Introduction to Flash Memory

ROBERTO BEZ, EMILIO CAMERLENGHI, ALBERTO MODELLI, AND ANGELO VISCONTI

Invited Paper

The most relevant phenomenon of this past decade in the field of semiconductor memories has been the explosive growth of the Flash memory market, driven by cellular phones and other types of electronic portable equipment (palm top, mobile PC, mp3 audio player, digital camera, and so on). Moreover, in the coming years, portable systems will demand even more nonvolatile memories, either with high density and very high writing throughput for data storage application or with fast random access for code execution in place. The strong consolidated know-how (more than ten years of experience), the flexibility, and the cost make the Flash memory a largely utilized, well-consolidated, and mature technology for most of the nonvolatile memory applications. Today, Flash sales represent a considerable amount of the overall semiconductor market.

Although in the past different types of Flash cells and architectures have been proposed, today two of them can be considered as industry standard: the common ground NOR Flash, that due to its versatility is addressing both the code and data storage segments, and the NAND Flash, optimized for the data storage market.

This paper will mainly focus on the development of the NOR Flash memory technology, with the aim of describing both the basic functionality of the memory cell used so far and the main cell architecture consolidated today. The NOR cell is basically a floating-gate MOS transistor, programmed by channel hot electron and erased by Fowler–Nordheim tunneling. The main reliability issues, such as charge retention and endurance, will be discussed, together with the understanding of the basic physical mechanisms responsible. Most of these considerations are also valid for the NAND cell, since it is based on the same concept of floating-gate MOS transistor.

Furthermore, an insight into the multilevel approach, where two bits are stored in the same cell, will be presented. In fact, the exploitation of the multilevel approach at each technology node allows the increase of the memory efficiency, almost doubling the density at the same chip size, enlarging the application range, and reducing the cost per bit.

Finally, the NOR Flash cell scaling issues will be covered, pointing out the main challenges. The Flash cell scaling has been demonstrated to be really possible and to be able to follow the Moore's law down to the 130-nm technology generations. The technology development and the consolidated know-how is expected to sustain the scaling trend down to the 90- and 65-nm technology nodes as forecasted by the International Technology Roadmap of Semiconductors. One of the crucial issues to be solved to allow cell scaling below the 65-nm node is the tunnel oxide thickness reduction, as tunnel thinning is limited by intrinsic and extrinsic mechanisms.

Keywords*—*Flash evolution, Flash memory, Flash technology, floating-gate MOSFET, multilevel, nonvolatile memory, NOR cell, scaling.

Manuscript received July 1, 2002; revised January 5, 2003.

The authors are with the Central Research and Development Department, Non-Volatile Memory Process Development, STMicroelectronics, 20041 Agrate Brianza, Italy (e-mail: roberto.bez@st.com).

Digital Object Identifier 10.1109/JPROC.2003.811702

I. INTRODUCTION

The semiconductor market, for the long term, has been continuously increasing, even if with some valleys and peaks, and this growing trend is expected to continue in the coming years (see Fig. 1). A large amount of this market, about 20%, is given by the semiconductor memories, which are divided into the following two branches, both based on the complementary metal–oxide–semiconductor (CMOS) technology (see Fig. 2).

- The volatile memories, like SRAM or DRAM, that although very fast in writing and reading (SRAM) or very dense (DRAM), lose the data contents when the power supply is turned off.
- The nonvolatile memories, like EPROM, EEPROM, or Flash, that are able to balance the less-aggressive (with respect to SRAM and DRAM) programming and reading performances with nonvolatility, i.e., with the capability to keep the data content even without power supply.

Thanks to this characteristic, the nonvolatile memories offer the system a different opportunity and cover a wide range of applications, from consumer and automotive to computer and communication (see Fig. 3).

The different nonvolatile memory families can be qualitatively compared in terms of flexibility and cost (see Fig. 4). Flexibility means the possibility to be programmed and erased many times on the system with minimum granularity (whole chip, page, byte, bit); cost means process complexity and in particular silicon occupancy, i.e., density or, in simpler words, cell size. Considering the flexibility-cost plane, it turns out that Flash offers the best compromise between these two parameters, since they have the smallest cell size

0018-9219/03$17.00 © 2003 IEEE

Absolute Exhibit No. 1046
Page 001

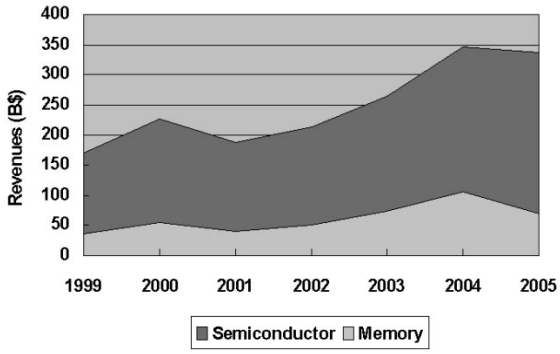

Fig. 1. Semiconductor market: revenues versus year. The bottom wave refers to the semiconductor memory amount.

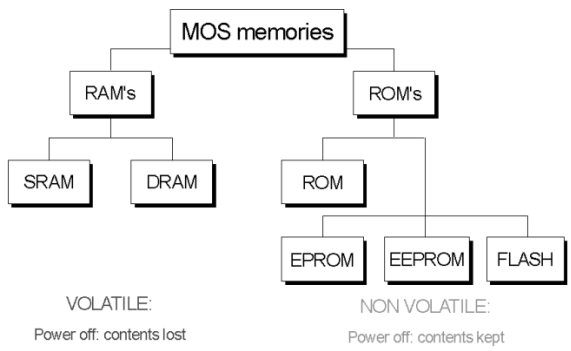

Fig. 2. MOS memory tree.

	Cell Phone	Consumer	Automotive	Computer & Communication
EPROM	Analog, Residential	Games, Set Top Box	Engine Mgt	HDD, Copiers, Fax, Switching
FLASH	Digital (GSM)	Set Top Box PDA	All Power Train Car Navigation, ABS, GPS	HDD, PC Bios, CDROM
EEPROM	Digital (GSM)	Audio, Video	All Car Body	PC SPD, Graphic boards, Printers

Fig. 3. Main nonvolatile memory applications.

(one transistor cell) with a very good flexibility (they can be electrically written on field more than 100 000 times, with byte programming and sectors erasing).

The most relevant phenomenon of this past decade in the field of semiconductor memories has been the explosive growth of the Flash memory market, driven by cellular phones and other types of electronic portable equipment (palm top, mobile PC, mp3 audio player, digital camera, and so on). Moreover, in the coming years, portable systems will demand even more nonvolatile memories, either with high density and very high writing throughput for data storage application or with fast random access for code execution in place.

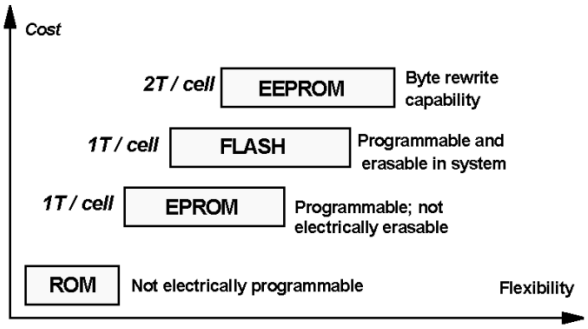

Fig. 4. Nonvolatile memory (NVM) qualitative comparison in the flexibility–cost plane. A common feature of NVMs is to retain the data even without power supply.

Based on these market needs, a well-known way to classify Flash products and the relative technologies is that of defining two major application segments:

- code storage, where the program or the operating system is stored and is executed by the microprocessor or microcontroller;
- data (or mass) storage, where data files for image, music, and voice are recorded and read sequentially.

Different type of Flash cells and architectures have been proposed in the past (see Fig. 5). They can be divided in terms of access type, parallel or serial, and in terms of the utilized programming and erasing mechanism, Fowler–Nordheim tunneling (FN), channel hot electron (CHE), hot-holes (HH), and source-side hot electron (SSHE). Among all of these architectures, today two can be considered as industry standard: the common ground NOR Flash [1]–[3], that due to its versatility is addressing both the code and data storage segments, and the NAND Flash, optimized for the data storage market [4], [5].

In the following, the basic concepts, the reliability issues, the evolution, and scaling trends will be presented only for the NOR Flash cell, but most of these considerations are also valid for the NAND since both of them are based on the concept of floating-gate MOS transistor.

II. NOR FLASH CELL

In 1971, Frohman-Bentchkowsky presented a floating gate transistor in which hot electrons were injected and stored [6], [7]. From this original work, the erasable programmable read only memory (EPROM) cell, programmed by CHE and erased by ultraviolet (UV) photoemission, has been developed. The EPROM technology became the most important nonvolatile memory in the 1980s. In the same period, the Flash EEPROM was proposed, basically an EPROM cell, with the possibility to be electrically erased [8]. The name Flash was given to represent the fact that the whole memory array could be erased in the same (fast) time.

