

NONVOLATILE MEMORY TECHNOLOGIES WITH EMPHASIS ON FLASH



IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Mohamed E. El-Hawary, *Editor in Chief*

R. Abari	T. Chen	R. J. Herrick
S. Basu	T. G. Croda	S. V. Kartalopoulos
A. Chatterjee	S. Farshchi	M. S. Newman
	B. M. Hammerli	

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*
Catherine Faduska, *Senior Acquisitions Editor*
Jeanne Audino, *Project Editor*

IEEE Components, Packaging, and Manufacturing Technology Society, Sponsor
CPMT Liaison to IEEE Press, Joe E. Brewer

IEEE Electron Devices Society, Sponsor
EDS Liaison to IEEE Press, Joe E. Brewer

Technical Reviewers
R. Jacob Baker, *Boise State University*
Ashok K. Sharma, *NASA/Goddard Space Flight Center*

NONVOLATILE MEMORY TECHNOLOGIES WITH EMPHASIS ON FLASH

A Comprehensive Guide
to Understanding and
Using NVM Devices

Edited by

Joe E. Brewer

Manzur Gill

IEEE Press Series on Microelectronic Systems
Stuart K. Tewksbury and Joe E. Brewer, *Series Editors*

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

IEEE Electron Devices Society, *Sponsor*



WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2008 by the Institute of Electrical and Electronics Engineers, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. All rights reserved.
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability of fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com

Library of Congress Cataloging-in-Publication Data is available.

ISBN 978-0471-77002-2

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

acteristics that define the ultimate retention potential of the approach. Natural decay tracking points to very large retention on the order of thousands or millions of years. Natural decay is so slow as to not be a factor in determining practical retention specifications. At the present time a typical unpowered retention specification is 10 years.

Defects associated with materials, details of device geometry, or aspects of circuit design can impact retention. Each of these three potential problem areas can result in the addition or removal of charge to/from the floating gate. Gate insulator or interpoly insulator defects are typical causes of retention degradation. Phenomena associated with ionic contamination or traps can also be contributing factors. Management of these issues is a function of the general state of the integrated circuit reliability arts and of the specific practices and equipment of given manufacturers.

1.3.4 Endurance

Achievement of nonvolatility depends on exploitation of some natural phenomena peculiar to a given technology approach. In most nonvolatile technologies the normal processes employed to write and/or read cells will result in stresses that eventually degrade the properties of the memory or disturb the contents of the memory. *Endurance* is the term used to describe the ability of a device to withstand these stresses, and it is quantified as a minimum number of erase–write cycles or write–read cycles that the chip can be expected to survive. For quite a number of years the industry has used 100,000 cycles as the minimum competitive endurance requirement.

Knowledge of the reliability physics of the memory technology is essential in order to develop meaningful endurance specifications. The endurance capabilities of a device are a function of both the intrinsic properties of the technology and the quality control of the production line. The impact of cycling stress on retention is a key aspect of endurance, and window closure or shift is a concern. The *window* for a Flash memory is the difference between the erased state threshold voltage and the programmed state threshold voltage. This is the difference that must be reliably detected by a readout sense amplifier.

Both retention and endurance pose difficult test and verification quandaries. Economics and the limits of the human life span make direct observation of retention periods of decades impractical. The process of testing endurance by cycling each part implies that the testing would consume much of the useful life time of the product.

1.4 FLASH MEMORY AND FLASH CELL VARIATIONS

Flash has been used as the example technology throughout the discussions above where a simple stacked gate transistor was used to illustrate several points. In practice the design of Flash cells has several flavors each of which has advantages and disadvantages. Figure 1.11 groups some of the existing Flash varieties by addressing method and by program and erase technique (see Figure 5 in Bez et al. [2]). As mentioned earlier, the common ground NOR used for both code and data storage, and the NAND used for mass storage, presently dominate the market.

Flash devices, cycles are counted on a block basis and erases on one block does not affect the reliability or performance of the others. The cycle count is relevant because the device's performance specifications, such as block erase time or byte program time, are only guaranteed to hold up to the maximum erase cycling specification. This is because erasing a Flash block puts wear on the oxide layer insulating the floating gate, which stores the information in the chip. A widely held misunderstanding of the cycling spec is that the device will fail or cease operating once the max cycling spec has been reached. This is not true; in most cases, the device will continue operating well after the max cycling spec has been exceeded. The spec signifies the amount of erases for which the manufacturer can guarantee the device's performance specs.

