generation of automotive electronics the programming of the Flash was controlled by the CPU running complex algorithms. The Flash memory was seen as a special memory area that could be erased and re-programmed by bulk. For this it was necessary to use a ROM whose content was the boot software, to be used in case the supply voltage was removed during the erasure of the Flash, to be able to restart the Flash programming.

The Flash also required dual power supplies, 5V and 12V, and the ability to control the programming voltage. If, during programming, the supply voltage varied from the optimum value, the component life in terms of the number of erase/program cycles could be severely reduced. The supply voltage regulation was controlled by software by monitoring the battery voltage and interrupting the operation by a hardware voltage regulator, if the voltage was not correct.

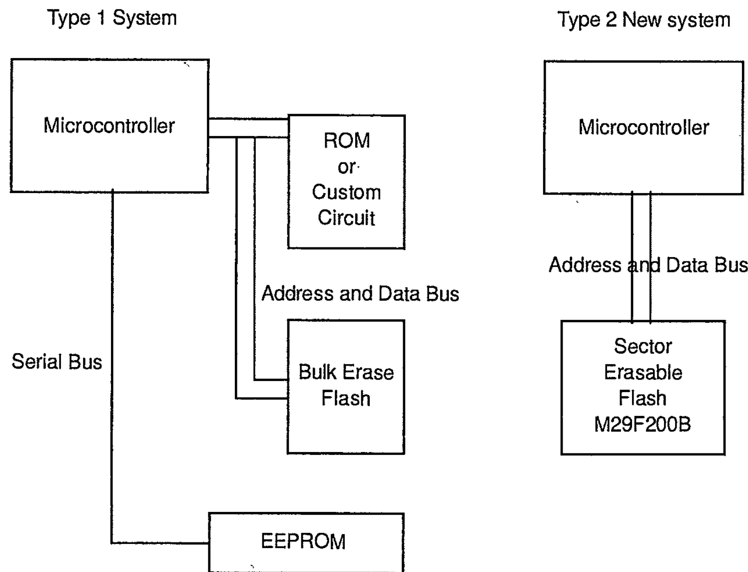


Figure 9.19 Evolution of the hardware architecture of automotive electronic systems.

During the last few years technical developments of Flash have allowed the introduction of block erase and single voltage supply operation, which has simplified the system design in three fundamental ways:

- elimination of the boot ROM;
- no control or regulation of the programming supply required;
- embedded programming algorithms.

The elimination of the boot ROM, substituting it with a sector of the Flash, has reduced the component costs needed for its management. The elimination of the dual power supply and the embedding of the programming algorithm, have reduced the overhead for the CPU of the programming process. The embedded algorithm also contributes to the security and reliability of the component, both for the software downloading and for write protection of sectors. The control of the write cycles and the generation of the programming voltages is under the control of the Flash and this has contributed to the extension of the erase/re-program cycle endurance to over 100,000 cycles.

Low power consumption and fast access times in line with the performance of the microprocessors and the possibility to suspend erase are important characteristics which encourage the use of Flash memory in automotive applications. For microprocessors with a bus frequency of 20MHz, for example, an access time of less than 70ns is needed for the Flash. The following diagram illustrates the evolution of the architecture of the automotive electronic systems through the introduction of block erasable Flash. The first system is one with Flash, EEPROM and boot ROM, while the second is the latest system which eliminates both the boot ROM and the EEPROM. The boot ROM used in the first system was often integrated in a custom circuit that, for other reasons, had the microcontroller address and data busses already available.

Flash Programming

The boot software is a new software module created for programming and re-programming the Flash. Re-programming of the board or downloading of the software is a process that makes possible not only the simple transfer of software from a file to the Flash, but also provides its traceability using specific identification of the software loaded, including any re-programming operations, thus always assuring the correct completion of the operation.

If we take as an example the ECU, the software to be programmed in the Flash consists of two main parts:

1. application control software, that is software that commands and controls all of the application;

APPLE INC.
EXHIBIT 1008 - PAGE 0523

2. parameters for adjustment or calibration, that is data with which the control software operates, which dedicates the software to the particular application.

For this application, a typical Flash partitioning is shown in Fig. 9.20 where the first sector is reserved for the boot block, with two small sectors for EEPROM emulation, followed by two parts reserved for the application software and the calibration data.

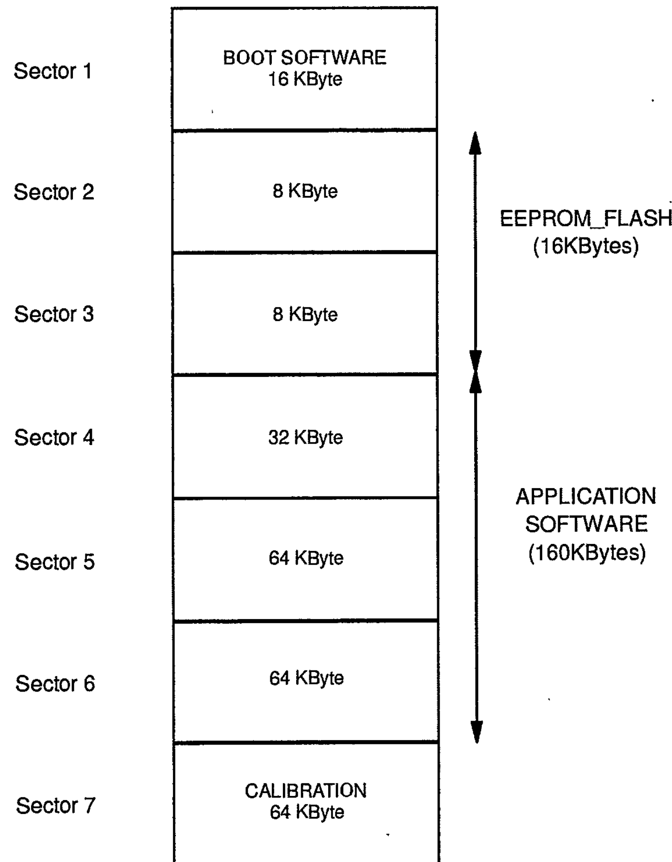


Figure 9.20 Example of mapping of the Flash in systems for engine control.

The downloading software in the boot block has the fundamental task of controlling the security of the conditions needed to make the erase and programming operations; in particular:

- state of the operation;

APPLE INC.
EXHIBIT 1008 - PAGE 0524

- protection against attempted pirating;
- protection against theft of the vehicle;
- control of the compatibility of the software with the type of board.

The control of the state of operation must block any re-programming when the board is executing its main function. For example, re-programming of the board is not allowed when the vehicle is in motion, but only during maintenance or diagnostics. The protection against attempted pirating protects the board against programming with non-authorized software. This protection is enabled with security mechanisms in the programming tool and on the board, for example by access keys or secret algorithms defined by the maker of the vehicle that permit downloading to the board only files with the correct encryption. The protection against theft is a special case of the pirating protection, to prevent re-programming of the board with software able to bypass the anti-theft present on modern ECUs (immobilizers). The control of compatibility guarantees a major security in operation, having the goal of blocking programming of a board with software which is incompatible and which could prevent or compromise the system's function.

The level of security achieved by products of the latest generation that use block erasable Flash is very high, and the risk of blocking permanently the board during downloading no longer exists thanks to the presence of the boot block that cannot be altered and is protected against accidental erasure. Using this boot block a system reliability level higher than the previous system using boot ROM and Flash is achieved.

Protection against unwanted writing and erasure. With current Flash there is a complete guarantee of avoiding unwanted erasure or writing operations, since in order to write to the Flash the following conditions must be satisfied:

1. write enable must be active;
2. the block protection must be disabled;
3. the correct sequence of programming must be given, which is an event very unlikely to occur in a random or casual way.