The first Flash product was presented in 1988 [9]. In terms of applications, initially Flash products were mainly used as an "EPROM replacement," offering the possibility to be erased on system, avoiding the cumbersome UV erase oper-

Absolute Exhibit No. 1046
Page 002

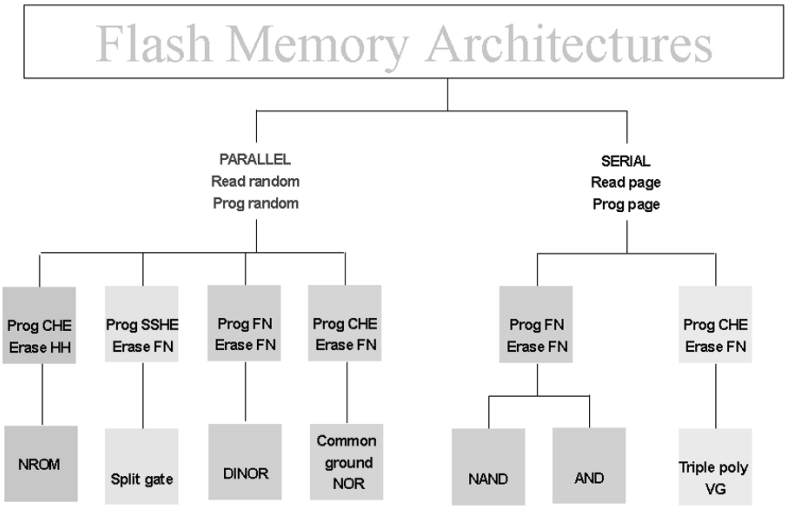

Fig. 5. The family tree of Flash memory cell architecture. The actual industry standard are: 1) The NOR for code and data storage application and 2) NAND only for data storage.

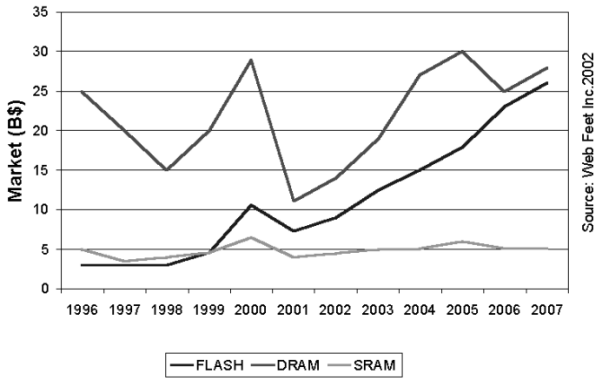

Fig. 6. Semiconductor memory market for the main memory, i.e., DRAM, Flash, and SRAM.

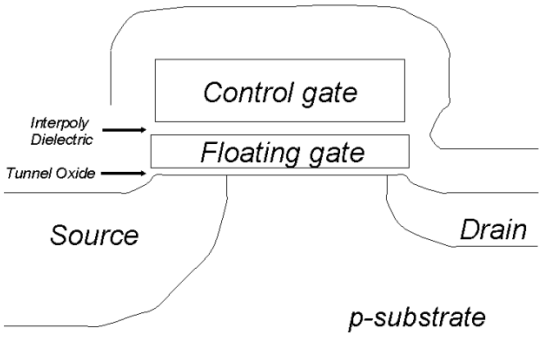

Fig. 7. Schematic cross section of a Flash cell. The floating-gate structure is common to all the nonvolatile memory cells based on the floating-gate MOS transistor.

ation. But the Flash market did not take off until this technology was proven to be reliable and manufacturable. In the late 1990s, the Flash technology exploded as the right nonvolatile memory for code and data storage, mainly for mobile applications. Starting from 2000, the Flash memory can be considered a really mature technology: more than 800 million units of 16-Mb equivalent NOR Flash devices were sold in that year.

In Fig. 6, the Flash market is reported and compared with the DRAM and SRAM one [10]. It can be seen that the Flash market became and has stayed bigger than the SRAM one since 1999. Moreover, the Flash market is forecasted to be above $20 billion in three or four years from now, reaching the DRAM market amount, and only smoothly following the DRAM oscillating trend, driven by the personal computer market. In fact, portable systems for communications and consumer markets, which are the drivers of the Flash market, are forecasted to continuously grow in the coming years.

In the following, we briefly describe the basics of the Flash cell functionality.

A. Basic Concept

A Flash cell is basically a floating-gate MOS transistor (see Fig. 7), i.e., a transistor with a gate completely surrounded by dielectrics, the floating gate (FG), and electrically governed by a capacitively coupled control gate (CG). Being electrically isolated, the FG acts as the storing electrode for the cell device; charge injected in the FG is maintained there, allowing modulation of the "apparent" threshold voltage (i.e., V_T seen from the CG) of the cell transistor. Obviously the quality of the dielectrics guarantees the nonvolatility, while the thickness allows the possibility to program or erase the cell by electrical pulses. Usually the gate dielectric, i.e., the one between the transistor channel and the FG, is an oxide in the range of 9–10 nm and is called "tunnel oxide" since FN electron tunneling occurs through it. The dielectric that separates the FG from the CG is formed by a triple layer of oxide–nitride–oxide (ONO). The ONO thickness is in the range of 15–20 nm of equivalent oxide thickness. The ONO layer as interpoly dielectric has been introduced in order to improve the tunnel oxide quality. In fact, the

Absolute Exhibit No. 1046
Page 003

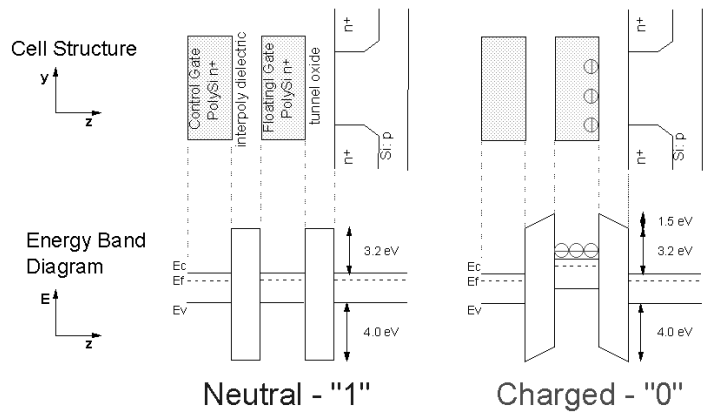

Fig. 8. Schematic energy band diagram (lower part) as referred to a floating gate MOSFET structure (upper part). The left side of the figure is related to a neutral cell, while the right side to a negatively charged cell.

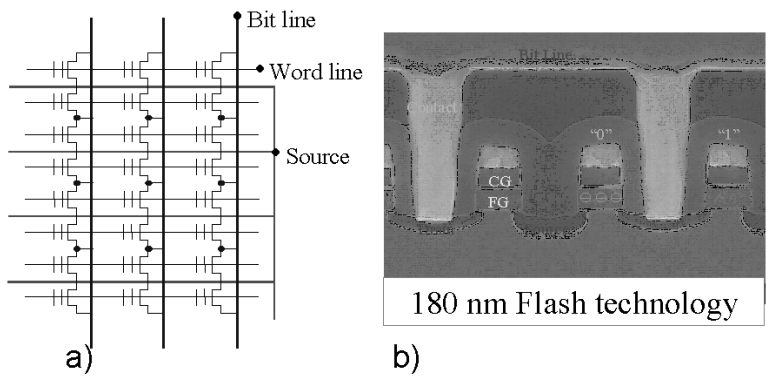

Fig. 9. (a) NOR Flash array equivalent circuit. (b) Flash memory cell cross section.

use of thermal oxide over polysilicon implies growth temperature higher than 1100 °C, impacting the underneath tunnel oxide. High-temperature postannealing is known to damage the thin oxide quality.

If the tunnel oxide and the ONO behave as ideal dielectrics, then it is possible to schematically represent the energy band diagram of the FG MOS transistor as reported in Fig. 8. It can be seen that the FG acts as a potential well for the charge. Once the charge is in the FG, the tunnel and ONO dielectrics form potential barriers.

The neutral (or positively charged) state is associated with the logical state "1" and the negatively charged state, corresponding to electrons stored in the FG, is associated with the logical "0."

The "NOR" Flash name is related to the way the cells are arranged in an array, through rows and columns in a NOR-like structure. Flash cells sharing the same gate constitute the so-called wordline (WL), while those sharing the same drain electrode (one contact common to two cells) constitute the bitline (BL). In this array organization, the source electrode is common to all of the cells [Fig. 9(a)].

A scanning electron microscope (SEM) cross section along a bitline of a Flash array is reported in Fig. 9(b), where three cells can be observed, sharing two by two the drain contact and the sourceline. This picture can be better understood considering the layout of a cell (see Fig. 10) and the two schematic cross sections, along the y direction (bitline) and the x direction (wordline). The cell area is given by the x pitch times the y pitch. The x pitch is given by the active area width and space, considering also that the FG must overlap the oxide field. The y pitch is constituted by the cell gate length, the contact-to-gate distance, half contact, and half sourceline. It is evident, as reported in Fig. 9(b), that both contact and sourceline are shared between two adjacent cells.