Erase cycling specifications have improved dramatically since the introduction of Flash. The first bulk-erase Flash devices were rated for 1000 cycles, but 10,000-cycle specs were soon commonplace. The addition of embedded write state machines to Flash devices brought tighter control of the erase process and made 100,000 cycles possible. Today, typical devices are rated for 100,000 to 1,000,000 cycles, but improved media management techniques (see Section 2.4.1.7) have decreased the number of erases used in typical data storage processes, so 100,000 cycles is generally enough for most applications.

SUSPENDS. One of the first issues with executing system code directly from Flash and expecting to write to that same Flash device is the *read while write* problem. This problem refers to the fact that a standard single-partition Flash device cannot be read while a write operation (program or erase) is in progress. Since the internal write state machine is manipulating the internal control signals and voltages to perform the write operation on the cells of the Flash array, those same Flash cells cannot be read at the same time. Because the system is usually architected such that CPU instructions are fetched from the Flash, a catch-22 situation is created as follows: Where does the CPU obtain the instructions it needs to control the Flash write operation when those same instructions will become unavailable as soon as the Flash write operation begins? This problem is magnified in a code+data application since writes of data to the Flash occur much more often than code modifications.

The standard method for dealing with the read-while-write issue is to copy the minimal code needed to write to the Flash into RAM and to execute from RAM during the write operation. However, the hole in this method is the inability to execute additional code from the Flash in the case that an interrupt occurs during the Flash write operation. To alleviate this constraint, write suspends were added to the Flash write state machine.

Write suspends allow the CPU to interrupt an in-progress program or erase operation, read from the Flash array, then resume the suspended operation. Recent Flash devices support erase suspend to read, program suspend to read, erase suspend to program, and even the nested erase suspend to program suspend to read. When executing the Flash write code from RAM, an interrupt handler can poll for interrupts and suspend the write operation to allow code execution in order to service an interrupt. This method of interrupt handling is sometimes called *software read-while-write*.

Suspends incur some performance penalties in the area of suspend latency and write performance. Suspend latency is the time the write state machine takes to

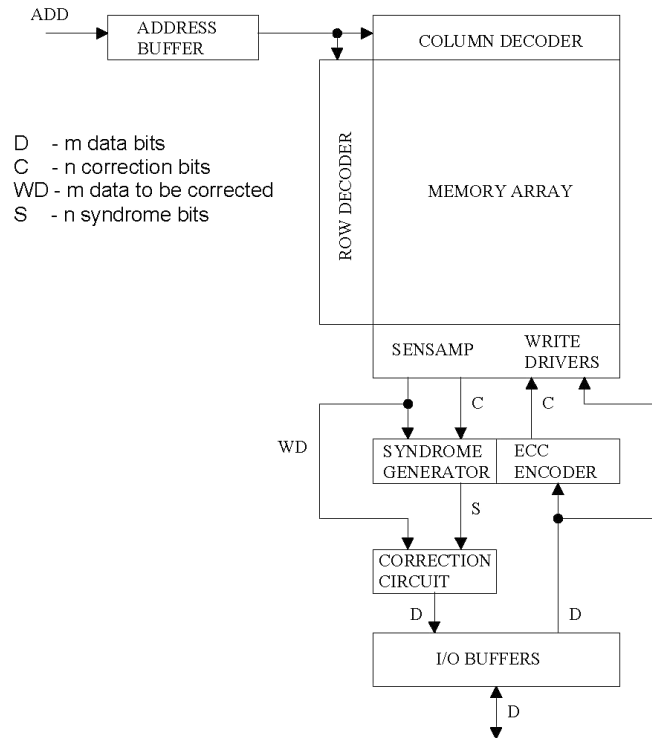


Figure 3.20. On-chip ECC for Flash memory.

3.5.1 On-Chip ECC and Endurance/Retention in Flash Memories

Error correction coding can be used in Flash memories to increase endurance and improve retention. After a certain number of erase/program operations a Flash memory can have defective cells due to oxide degradation or charge trapping. As soon as this happens, and provided that the ECC circuit is able to repair the amount of failing bits, the repair will be automatically performed. The apparent endurance of the chip is thus extended to a greater number of cycling operations.

