# Scanning the Issue

## Special Issue on Flash Memory Technology

The objective of this special issue is to present a comprehensive overview of one of the most innovative and complex types of high-tech, nonvolatile memories in use today: Flash memories.

In today's world, computers are everywhere, and the electronic memory that we often associate most with computers is the so-called dynamic RAM (DRAM); it is a technically complex memory, sized in ultracompact packages, but it is "volatile," i.e., it loses stored information when power is removed. It is obvious that human beings need alternative electronic memories to use in electronic devices. After all, human memory (both the one embedded in our DNA and the more dynamic one in our brain) is nonvolatile. We do not forget—usually—and this is one of the strengths of the Flash memories discussed in this issue.

Electronics continue to evolve toward superior functionality and greater miniaturization. The integration of functions with basic components has been a reality for decades, but present and future availability of more and more sophisticated technologies will lead to further integration of single-silicon chips yielding more complex electronic systems; these systems, just like present ones, would not exist without nonvolatile memories. And there are several types of nonvolatile memories.

Nonvolatile memories can be grouped under different families: ROM, electrically programmable ROM (EPROM), one-time programmable (OTP), electrically erasable programmable ROM (EEPROM) and Flash. This special issue focuses on Flash; this type is considered by many to be the most interesting, glamorous, economically attractive, and strategic family of them all, but also the most complex type available. The unique peculiarity of Flash memory is that such nonvolatility and reprogrammability, after electrical erase, are associated with very high performance and low cost. Any electronic system which embeds Flash memories is more flexible—code and data can be modified by the system itself autonomously and on the fly. In the electronic world, Flash memories represent an "enabling technology."

### I. History and Complexity

Since their introduction in the early 1990s, these products have experienced a continuous evolution. From the first simple Flash memories, used to emulate EPROM devices, many changes have occurred: circuits have grown in complexity and performances have improved in terms of speed, lower power consumption, lower supply voltage requirements, reliability, and, last but not least, smaller size. A static situation is far from being achieved at this time. Extreme flexibility of design application is the base for continuous new evolution, sometimes standardized, sometimes tailored to specific applications. Therefore, this is also an enabling technology in the sense that future limits are beyond our current expectations and limited only by our imagination.

Some say that Flash memories are not just memories, they are "complex systems on silicon"; Flash memories are challenging to design, because a wide range of knowledge in electronics is required (both digital and analog), and they are difficult to manufacture. "Muscles" of technology are not enough: physics, chemistry, and other fields must be integrated; and conditions must be carefully monitored and controlled in the manufacturing process. A minimum imperfection or a design marginality with respect to technology or process could destroy the nonvolatile features to a Flash memory.

From the early 1970s, when they were invented, Flash memories have led the widespread use of semiconductors. They have guided the technological development from the foundation of the technology of integrated circuits to current submicrometer production. They will still guide the technology for years to come.

### II. Market and Applications

Memories demand enormous investments in R&D, but they also reward with enormous potential market values. Flash memory markets (considered the most important market segment among nonvolatile memories) are expected to progress at a very fast pace, and to gain second place in the overall memory market. This is due to the optimization of cost/performance tradeoffs, and in particular to the inherent flexibility and versatility of this memory, which brings benefits in many applications. The leading application is in multimedia systems, which require memories that are increasingly larger in size, and demand ever-increasing performance characteristics. Telecommunications, computers, and automotive and consumer electronics are some