B. Reading Operation

The data stored in a Flash cell can be determined measuring the threshold voltage of the FG MOS transistor. The best and fastest way to do that is by reading the current driven by the cell at a fixed gate bias. In fact, as schematically reported in Fig. 11, in the current–voltage plane two cells, respectively, logic "1" and "0" exhibit the same transconductance curve but are shifted by a quantity—the threshold voltage shift (ΔV_T)—that is proportional to the stored electron charge Q.

Hence, once a proper charge amount and a corresponding ΔV_T is defined, it is possible to fix a reading voltage in such

Absolute Exhibit No. 1046
Page 004

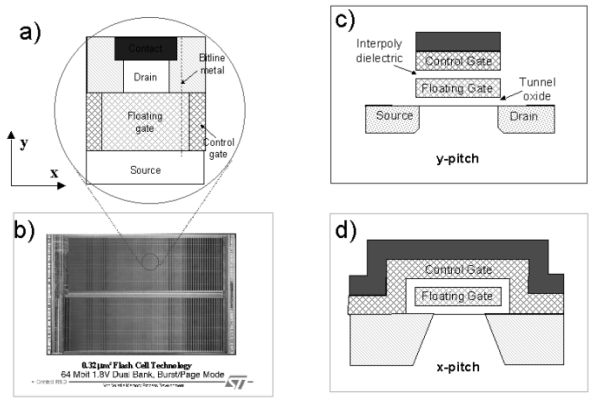

Fig. 10. The NOR Flash cell. (a) Basic layout. (b) Updated Flash product (64-Mb, 1.8-V Dual bank). (c) and (d) are, respectively, the schematic cross section along bitline (y pitch) and wordline (x pitch).

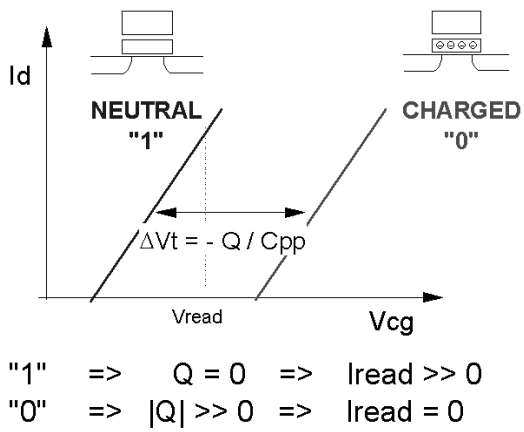

Fig. 11. Floating-gate MOSFET reading operation.

a way that the current of the "1" cell is very high (in the range of tens of microamperes), while the current of the "0" cell is zero, in the microampere scale. In this way, it is possible to define the logical state "1" from a microscopic point of view as no electron charge (or positive charge) stored in the FG and from a macroscopic point of view as large reading current. Vice versa, the logical state "0" is defined, respectively, by electron charge stored in the FG and zero reading current.

C. Writing Operation

Considering Fig. 8, the problem of writing an FG cell corresponds to the physical problem of forcing an electron above or across an energy barrier. The problem can be solved exploiting different physical effects [11]. In Fig. 12, the three main physical mechanisms used to write an FG memory cell are sketched.

– The CHE mechanism, where electrons gain enough energy to pass the oxide–silicon energy barrier, thanks to the electric field in the transistor channel between source and drain. In fact, the electron energy distribution presents a tail in the high energy side that can be modulated by the longitudinal electric field.

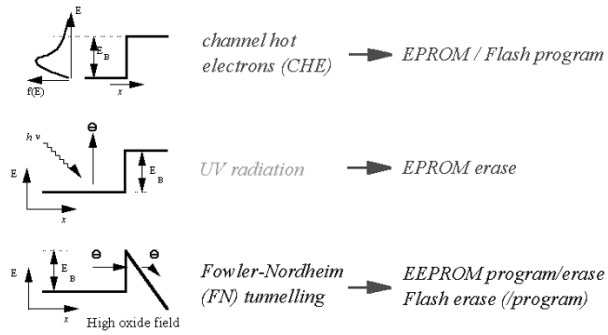

Fig. 12. Writing mechanism in floating-gate devices.

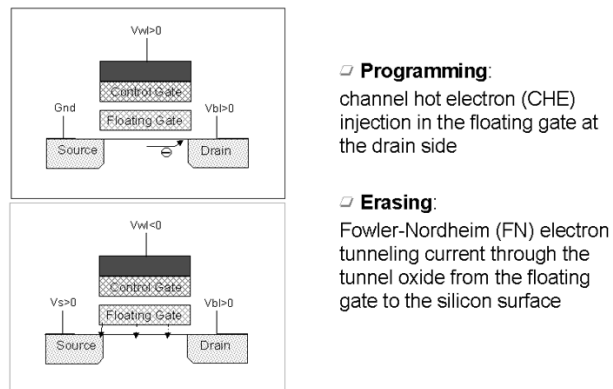

Fig. 13. NOR Flash writing mechanism.

– The photoelectric effect, where electrons gain enough energy to surmount the barrier thanks to the interaction with a photon with energy $h\nu$ larger than the barrier itself. For silicon–dioxide, this corresponds to UV radiation. This mechanism is the one originally used in EPROM's products to erase the entire device.

– The Fowler–Nordheim electron tunneling mechanism is a quantum-mechanical tunnel induced by an electric field. Applying a strong electric field (in the range of 8–10 MV/cm) across a thin oxide, it is possible to force a large electron tunneling current through it without destroying its dielectric properties.

A NOR Flash memory cell is programmed by CHE injection in the FG at the drain side and it is erased by means of the FN electron tunneling through the tunnel oxide from the FG to the silicon surface (see Fig. 13).

III. Reliability

Many issues have to be addressed when, from the theoretical model of a single cell, a Flash product has to be realized, integrating millions of cells in an array. Nonvolatility implies at least ten years of charge retention, and the data must be stored in a cell after many read/program/erase cycles. The confidence in Flash memory reliability has grown together with the understanding of the single memory cell failure mechanisms.

Absolute Exhibit No. 1046
Page 005

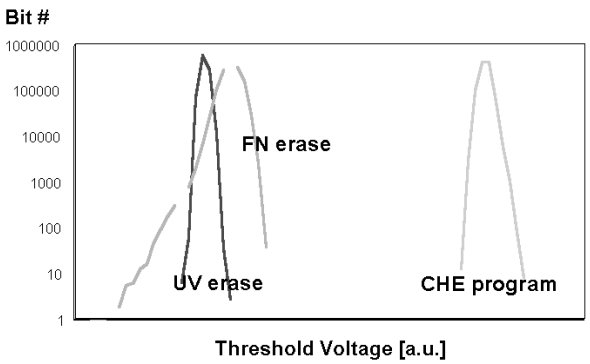

Fig. 14. Threshold voltage distribution of a 1-Mb Flash array after UV erasure, after CHE programming, and after FN erasure.

The high degree of testability [12] allows the detection at wafer level of latent defects which may cause single-cell failures related to programming disturbs, data retention, and oxide defects [13], thus making Flash one of the most reliable nonvolatile memories.

A. Threshold Voltage Distribution

When dealing with a large array of cells, e.g., from tens of thousands to one million, it is very important to understand the type of dispersion given by the large set of cells. The best way to do it is to compare the threshold voltage distribution of the whole array, considering it after UV erasure—that can be considered as the reference state—after CHE programming and after FN erasing.

Fig. 14 shows typical distributions of cell threshold voltages in a large memory array. The UV-erased distribution is pretty narrow and symmetrical. A more accurate analysis would reveal a Gaussian distribution due to random variations 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 it is still symmetrical. The enlargement occurs because most of the parameters that cause V_T dispersion of UV-erased cells also impact the threshold shift of programmed cells.

The distribution of threshold voltages after electrical erase is much wider and heavily asymmetrical. A more detailed analysis would show that the bulk of the distribution is again a Gaussian with a standard deviation larger than the one of programmed cells. Cells in this part of the distribution are referred to as "normal" cells. But there is also an exponential tail at low V_T, composed of cells that erase faster than the average, also called "tail" cells.

The dispersion of threshold voltages of normal cells is due to coupling ratio variations, and it has been accurately modeled [14]. Instead, the understanding of the tail cells, although of key importance, is more difficult. In fact, as these cells erase faster than normal cells with the same applied voltage, one should assume that they are somehow "defective." However, they are just too numerous for being associated with extrinsic defects.

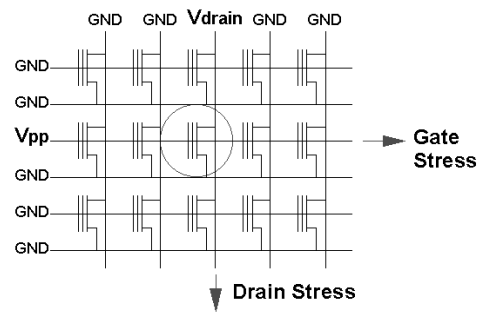

Fig. 15. Schematic of a Flash array, showing row and column disturbs occurring when the cycled cell is programmed.