After a certain lifetime a Flash memory can have a degradation of the cell thresholds due to charge loss or gain and, therefore, can undergo hard or random errors. On-chip ECC can correct these errors as soon as they happen in the application. The apparent data retention of the Flash memory is thus improved.

Redundancy is not useful to fix errors due to endurance or data retention because it can only be set up for fixed locations one time at the factory.

Flash memory ECC can be linked to circuitry to facilitate the performance of tests to screen chips that are weak for endurance [54] or retention. A Flash memory can in fact be provided with an internal address counter and an *internal repairs*

1. An embedded Flash memory is designed as a module that can readily be integrated into the host logic device. A stand-alone Flash memory is a self-contained chip that has I/O pads or pins to interface with other logic devices.
2. A Flash module for embedded applications must meet and be optimized for special requirements on array size, memory configuration, speed, and particular bus protocol for the host logic device and is considered a custom or semicustom design. Stand-alone Flash memory products typically have a common set of features and specs.
3. A Flash memory module designed for embedded applications must provide special test features, such as scan path or joint test action group (JTAG) interface capabilities called for by the host logic device. A stand-alone Flash memory may not have these requirements.
4. A Flash memory module for embedded applications is typically designed to support extensive reuse of intellectual property (IP) for multiple product families and multiple products within a product family. For this reason, a Flash memory module must be designed with a built-in memory generator or compiler to provide ease for designing new derivatives. A stand-alone Flash chip is typically designed and optimized for one specific array configuration.
5. Design of an embedded Flash memory can be rather logic intensive due to the required memory control block (MCB) and bus interface unit (BIU). For this reason, behavioral model, synthesis, and automated place-and-route design approaches are employed to shorten the design cycle time. A stand-alone Flash memory on the other hand may rely extensively on custom design to minimize silicon area and maximize speed.
6. The embedded Flash memory module typically shares the same tester with the host logic device, which costs more than a dedicated memory tester for stand-alone Flash memory because of high speed and high pin-count requirements for the host logic device. For this reason design for test time reduction is more critical for the embedded Flash memory than for the stand-alone Flash memory.
7. An embedded Flash module core once designed is to be reused by many different design groups from different companies, with varied degrees of knowledge and expertise in Flash memory technology and, therefore, good portability, good documentation, and ease of use are critical to the Flash memory core design. For the stand-alone Flash memory, there is typically one design for a given generation of technology and little reuse, if any.
8. An embedded Flash memory has less direct accessibility, observability, and controllability than stand-alone Flash memory due to lack of accessible I/O pins. Also, supply noise and substrate operating temperature may increase for an embedded Flash memory product due to integration of other modules on the same chip.
9. In order to serve as wide a range of product families as possible, the embedded Flash core (FMC) must be designed to meet the highest performance and density required among the product families to be served. Also, performance margins required for the embedded Flash memory have to be

higher than that for the stand-alone Flash memory because, unlike the stand-alone flash memory that can be specified for multiple speed grades, there is typically only one speed grade for the host logic device.

10. An embedded Flash memory module may involve special requirements of very high speed operation, very low voltage operation, or ease of design and manufacturing not available for a stand-alone Flash memory. For this reason, selection of the Flash memory cell, Flash memory process, and Flash array architecture for embedded application can be different from the stand-alone Flash memory and are optimized for the particular application space.

9.5.2 Flash Module Design for Embedded Applications

As previously stated, a Flash memory module for embedded applications consists of a Flash memory core (FMC) and a bus interface unit (BIU). The FMC comprises two key parts, a Flash memory array (FMA) and a memory control block (MCB). The FMA contains all the basic array elements that include cells, decoders, sense amplifiers, and program data buffers. The MCB contains all the essential control logic, register sets, state machines, high-voltage charge pumps, voltage regulators, level shifters, voltage switches, and I/O buffers required to support proper FMA operations. A simplified block diagram of the Flash module is shown in Figure 9.13, and Figure 9.14 shows a simplified FMA.

Flash memory core operations include array program, array erase, program/erase verify read, regular read, and any special test operations for device characterization or quality/reliability enhancement. The BIU, as the name implies, provides all interface needs between the Flash memory core and the host logic device, or specifically the bus of the host logic device such as MCU or DSP, to which the Flash memory module is connected. FMC is the essential portion of an embedded Flash memory module and contains all basic elements required for a stand-alone Flash memory. Figure 9.15 is a block diagram for the FMC showing key circuit blocks required for the MCB. The FMC is typically designed to be portable, reusable, and easily reconfigurable for other embedded Flash memory applications.

