Flash Memory Cells—An Overview

PAOLO PAVAN, MEMBER, IEEE, ROBERTO BEZ, PIERO OLIVO, AND ENRICO ZANONI, SENIOR MEMBER, IEEE

The aim of this paper is to give a thorough overview of Flash memory cells. Basic operations and charge-injection mechanisms that are most commonly used in actual Flash memory cells are reviewed to provide an understanding of the underlying physics and principles in order to appreciate the large number of device structures, processing technologies, and circuit designs presented in the literature. New cell structures and architectural solutions have been surveyed to highlight the evolution of the Flash memory technology, oriented to both reducing cell size and upgrading product functions. The subject is of extreme interest: new concepts involving new materials, structures, principles, or applications are being continuously introduced. The worldwide semiconductor memory market seems ready to accept many new applications in fields that are not specific to traditional nonvolatile memories.

Keywords**—***Charge carrier processes, read-only memory, semiconductor memory.*

I. INTRODUCTION

Complementary metal-oxide-semiconductor (CMOS) memories can be divided into two main categories: random access memories (RAM's), which are volatile, i.e., they lose stored information once the power supply is switched off, and read-only memories (ROM's), which are nonvolatile, i.e., they keep stored information also when the power supply is switched off. Nonvolatile memory market share has been continuously growing in the past few years, and further growth in the near future is foreseen, especially for Flash memories (in which a single cell can be electrically programmable and a large number of cells—called a block, sector, or page—are electrically erasable at the same time) due to their enhanced flexibility against electrically programmable read-only memories (EPROM's), which are electrically programmable but erasable via ultraviolet (UV) exposure. Electrically erasable and programmable read-only memories (EEPROM's), which are electrically erasable and programmable per single byte, will be manufactured for specific applications only, since they use larger areas and, therefore, are more expensive. This is a conservative scenario, since there might be changes due to technology evolution. For example, at the end of the 1980's, Flash memories were supposed to replace EPROM's rapidly in every application. This did not happen, mainly due to early Flash reliability problems. On the other hand, the realization of new generations of Flash memories that can be erased by blocks of different sizes, emulating EEPROM's in some applications, and with single power supply widens the field of applicability for Flash memories and encourages new uses. In 1996, it was forecasted [1] that the memory market is going to be about half of the total integrated circuit market by the year 2000. Dynamic random-access memories (DRAM's) represent and will represent the main portion of the memory market (Fig. 1). Nonvolatile memories (NVM's) will account for 12% of the total available market, and Flash memory cells are forecast to be more than 50% of the year 2000 NVM market.

There are two major applications for Flash memories that should be pointed out. One application is the possibility of nonvolatile memory integration in logic systems—mainly, but not only, microprocessors—to allow software updates, store identification codes, reconfigure the system on the field, or simply have smart cards. The other application is to create storing elements, like memory boards or solid-state hard disks, made by Flash memory arrays which are configured to create large-size memories to compete with miniaturized hard disks. Since 1993, 40-Mbyte solid-state hard disks have been produced [2]. They are based on 16-Mb Flash memory boards but are very expensive. This scenario is not foreseen to change in the next few years until 64-Mb Flash memories optimized for this specific application will be available. Solid-state disks are very useful for portable applications, since they have small dimensions, low power consumption, and no mobile parts, therefore being more robust. Flash memories combine the capability of nonvolatile storage with an access time comparable to DRAM's, which allows direct execution of microcodes. If this is going to happen, Flash memories will

Manuscript received January 28, 1997; revised May 8, 1997. This work was supported in part by the European Commission Standards, Measurement and Testing Programme in the framework of the "PROPHECY" project, in part by Italian MURST (Ministry of University and Scientific Research), and in part by CNR (National Council of Research).

P. Pavan is with the Dipartimento di Scienze dell'Ingegneria, Università di Modena, 41100 Modena Italy.

R. Bez is with Central R&D, SGS-Thomson Microelectronics, 20041 Agrate Brianza (MI) Italy.

P. Olivo is with the Dipartimento di Ingegneria, Università di Ferrara, 44100 Ferrara Italy.

E. Zanoni is with the Dipartimento di Elettronica e Informatica, Università di Padova, 35131 Padova Italy.

Publisher Item Identifier S 0018-9219(97)05720-4.