It is important to mention at this point the protocol used on the interface serial line for Flash programming, as it is its implementation on our boards that provides superior protection against undesired writes.

Protocols used for automotive systems. In the automotive world there are more and more requests to implement a standard of communication between

APPLE INC.
EXHIBIT 1008 - PAGE 0525

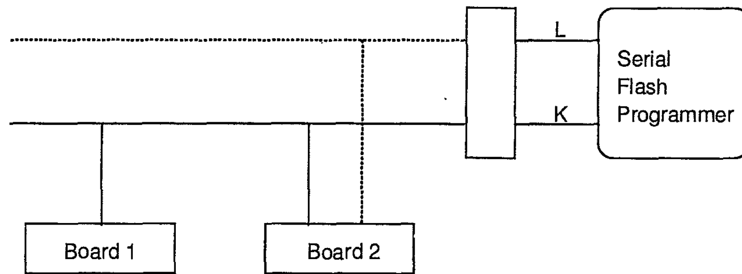


Figure 9.21 Possible configurations of the serial lines for an automotive system.

boards. The standards ISO CARB 9141-2 and SAE J1979 specify the connection between electronic components in a vehicle and test equipment using two lines K and L. These connections allow communications for diagnostics, test and maintenance. The K line is bi-directional, and normally used during communications for the exchange of data and commands. The L line is used for initialization of the communications with messages at 5 baud, but is optional as the same message is sent on line K [1].

Protocols of last generation, like "Keyword 2000" [4, 5, 6] offers a higher protection against accidental writing due to the concept of a diagnostic session. The diagnostic session is a specific state of the protocol in which only certain operations are permitted. In order to start Flash writing the diagnostic session for Flash programming must be entered, which requires the recognition of an access key by a security access block. Only after these two exchanges does the processor permit writing of the Flash by enabling the write control signal output from the microcontroller. This new generation of protocols in addition allows downloading of part of the software, for example only the application software or only the calibration data. This performance of the protocol has been only made possible by the availability of block erase Flash. The following section outlines how the availability of Flash changes the way of operating during the fundamental phases of the life cycle of the board.

DESIGN AND DEVELOPMENT. The reduction of the operations needed to program a prototype translates directly into a saving of time and thus the reduction of the cost of development.

The possibilities to program the Flash are many, but none requires action at the hardware level as is the case for EPROM. To re-program a board, it is not necessary to open the equipment, but just to download through the serial line accessible at the external connector.

This applies both in the laboratory and in the vehicle installation. The reprogramming can be done also with the board mounted in the vehicle, and if

APPLE INC.
EXHIBIT 1008 - PAGE 0526

required, also by the manufacturer's designers during construction, experimentation and adjustment of the product.

The download to the Flash can also be made from emulators, both classical JTAG [3] types or BDM [2], which facilitates considerably the software development in the laboratory.

MANUFACTURING. The first programming of the board can be made, as for a normal EPROM, using a PROM-programmer off-line (the memory can be programmed before mounting on the board). The Flash can be pre-programmed with definitive software that includes the boot software, the application and the calibration data. The board thus leaves the factory ready for installation in the vehicle without further intervention.

The use of Flash memory moreover allows re-programming by the car maker or authorized service centers after delivery, thanks to the availability of the diagnostic line. In this way it becomes possible for the vehicle manufacturer to personalize the board by software directly on the production line or at service centers. The vehicle manufacturer has in this way the advantage of not needing to forecast the number of boards for a particular model of vehicle, but to configure the product for each single one, as is done today for example for the paintwork.

AFTER SALES SERVICE. The possibility to re-program the Flash opens also new opportunities for the after sales service phase. Whenever needed, it is possible to update the software of the board to correct for defects or provide improvements in functionality of the product, or to add options not available at the time of purchase of the vehicle.

In this case also there are significant savings in the cost of updating the product (updating the software does not require substitution of the board) especially in the case of the correction of software defects, for which the whole operation requires less than a few minutes and can be done directly at a service center.

Management of EEPROM FLASH

Adaptive parameters. Different strategies for the control of the motor necessitate the calculation and recording of parameters throughout the use of the board. These parameters are called adaptive as their values may be changed to adapt to the specific vehicle on which the board is mounted. The object is to obtain the best performance possible by taking into account the spread in the characteristics of the elements making up the complete system and the change of characteristics during its use. These parameters must be modifiable

APPLE INC.
EXHIBIT 1008 - PAGE 0527

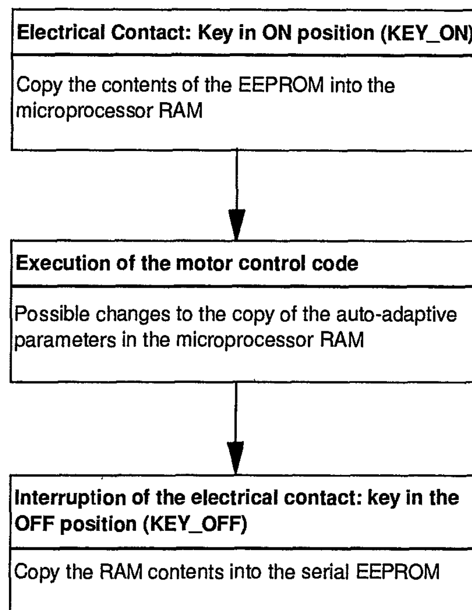


Figure 9.22 Management scheme for the adaptive solution using serial EEPROM.

at run-time and cannot be regarded as simple calibration data that are changed only on re-programming, described before.

A Typical Solution Using an EEPROM Memory. A common solution to this need is to use a serial EEPROM memory. In this way it is possible to modify directly the value of a parameter in the EEPROM or to work with a data image in the RAM of the microprocessor and subsequently update the new values. This solution is preferable in terms of the execution time and allows changes in parameter values without impacting on the real-time execution of the board. Fig. 9.22 shows the adaptive solution using the following logic: copy the previous values saved into RAM, modify them, then write to the EEPROM.

An Alternative Solution Using a Paged Flash Memory. With the objective of reducing the costs of the boards, the use of a paged Flash memory can allow the elimination of the EEPROM component. The EEPROM function is replaced by a use of the Flash solution which we will call EEPROM_FLASH.

The basic principle is to use two sectors of the Flash as a circular buffer of N zones. Where N is the whole number resulting from the size in Kilobytes of the two zones divided by the size, also in Kilobytes, of the amount of data to be

APPLE INC.
EXHIBIT 1008 - PAGE 0528

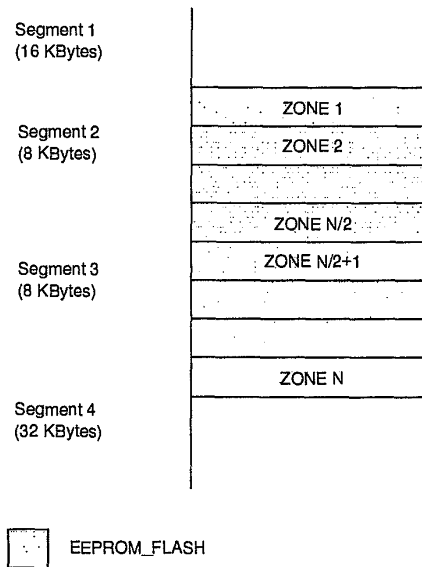


Figure 9.23 Mapping of the zone N onto sectors 2 and 3 of the Flash.

managed. Taking for example the Flash memory M29F200B [7], the possible memory map is shown in Fig. 9.23.

If the size of the data to be saved in EEPROM is 2 Kbytes, there will be in this example $N = 8$ logical zones in the two sectors chosen.