While the FMC can be a common design for many applications, the BIU is typically application or product family specific and must be designed to meet

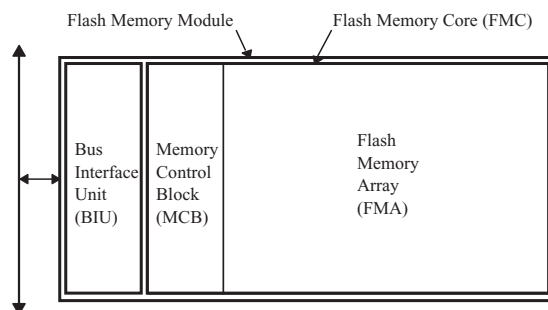


Figure 9.13. Embedded Flash memory module consists of bus interface unit (BIU) and Flash memory core (FMC) that contains Flash memory array (FMA) and memory control block (BCB).

floating gate. For single-bit/cell devices, the transistor either has little charge (<5000 electrons) on the floating gate and thus stores a 1 or it has a lot of charge (>30,000 electrons) on the floating gate and thus stores a 0. When the memory cell is read, the presence or absence of charge is determined by sensing the change in the behavior of the memory transistor due to the stored charge. The stored charge is manifested as a change in the threshold voltage of the memory cell transistor. Figure 12.5 illustrates the threshold voltage distributions for a half-million cell ($\frac{1}{2}$ -Mc) array block. After erasure or programming, the threshold voltage of every memory cell transistor in the $\frac{1}{2}$ -Mc block is measured, and a histogram of the results is presented. Erased cells (data 1) have threshold voltages less than 3.1 V, while programmed cells (data 0) have threshold voltages greater than 5 V.

The charge storage ability of the Flash memory cell is a key to the storage of multiple bits in a single cell. The Flash cell is an analog storage device not a digital storage device. It stores charge (quantized at a single electron) not bits. By using a controlled programming technique, it is possible to place a precise amount of charge on the floating gate. If charge can be accurately placed to one of four charge states (or ranges), then the cell can be said to store 2 bits. Each of the four charge states is associated with a 2-bit data pattern. Figure 12.6 illustrates the threshold voltage distributions for a $\frac{1}{2}$ -Mc block for 2-bit-per-cell storage. After erasure or precise

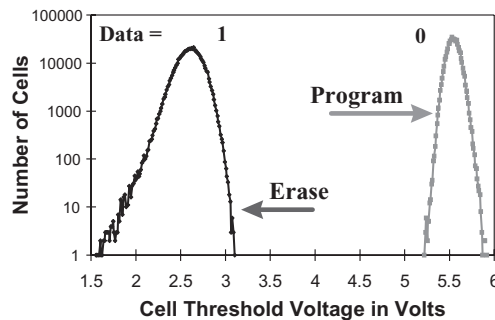


Figure 12.5. Single-bit/cell array threshold voltage histogram.

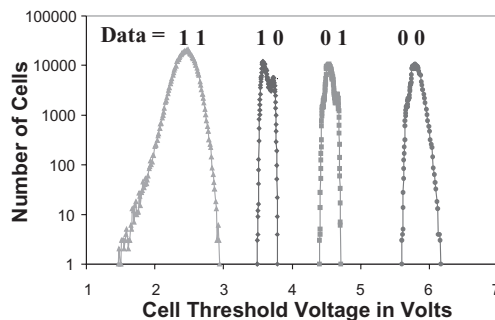


Figure 12.6. Two-bit/cell array threshold voltage histogram.

programming to one of three program states, the threshold of each of the $\frac{1}{2}$ -Mc blocks is measured and plotted as a histogram. Note the precise control of the center two states, each approximately 0.3 V (or 3000) electrons in width.

Higher bit/cell densities are possible by even more precise charge placement control. Three bits per cell requires eight distinct charge states; 4 bits per cell requires 16 distinct charge states. In general, the number of states required is equal to 2^N where N is the desired number of bits.

The ability to precisely place charge on the floating gate and at some later time sense the amount of charge that was stored has required substantial innovations and extensive characterization and understanding of cell device physics, memory design, and memory test. The aspects are discussed later in this chapter.