0018–9219/97$10.00 © 1997 IEEE

INTEL-1020
8,010,740

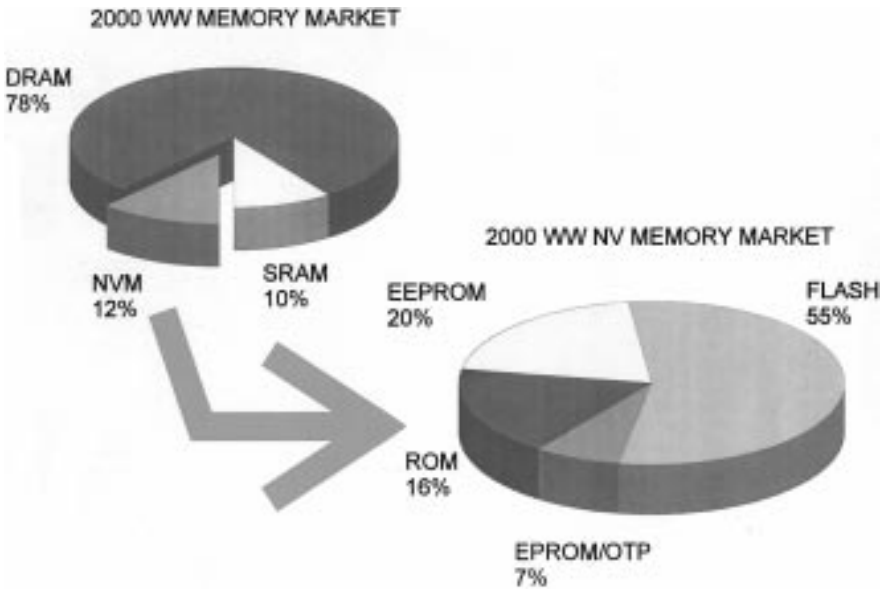

Fig. 1. Forecast for year 2000 worldwide memory market and NVM market. (Source: WSTS/SGS-Thomson, May 1996 [1].)

compete for the same market share with DRAM. Moreover, Flash memories can find interesting applications in personal computer program management: many programs can be stored in Flash chips, without being continuously loaded and unloaded from hard disk partitions, and directly executed.

In the technical sessions on Flash memory technology at the 1990–1996 IEEE International Electron Device Meetings, very advanced technical schemes were proposed for very-high-density Flash memory generations. New concepts have been disclosed for 256-Mb cell processing, and multilevel programming techniques have been emphasized to improve packing density. New operating schemes have been presented to improve scalability and reliability issues.

This paper is a general review of mainstream Flash memory cell technologies. The next section will give an overview on the basic operating principles of a schematic cell, which is commonly depicted as a standard metal-oxide-semiconductor (MOS) transistor with a floating gate (FG) surrounded by a dielectric. Section III reviews the principles of charge-injection mechanisms, namely, hot-electron injection and tunneling. The main concepts described will be referred to throughout the paper. Section IV starts with a thorough overview of the industry-standard Flash cell, which is based on the double-poly stacked-gate cell, and then gives an overview of basic reliability issues inherent to the cell structure itself. Scaling issues are also briefly addressed. Section V gives an overview of the latest Flash structures presented in the literature.

II. Nonvolatile Memories

There is a widespread variety of NVM's, and they all show different characteristics according to the structure of the selected cell and the complexity of the array organization. They all have performance that can go from those of ROM memories, which cannot be reconfigured, to those of information alterability with almost the same flexibility of RAM memories.

A. Programmable ROM—Information Storage and Access

The need for information alterability always contrasts with the need for good data retention. Cells with different characteristics have different applications according to the relevance that the device functional parameter has (absorbed power, programming/erasing speed and selectivity, capacity, and so forth).

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

One solution is to have a transistor with a threshold voltage that can change repetitively from a high to a low state, corresponding to the two states of the memory cell, i.e., the binary values ("1" and "0") of the stored bit. Cells can be "written" into either state "1" or "0" by either "programming" or "erasing" methods. One of the two states is called "programmed," the other "erased." In some kinds of cells, the low-threshold state is called "programmed"; in others, it is called "erased." Though this may induce some confusion, the different terms are related to the different organizations of the memory array. In fact, if a datum has to be stored in a bit of the memory, there are different procedures.

1) The whole memory is erased (i.e., all the cells are driven in the same conductive or nonconductive state) and, after this, the information is programmed in the selected bit; the rest of the array is reprogrammed.
2) Only the byte that includes the bit to change is erased and then reprogrammed with the new information.

3) Only the bit that has to be changed is addressed; the value to be stored is compared with the already stored one and written only if different.

When memories are organized as in cases 1) and 2), there is only one operation that can be performed bit by bit, called "program." The other operation, which is performed on the whole array or on a part of it, is the "erase" operation. When memories are organized as in case 3), both operations can be performed bit by bit but "program" needs a much more complicated array organization.

The "read" operation is performed by applying to the cell a gate voltage that is between the values of the thresholds of the erased and programmed cells and senses the current flowing through the device.

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

$$V_T = K - \bar{Q}/C_{\text{ox}} \tag{1}$$

where K is a constant that depends on the gate and substrate material, doping, and gate oxide thickness, $\bar{Q}$ is the charge weighted with respect to its position in the gate oxide, and C_{ox} is the gate oxide capacitance.