Digital Object Identifier 10.1109/JPROC.2003.811701

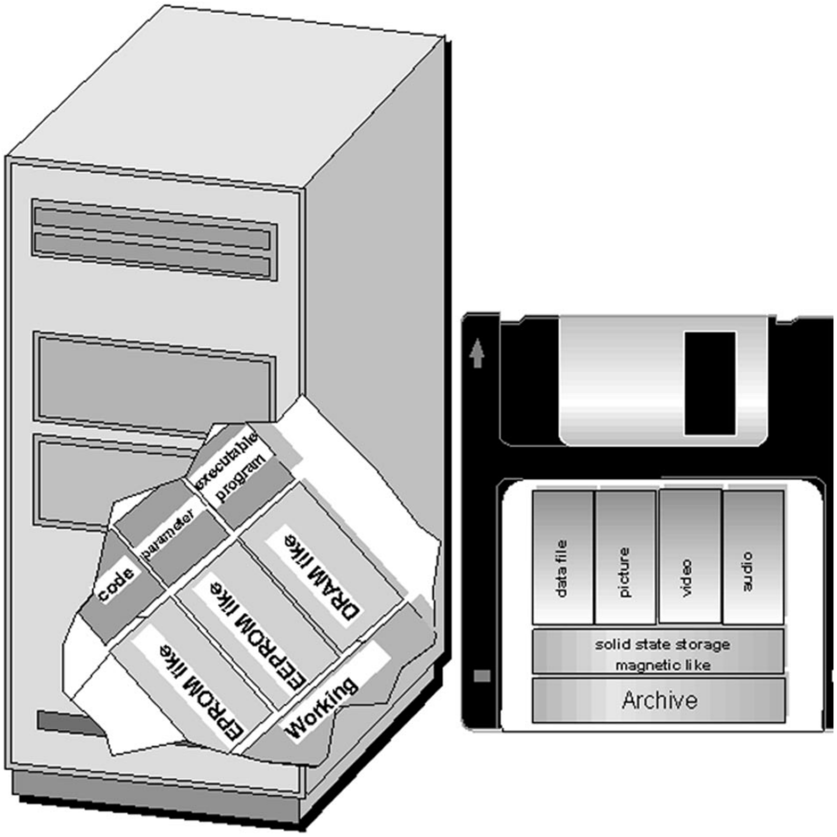

**Fig. 1.** Flash segmentation. Dimensions do not reflect relative market sizes.

additional areas where these memories facilitate numerous emerging new applications.

The use of Flash memories in electronic systems can be divided into two main categories: 1) memories which store all of the required information for system operation; and 2) memories which contain information not necessarily elaborated on by the system itself. The former (working memory) can be used to store microcontroller code, application parameters, and the program of a processor; they represent today the major share of usage of nonvolatile memories. While historically ROM, EPROM, and EEPROM have played the same role, today's Flash memories are increasingly being used for this application because of their flexibility.

The latter, archive memory, is currently used to store data, and it will be used more and more in the future to store voice, pictures, video, etc. This market segment could not be completely served by other first- or second-generation nonvolatile memories (i.e., ROM and EPROM), and it is now a dream coming true in Flash memory applications (see Fig. 1).

Flash memory flexibility and the possibility of tailoring to applications have caused since their introduction a plethora of differentiations technologically, architecturally and performance-wise.

## III. Flash Market Stream

In the field of "working memories," where nowadays NOR architecture is almost exclusively used, memory architecture is differentiated in segmentation and granularity, while for mass storage memories the winning architecture between NOR (eventually multilevel) and NAND has not emerged yet. The entire Flash memory market of the Flash can be divided into three large categories (see Table 1).

**Table 1**
Flash Market Stream

| | Key applications | Density | Key features | Comments |
|---|---|---|---|---|
| **STANDARD** | Auto, Hard Disk Drivers, PC Bios, Networking, Modems, Set Top Box … | From 1 Mb to 16 Mb | 5 V => 3 V<br>Access Time | "Eprom like"<br>NOR type |
| **MOBILE** | Cellular Phones<br>Pagers | Medium<br>4, 8 Mb up to 128 Mb | Low energy: 3 V => 1.8 V<br>Chip size<br>Package | Market driver<br>NOR type |
| **MASS STORAGE** | Digital Cameras, Voice recorders, MP3 players, PDA | High<br>From 128 Mb to 1 Gb | Very low $/bit<br>High program throughput | Bit integration |

1) Flash for portable applications, featuring very high performance and, above all, low voltage and low power consumption. This is the leading edge of technology, since emerging complex multimedia systems ask for larger memory size.