12.3.2 Evolution of MLC Memory Technology Development

This section will outline the development of the Intel StrataFlash memory technology from conception in 1992 to production in 1997, highlighting the key innovations along the way. The 64-Mbit product initially introduced in 1997 differs markedly from the 1992 view of what a 2-bit/cell product might look like. Today, MLC products have become the mainstream Flash memory technology. The learning that has occurred over the years has enabled the development of a 2-bit/cell memory device that functionally looks identical to a 1-bit/cell device, far exceeding the capability that was considered possible when development began.

12.3.3 Multilevel Cell Concept

Storage of analog data in a floating-gate memory device is not a new concept. It was suggested as early as 1971 for erasable programmable ROM (EPROM) devices [1] and was implemented on E²PROM devices for use in neural networks, voice recorders, and toys as early as 1982. These analog storage applications can tolerate a high error rate and thus do not place stringent requirements on the memory reliability or accuracy. Neural networks are, by their nature, fault tolerant. Voice storage and simple talking toys can tolerate a few lost bits without any audible impact. These high error rate lossy memories are generally not usable for mainstream digital storage and thus have had limited acceptance. The goal of the MLC program was to produce a 2-bit/cell digital storage technology capable of penetrating the larger nonvolatile memory market, enabling the growth of new digital Flash memory applications.

12.3.3.1 The 1992 View of MLC. In the early 1990s, Flash memory was considered as a potential replacement for hard disks at lower densities for applications that require small, rugged, and low-power storage. One of the main issues for use of Flash in this application was the high cost of the Flash memory as compared to that of magnetic storage. A lower cost Flash memory was required. The hard disk requirements are much relaxed over silicon memory due to the inclusion of error correction in the hard disk subsystem, the block transfer of data (no byte access), and the relative low read performance. Multilevel cell technology appeared to be an ideal solution for the solid-state disk, addressing the lower cost through 2-bit-per-cell (and later 3 or 4 bits/cell) technology. The use of error correction and the large block transfer of data in the solid-state disk would address any reliability

issues with multilevel storage. The Intel MLC program was thus started with a goal of a high-density, low-cost, solid-state disk.

The basic techniques for accurate charge placement and sensing were developed in the lab and implemented into a 32-Mbit silicon test chip. During this time frame, the three major challenges for multibit storage were identified:

- *Precise Charge Placement:* The Flash memory cell programming must be very accurately controlled, requiring a detailed understanding of the physics of programming as well as the control and timing of the voltages applied to the cell.
- *Precise Charge Sensing:* The read operation of an MLC memory is basically an analog to digital conversion of the analog charge stored in the memory cell to digital data, a concept new to memory devices.
- *Stable Charge Storage:* Meeting the data retention goals would require the stored charge to be stable with a leakage rate of less than one electron per day.

The 32-Mbit test chip clearly demonstrated the ability to store multiple bits in a single cell. Based on the functionality of this device, the MLC technology was announced in 1994.

12.3.3.2 The First MLC “Product.” With the knowledge gained from the 32-Mbit test chip, the first attempt at a 2-bit/cell storage product was started. This device was aimed at the solid-state disk goal. The solid-state disk system would include error correction and would generate nonstandard voltages to interface to the 2-bit/cell memory device. A special dc-to-dc (direct current) voltage converter was commissioned that would generate $12\text{ V} \pm 1\%$ and $5.5 \pm 1\%$. The MLC part required these precise supply voltages to perform the accurate program and read operations. Also designed to be integrated with the other control logic of the solid-state disk was an error corrector. A paper based on this 32-Mbit MLC memory was presented at the prestigious International Solid State Circuits Conference (ISSCC) in 1995 [2], winning the best paper of the conference award. The 32-Mbit chip became the workhorse for the MLC technology development effort, demonstrating the ability of MLC to meet stringent reliability requirements and to produce yield equivalent to single bits per cell Flash memories. It was also used to develop the MLC testing and to debug the manufacturing process for test and packaging.