At the moment of "KEY_ON" a search is made for the last zone written to copy the values to the internal RAM of the microprocessor, while at "KEY_OFF" the modified values are saved in the next zone of the buffer. Fig. 9.24 shows logic for the auto-adaptive system using this EEPROM_FLASH.

At the end of the search for the last zone written and to guarantee the validity of the data chain, each zone has an associated status byte that is updated for the two zones of the EEPROM_FLASH used for reading and writing: corresponding to each state transition for a zone (copy into RAM finished, start and end of writing in the Flash, etc.), this byte changes its value. In the case that the microprocessor is reset, at the next "POWER_ON" the memory state can be found and the correct zone read.

To assure the circular management of the buffer, referring to Fig. 9.23, when writing to the last zone ($N/2$) of the second sector, the third sector is erased and when writing to the third sector (N) the second is erased. The use of at least two sectors of the Flash is unavoidable for security reasons especially during erasure. If only one sector were used, if the board was reset for whatever reason just after the erasure of the sector the values for the auto-adaptive system would

APPLE INC.
EXHIBIT 1008 - PAGE 0529

be lost. Using the EEPROM_FLASH not only reproduces the functionality of the serial EEPROM, but gives the possibility to review the previous six zones written thus providing a history of the evolution of the values for the auto-adaptive parameters.

Applications

The use of Flash memory is today made possible by the technology progress over the last few years, above all in the number of erase/write cycles for the Flash. This aspect is critical since with the EEPROM_FLASH design the Flash is written every time the key is turned off at "KEY_OFF" and the sector is erased at every N "KEY_OFF's". Evidently if,

- Na = the number of turn-on/turn-off's of the engine for which the manufacturer requires the correct management of the recording of the auto-adaptive parameters (of the order of every tens of miles),
- Nc = the number of cycles of erase/write cycles guaranteed for the component,

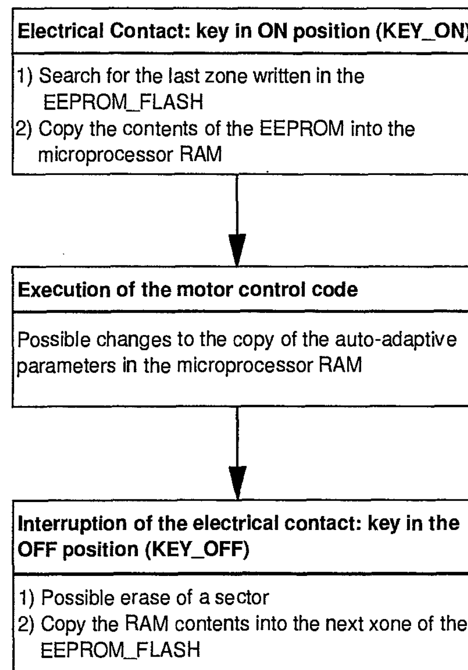


Figure 9.24 Management logic for the auto-adaptive system using EEPROM_FLASH.

APPLE INC.
EXHIBIT 1008 - PAGE 0530

Index

A
AC read/command interface, 469–470
Address sequences, in product characterization, 475–476
Address transition detector (ATD), 296
Algorithms, in control logic, 343–344
Alternate metal ground (AMG) cell
  in binary flash cells, 123–127
  in multilevel flash cells, 144
AND cell and array
  in binary flash cells, 114–118
  in multilevel flash cells, 144
Application maturity, 494–495
Architecture
  of AMG cells, 125f
  array architecture summary in, 130, 131t
  of binary flash cells, 130–131
  of AND cells, 118f
  of common ground array, 142–144
  of control logic, 337–350 (*See also* Architecture, of control logic)
  of DINOR cells, 110f, 113f, 142–144
  of drain and source, 67–68
  of EEPROM (Electrically Erasable Programmable Read Only Memories), 12–13, 129–130
  of embedded algorithms, 337–350 (*See also* Architecture, of control logic)
  erase operation circuitry in, 327–360 (*See also* Architecture, erase operation circuitry in)
  error correction codes in, 356–360
  flash architecture overview in, 241–257 (*See also* Flash architecture, overview of)
  of flash array family tree, 99f
  of flash memory, general scheme in, 253f
  of memory NOR, 16–18
  of multilevel cells, 142–146, 368–373
  of NAND cells, 24f
  program operation circuitry in, 314–327 (*See also* Architecture, of program operation circuitry)
  of read path decoding, 257–270 (*See also* Read path decoding)
  of read path input and output buffers, 270–280 (*See also* Read path buffers)
  of read path sensing techniques, 280–313 (*See also* Read path sensing techniques)
  redundancy in, 350–356 (*See also* Redundancy)
  in scaling issues, 81–83
  of split gate virtual ground cell, 121f, 122f
Architecture, erase operation circuitry in, 327–360
  charge pumping in, 332–335
  critical node slow discharge in, 330–331
  double supply voltage approach in, 329–331
  single supply voltage approach in, 331–337
  source erase circuitry in, 329–330
  source switch in, 336–337
  voltage regulators in, 335–336
Architecture, of control logic, 337–350
  buffers in, 341
  command interpreter in, 340–341
  command write in, 339, 340f
  data pins in, 339
  embedded algorithms in, 343–344
  erase flow in, 346–347
  fundamental circuits in, 340–342
  program flow in, 344–346
  program/erase controller in, 342–343
  read path in, 342
  slow operations in, 337–339
  testability issues in, 348–350

APPLE INC.
EXHIBIT 1008 - PAGE 0531

Architecture, of program operation circuitry, 314–327
  cell voltages in, 314–315
  drain voltage regulation in, 317–320
  gate voltage regulation in, 320–327
  program path in, typical, 315–316
Array architectures, of multilevel flash cells, 368–373
Array classification, of binary flash cells, 96–98
Array distribution, on reliability, 401–409
Array efficiency, 60–61
Array ground line (AGL), in NAND architecture, 372–373
ASIC (Application-Specific-Integrated-Circuits), 41
ASP (Application-Specific-Products), 41
Automotive market, 515–525
  after sales service in, 521
  applications in, 524–525
  design and development in, 520–521
  EEPROM flash management in, 521–524
  flash programming in, 517–520
  manufacturing in, 521
  protocols in, 519–520
  software downloading in, 516–517
Average energy model, 190–191

B

Band diagram, of MOS structure, 168f
Band structure transition, at silicon-silicon dioxide interface, 160f
Band-gap reference, in gate voltage regulation, 325f
Band-to-band tunneling (BBT), 54–56
  in MOSFET's, 170–174
Barrier height
  in BBT tunneling, 174f
  in tunneling phenomenon, 164f
Biasing
  in read mode, 282–283
  in row decoders, 260–261
Binary flash cells, 93–136
  alternate metal ground (AMG) cell and array in, 123–127
  array architecture summary in, 130–131
  array classification in, 96–98
  AND cell and array in, 114–118
  cell design complexity in, 94–95
  DINOR cell and array in, 108–114
  EEPROM based flash architecture in, 129–130
  figures of merit in, 93–94
  NAND cell and arrays in, 104–108
  NOR cell in, 99–104
  scaling in, 131–133
  scaling in, of internal voltage, 133–135
  source injection concepts in, 127–129
  split gate virtual ground cell and array in, 118–123
Binary search sensing, 379
BIOS market, 508–515
  IBM AS/400 in, 511–512
  IBM RS/6000 in, 510–511
  IBM S/390 in, 512–513
  input/output devices in, 514
  network computer in, 513–514
  networking hardware in, 514
  personal computer in, 509
Bit density, *vs.* production year, 59f
Bit line(s)
  to bit line coupling, 143–144
  selection signals for, 357f, 358f
Boltzmann transport equation, 178–179
Bonding wires, 275–277
Boost concept, in read path decoding, 267–270
Boron-phosphosilicate film (BPSG), 68–69
Breakdown lifetime evaluation model, 215–217
Buffers, for read path
  input, 270–272
  output, 272–275
Buffers, in control logic, 341
Building blocks, of flash memory operations, 253–257
Burn-in, 449
Byte pin, 251–252