Different models have been presented with the aim to explain the tail cells. For example, a distribution in the polycrystalline structure of the FG, with a barrier height variation at the grain boundaries, would give rise to a local enhancement of the tunnel barrier [15]. Another model explains the tail cells as due to randomly distributed positive charges in the tunnel oxide [16]. This model is solidly based on the well-known existence of donor-like bulk oxide traps and on calculations that show the huge increase of the tunnel current density caused by the presence of an elementary positive charge closed to injecting electrode.

Independently from a consolidated model, it can be stated that the exponential tail of the erased distribution is mostly related to structural imperfections, i.e., intrinsic defects, and it can be minimized by process optimization (for example, working on silicon surface preparation, tunnel oxidation, FG polysilicon optimization) but not eliminated. Flash products must be designed taking into account the existence of such a tail.

B. Program Disturb

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. Two types of program disturbs must be taken into account: row and column disturbs, also referred as gate and drain stress, as schematically reported in Fig. 15, representing a portion of a cell array.

Row disturbs are due to gate stress applied to a cell while programming other cells on the same wordline. If a high voltage is applied to the selected row, all the other cells of that row must withstand the gate stress without losing their data. Depending on the data stored in the cells, data can be lost either by a leakage in the gate oxide or by a leakage in the interpoly dielectric.

Column disturbs are due to drain stress applied to a cell while programming other cells on the same bitline. Under this condition, programmed cells can lose charge by FN tunneling from the FG to the drain (soft erasing). The program disturb depends on the number of cells along bitline and wordline and then depends strongly on the sector organization. The most effective way to prevent disturb propagation is to use block select transistor in a divided bitline and wordline

Absolute Exhibit No. 1046
Page 006

organization to completely isolate each sector. Program disturb really could be a critical issue in Flash memory, and cells and circuits must be designed with safety margins versus the stress sensitivity.

C. Data Retention

As in any nonvolatile memory technology, Flash memories are specified to retain data for over ten years. This means the loss of charge stored in the FG must be as minimal as possible. In updated Flash technology, due to the small cell size, the capacitance is very small and at an operative programmed threshold shift—about 2 V—corresponds a number of electrons in the order of 10^3 to 10^4. A loss of 20% in this number (around 2–20 electrons lost per month) can lead to a wrong read of the cell and then to a data loss.

Possible causes of charge loss are: 1) defects in the tunnel oxide; 2) defects in the interpoly dielectric; 3) mobile ion contamination; and 4) detrapping of charge from insulating layers surrounding the FG.

The generation of defects in the tunnel oxide can be divided into an extrinsic and an intrinsic one. The former is due to defects in the device structure; the latter to the physical mechanisms that are used to program and erase the cell. The tunnel oxidation technology as well as the Flash cell architecture is a key factor for mastering a reliable Flash technology.

The best interpoly dielectric considering both intrinsic properties and process integration issues has been demonstrated to be a triple layer composed of ONO. For several generations, all Flash technologies have used ONO as their interpoly dielectric.

The problem of mobile ion contamination has been already solved on the EPROM technology, taking particular care with the process control, but in particular using high phosphorus content in intermediate dielectric as a gettering element. [17], [18]. The process control and the intermediate dielectric technology have also been implemented in the Flash process, obtaining the same good results.

Electrons can be trapped in the insulating layers surrounding the floating gate during wafer processing, as a result of the so-called plasma damage, or even during the UV exposure normally used to bring the cell in a well-defined state at the end of the process. The electrons can subsequently detrap with time, especially at high temperature. The charge variation results in a variation of the floating gate potential and thus in cell V_T decrease, even if no leakage has actually occurred. This apparent charge loss disappears if the process ends with a thermal treatment able to remove the trapped charge.

The retention capability of Flash memories has to be checked by using accelerated tests that usually adopt screening electric fields and hostile environments at high temperature.

D. Programming/Erasing Endurance

Flash products are specified for 10^5 erase/program cycles. Cycling is known to cause a fairly uniform wear-out of the cell performance, mainly due to tunnel oxide degradation, which eventually limits the endurance characteristics [19]. A typical result of an endurance test on a single cell is shown in Fig. 16. As the experiment was performed applying constant pulses, the variations of program and erase threshold voltage levels are described as "program/erase threshold voltage window closure" and give a measure of the tunnel oxide aging. In real Flash devices, where intelligent algorithms are used to prevent V_T window closing, this effect corresponds to a program and erase times increase (see Fig. 17).

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, which are mechanisms specific to hot-electron degradation.

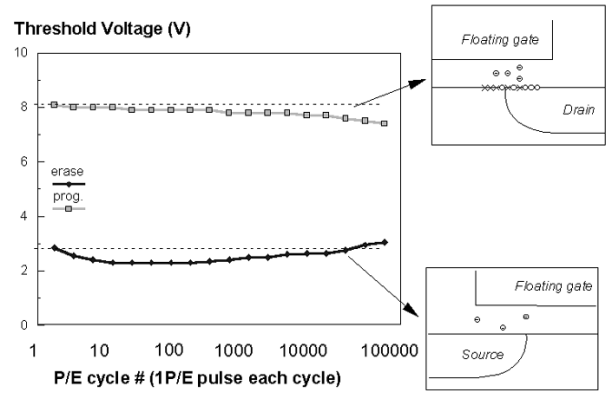

Fig. 16. Threshold voltage window closure as a function of program/erase cycles on a single cell.

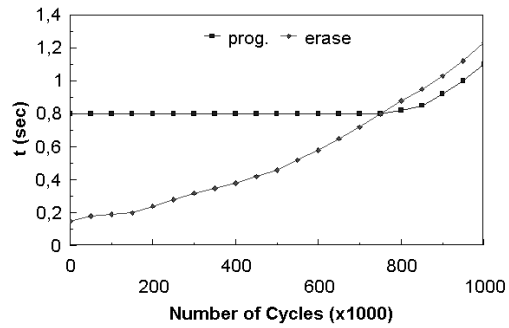

Fig. 17. Program and erase time as a function of the cycles number.

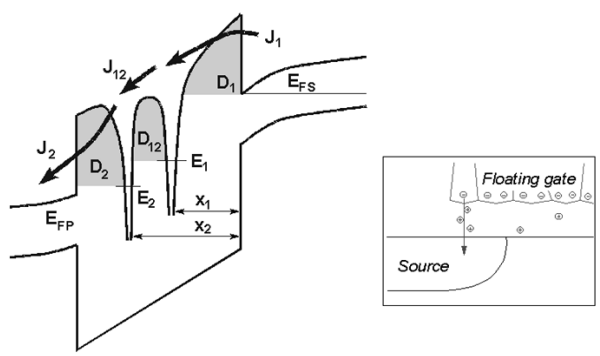

Fig. 18. Anomalous SILC modeling. The leakage is caused by a cluster of positive charge generated in the oxide during erase (left-hand side). The multitrap assisted tunneling is used to model SILC: trap parameters are energy and position.

Absolute Exhibit No. 1046
Page 007

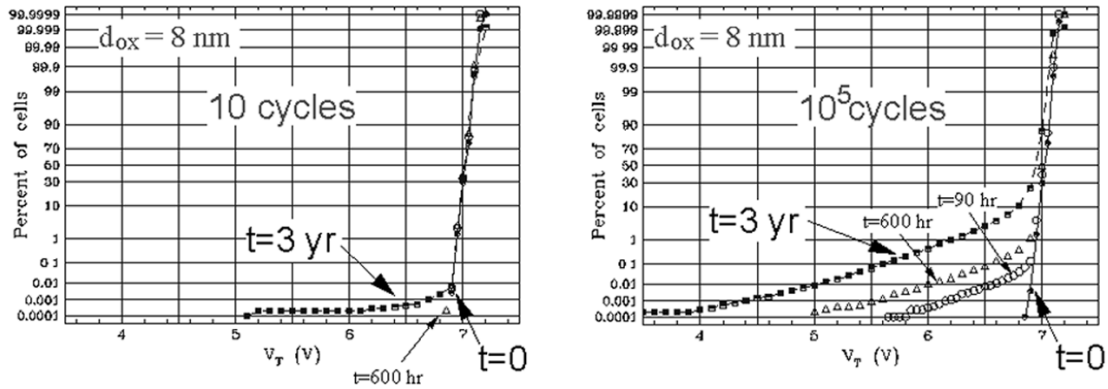

Fig. 19. Data retention tests at room temperature.

The evolution of the erase threshold voltage reflects the dynamics of net fixed charge in the tunnel oxide as a function of the injected charge. 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 a generation of negative traps.

Cycling wear-out can be reduced by proper device engineering and by optimization of the tunnel oxide process. However, once process and product are qualified for a given endurance specification, no major problems should come from lot-to-lot variation.