2) Flash for standard architecture. Most markets require EPROM-like memories; a unique standard is not mandatory, as for EPROM, but different standards are required by different applications. These differences are realized by adding/removing part of the control/command circuitry, varying memory matrix organization or supply voltage, or integrating more or less logic.
3) Flash for mass storage. These are the memories of the future, whose main features are highest density and low cost.

Moreover, Flash memory integration is becoming a reality in order to realize the so-called system on silicon.

It has been said that the Flash memory is the building block to obtain this further technological innovation. The level of integration varies as a function of the goals that companies want to pursue.

Undoubtedly, when a high-performance system does not require either very high density nonvolatile memory or very low cost, integration of logic circuits with nonvolatile memories will be the winning factor; therefore, it will be a winning factor for a company to be able to dominate all the technologies, from CMOS to power to nonvolatile memories.

On the other hand, large size and cost-effective memories might require efforts that are not compatible with total integration issues, even if partial circuit integration is applied. In any case, the previously mentioned flexibility of Flash memories will allow those companies that can master this complex technology to gain a competitive advantage.

## IV. Telecommunications, the Internet, and Beyond

Currently, the primary and most rapidly growing market is telecommunications, led by the explosion of mobile phones, where a Flash memory is used for both the codes for the protocols which implement functionality and man/machine interface, i.e., the interface between keyboard, display, and protocols. Memory size enormously varies as a function of languages and graphic requirements and, in general, of all the supported functions. Flash memories on a mobile phone can easily exceed 128 Mb.

The telecommunication world also includes other important users for high-density Flash memories, from modems to network routers, whose usage is growing following Internet expansion. Inside personal computers, the usual applications for nonvolatile memories (Flash, nowadays) are the Basic Input/Output System (BIOS) and Disk Drive Operating System; in the future, thanks to the widespread use of personal digital assistants (PDAs), applications will include Flash cards. Another important application for high-density Flash memories is the TV set-top box, whose usage is growing thanks to the expansion of both satellite and cable digital television; a big impulse will also come from virtual integration of the TV set with other appliances in the "house of tomorrow." Digital still cameras represent another important application.

Last but not least (indeed, this field was the first to apply Flash memories) is the automotive industry. Starting from applications related to engine control, Flash memory usage is expanding to the gearbox, the brakes, the global positioning system, the mobile phone, etc. The more electronic features introduced in the car, the more Flash memories (in their most sophisticated architecture) become fundamental in product architecture.

Flash memory is the demonstration, in the memory arena, of the pervasive use of new and varied applications of electronics in every aspect of our lives. Every new application scenario can utilize this flexible and powerful memory technology, either as a stand-alone component or integrated as the enabling factor of the whole silicon integration. We hope that the papers in this issue will bring this important and useful technology into better focus for the readers of this journal.

## V. Description of Papers

### Introduction to Flash Memory

*(Invited Paper)*

*Bez, Camerlenghi, Modelli, and Visconti*

The special issue on Flash memory technology begins with "Introduction to Flash Memory,"where Bez *et al.* discuss the fundamentals of the floating-gate MOS transistor. In NOR architecture, programs are based on channel hot electrons, while erase is performed using Fowler–Nordheim (FN) tunneling. Reliability issues are considered also: program disturbance and data retention, as well as endurance, are addressed from the single bit to a multilevel approach. After this introductory section, memory evolution and technology scaling trends are presented. Considering applications (EPROM replacement, code storage, data storage) and power supply (single or double), three main generations have been identified starting from 1990. Of course, memory evolution from one generation to the other has been realized thanks to technology: from local oxidation of silicon (LOCOS) isolation at 0.8 $\mu$m to shallow trench isolation (STI) at 0.13 $\mu$m, passing through the triple well technology, a lot of improvements have been made during the last decade.