C

Capacitive coupling ratio, of MOS transistor, 46
Capacitors
  ferroelectric, 135
  thin oxide, I-V characteristics of, 210f
Carrier distribution
  homogenous electric fields on, 182–184
  model of, 191
  thermal equilibrium on, 181–182
Carrier heating
  gate length on, 189–190
  models of, 190–191
Carrier transport, fundamentals of, 176–185
  Boltzmann transport equation in, 178–179

APPLE INC.
EXHIBIT 1008 - PAGE 0532

carrier distribution in, homogenous electric fields on, 182–184
carrier distribution in, thermal equilibrium on, 181–182
distribution function in, 176–178
effective temperature model in, 184–185
scattering in, 180–181
Carriers, electronic properties of, 155–161
as classical particles, 156–157
crystal electrons in, 155–156
interface traps in, 159–161
oxide in, 159–161
silicon dioxide in, 157–159
silicon in, 157
silicon-silicon dioxide interface in, 159
Cascade schematic, in read path sensing, 284f, 285f, 286f
Cell array stress test (CAST), 405–408, 409f
Cell bias, during programming, 48f
Cell layout
cell size future projections in, 132f
in scaling, 81
Cell operation, physical aspects of, 153–223
carrier transport in, 176–185 (*See also* Carrier transport, fundamentals of)
carriers in, electronic properties of, 155–161 (*See also* Carriers, electronic properties of)
hot carrier effects in MOSFET's in, 185–207 (*See also* MOSFET's, hot carrier effects in)
MOS structures in, 155–161 (*See also* Carriers, electronic properties of)
oxide and interface degradation in, hot carrier injection on, 217–223 (*See also* Oxide degradation, hot carrier injection on)
oxide degradation in, high field stress on, 207–217 (*See also* Oxide degradation, high field stress on)
tunneling phenomenon in, fundamentals of, 161–165 (*See also* Tunneling phenomenon, fundamentals of)
tunneling phenomenon in, MOSFET's in, 165–176 (*See also* MOSFET's, tunneling phenomenon in)
Cell scaling, 27
Cell structure, definition of, 64–68
Cell voltages. *See* Voltage(s)
Channel doping, 61–63, 68
Channel hot-electron (CHE) injection/programming
in BBT-TBT, 174
device operation and, 205–206
disadvantage of, 97
drain available, 201–202
in EPROM, 8
on gate current, 199–201
in industry standard cell, 47–49
mechanism of, 139f
in MOSFET's, 199–201
in NOR cell architecture, 369–371
oxide traps and interface state generation from, 420
secondary generated, 202–203
in self-converging programming, 388
substrate, 203–205
vs. FN tunneling, 7, 365–366
Channel length, on programming curves, 49f, 50–51
Charge injection mechanisms, 7
Charge loss, control gate/floating gate overlap in, 69f
Charge pumps
in erase operation circuitry, 332–335
in NOR organization, 20
on programming efficiency, 51–52
Charge retention
improvement of, 9
in scaling issues, 81
Chemical Mechanical Polishing (CMP), 71, 72f
Chip defects, 351. *See also* Defects
Chip enable pin, 251
Classification
of flash architecture, 242–243
of flash cells/arrays, 96–98
of market segments, 485f, 486–487
Clock interaction, on output node, 303f
CMOS
for flash memory, 58–60
floating gate potential of, 42–46
as pass switch, in read path decoding, 261–262
Collisions, in Boltzmann transport equation, 178–179
Column decoder, 263–264
in drain voltage regulation, 317–318, 319f
Command interpreter, 340–341
Command write, in control logic, 339, 340f
Commodity, *vs.* non-commodity products, 493–495
Common ground array
fabrication steps for, 101f
in multilevel flash memory, 142–144
schematic layout of, 100f
voltage levels in, 102f
vs. AMG array, 126f

APPLE INC.
EXHIBIT 1008 - PAGE 0533

Communication tools, 501. *See also* Mobile phone market
Contact technology, 70–71
Control gate, 4–5
Control pins, 250–252
Coupling capacitance, in erase operation circuitry, 331f
Critical node slow discharge, 330–331
Customer/supplier relationship, 495–496
CVD technique, 66

D

Damascene technique, 73f
Data address, in read path, 254
Data loss, single bit, 426–430
Data pins, in control logic architecture, 339
Data retention, in NOR organization, 23
Data storage. *See also* Mass storage
- bilevel *vs.* multilevel concept in, 362–364
- corruption of, 411
- multilevel, reliability of, 436–438
- in multilevel flash cells, 364–367
- program disturbs in, 411
- read disturb on, 438

DC tests, 467
DC-programming disturbs, 77–78, 79f
Decoders
- column, 263–264
- hierarchical, 264–266
- row, 259–263

Defects
- of chip, 351
- detection of, 464–465 (*See also* Fault repair; Testing)
- of tunnel oxide, 401–409

Degradation
- gain degradation in, 430–436
- of oxide (*See* Oxide degradation)

Density, of silicon and silicon dioxide, 158f
Depleted bits, in read path sensing, 305–307
Depletion test, 455–456
Depletion-mode operation, of MOS transistor, 45
Deposition
- of polysilicon, 66
- spin on glass (SOG) technique in, 71

Design, on MOSFET's carrier heating, 189–190
Design complexity, in binary flash cells, 94–95
Device operation, in MOSFET's, 205–206
Diagnosis, in fault repair, 464–465
Die cost, 463
Dielectrics, interlevel, 68–69
Differential sensing, 285–288
- active loads for, 289f
- offset current in, 289–293
- in read path, 254–255
- semiparallel, 293–294

DINOR (divided bit-line NOR), in binary flash cells
- array device cross-section of, 109f
- erase and programming time in, 112f
- merit summary of, 114t
- schematic layout of, 110f, 113f
- triple well process in, 111
- truth table for, 114t
- word line pitch scaling in, 111–112

DINOR (divided bit-line NOR), in multilevel flash cells, 142–144
Dirac distribution, 156
Direct memory access, testing of, 451–452
Distribution function
- in carrier transport, 176–178
- for MOSFET's hot carrier effects, 198

Disturb margins, in multilevel cells, 139–141
Disturbs. *See* specific types e.g., Program disturbs
Divided bit-line NOR (DINOR). *See* DINOR (divided bit-line NOR)
Doping, 61–64
- of channel, 68
- concentrations of, for BBT-TBT, 172–173

Double supply memory array, 249f
Double supply voltage approach, 329–331
Drain architecture, 67–68
Drain available hot electron injection, in MOSFET's, 201–202
Drain coupling, in floating gate transistor, 45f
Drain current, monitoring of, 387–388
Drain stress, 77–78
- in gain degradation, 434, 435f
- in multilevel cells, 140

Drain voltages
- FN tunnel programming and, 385–387
- on programming curves, 50f
- on read path sensing, 281–282
- regulation of, in program operation circuitry, 317–320

DRAM (Dynamic Random Access Memory), 2
- ferroelectric nonvolatile DRAM in, 14f
- market ramp up time for, 132f
- as market segment, 487

Drift velocity, 182

APPLE INC.
EXHIBIT 1008 - PAGE 0534

Dynamic inhibit concept, for NAND array, 106–107
Dynamic Random Access Memory (DRAM). *See* DRAM (Dynamic Random Access Memory)