Actually, endurance problems are mostly given by single-cell failures, which present themselves like a retention problem after program/erase cycles. In fact, 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 from the current–voltage curves from the theoretical FN characteristics at low field, is known as stress-induced leakage current (SILC). SILC is clearly attributed to stress-induced oxide defects and, as far as a conduction mechanism, it is attributed to a trap assisted tunneling (see Fig. 18). The main parameters controlling SILC are the stress field, the amount of charge injected during the stress, and the oxide thickness. For fixed stress conditions, the leakage current increases strongly with decreasing oxide thickness [20]–[22].

The effect of cycling on data retention cannot be referred to in the typical cell, but must be studied considering a wide array of cells, looking in particular to the tail distribution. In Fig. 19, we report the results of retention test on a 1-Mb array of cells with 8-nm tunnel oxide in order to enhance the SILC defects in single cells. Retention tests have been performed on arrays cycled 10 and 10^5 times [23]. As can be seen, the amount of cells that lose charge after three years are much more in the case of longer endurance. Data retention after cycling is the issue that definitely limits the tunnel oxide thickness scaling. For very thin oxide, below 8–9 nm, the number of leaky cells becomes so large that even error-correction techniques cannot fix the problem.

IV. Multilevel Concept

An attractive way to speed up the scaling of Flash memory is offered by the multilevel (ML) concept [24]. The idea is based on the ability to precisely control the amount of charge stored into the floating gate in order to set the threshold voltage V_T of a memory cell within any of a number of different voltage ranges, corresponding to different logical levels. A cell operated with 2^n different levels is capable of storing n bits, the case $n = 1$ being the conventional single-bit cell.

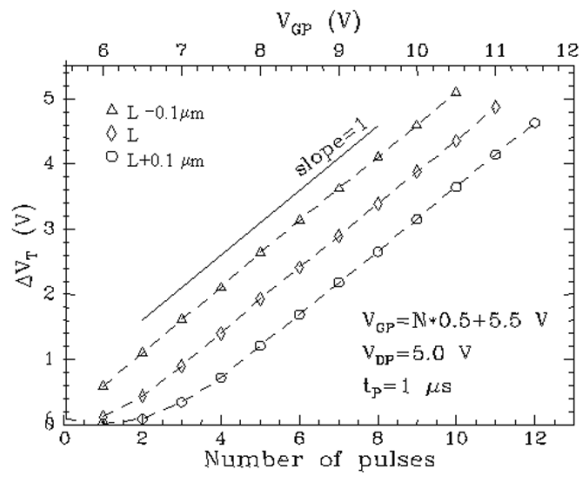

Fig. 20. $D\,V_T$ as a function of the pulse number for three different channel lengths (the upper axis also shows the gate voltage at each programming step).

Three main issues must be afforded when going from conventional to ML Flash [25]. A high programming accuracy is required to obtain narrow V_T distributions; reading operation implies multiple, either serial or parallel, comparison with suitable references to determine the cell status, requiring accurate and fast current sensing; V_T window and read voltage are larger while read margins are smaller than the single-bit case, this for allocating all levels, requiring improved reliability and/or error-correction circuitry. These key points will be discussed with reference to a common-ground NOR architecture.

A. Multilevel Flash Programming

CHE programming has been shown to give, under proper conditions, a linear relationship with unit slope between programming gate voltage V_{GP} and V_T variation [26], independently of cell parameters (see Fig. 20). Very tight V_T distri-

Absolute Exhibit No. 1046
Page 008

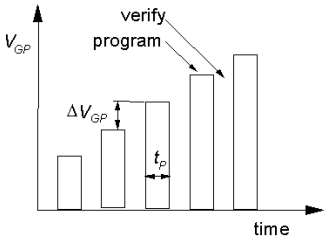

Fig. 21. Schematic of the control-gate voltage pulses.

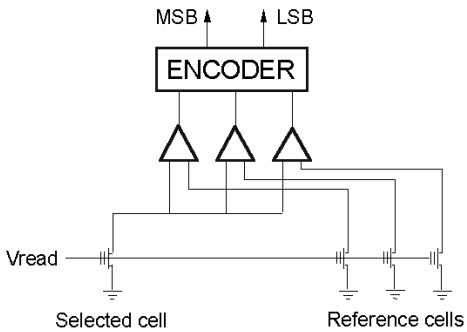

Fig. 22. Parallel multilevel sensing architecture. MSB = most significant bit; LSB = less significant bit.

butions can be obtained by combining a program-and-verify technique with a staircase V_{GP} ramp (see Fig. 21). In fact, this method should theoretically lead to a V_T distribution width for any state not larger than ΔV_{GP}. Indeed, neglecting any error due to sense amplifier inaccuracy or voltage fluctuations, the last programming pulse applied to a cell will cause its threshold voltage to be shifted above the program verify decision level by an amount at most as large as ΔV_{GP}. It follows that by decreasing ΔV_{GP}, it is possible to increase the programming accuracy. Obviously, this is paid in terms of a larger number of programming pulses together with verify phases and, therefore, with a longer programming time. Hence, the best accuracy/time tradeoff must be chosen for each case considering the application specification.

However, high programming throughput, equal to 1-b/cell devices, is normally achieved via a large internal program parallelism, which is possible because cells need a low programming current in ML staircase programming. To do that, ML devices operate with a program write buffer, whose typical length is 32–64 bytes, i.e., 128–256 cell data length.

Also, evolution to 3–4 b/cell will not have an impact on programming throughput. In fact, program pulses and verify phases increase proportionally with the number of bits per cell, thus keeping roughly constant the effective byte programming time.

Despite a not-negligible programming current, another advantage in using CHE programming for multilevel devices is to avoid the appearance of erratic bits that instead can be a potential failure mode affecting FN programming. In fact, erratic bit behavior was observed in the FN erase of standard NOR memories [27] but, for its nature, it should be present in every tunneling process [28].

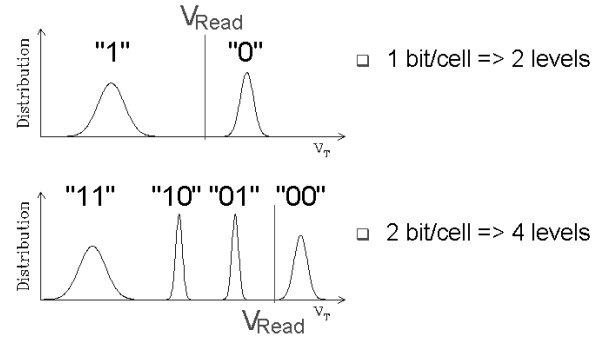

Fig. 23. Threshold voltage distribution for 2 b/cell compared with the standard 1 b/cell.

B. Reading Operation

In order to have a fast reading operation in the NOR cell, a parallel sensing approach can be used [29]. The cell current, obtained in reading conditions, is simultaneously compared with three currents provided by suitable reference cells (see Fig. 22). The comparison results are then converted to a binary code, whose content can be 11, 01, 10, or 00, due to the multilevel nature. In Fig. 23, we report the threshold voltage distribution of a 2-b/cell memory. The 11, 10, and 01 cell distribution will give rise to a different current distribution, measured at fixed V_{read}, while the 00 cell distribution does not drain current as well as the programmed level of a standard 1-b/cell device. High read data rate, via page or burst mode, is normally supported by large internal read parallelism.

A parallel sensing approach does not seem transferable to 3- or 4-b/cell generations because of the exponential increase, $2^n - 1$, in comparators number, respectively 7 or 15 per cell, that means exponential increase in sensing area and current consumption. At this moment, a serial sensing approach, e.g., dichotomic, or a mixed serial-parallel is considered the more suitable approach. Serial sensing is also useful for a 2-b/cell device when high-speed random access is not necessary, e.g., in Flash Cards applications.

C. Data Retention

One of the main concerns about multilevel is the reduced margin toward the charge loss, compared with the 1-b/cell approach. We can basically divide the problem of data retention into two different issues.

The first is related to the extrinsic charge loss, i.e., to a single bit that randomly can have different behaviors with respect to the average and that usually form a tail in a standard distribution. It is well known that extrinsic charge loss strongly depends on tunnel oxide retention electric field and that this issue can become more critical if an enhanced cell threshold range has to be used to allocate the 2^n levels [30]. This problem is usually solved with the introduction of the error correction code (ECC), whose correction power must be chosen as a function of the technology and of the specification required to the memory products.

The second one is related to the intrinsic charge loss, i.e., to the behaviors of the Gaussian part of a cell distribution,

Absolute Exhibit No. 1046
Page 009

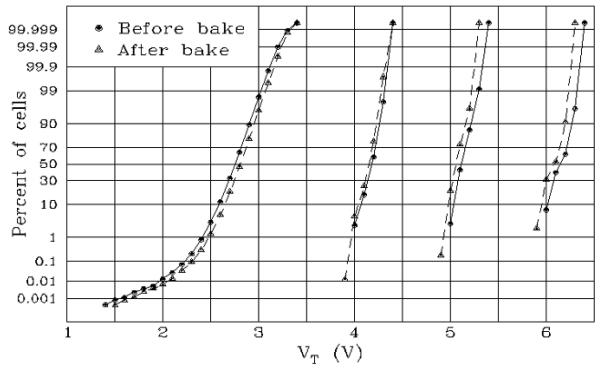

Fig. 24. Shift in the threshold voltage distribution after 500 h bake at 250 °C.