12.3.3.3 Question of Reliability. The primary concern for MLC was the reliability of the storage of the multiple charge states. Charge states would be separated by a few thousand electrons in an MLC device, and a loss of one electron per day from the floating gate could result in a bit error after 10 years of storage. To understand the detailed physics of charge storage, a large experiment was started to monitor the charge storage behavior of 200 billion cells (2×10^{11} cells). This massive experiment could resolve changes in the stored charge of as small as 100 electrons on all of the cells under evaluation. The rate of charge loss was accelerated through the use of elevated temperatures. The knowledge gained and models developed based on this experiment have resulted in changes to the design of both the product and the process, allowing removal of the error correction requirement for 2 bits per cells [3]. This data fundamentally changed the direction of the multibit storage program.

12.3.3.4 Removing the Constraints. Toward the end of 1995, the MLC project had grown from a small research effort to a full-blown program. Almost 2 years worth of reliability data was showing excellent performance, indicating that the error corrector was not required. The 32-Mbit device had demonstrated the viability of the circuit techniques and the device physics used for the precision program and read operations. Moreover, the yield was looking excellent, and the manufacturing issues were understood. Test circuits had demonstrated the ability to provide the required voltages and voltage regulation on the memory chip, eliminating the need for the external dc-to-dc converter. It became clear that the project could accomplish much more than the initial vision of a solid-state disk. The team believed that it was possible to remove the two major requirements initially envisioned for MLC: error correction and precision external power supplies. The solid-state disk market, while developing, had not reached the desired volume levels. The decision was made to not take the 32-Mbit device to production and focus on the design of an MLC 2-bit/cell part with functionality substantially equivalent to the standard 1-bit/cell products.

12.3.3.5 The 1997 View of MLC. The first 2-bit/cell Intel StrataFlash memory device was introduced in September of 1997, a 64-Mbit device. This device has functionality that is largely equivalent to the standard 1-bit/cell Flash products. A highlight comparison of the Intel StrataFlash memory features to an Intel 16-Mbit single-bit/cell product is shown in Table 12.1.

Read performance was in line with expectations for memories of 32- and 64-Mbit densities with about a 20% increase in read access time for a doubling of memory density. Two bits/cell doubles the erase block size as compared to 1 bit/cell since each cell now stores twice as much data. The power supply was maintained at the 5 V industry standard used at that time. The 2 bits/cell write performance was maintained equivalent to 1 bit/cell, even with the more complex (and slower) precision write algorithm, through the use of an 8-byte write buffer and a higher write bandwidth into the array. The 10,000 erase/write endurance specification was more than acceptable for virtually all Flash applications and easily justified by the reduced cost.

TABLE 12.1. Comparison of 1-Bit/Cell and 2-Bit/Cell Product Features

	1-Bit/CellFlash Memory	Intel StrataFlash 2-Bit/Cell Memory	
Density	16Mbits	32Mbits	64Mbits
Read speed	100ns	120ns	150ns
Block size	64kbytes	128kbytes	
Architecture	$\times 8$	$\times 8 / \times 16$	
V_{cc} power supply ($\pm 10\%$)	5 V	5 V	
V_{pp} (program/erase voltage)	5 V or 12 V	5 V	
Effective write speed	6 μ s/byte	6 μ s/byte	
I_{ccr} (read current)	35 mA	55 mA	
$I_{ppw} + I_{ccw}$ (write current)	75 mA	90 mA	
Endurance	100,000 cycles	10,000 cycles	
Operating temperature	Extended	Commercial	

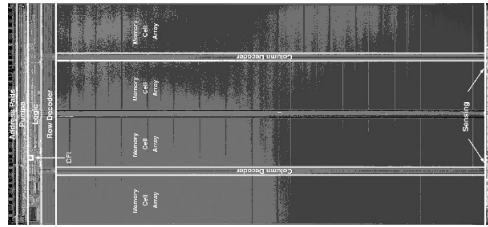


Figure 12.7. Intel StrataFlash 64-Mbit memory.

The 64-Mbit device integrated all of the knowledge gained from the two previous test vehicles and advanced beyond them with the introduction of precision internal voltage regulation and internal test capability. The 64-Mbit 2 bit/cell Intel StrataFlash memory was less than 5% larger than the 32-Mbit 1 bit/cell device on the 0.4 μ m ETOX Flash memory process, delivering on the promise of 2 $\times$ the bits in 1 $\times$ the space and setting a new cost paradigm for Flash memory devices. A photomicrograph of the 64-Mbit Intel StrataFlash memory is shown in Figure 12.7.