E

EEPROM (Electrically Erasable Programmable Read Only Memory), 9–15
  architecture of, 12–13
  architecture of, in binary flash cells, 129–130
  ferroelectric memory in, 13–15
  floating gate thin oxide memory (FLOTOX) in, 9–10
  as market segment, 486
  textured polysilicon cells in, 10–12
Electric fields, on carrier distribution, 182–184
Electrical behavior
  of erase efficiency, 59f
  of floating gate device, 42–46
Electrical erase. *See* Erase, electrical
Electrical model, of floating gate device, 5
Electrically Erasable Programmable Read Only Memory (EEPROM). *See* EEPROM (Electrically Erasable Programmable Read Only Memory)
Electron injection
  hot (*See* Channel hot-electron (CHE) injection/programming)
  secondary, 135
Electrons, behavior of
  as classical particles, 156–157
  in crystals, 155–156
  envelope function of, 161f, 162
  in MOSFET's, carrier heating on, 186–188
Embedded controller, in NOR organization, 23–24
Embedded flash memory, 26–27
Embedded market, *vs.* mass storage market, 92
Endurance, 3–4
  gate oxide degradation on, 418–421
  intrinsic, 418–421
  in NOR organization, 23
  of program/erase efficiency, 415–416
  in reliability, 75
Energy band diagram, of floating gate transistor, 43f
EPROM (Erasable Programmable Read Only Memory), 1–2, 7–9
  device schematic for, 95f
  as flash memory application classification, 242
  as market segment, 486
  in read path sensing, 304–305
  threshold voltage of, in read path sensing, 287f
Erasable Programmable Read Only Memory (EPROM). *See* EPROM
Erase, electrical
  in erase operation circuitry, 327–328
  flash cell *vs.* EPROM in, 94–95
  vs. UV erase, 401, 402f
Erase, parallel, 457
Erase disturbs, in reliability, 79–81
Erase function
  in DINOR array, 112f
  erase bias in, 171f
  in flash cell classification, 96–98
  gate voltages on, 59f
  of industry-standard flash memory, 53–58
  in memory NOR, 18–20
  negative gate *vs.* source in, 403f
  of NOR cell, 18, 19f
  over-erasing in, 409–411
  oxide thickness on, 57
  in split gate virtual ground, 121f
  tail bits in, 403–405
Erase path
  building blocks of, 256–257
  flowchart of, 258f
Erase/program performance, in production testing, 470
Erratic erase phenomenon, 423–426
Error correction, in fault repair, 462
Error correction codes, 356–360
  unrecoverable events curves in, 360f
ETOX, 15
EWS, in parallel testing, 473–474
Exponentially ramped current stress (ERCS), 405

F

Fabrication steps
  for AND array, 117f
  for common ground NOR array, 101f
Failure analysis
  direct memory access testing for, 451–452
  functional failures in, 468–469
FAMOS (Floating gate Avalanche-injection MOS), 7–9
Fault repair, 460–466
  diagnosis and repair in, 464–465

APPLE INC.
EXHIBIT 1008 - PAGE 0535

error correction in, 462
redundancy in, 462–464
redundancy in, test tools for, 465–466
Ferroelectric capacitor, 135
Ferroelectric memory, for EEPROM, 13–15
FG charge, and gate current, 48–49
Figures of merit
for AMG array, 127t
for binary flash cells, 93–94
for DINOR array, 114t
for industry-standard cell, 103t
for market applications, 490t
for NAND array, 93–94
for split gate array, 123t
Final passivation, 74
Finite state machine, 344
Flash architecture, overview of, 241–257
application classification in, 242–243
erase path building blocks in, 256–257
NOR cell operation and array organization in, 243–250
program path building blocks in, 256
read path building blocks in, 253–256
user interface in, 250–252
Flash memory, 16 Mbit
implantation steps for, 62t
implantation steps for, dose and energy ranges in, 68t
Flash memory, future of, 27–33
non volatile memory market development in, 29–33
technology evolution in, 27–29
Flash memory cells
binary cells in, 93–136 (*See also* Binary flash cells)
carrier heating in, 186–189
configurations of, 40t
density of, evolution curve for, 131f
device schematic for, 95f
embedded, 26–27
gate current density in, 201f
low voltage hot carrier effects on, 206–207
multilevel cells in, 137–147, 361–391 (*See also* Multilevel flash cells)
operations overview of, 253–257
program disturbs on, 413–414
reliability of (*See* Reliability)
testing of (*See* Testing, of flash memory)
Flash memory cells, industry-standard in, 37–83
array efficiency in, 60–61
band-to-band tunneling in, 54–56
basic structure of, 42–46
cell bias in, during programming, 48f
cell structure definition in, 64–68
CMOS process in, 58–60
common ground NOR array in, 100f
coupling ratios in, 51–52
critical technology steps in, 102
doping for, 61–63
double polysilicon stacked gate in, 39f
drain voltages in, on programming curves, 50f
dual *vs.* single voltage in, 54–56
erase function of, 53–58
figures of merit summary for, 103t
generic, schemata of, 16f
implantation steps for, 62t
interconnections in, 70–73
interlevel dielectrics in, 68–69
isolation in, 60–61, 63f
process flow for, 60t
programming curves of, channel length on, 49f
programming of, 47–53
programming of, efficiency in, 51–52
read function of, 46–47
scaling issues in, 81–83
technology and process in, 58–74
temperature on, 51–52
threshold voltages in, hot electron injection on, 47–49
T-shaped staked gate in, 82, 83f
voltages in, on erase curves, 58
yield and reliability in, 74–81
Flash read, low voltage, 308–311
Floating gate device, 4–6
in AND array, 117
avalanche-injection MOS (FAMOS) as, 7–9
charge injection mechanism for, 7
floating gate potential in, 42–46
in industry standard cell, 39–40
Floating gate thin oxide memory (FLOTOX), 9–10
Floating gate transistor
I-V characteristics of, 45f
schematic cross section of, 44f
FLOTOX (floating gate thin oxide memory), 9–10
FN tunneling. *See* Fowler-Nordheim (FN) tunneling
Fowler-Nordheim (FN) programming, in NOR cell architecture, 371
Fowler-Nordheim (FN) tunneling
in AND arrays, 115–116
in band-to-band tunneling, 56f
in DINOR arrays, 110–111

APPLE INC.
EXHIBIT 1008 - PAGE 0536

in flash memories, 7
in flash memory evolution, 28
in MOSFET's, 166–167
in multilevel cells, 140–142
in NAND arrays, 25–26
vs. CHE, 142t, 365–366
Functional model, of testing, 446
Functional tests, in production testing, 468–469

G
Gain degradation, 430–436
Gate currents
measurement of, 53
in MOSFET's, hot carrier effects on, 199–206
in MOSFET's, in BBT-TBT regime, 173f
in relation to substrate, 55
Gate Induced Drain Leakage current (GIDL), 171
Gate length, on carrier heating, 189–190
Gate oxide, 41
degradation of, on memory endurance, 418–421
direct tunneling through, 166–167
Gate stress, 77–78
Gate voltage(s). *See also* Threshold voltage(s); Voltage(s)
band-gap reference in, 325f
of floating gate, 5–6
level shifter in, 324f
regulation of, in program operation circuitry, 320–327
on serial sensing, 378–379
in staircase ramp programming, 384–385
vs. snap-back triggering, 314f
Ground, common *vs.* virtual, 97–98