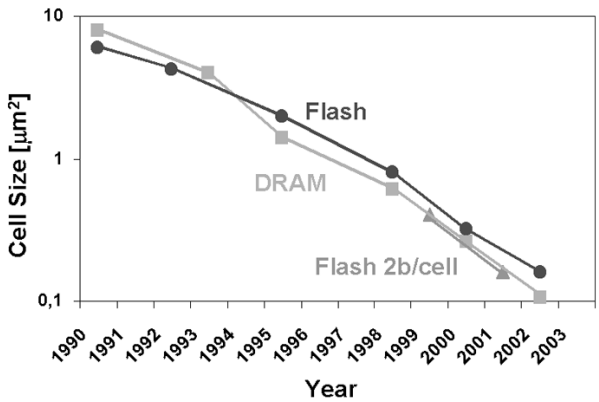

Fig. 25. DRAM and Flash cell size reduction versus year. The scaling has been of about a factor 30 in ten years.

that must be characterized and defined as a function of the different level distributions. In order to study the data retention on multilevel memories, usually tests at high-temperature bake on programmed cells are performed. A result of data retention after bake (500 h, 250 °C) is shown in Fig. 24, on one million cells [31].

The maximum V_T shift, which occurs for the uppermost level, is about 0.1 V. This means the spacing between levels is reduced by a very small amount. It is interesting to note that the three programmed levels are shifted by an amount proportional to their respective programmed ΔV_T, so that the spacing between adjacent levels is reduced by only a fraction of the observed maximum shift.

V. Evolution and Scaling Trend

The Flash memories were commercially introduced in the early 1990s and since that time they have been able to follow the Moore law or, better, the scaling rules imposed by the market. Fig. 25 reports in a logarithmic scale the Flash cell size as a function of time, from 1992 to 2002. It turns out that the reduction of the cell size has been about a factor 30 in those ten years, closely following the scaling of the DRAM, today still considered as the reference memory technology that sets the pace to the technology node evolution. More specifically, the NOR Flash cell has scaled from 4.2 μm^2 for the 0,6-μm technology node to the present cell size of 0.16 μm^2 at the 0.13-μm node.

Moreover, considering the multilevel approach for the Flash cell with the capability to store two bits in the same cell, as presented in Section IV, not only the scaling trend but even the bit size itself is well aligned with the DRAM one.

Together with the Flash cell scaling, there has also been an evolution of the Flash product specification and application. Three main generations can be considered, well differentiated as a technology node, process complexity, and specification.

– *First generation (1990–1997).* The Flash applications were mainly "EPROM replacement." The products were characterized by a single array (bulk), with memory density from 256 kb to 2 Mb. The program and erase algorithms were controlled externally and all the product were dual voltage: 12 V for the write operations and 5 V for the power supply. Cycling specification was limited to 10^4.

– *Second generation (1995–2000).* The Flash memory has become the right nonvolatile memory technology for code storage application, where software updates must be performed on the field. In particular, portable systems, mainly cellular phones, were strongly interested in this feature.

  The cellular phone applications brought a lot of innovations:

  - The density was increased from 1 to 16 Mb and sectors were introduced, instead of a single (bulk) array, in order to allow different use of the memory (some sectors can be used to store code while others to store data, with different requirements in terms of cycling). Sector density was from 10 to 256 kb.
  - A single voltage supply pin (5 or 3 V according to the system specification) substituted the two high-voltage and low-voltage pins previously used. The need to be programmed on field, without the possibility to have the high voltage from an external pin, has developed the technology to internally generate the writing voltages using charge-pump techniques. A high-voltage supply is sometimes still used, but limited to the first programming operation in the system manufacturing line, to improve the throughput.
  - Algorithms to perform all the operation on the array—reading programming and erasing—were embedded into the device in order to avoid the need for an external microcontroller.
  - 10^5 writing cycles were introduced as a specification. More than effectively needed by the system, this high endurance is the result of a highly reliable technology.

– *Third generation (from 1998 on).* The portable system specifications push toward Flash memory products that look more and more like an application-specific memory. Obviously, the density is one of the most important parameters, and devices well

Absolute Exhibit No. 1046
Page 010

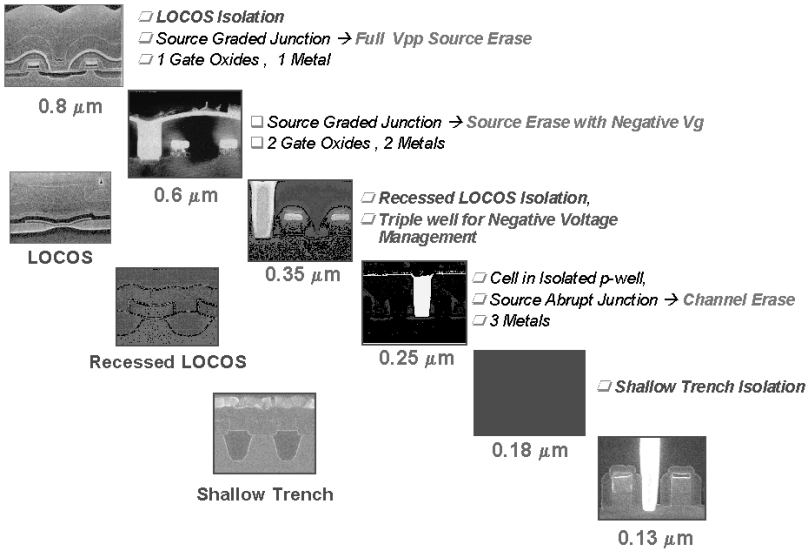

Fig. 26. NOR Flash technology and architecture evolution.

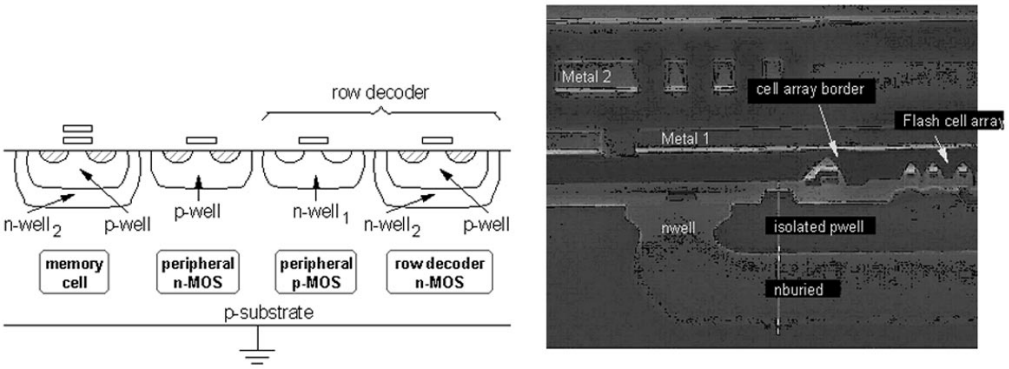

Fig. 27. Triple well structure cross section: schematic (left side) and SEM (right side).

beyond 64 Mb will be realized entering the Flash in the gigabit era. The sectorization is becoming more complex, and dual or multiple bank devices have already been presented. In these devices, different groups of sectors (banks) can be differently managed: at the same time one sector belonging to a bank can be read while another one, inside a different bank, can be programmed or erased. Also, following the general trend of reducing the power supply, the device supply is scaling to 1.8 V (with the consequent difficulties of internally generating high voltages starting from this low supply voltage value) and will go down to 1.2 V. Another issue, becoming more and more important, is the high data throughput, in particular considering the density increase. Burst mode is often used in order to speed up the reading operation and quickly download the software content, reaching up to 50 MB/s.

The introduction of the different generation as well as the reduction of the cell size has been made possible by the developments of Flash technology and process, and of cell architecture.

For what concerns the process architecture, all the main technology steps that have allowed the evolution of the CMOS technology have also been used for Flash. In Fig. 26, the different cell cross sections as a function of the different technology node are reported. For every generation, the main innovative introduced steps are pointed out. It turns out that the evolution of the different generations has been sustained by an increased process complexity, from the one gate oxide and one metal process with standard local oxidation of silicon isolation at the 0.8-μm technology node, to the two gate oxides, three metals, and shallow trench isolation at the 0.13-μm node. In between is the introduction of tungsten plug, of self-aligned silicided junctions and gates, and the wide use of chemical mechanical polishing steps. But one of the most crucial technologies for Flash evolution was the high-energy implantation development that has allowed the introduction of the triple well architecture (see Fig. 27). With this process module, further development of the single-voltage products has been possible, allowing the easy management of the negative voltage required to erase the cell and, furthermore, the possibility to completely change the erasing scheme of the cell.