### Survey on Flash Technology with Specific Attention to the Critical Process Parameters Related to the Manufacturing

*(Invited Paper)*

*Ginami, Canali, Fattori, Girardi, Scintu, Tarchini, and Tricarico*

In this paper, Ginami *et al.* discuss the Flash technology history. High-density memories and low cost per megabit are based on size reduction of cell and transistor dimensions. A survey on technology nodes from 1.2 to 0.13 $\mu$m is here presented. Lithography issues in terms of the photo resistive material (resist), equipment, and mask are addressed as well as wet and dry etching for resist removal. To reduce the pitch of the cell, the distance between contiguous active areas must be minimized; besides lithography, insulation techniques, from LOCOS to STI, have to be considered. Gate oxide as well as source and drain junctions are essential for transistor performances; tunnel oxide is fundamental for erased and programmed cells distributions. On-chip interconnections and insulation problems between different metal layers are analyzed in the last part of the paper.

## VI. Flash Memory Design Issues

### An Overview of Flash Architectural Developments

*(Invited Paper)*
*Campardo, Scotti, Scommegna, Pollara, and Silvagni*

This paper focuses on matrix architectures, which are some of the most complex items in a memory chip design: chip dimensions must be as small as possible to allow the required memory performances. During the last decade, program and erase mechanisms have been changed as well as market requests in term of sector size, speed, and supply voltage. As a consequence, innovative architectural solutions have been developed and are here presented. Then a brief excursus on “test modes” is presented: they are a way to enter inside the Flash in order to test the memory cells and their relationship with the internally generated voltages. Memory testing issues, such as test time reduction at electrical wafer sort (EWS) and final test (FT) against failure coverage, complete the paper.

### The Flash Memory Read Path: Building Blocks and Critical Aspects

*(Invited Paper)*
*Micheloni, Crippa, Sangalli, and Campardo*

In this paper, all the different blocks of the read path are considered, from input to output pads. Technology and power supply scaling as well as multilevel storage are the driving forces in the nonvolatile memory evolution. At every step the read operation becomes more critical as the distance between zero and one distributions is reducing. When an address changes, a circuit named Address Transition Detector (ATD) tells the memory that it is time to access its content. Row and column decoders bias the selected cells; the issue of generating an adequately high read voltage from the available $V_{\mathrm{DD}}$ is then addressed, and a “one shot” boost technique is discussed. Once biased, the cell current is ready to be sensed and converted in digital data. No doubt the sense amplifier represents the heart of the read path; design solutions from EPROM to multilevel Flash are presented, taking the $V_{\mathrm{DD}}$ range into account to ensure correct sensing in any condition. The last part of the paper is devoted to the multilevel approach: from 2 to 4 b/cell; also, state-of-the-art solutions are presented and discussed at the transistor level.

### High-Voltage Management in Single-Supply CHE NOR-Type Flash Memories

*(Invited Paper)*
*Motta, Ragone, Khouri, Torelli, and Micheloni*

This is the third paper in this design section. Flash memories contain a voltage generator system that produces the voltages greater than the power supply $V_{\mathrm{DD}}$ [high voltage (HV)]. Program, erase, and read HV paths are described considering hierarchical row and columns decoders. Mass storage memories have a large number of sectors with associated HV switches; a hierarchical biasing approach for area occupation reduction is presented. HVs are produced by on-chip charge pumps, which are characterized by a high-output voltage ripple. To get the necessary precision, a system of voltage regulators is developed and discussed with particular attention to multilevel memories. Finally, the standby management is considered. During standby mode, the device is put in a low-power condition, but it must be able to return to active mode with no penalty on access time. Of course, this is a challenging issue because all the voltage regulators must be turned off to reduce power consumption.

### An Overview of Logic Architectures Inside Flash Memory Devices

*(Invited Paper)*
*Silvagni, Fusillo, Ravasio, Picca, and Zanardi*

This paper by Silvagni *et al.* concludes the design issues section. Logic blocks are becoming more and more important inside the memory chip, in terms of complexity and area occupation. First of all, program and erase algorithms require a dedicated controller (it could be either a finite-state machine or a microcontroller) to handle all the different phases. Moreover, data storage memories impose the use of column, row, and sector redundancy for yield enhancement with the related control logic. It must also be considered that many applications require fast memory access time together with high read throughput. In synchronous burst mode, an external clock is provided to the device; after a first latency time, the memory is able to give out new data on every active clock edge. Pipelined architecture is essential for burst access, and dedicated control logic has to be designed taking into account skew problems between internally generated synchronous signals.