H
Hamming code, 357–360
Hierarchical decoder, 264–266
High field stress, on oxide degradation, 207–217. *See also* Oxide degradation, high field stress on
Hole impact ionization coefficients, for MOSFET's, 193f
Hole trapping, in erratic erase phenomenon, 423–424
Holes, in gain degradation, 432
Homogenous electric fields, on carrier distribution, 182–184
Hot carrier effects, in MOSFET's, 185–207. *See also* MOSFET's, hot carrier effects in
Hot-electron injection/programming. *See* Channel hot-electron (CHE) injection/programming
Hysteresis curve
of ferroelectric capacitor, 14f
of input buffer inverter, 272f

I
IBM AS/400, 511–512
IBM RS/6000, 510–511
IBM S/390, 512–513
Image force corrections, in MOS structures, 168–169
Impact ionization, on MOSFET's hot carrier effects, 192–194, 195f, 196f
Industry-standard device. *See* Flash memory cells, industry-standard in
Injection probability, for MOSFET's hot carrier effects, 198
Input buffers, of read path, 254, 270–272
Input/output devices, 514
Interband transitions, mechanisms of, 17
Interconnection structure, 70–73
Interface traps, 159–161
Interlevel dielectrics, 68–69
Inter-particle collisions, in Boltzmann transport equation, 178
Interpoly dielectric, 66
Intrinsic threshold, 49–50
Inverters, for row decoders, 259–261
Isolation, in flash memory process, 60–61, 63f
I-V characteristics
of floating gate device, 6
of floating gate device, with no FG stored charge, 48f
of floating gate transistor, 45f
of thin oxide capacitors, 210f

J
Junction breakdown, in MOSFET's, 194–195

L
Lattice-particle collisions, in Boltzmann transport equation, 178
Leakage current, parasitic, 60
Level shifter, in gate voltage regulation, 324f
Lifetime evaluation model

APPLE INC.
EXHIBIT 1008 - PAGE 0537

for hot carrier injection oxide degradation, 221–223
for SILC, 214–215
Line biasing, in memory NOR, 20
Local self boosting (LSB) technique, 372
LOCOS isolation, 61–62
Logic design, testing of, 457–460
Low pressure chemical vapor deposition (LPCVD), 9
Low voltage flash read, 308–311

M

Market/marketing
automotive market in, 515–525 (*See also* Automotive market)
BIOS in, 508–515 (*See also* BIOS market)
customer/supplier relationship in, 495–496
DRAM ramp up times in, 132f
embedded *vs.* mass storage in, 92
growth of, 484f
market applications in, survey of, 500–503
market development in, 496–498
market share in, by application, 31f
mobile phones in, 503–506 (*See also* Mobile phone market)
MOS memory market size in, 30f
of non volatile memory, 1–3, 29–33
price in, 497
product development history in, 482–483
relationship marketing in, 499–500
segmentation in, 483–495 (*See also* Market, segmentation of)
supply shortage in, 497
terminal manufacturers in, 506–508
Market, segmentation of
application diversity in, 488–489, 490f, 491f, 492
application maturity in, 494–495
classifications in, 485f, 486–487
commodity *vs.* non-commodity issues in, 493–495
communication in, 498
integration in, 489–490, 491f
product maturity in, 493–494
segment dynamics in, 494
windowed markets in, 492–493
Mass storage
flash memory application in, 242
market for, *vs.* embedded market, 92
NAND arrays in, 24–26
split gate concept in, 122
Maxwell-Boltzmann distribution, 181–182, 184
Memory
multilevel flash, 361–391 (*See also* Multilevel flash cells)
non volatile (*See* Non volatile memory)
sectors of, program disturbs on, 412–413
Memory access times, in read path output buffers, 273
Memory array, double supply, 249f
Memory function, in flash cell classification, 96–98
Memory NOR, line biasing in, 20
Merit summary. *See* Figures of merit
Metal ground cell. *See* Alternate metal ground (AMG) cell
Metallization, 72–73
Metal-Nitride-Oxide-Silicon cells (MNOS). *See* MNOS (Metal-Nitride-Oxide-Silicon)
Microelectronic systems, non volatile memory in, 1–3
Miniboost concept, in read path decoding, 268–270
MNOS (Metal-Nitride-Oxide-Silicon), 4
concept of, 136, 137f
in EEPROM, 9
Mobile phone market, 503–506
customer software flash programming in, 505–506
flash operation in, 504–505
memory usage needs of, 503–504
production software flash programming in, 505
Monte Carlo method, for carrier distributions, 182–183
Moore's law, 58–59
MOS structures
electronic properties of, 155–161 (*See also* Carriers, electronic properties of)
in MOSFET's tunneling phenomenon, 167–170
oxide breakdown in, 211–213
MOS transistor
I-V characteristics of, 45–46
threshold voltage of, 4
MOSFET's, hot carrier effects in, 185–207
average energy model in, 190–191
carrier distribution model in, 191
carrier heating in, 186–189
carrier heating in, design on, 189–190
channel hot electron injection in, 199–201
device operation in, 205–206
distribution function in, 198
drain available hot electron injection in, 201–202

gate current in, 199–206
impact ionization in, 192–194
injection probability in, 198
low voltages on, 206–207
secondary generated hot electron injection in, 202–203
SiO_2 hot carrier injection in, 197–198
substrate current in, 194–197
substrate hot electron injection in, 203–205
typical phenomenon in, 185
MOSFET's, tunneling phenomenon in, 165–176
band-to-band (BBT) tunneling in, 170–174
BBT/TBT modeling in, 174–176
Fowler-Nordheim in, 166–167
MOS structures in, tunnel current modeling of, 167–170
oxide gate direct tunneling in, 166–167
trap-to-band (TBT) tunneling in, 170–174
MOSFET's transistor, 39–40
Multilevel approach, 362–364
Multilevel flash cells, 137–147, 361–391
array architectures in, 368–373
array concept summary for, 146
common ground array architecture in, 142–144
development of, 27–28
DINOR architecture in, 142–144
fundamentals of, 137–138
multilevel approach in, 362–364
multilevel sensing in, 373–382 (*See also* Sensing, multilevel)
multilevel storage issues in, 364–367
NAND architecture in, 371–373
NAND array in, 145
NOR architecture in, with CHE programming, 369–371
NOR architecture in, with FN programming, 371
performance summary of, 390t
program-and-verify approach in, 384–387
programming mechanisms in, 138–142, 382–384
scaling in, 146–147
self-controlled programming approach in, 387–389
virtual ground in, 144

N

NAND cells and arrays, 24–26
architecture of, 24f
in binary flash cells, 104–108
matrix information access in, 243
multilevel flash cell architecture in, 371–373
in multilevel flash cells, 145
in scaling issues, 83
N-channel devices, in hot carrier injection oxide degradation, 217–218
Network computer, 513–514
Networking hardware, 514
Nitradation, 65–66
Node equalization, in reading speed-up, 296
Noise, in read path buffers, 275–277
Non volatile memory
in circuit integration, 41
evolution of, 3–4
market development in, 29–33
in scaling issues, 82
in semiconductor market, 1–3
NOR cells
in binary flash cells, 99–104
depleted bit modification in, 305, 307f
multilevel flash cell architecture in, with CHE programming, 369–371
multilevel flash cell architecture in, with FN programming, 371
NOR cells, organization of, 16–24, 243–250
data retention in, 23
embedded controller in, 23–24
endurance in, 23
erase operation in, bulk, 248f
erase operation in, negative gate, 246f
erase procedure in, 18, 19f, 245–248
flash memory array organization in, 247f
line biasing in, 20
memory array in, double supply, 249f
memory array in, single supply, 250f
program disturbs in, 21–22
program operation in, 244, 245f
read disturbs in, 22–23
read operation in, 244f
read path structure in, 17f
in scaling issues, 82–83
voltage requirements in, 20–21
Numerical analysis, for BBT, 176

O

Offset current, in differential sensing, 289–293
Operating conditions
building blocks in, 253–257
of industry-standard cell, 46–58
Output buffers, of read path, 272–275