In fact, as reported in Fig. 28, the cell programming and erasing applied voltages have been changed as a function of the different generation, always staying inside the CHE programming and the FN erasing. The first generation of cells

Absolute Exhibit No. 1046
Page 011

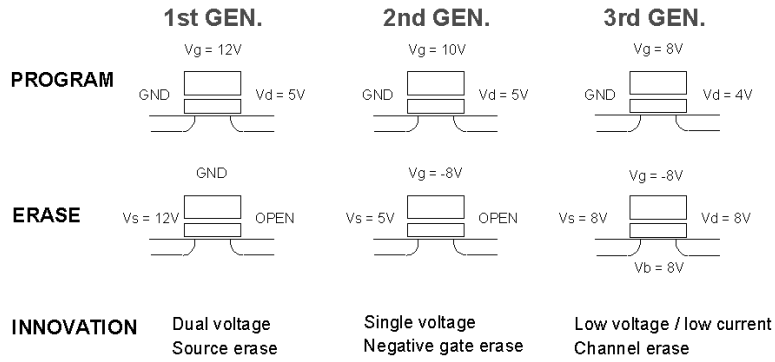

Fig. 28. NOR Flash cell evolution.

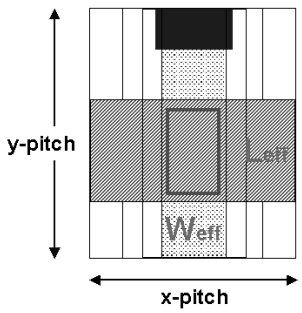

Fig. 29. NOR cell scaling. The basic layout has remained unchanged through different generations.

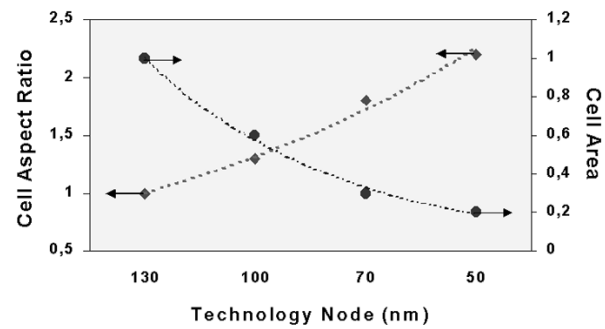

Fig. 30. NOR Flash cell scaling trends for cell area (right y axis) and cell aspect ratio (left y axis). Both values are normalized to the 130-nm technology node.

was erased, applying the high voltage to the source junction and then extracting electrons from the FG-source overlap region (source erase scheme). This way was too expensive in terms of parasitic current, as the working conditions were very close to the junction breakdown. Moving to the second generation with the single-voltage devices, the voltage drop between the source and the FG was divided, applying a negative voltage to the control gate and lowering the source bias to the external supply voltage (negative gate source erase scheme).

Finally, with the exploitation of the triple well also for the array, the erasing potential is now divided between the negative CG and the positive bulk (the isolated p-well) of the array, moving the tunneling region from the source to the whole cell channel (channel erase scheme). In this way, electrons are extracted from the FG all along the channel without any further parasitic current contribution from the source junction, consequently reducing the erase current amount of about three orders of magnitude; the latter being a clear benefit for battery saving in portable low-voltage applications.

The NOR Flash cell is forecasted to scale again following the International Technology Roadmap of Semiconductors (ITRS) [32]. The introduction of the 130-nm technology node has occurred in 2002–2003 with a cell size of 0.16 μm^2 [33], following the 10-λ^2 golden rule for the cell area scaling, where λ is the technology node. The representation of the memory cell size in terms of number of λ^2 is a usual way to compare different technology with the same metric; for example, the DRAM cell size is today quoted to stay in the range of 6–8 λ^2.

The next technology step for the NOR Flash will be the 90-nm technology node in 2004–2005. The cell size is expected to stay in the range of 10–12 λ^2, translating to a cell area of 0.1–0.08 μm^2. As reported again in Fig. 29, the cell basic layout and structure has remained unchanged through the different generations. The area scales through the scaling of both the x and y pitch. Basically, this must be done contemporarily reducing the active device dimensions, effective length (L_{eff}) and width (W_{eff}), and the passive elements, such as contact dimension, contact to gate distance, and so on.

For future generation technology nodes, i.e., the 65 nm in 2007 and the 45 nm in 2010, as forecasted by ITRS, the Flash cell reduction will face challenging issues. In fact, while the passive elements will follow the standard CMOS evolution, benefiting from all the technology steps and process modules proposed for the CMOS logic (like advanced lithography for contact size, cupper for metallization in very tight pitch), the active elements will be limited in the scaling. In particular, the effective channel length will be limited by the possibility to further scale the active dielectric, i.e., the tunnel oxide and the interpoly ONO. As already presented in Section III, the tunnel oxide thickness scaling is limited by intrinsic issues related to the Flash cell reliability, in particular the charge retention one, especially after many writing cycles. Although the direct tunneling, preventing the ten-year retention time, occurs at 6–7 nm, SILC considerations push the tunnel thickness limit to no less than 8–9 nm. Moreover, the effective

Absolute Exhibit No. 1046
Page 012

width reduction could be limited by the read current reduction, strongly proportional to the W_{eff}, then impacting the access time.

Scaling the technology node, while the cell x pitch will more and more approach the 2 λ, the y-pitch scaling will be limited by the cell gate scaling. Hence, for the 65- and 45-nm technology nodes, it is expected to have smaller cell size but with an increased number of λ^2, from 10 to 14. In particular, the cell aspect ratio, i.e., the y pitch over the x pitch, will continue to rise, due to the slowdown of the L_{eff} reduction. Fig. 30 reports the cell area and the cell-aspect ratio versus the technology node, both normalized at 130 nm. As can be observed, the cell area will be roughly one half at 90 nm (same number of 10–12 λ^2) and will decrease, but with slower trend at 65 and 45 nm. The cell-aspect ratio will continue to increase, almost doubling the one at 130 nm when the technology node will reach 45 nm.

VI. Summary

With more than ten years of consolidated know-how and thanks to its flexibility and cost characteristics, Flash memory is today a largely utilized, well-consolidated, and mature technology for nonvolatile memory application.

Flash sales represent a considerable amount of the overall semiconductor market. In particular, the NOR Flash is today the most diffused architecture, being able to serve both the code and the data storage market.

The cell is basically a floating-gate MOS transistor, programmed by CHE and erased by Fowler–Nordheim tunneling. The main reliability issues, like charge retention and endurance, have been extensively studied and the physical mechanism well understood in such a way to guarantee the present specification requirements.

The Flash cell scaling has been demonstrated to be really possible and to be able to follow the Moore's law down to the 130-nm technology generations. The technology development and the consolidated know-how will sustain the scaling trend down to the 90- and 65-nm technology nodes as forecasted by the ITRS.

One of the crucial issues to be solved to allow cell scaling below the 65-nm node is the tunnel oxide thickness reduction, as tunnel thinning is limited by intrinsic (direct tunneling) and extrinsic (SILC) mechanisms.

At each technology node, the multilevel approach will increase the memory efficiency, almost doubling the density at the same chip size, enlarging the application range, and reducing the cost per bit.

References

[1] S. Lai, "Flash memories: Where we were and where we are going," in *IEDM Tech. Dig.*, 1998, pp. 971–973.

[2] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells—An overview," *Proc. IEEE*, vol. 85, pp. 1248–1271, Aug. 1997.

[3] P. Pavan and R. Bez, "The industry standard Flash memory cell," in *Flash Memories*, P. Cappelletti *et al.*, Ed. Norwell, MA: Kluwer, 1999.

[4] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, "New ultra high density EPROM and Flash with NAND structure cell," in *IEDM Tech. Dig.*, 1987, pp. 552–555.

[5] S. Aritome, "Advance Flash memory technology and trends for file storage application," in *IEDM Tech Dig.*, 2000, pp. 763–766.

[6] D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.*, vol. 18, pp. 332–334, 1971.

[7] ——, "FAMOS – A new semiconductor charge storage device," *Solid State Electron.*, vol. 17, pp. 517–520, 1974.

[8] S. Mukherjee, T. Chang, R. Pang, M. Knecht, and D. Hu, "A single transistor EEPROM cell and its implementation in a 512 K CMOS EEPROM," in *IEDM Tech. Dig.*, 1985, pp. 616–619.

[9] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, and S. Wells, "An in-system reprogrammable 256 K CMOS Flash memory2," in *ISSCC Conf. Proc.*, 1988, pp. 132–133.

[10] Webfeet Inc., "Semiconductor industry outlook," presented at the 2002 Non-Volatile Memory Conference, Santa Clara, CA.

[11] L. Selmi and C. Fiegna, "Physical aspects of cell operation and reliability," in *Flash Memories*, P. Cappelletti *et al.*, Ed. Norwell, MA: Kluwer, 1999.

[12] G. Casagrande, "Flash memory testing," in *Flash Memories*, P. Cappelletti *et al.*, Ed. Norwell, MA: Kluwer, 1999.