### Current Criticalities and Innovation Perspectives in Flash Memory Design Automation

*(Invited Paper)*
*Conci, Faldarini, Fumagalli, Girardi, Pesare, Tecli, and Zucchinali*

This paper deals with computer-aided engineering, i.e., tools and design methodologies developed for helping designers in getting project success in a reduced time. The memory market is very competitive, and time-to-market can be considered a key competitive advantage. On the other hand, memory designers have to handle systems that are gaining complexity every day; size and functionalities of both analog and digital circuitry have to increase in order to fit the requests coming from new applications. Planning decisions early in the design cycle (floor planning) is mandatory in a complex design in order to predict physical layout topology. Designers usually develop analog circuits using SPICE-like simulations; digital design is hardware description language oriented, gaining advantage from standard cell libraries and automatic place and route.

Once all the blocks (analog and digital) are ready, the final step of design flow is the full chip physical and electrical verification. Geometrical rules, together with the correspondence between schematic and layout, must be checked. The last part of the paper is devoted to package issue; in fact, high-frequency applications and multichip modules require accurate simulations of die/package/board interfaces.

## VII. Next Generation Multilevel Flash Memories

### Program Schemes for Multilevel Flash Memories

*(Invited Paper)*
*Grossi, Lanzoni, and Riccò*

Program issues in multilevel storage are discussed by Grossi *et al.* in this paper. Multilevel operation in Flash memories implies accurate control of charge transfer to the cell floating gate (FG) in order to get tight threshold voltage distributions. Program and verify (P&V) is the easiest algorithm to reach this goal: small program steps are performed, each followed by a read operation to decide whether or not to proceed with another step. P&V, however, is inherently time consuming, particularly when high accuracy requires a large number of steps, and a number of alternatives have been proposed in order to improve the program throughput. Also, in this context, the physical mechanisms used to transfer electrons to the cell FG can be considered as a design variable, since both schemes based on channel hot-electron injection and FN tunneling have been studied. The paper critically reviews the most significant program schemes presented in the literature and identifies application-specific competitive advantages.

### On-Chip Error Correcting Techniques for New-Generation Flash Memories

*(Invited Paper)*
*Gregori, Cabrini, Khouri, and Torelli*

The next paper by Gregori *et al.* discusses error correction. Data storage applications require high-density memories, but the progressive increase in single chip cells contributes to reliability decrease in terms of data retention, read disturbance, and so on. This is in contrast with the multilevel approach, where the distance between adjacent threshold voltage distributions is reduced. A way to overcome these problems is to use an error correction code (ECC) inside the memory chip. It is clear that among all the possible choices in terms of correcting codes, the less time-consuming ones must be selected. Access time is a key parameter for a NOR memory. Of course, ECC is also an area-consuming approach: "parity check" cells impact can reduce the area advantage of multilevel storage. The paper focuses on ECC for multilevel memories. The ECC design procedure is analyzed, including the choice of the best code to be used to assure the requested reliability. Coding/decoding networks design and optimization are also presented.

## VIII. Reliability and Low-Cost Alternatives

### Overerase Phenomena: An Insight Into Flash Memory Reliability

*(Invited Paper)*
*Chimenton, Pellati, and Olivo*

Reliability aspects in NOR Flash memories are discussed by Chimenton *et al.* in "Overerase Phenomena: An Insight Into Flash Memory Reliability." Among the different issues in the Flash reliability arena (data retention, program and read disturbs, etc.), the authors focus on cells that are erased faster than the others, named tail and fast cells. A very interesting point is the erratic nature of tail bits: during cycling, cells belonging to the normal part of the distribution can shift in the tail after an erase operation. Of course, this behavior imposes difficulties in screening these bits. Erasing schemes and identification algorithms of fast and tail bits are presented. Experimental results show the creation and evolution of the tail distribution, the characteristics of erratic threshold jumps during erasing, and the fast bits identification under constant voltage and constant field stresses.