APPLE INC.
EXHIBIT 1008 - PAGE 0539

Output enable pin, 251
Output node, clock interaction on, 303f
Over-erasing, on reliability, 409–411
Oxidation process, requirements of, 65–66
Oxide. *See also* Tunnel oxide
electronic properties of, 159–161
thickness of, on erase function, 57
thickness of, on scaling limits, 55
thin, high stress field on, 427
thin, in floating gate thin oxide memory (FLOTOX), 9–10
thin, on capacitors I-V characteristics, 210f
Oxide degradation, high field stress on, 207–217
breakdown lifetime evaluation model in, 215–217
oxide breakdown in, 211–213
oxide wear-out in, 207–210
SILC in, 210–211
SILC in, lifetime evaluation model of, 214–215
SILC in, oxide wear-out on, 207–210
Oxide degradation, hot carrier injection on, 217–223
homogenous degradation in, 217
lifetime evaluation models in, 221–223
n-channel devices in, 217–218
non-homogenous degradation in, 218–220
p-channel devices in, 218
Oxide gate direct tunneling, in MOSFET's, 166–167
Oxide stress, testing of, 446–447
Oxide/nitride/oxide (ONO), 41
in interpoly dielectric, 66

P

Parallel erase, as test feature, 457
Parallel programming, as test feature, 457
Parallel sensing, 293f
Parallel test systems
at EWS, 473–474
final test in, 473
tester structure in, 471–473
Parasitic leakage current, 60
Particle flow, in Boltzmann transport equation, 179f
Pass switch, in read path decoding, 261–262
Passivation, final, 74
Payback curve, *vs.* testing, 444f
P-channel devices, in hot carrier injection oxide degradation, 217–218
Personal computer, 509
Personal Digital Assistants (PDA's), 29
Physical aspects, of cell operation, 153–223. *See also* Cell operation, physical aspects of
Pins, in flash memory user interface, 250–252
Planarization, 71
Poissonian approximation, 351–352
Polycide deposition, in AMG cell, 124f
Polysilicon cells, textured, 10–12
Polysilicon deposition, 66
Potential energy profiles, in tunneling phenomenon, 163f
Power consumption
in erase operation circuitry, 335
in NOR organization, 21
Power down/reset pin, 251
P-pocket, 69f
in reading disturbs, 76
Precharge technique, 298–300
Predecoding, 259
Price, 497
Process steps. *See* Fabrication steps
Process yield, 350–352
Product characterization, 474–478
address sequences in, 475–476
objectives of, 474–475
statistical sampling in, 478
Product cost, testing on, 443–444
Product life cycle, testing on, 444–445
Product maturity, 493–494
Production testing, 466–478
AC read/command interface in, 469–470
DC tests in, 467
erase/program performance in, 470
functional tests in, 468–469
objectives of, 445
reliability in, 470
wafer probe testing in, 466, 467f
Program and verify, 365
Program disturbs
in NAND architecture, 372
in NOR organization, 21–22
on reliability, 411–414
Program loads, drain voltage regulation on, 318–319, 320f
Program operation circuitry, 314–327. *See also* Architecture, of program operation circuitry
Program path
building blocks of, 256
flowchart of, 257f
typical, 315–316
Program verify, 315–316

APPLE INC.
EXHIBIT 1008 - PAGE 0540

in multilevel flash cells, 384–387
Program/erase controller, 342–343
Program/erase cycle failure modes, 418–435
erratic erase phenomenon in, 423–426
gain degradation in, 430–436
memory cell intrinsic endurance in, 418–421
single bit data loss in, 426–430
single bit failure in, 422–423
tail bits in, 422
Program/erase endurance, on reliability, 415–416
Programming
of industry-standard cell, 47–53
parallel, as test feature, 457
Programming disturbs, in reliability, 77–78, 79f, 80f
Programming function
in DINOR array, 112f
in flash cell classification, 96–98
in memory NOR, 17–18
in multilevel flash cells, 138–142
Pseudo-microcontroller, 344
Pull down buffers, 279f
PZT (lead zirconate titanate), in ferroelectric memory, 13

Q

Quality, *vs.* testing, 445

R

Read disturbs
in reliability, 76
on reliability, 414
schemata of, 22f
Read function
of industry-standard cell, 46–47
in memory NOR, 16–17
multilevel sensing in, 373–374
of NOR cell, 244f
Read only memory (ROM), 95f
Read path
building blocks of, 253–256
in control logic, 342
decoding of, 257–270 (*See also* Read path decoding)
input and output buffers of, 270–280 (*See also* Read path buffers)
in product characterization, 475
sensing techniques for, 280–313 (*See also* Read path sensing techniques)
Read path buffers, 270–280
high voltage tolerance in, 277–280
input buffer in, 270–272
noise issues in, 270f, 275–277
output buffer in, 272–275
Read path decoding, 257–270
boost concept in, 267–268
CMOS pass switch in, 261–262
column decoder in, 263–264
hierarchical decoder in, 264–266
low V_{CC} problems in, 266–267
miniboost concept in, 268–270
predecoding in, 259
row decoder in, 259–263
Read path sensing techniques, 280–313
cascade biasing schematic in, 284f, 285f
depleted bits in, 305–307
differential semiparallel sensing in, 293–294
differential sensing in, 285–288
differential sensing in, offset current in, 289–293
from EPROM to flash in, 304–305
low voltage flash read in, 308–311
overview of, 281–285
read fail in, 306f
reading speed-up techniques in, 295–304
reference problems in, 312–313
single end converter in, 283f
Read through concept, in NAND cell/array, 104–106
Ready/busy pin, 251
Redundancy, 350–360
case example in, 354–356
in fault repair, 462–464
process yield in, 350–352
redundancy columns in, 356f, 357f
static redundancy in, 352–353
test tools for, 465–466
vs. error correction codes, 356–360
wafer yield in, 353–354
Reference cell, in read path sensing, 312–313
Relaxation time approximation, 179–181
Reliability, 74–81, 399–439
carrier heating in, 185
data retention in, 416–417
electron retention in, 75
endurance in, 75
EPROM/EEPROM comparison in, 399–401
erasing disturbs in, 79–81
low voltage hot carrier effects on, 206–207
memory array distribution on, 401–409
of multilevel storage, 436–438
over-erasing in, 409–411

APPLE INC.
EXHIBIT 1008 - PAGE 0541

physical aspects of (*See* Cell operation, physical aspects of)
program disturbs in, 411–414
program/erase cycle failure modes in, 418–435 (*See also* Program/erase cycle failure modes)
program/erase endurance in, 415–416
programming disturbs in, 77–78, 79f, 80f
read disturbs in, 76, 414
testing for, 417–418
tunnel oxide defects on, 401–409
vs. testing, 445
Repair, of faults. *See* Fault repair
Resistive equalization, in reading speed-up, 304f
Retention, 3–4
in reliability, 75
Retrograde well, 63f
ROM, device schematic for, 95f
Row decoder, 259–263
in erase operation circuitry, 336