[13] P. Cappelletti and A. Modelli, "Flash memory reliability," in *Flash Memories*, P. Cappelletti *et al.*, Ed. Norwell, MA: Kluwer, 1999.

[14] K. Yoshikawa, S. Yamada, J. Miyamoto, T. Suzuki, M. Oshikiri, E. Obi, Y. Hiura, K. Yamada, Y. Ohshima, and S. Atsumi, "Comparison of current Flash EEPROM erasing methods: Stability and how to control," in *IEDM Tech. Dig.*, 1992, pp. 595–598.

[15] S. Maramatsu, T. Kubota, N. Nishio, H. Shirai, M. Matsuo, N. Kodama, M. Horikawa, S. Saito, K. Arai, and T. Okazawa, "The solution of over-erase problem controlling poly-Si grain size—Modified scaling principles for Flash memory," in *IEDM Tech. Dig.*, 1994, pp. 847–850.

[16] C. Dunn, C. Kay, T. Lewis, T. Strauss, J. Screck, P. Hefley, M. Middendorf, and T. San, "Flash EEPROM disturb mechanism," in *Proc. Int. Rel. Phys. Symp.*, 1994, pp. 299–308.

[17] G. Crisenza, G. Ghidini, S. Manzini, A. Modelli, and M. Tosi, "Charge loss in EPROM due to ion generation and transport in interlevel dielectrics," in *IEDM Tech. Dig.*, 1990, pp. 107–110.

[18] G. Crisenza, C. Clementi, G. Ghidini, and M. Tosi, "Floating gate memories," *Qual. Reliab. Eng. Int.*, vol. 8, pp. 177–187, 1992.

[19] P. Cappelletti, R. Bez, D. Cantarelli, and L. Fratin, "Failure mechanisms of Flash cell in program/erase cycling," in *IEDM Tech. Dig.*, 1994, pp. 291–294.

[20] D. Ielmini, A. Spinelli, A. Lacaita, and A. Modelli, "Statistical model of reliability and scaling projections for Flash memories," in *IEDM Tech. Dig.*, 2001, pp. 32.2.1–32.2.4.

[21] D. Ielmini, A. S. Spinelli, A. L. Lacaita, L. Confalonieri, and A. Visconti, "New technique for fast characterization of SILC distribution in Flash arrays," in *Proc. IRPS*, 2001, pp. 73–80.

[22] D. Ielmini, A. S. Spinelli, A. L. Lacaita, R. Leone, and A. Visconti, "Localization of SILC in Flash memories after program/erase cycling," in *Proc. IRPS*, 2002, pp. 1–6.

[23] A. Modelli, "Reliability of thin dielectrics for nonvolatile applications," *Microelectron. Eng.*, vol. 48, pp. 403–408, 1999.

[24] B. Riccò, G. Torelli, M. Lanzoni, A. Manstretta, H. E. Maes, D. Montanari, and A. Modelli, "Nonvolatile multilevel memories for digital applications," *Proc. IEEE*, vol. 86, pp. 2399–2423, Dec. 1998.

[25] A. Modelli, R. Bez, and A. Visconti, "Multi-level Flash memory technology," in *2001 Int. Conf. Solid State Devices and Materials (SSDM)*, Tokyo, Japan, 2001, pp. 516–517.

[26] C. Calligaro, A. Manstretta, A. Modelli, and G. Torelli, "Technological and design constraints for multilevel Flash memories," in *Proc. 3rd IEEE Int. Conf. Electronics, Circuits, and Systems*, 1996, pp. 1003–1008.

[27] T. C. Ong *et al.*, "Erratic erase in ETOXTM Flash memory array," in *Tech. Dig. VLSI Symp. Technology*, vol. 7A-2, 1993, pp. 83–84.

[28] A. Chimenton, P. Pellati, and P. Olivo, "Analysis of erratic bits in FLASH memories," in *Proc. IRPS*, 2001, pp. 17–22.

[29] G. Campardo *et al.*, "40-mm^2 3-V-only 50-MHz 64-Mb 2-b/cell CHE NOR flash memory," *IEEE J. Solid-State Circuits*, vol. 35, pp. 1655–1667, Nov. 2000.

[30] H. P. Belgal *et al.*, "A new reliability model for post-cycling charge retention of Flash memories," in *Proc. IRPS*, 2002, pp. 7–20.

[31] A. Modelli, A. Manstretta, and G. Torelli, "Basic feasibility constraints for multilevel CHE-programmed Flash memories," *IEEE Trans. Electron Devices*, vol. 48, pp. 2032–2042, Sept. 2001.

[32] *International Technology Roadmap for Semiconductors*, 2001 ed.

[33] S. Keeney, "A 130 nm generation high-density ETOX Flash memory technology," in *IEDM Tech. Dig.*, 2001, pp. 2.5.1–2.5.4.

Absolute Exhibit No. 1046
Page 013

Roberto Bez was born in Milan, Italy, in 1961. He received the Ph.D. degree in physics from the University of Milan in 1985.

In 1986, he joined the VLSI Process Development Group of STMicroelectronics, Agrate Brianza, Italy, where he worked on nonvolatile memory process architectures. Until 1989, he was engaged in the electrical characterization and modeling of nonvolatile memory cells, contributing to the development of original device models. From 1989 to 1993 his work focused on the development of Flash memory, studying the device physics related to the programming/erasing mechanisms and participating to the process architecture definition. Then he was Project Leader of the Flash memory device process development for single power supply application from 1994 to 1997, and for multilevel products since 1998. Currently, he is Section Manager in the Non-Volatile Memory Process Development Group of the Central Research and Development Department of STMicroelectronics. He has authored many papers, conference contributions, and patents on topics related to nonvolatile memories. He was lecturer in Electron Device Physics at the University of Milan and in Non-Volatile Memory Devices at the University of Padova and Polytechnic of Milan. He is a member of the Symposium on VLSI Technology Technical Program Committee.

Emilio Camerlenghi was born in Bergamo, Italy, in 1959. He received the Ph.D. degree in physics (curriculum in solid state physics) from the University of Milan, Milan, Italy, in 1984.

In 1985, he joined the Central Research and Development Department, VLSI Process Development Group of STMicroelectronics, Agrate Brianza, Italy, working on nonvolatile memories process architectures. Until 1989, he was engaged in development of the 1.2-μm EPROM technology, with the main objective of studying the memory cell hot-electron programming physics and developing the memory cell device. From 1990 to 1992, he became responsible of the development of the new generation 0.6-μm EPROM process architecture. In 1992, he joined the Flash development team, where he was Project Leader of the 0.6-μm Flash technology, designed to realize both double (5–12 V) and single (5 V only) power supply devices. In 1995, he was in charge of leading the ST part of an advanced project (a codevelopment between STM and a U.S. company) whose target was to demonstrate the functionality of an innovative Flash memory virtual-ground architectural concept. In the 1997–1998 years, he was appointed to lead the development of the 0.25-μm Flash process architecture for low-voltage power supply applications. Since 1998, he has been Section Manager of High Performance Flash Memory in the Non-Volatile Memory Process Development Group of the Central Research and Development Department, STMicroelectronics; under his responsibility, the 0.18-μm, 0.15-μm generations were developed and qualified, while the 0.13-μm technology is at present in the qualification phase. He has authored many conference papers and patents on nonvolatile memory related topics. He is currently a member of the IEDM conference subcommittee on "Integrated Circuits and Manufacturing."

Alberto Modelli was born in Milan, Italy, in 1953. He received the Ph.D. degree in physics from the University of Milan in 1978.

In 1978, he joined the Device Physics Laboratory of the Research and Development Department, STMicroelectronics, Agrate Brianza, Italy, where he initially worked on the development of silicon solar cells and later on the physics and electrical characterization of the Si/SiO2 system. In 1994, he joined the Non-Volatile Memory Process Development Group of STMicroelectronics, where he has been working on the reliability of flash memories. Since 1996, he has been in charge of multilevel flash development. He is author or coauthor of over 40 publications, one book, and five patents on the above-mentioned topics.

Angelo Visconti was born in Como, Italy, in 1966. He received the Ph.D. degree in physics, *cum laude*, from the University of Milan, Como, in 1997. His thesis was on the feasibility of optical temporal solitons in quadratic nonlinear materials.

Beginning in 1987, he worked for ten years as a hardware and software designer and project leader for industrial automation systems. In 1997, he joined the Central Research and Development Department of STMicroelectronics, Agrate Brianza, Italy, in the Non-Volatile Memory Process Development Group. His first activities were about the study and characterization of channel erase and programming currents in Flash cells. Afterward, he was involved in the development of a 0.18-μm CMOS high-density Flash memory process. His interests are the characterization, reliability, and multilevel applications of Flash cells. He is author or coauthor of several publications and patents in the nonvolatile memory field and nonlinear optical field. He is currently a Lecturer of Non-Volatile Memory Devices at the University of Padova, Padova, Italy.

Absolute Exhibit No. 1046
Page 014