### The HIMOS Flash Technology: The Alternative Solution for Low-Cost Embedded Memory

*(Invited Paper)*
*Van Houdt, Wellekens, and Haspeslagh*

The subject of this paper is the high-injection MOS (HIMOS) cell, which is made up of a basic split-gate cell plus a program gate capacitor; as a result, two external gates are available. High-efficiency program operation and low power consumption are achieved using source-side injection. The erase mechanism is based on FN tunneling. A limited number of masks (usually between two and five) are necessary over a digital CMOS process; this is why HIMOS can be considered a valid solution for embedded applications.

As it appears from the list of the papers included in this special issue, Flash memories cover a wide variety of topics, from pure design to technological aspects, passing through reliability issues. NOR architecture is the main subject in all the papers; of course, different matrix architectures should be analyzed together with many other aspects. We hope, however, that this overview will provide a solid basis to enter the nonvolatile world.

The Guest Editors would like to thank the reviewers who volunteered their time to provide feedback to authors. Special thanks go to J. Calder, Managing Editor of this journal, for his assistance, and to M. Guedj for the fruitful marketing discussions.

Giovanni Campardo, *Guest Editor*
STMicroelectronics
Agrate Brianza, Italy

Rino Micheloni, *Guest Editor*
STMicroelectronics
Agrate Brianza, Italy

**Giovanni Campardo** (Member, IEEE) was born in Bergamo, Italy, in 1958. He received the Laurea degree in nuclear engineering from the Politecnico di Milano, Milan, Italy, in 1984 and the Laurea degree in physics from the Università Statale di Milano, Milan, Italy, in 1997.

In 1984, he was at CISE, Milan, Italy, working in the field of lasers. From 1984 to 1988, he was a Project Leader in the VLSI division of SGS (now STMicroelectronics), Agrate Brianza, Italy, where he designed the family of EPROM nMOS devices (512, 256, 128, and 64 k) and a lookup-table-based EPROM digital filter in CMOS technology. From 1988 to 1992, he was an ASIC Designer at FIBER, Bergamo, Italy, realizing four devices. In 1992, he rejoined STMicroelectronics as a Project Leader, concentrating on Flash memory design for the microcontroller division. Here, he has realized a Flash + SRAM memory device for automotive applications and two embedded Flash memories (256 k and 1 M) for the ST10 microcontroller family. Since 1994, he has been responsible for Flash memory design in the Memory Division, STMicroelectronics, where he has realized two double-supply Flash memories (2 and 4 M) and the single-supply 8 M at 1.8 V. He was the Design Manager for the 64-M multilevel Flash project. Unitl late 2001, he was the Product Development Manager for the Mass Storage Flash Devices in the STMicroelectronics Flash Division, realizing the 128-M multilevel Flash and a test pattern to store more than 2 b/cell. He currently has the responsibility of building a team to develop 3-D integration for system-in-package solutions in the ST Wireless Flash Division. He has authored or coauthored more than 70 patents and several papers, and is coauthor of *Flash Memories* (Norwell, MA: Kluwer, 1999) and author of *Design of Nonvolatile Memories* (Milan, Italy: Franco Angeli, 2000). He was also Lecturer in the Electronic Laboratory course at the Università Statale di Milano.

**Rino Micheloni** (Member, IEEE) was born in San Marino in 1969. He received the Laurea degree (*cum laude*) in nuclear engineering from the Politecnico di Milano, Milan, Italy, in 1994.

In 1994, he was with Italtest, Settimo Milanese, Italy, working on industrial nondestructive testing reliability. In 1995 he joined the Memory Product Group, STMicroelectronics, Agrate Brianza, Italy, where he worked on a 8-Mb 3-V-only Flash memory, especially on the analog circuitry of the read path. He was the Project Leader of a 64-Mb four-level Flash memory; after that, he designed a 0.13-$\mu$m test chip exploring architectural solutions for Flash memories storing more than 2 b/cell. He is currently the Product Development Manager of the multilevel products in the Flash Storage Media Division. He is author or coauthor of more than 50 patents on nonvolatile memories and 20 papers on the same subject. He is coauthor of *Flash Memories* (Norwell, MA: Kluwer, 1999).