S

Saturation region, of MOS transistor, 45–46
Scaling, 81–83
in binary flash cells, 131–135
limits of, oxide thickness on, 55
in multilevel flash cells, 146–147
Scattering
in carrier transport, 180–181
in MOSFET's, 194
Schemata
of erase operation threshold voltages, 19f
of floating gate, generic, 5f
of NOR organization, 17f
of program disturbs, 21f
of read disturbs, 22f
Schrödinger equation, 155–156
Secondary generated hot electron injection, in MOSFET's, 202–203
Sectorization, in NOR array, 248, 249f
Self aligned field isolation, in AND array, 116f
Self-controlled programming, in multilevel flash cells, 387–389
Self-converging programming, 388–389
Semiconductor Industry Association (SIA), on interconnections, 70t
Semiconductor market. *See* Market/marketing
Semi-parallel sensing, 293–294
schematic of, 297f
Sense amplifier behavior, 454f
Sensing
differential, 285–288
parallel, 293f
for read path, 280–313 (*See also* Read path sensing techniques)
semi-parallel, 293–294, 297f
Sensing, multilevel, 373–382
mixed sensing in, 379–382
parallel sensing in, 377
performance summary in, 382t
serial sensing in, 377–379
signal production and recognition in, 374–376
Sequencer, 344
Serial sensing, 377–379
Shumpeter principle, 499
Signal, production and recognition of, 374–376
Silicon
electron scattering mechanisms in, 180f
electronic properties of, 157
Silicon dioxide
electronic properties of, 157–159
hot carrier injection into, 197–198
Silicon-silicon dioxide interface, 159
Single bit data loss, 426–430
Single bit failure, 422–423
Single end converter, in read path sensing, 283–285
Single supply voltage approach, 331–337
charge pumping in, 332–335
source switch in, 336–337
voltage regulators in, 335–336
Slow discharge, of critical node, 330–331
Snap-back triggering, 314–315
SNOS (Silicon-Nitride-Oxide-Semiconductor), 9
Software downloading, for automotive market, 516–517
Source architecture, 67–68
Source discharge timing, in erase operation circuitry, 331f
Source erase circuitry, 329–330
Source injection, for binary flash cells, 127–129
Source junction profile, 55
Source line elements, in erase operation circuitry, 330f
Source series resistance, on programming speed, 51–52
Source switch, in erase operation circuitry, 336–337
Spin on glass (SOG) deposition, 71

APPLE INC.
EXHIBIT 1008 - PAGE 0542

Split gate virtual ground, in binary cells
  cell layout in, 121f, 122f
  triple poly split gate cross section in, 120f
Split gate virtual ground, in binary flash cells, 118–123
Staircase gate voltage ramp programming, 436–437
Static redundancy, 352–353
Statistical sampling, 478
Storage, of data. *See* Data storage; Mass storage
Stress induced leakage currents (SILC), 210–211
  oxide wear-out in, 207–210
  in single bit data loss, 429–430
Stress modes, in testing, 454
Substrate, on gate currents, 55
Substrate current, in MOSFET's hot carrier effects, 194–197
Substrate hot electron injection, in MOSFET's, 203–205
Successive approximation register (SAR), 379, 380f
Supply shortage, 497
Surface planarization, 71

T

Tail bits
  in erase function, 403–405
  in program/erase cycle failure, 422
Temperature
  on programming speed, 51–52
  on substrate current, 196–197
Temperature model, for carrier transport, 184–185
Terminal manufacturers, 506–508
Testing
  cell array stress test (CAST) in, 405–408, 409f
  of control logic, 348–350
  of reliability, 417–418
Testing, of flash memory
  conceptual test flow in, 448–449
  depletion test in, 455–456
  direct memory access in, 451–452
  fault repair in, 460–466 (*See also* Fault repair)
  functional model in, 446
  logic design in, 457–460
  low Vt test in, 456
  oxide stress in, 446–447
  parallel test systems in, 470–474
  parameters of, 447–448
  product characterization in, 474–478
  on product cost, 443–444
  on product life cycle, 444–445
  production testing in, 466–478 (*See also* Production testing)
  production testing in, objectives of, 445
  stress modes in, 454
  test productivity in, 457
  vs. quality and reliability, 445
  Vt measurement in, 452–454
Textured polysilicon cells, 10–12
Thermal equilibrium, on carrier distribution, 181–182
Threshold distributions
  for CAST, 408f
  for erased flash cells, 305f
  for standard *vs.* multilevel flash memory, 254
Threshold voltage(s)
  of erase operation, 19f
  erase procedures on, 80f
  of flash memory transistor, 362–363
  of floating gate, 5–6
  hot electron injection on, 47–49
  of MOS transistor, 4
  of negative gate *vs.* source erase, 403f
  in NOR architecture, 370f
  shift of, on read function, 46–47
  of UV *vs.* electrical erase, 401, 402f
  window closure in, 76f
Transconductance, of MOS transistor, 46
Transmission coefficient, in tunneling phenomenon, 162–165
Traps. *See also* Interface traps
  elastic, in SILC mechanisms, 212f
  in erratic erase phenomenon, 423
Trap-to-band tunneling (TBT), in MOSFET's, 170–174
Triple poly architecture, in multilevel flash cells, 144
Triple poly split gate, in binary flash cells, 120f
Triple well process, in DINOR, 111
Triple well structure, 64f, 65f
  pull down buffer in, 279f
Truth table
  for AMG array, 126t
  for DINOR array, 114t
  for NAND array, 107t
  for split gate array, 123t
Tunnel oxide
  defects of, on reliability, 401–409
  degradation of, 75
  growth of, 64–66

APPLE INC.
EXHIBIT 1008 - PAGE 0543

Tunneling
band-to-band, 54–56
band-to-band, in MOSFET's, 170–174
trap-to-band, in MOSFET's, 170–174
Tunneling current, 165
enhancement of, 423–425
of MOS structures, 167–170
Tunneling phenomenon, fundamentals of, 161–165
basic concepts of, 161–162
transmission coefficient in, 162–165
tunneling current in, 165
WKB approximation in, 161–162
Tunneling phenomenon, in MOSFET's, 165–176. *See also* MOSFET's, tunneling phenomenon in

U
Ultra-violet (UV) radiation
for EPROM erasure, 8
vs. electrical erasure, 401, 402f
Unbalanced loads, in differential sensing, 293f
User interface, in flash architecture, 250–252

V
Virtual ground array, in multilevel flash memory, 144
Volatility, in scaling, 82
Voltage(s). *See also* Threshold voltage(s)
in common ground NOR array, 102f
on control pins, 252
drain, in program operation circuitry, 317–320
of EPROM, 8
on erase curves, 58
on erase function efficiency, 59f
gate, in program operation circuitry, 320–327 (*See also* Gate voltage(s))
low, in flash memory evolution, 28
low, in hot carrier effects, 206–207
in MOS structure tunnel currents, 167–170
in NAND architecture, 25f
in NOR organization, 20–21
in program operation circuitry, 314–315
on read path buffers, 277–280
in read path decoding, 266–267
on read path output buffers, 273–275
in scaling issues, 81–82
scaling of, in binary flash cells, 133–135
on thin gate oxide tunneling current, 170f
on write disturbs, 12–13
Voltage regulators, in erase operation circuitry, 335–336
VPCX values, in gate voltage regulation, 321–323
Vt measurement, in testing, 452–454, 456

W
Wafer probe testing, 448–449
in production testing, 466, 467f
Wafer sort testing, 448–449
Wafer yield, 353–354
Wave packet, 156–157
Well doping, 61–63, 64f
Wentzel-Kramers-Brillouin (WKB) calculations, 80–81
in tunneling phenomenon, 161–162
Window closure, in threshold voltages, 76f
Windowed markets, 492–493
Word line pitch scaling, in DINOR array, 111–112
Word lines
level of, in NAND array, 105f
in read path decoding, 264–266
Write disturbs, in NOR-EEPROM, 12–13
Write enable pin, 251
Write functions
CHE *vs.* FN tunneling in, 365
control logic command write in, 339, 340f
in memory NOR, 17–24
Write protect pin, 251

Y
Yield
error correction on, 462
process, 350–352
redundancy on, 462–464
for wafer, 353–354

Z
Zener tunneling, 171

APPLE INC.
EXHIBIT 1008 - PAGE 0544