12.4 VIEW OF MLC TODAY

Today, further advances in MLC capabilities have resulted in the functionality of MLC products to be indistinguishable from their single-bit-per-cell counterparts. Random byte read access times below 90nS, V_{cc} of 1.8V, and flexible read-while-write features on state-of-the-art lithography are the mainstream for both products [4].

12.4.1 Multilevel Cell Key Features

The concept of MLC is ideally suited to the Flash memory cell. The cell operation is governed by electron charge storage on an electrically isolated floating gate. The amount of charge stored modulates the Flash cell's transistor characteristic. MLC requires three basic elements: (1) accurate control of the amount of charge stored, or placed, on the floating gate such that multiple charge levels, or multiple bits, can be stored within each cell, an operation called placement; (2) accurate measurement of the transistor characteristics to determine which charge level, or data bit, is stored, an operation called sensing; and (3) accurate charge storage, such that the charge level, or data bit, remains intact over time, an operation called retention. These elements are achieved by exploiting stable device operation regions and by the direct cell access of the ETOX Flash memory array.

12.4.2 Flash Cell Structure and Operation

An explanation of MLC first requires a review of the Flash memory cell. Key points relevant to MLC operation are covered here. The ETOX Flash memory cell and products [5] have a long manufacturing history, having evolved in the late 1980s from EPROMs, which had been an industry standard from the early 1970s.

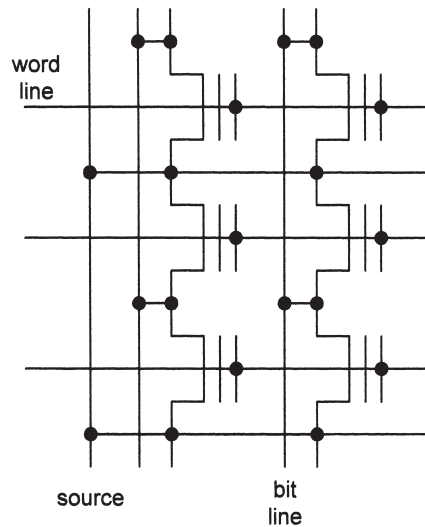


Figure 13.104. NOVORAM/FGRAM architecture.

13.8.4 NOVORAM/FGRAM Summary

Because of its high density (at comparable speed) NOVORAM/FGRAM may present a serious challenge for DRAM and SRAM even at the present technological level, since its nonvolatility is very important for many electronics applications. However, the most important feature of such memories is high scalability, limited mostly by the exponentially growing sensitivity of readout MOSFETs to fabrication spreads. The use of advanced double-gate SOI MOSFETs may push this limit to physical gate length ~ 5 nm. Apparently, an extension of this limit is possible using individual compensation of the threshold gate voltage of each readout MOSFET using an additional floating gate [Fig. 13.103(b)].

13.9 PHASE CHANGE MEMORIES

Joe E. Brewer, Greg Atwood, and Roberto Bez

13.9.1 Introduction

Phase change memory (PCM) technology is currently (2006) the subject of intense research and development by multiple companies and is a promising future alternative to Flash. In terms of cell size, die size, and cost it is approximately on a par with

Flash. It offers fast random read performance ~ 50 ns, fast write performance ~ 100 ns, and good data retention >10 years. PCM is a “direct write” technology. There is no need to erase a block of data and then program it. The technology can be made to fit into a CMOS fabrication process. It exceeds Flash in bit granularity and in endurance. In the long term it appears to have good scalability, and it has multilevel potential.

Phase change memory is certainly not a new concept. For example, the September 28, 1970, issue of *Electronics Magazine* [184] contained the die picture shown in Figure 13.105. This was a 1970 chip with a 256-bit capacity that measured 122×131 mils. Reset was accomplished by <200 mA at <25 V for 5μ s. Set required 5 mA at ~ 25 V for 10 ms. The read mode current was 2.5 mA with a supply <5 V. The history of the technology can be traced back to multiple patents filed by Stanford R. Ovshinsky beginning in the 1960s.

This primitive circa 1970 chip did not prove to be producible or competitive with the emerging floating-gate and MNOS technologies. Soon interest waned and the approach was abandoned. Later phase change technology experienced a rebirth in a new context—optical memory. Chalcogenide films were found to offer advantages when used as the memory medium for rotating disks. The rewritable phase change compact disk (CD-RW), the rewritable digital versatile disk (DVD-RW), the random-access memory digital versatile disk (DVD-RAM), and the blue-ray disk are commercially successful examples.

The optical disk in many ways is a simpler platform for investigation of the phase change phenomena. The active bit storage regions are not individually contacted. It was not necessary to first develop compatible electrode contacts and devise suitable heating elements. The medium is programmed by a high-power laser beam and interrogated by a lower power laser combined with an optical sensor. The primary issue of fast reliable and reversible phase conversion could be examined without the distraction of dealing with these other issues. The disk business provided

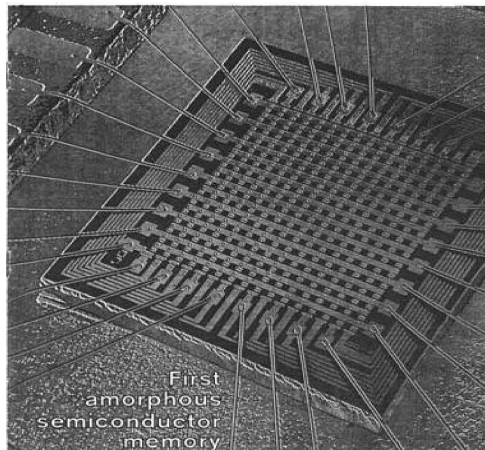


Figure 13.105. Photo titled “First Amorphous Semiconductor Memory.” (From [184].)

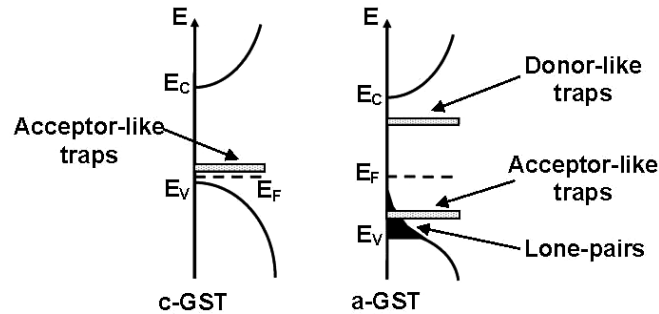


Figure 13.115. Band diagrams of c-GST and a-GST proposed by [194].

Comparatively speaking, the c-GST state offers relatively high mobility and high carrier concentration while the a-GST state has relatively low mobility and low carrier concentration. In the c-GST state lattice defects are present in concentrations as high as 15% [189]. The defects act as shallow acceptorlike traps causing the material to be a defect semiconductor that exhibits degenerate P-type conductivity with a bandlike mobility [194]. The optical bandgap of the c-GST is 0.5 eV [189].

When the c-GST is melted, the lattice and the lattice defects disappear, and the resulting structural disorder produces different kinds of electronic states. The most numerous are spatially localized C_2^0 lone-pair valence band tail states that cause a very low trap-limited hole mobility. In addition there are valence alternation pair (VAP) states believed to be caused by special defect bonding configurations between the nonbonding orbitals on the chalcogen atoms [189]. These states are the donorlike and acceptorlike traps shown in Figure 13.115. The donorlike traps are described as being C_3^+ states, and the acceptorlike traps are C_1^- states [194].

Pirovano et al. [194] describes the conduction processes for a-GST in the following way. When the voltage is low, the conduction is ohmic. As the electric field increases, impact ionization takes place, and the current rises exponentially due to secondary hole generation. Initially, most of the secondary electrons get trapped in the donorlike traps. Further increase in the voltage begins to fill up the traps and the free-electron density increases. The electron quasi-Fermi level is forced to move closer to E_C . Just beyond V_{th} impact ionization dominates over carrier recombination, the traps are filled and a voltage snap-back occurs. Now higher current can flow at a lower voltage. Impact ionization still takes place, but at a reduced multiplication rate. A much higher free-carrier density now sustains the current.

In Flash and many other memory technologies it is necessary to perform two steps to establish the state of the individual cells. Typically, an erase operation is followed by a program operation. In contrast, PCM is a direct write technology. No matter what the original state of the cell is, the final state of the cell can be established by the magnitude of current passed through the cell. For PCM it is appropriate to speak of a “read” operation and a “write” operation, but no doubt the legacy of Flash will lead to maintaining the terminology of “programming” as opposed to “write.”

Figure 13.116 provides an alternative view of the writing process. Here resistance for a particular cell is plotted against the programming current. The procedure