As can be seen, the threshold voltage of the memory cell can be altered by changing the amount of charge present between the gate and the channel, i.e., changing $\bar{Q}/C_{\text{ox}}$. There are many ways to obtain the threshold voltage shift. Two are the most common solutions used to store charge.

1) In traps that are present in the oxide, more precisely at the interface between two dielectric materials. The most commonly used interface is the silicon oxide/nitride interface. Devices obtained in this way are called metal-nitride-oxide-silicon (MNOS) cells [3], [4].
2) In a conductive material layer between the gate and the channel and completely surrounded by insulator. This is the FG device.

In any case, endurance (capability of maintaining the stored information after erase/program/read cycling) and retention (capability of keeping the stored information in time) are the two parameters that describe how "good" and reliable a cell is.

MNOS devices are not used anymore in consumer electronics due to their low endurance and retention. FG devices are at the basis of every modern NVM, particularly for Flash applications.

B. FG Device

The basic concepts and the functionality of an FG device are easily understood if it is possible to determine the FG potential. The schematic cross section of a generic FG device is shown in Fig. 2; the upper gate is the control gate and the lower gate, completely isolated within the gate dielectric, is the FG. The FG acts as a potential well (see Fig. 3). If a charge is forced into the well, it cannot move from there without applying an external force: the FG stores charge.

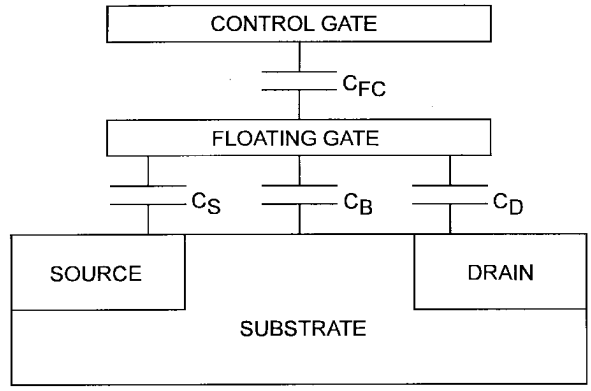

Fig. 2. Schematic cross section of an FG transistor.

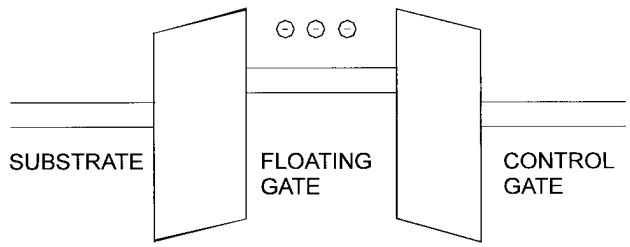

Fig. 3. Energy band diagram of an FG transistor.

The simple model shown in Fig. 2 helps in understanding the electrical behavior of an FG device. C_{FC}, C_S, C_D, and C_B are the capacitances between the FG and control gate, source, drain, and substrate regions, respectively. Consider the case when no charge is stored in the FG, i.e., $\bar{Q} = 0$

$$\begin{aligned}\bar{Q} = 0 = C_{\text{FC}}(V_{\text{FG}} - V_{\text{CG}}) + C_S(V_{\text{FG}} - V_S) \\ + C_D(V_{\text{FG}} - V_D) + C_B(V_{\text{FG}} - V_B)\end{aligned} \tag{2}$$

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

$$V_{\text{FG}} = \alpha_G V_{\text{GS}} + \alpha_D V_{\text{DS}} + \alpha_S V_S + \alpha_B V_B. \tag{3}$$

It should be pointed out that (3) shows that the FG potential does not depend only on the control gate voltage but also on the source, drain, and bulk potentials. If the source and bulk are both grounded, (3) can be rearranged as

$$V_{\text{FG}} = \alpha_G\left(V_{\text{GS}} + \frac{\alpha_D}{\alpha_G}\cdot V_{\text{DS}}\right) = \alpha_G(V_{\text{GS}} + f\cdot V_{\text{DS}}) \tag{4}$$

where

$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_{\text{FC}}}. \tag{5}$$

Device equations for the FG MOS transistor can be obtained from the conventional MOS transistor equations by

replacing MOS gate voltage $V_{\rm GS}$ with FG voltage $V_{\rm FG}$ and transforming the device parameters, such as threshold voltage V_T and conductivity factor β, to values measured with respect to the control gate. If we define for $V_{\rm DS}=0$

$$V_T^{\rm FG} = V_T(\text{floating gate}) = \alpha_G V_T(\text{control gate}) = \alpha_G V_T^{\rm CG} \tag{6}$$

and

$$\beta^{\rm FG} = \beta(\text{floating gate}) = \frac{1}{\alpha_G}\beta(\text{control gate}) = \frac{1}{\alpha_G}\beta^{\rm CG} \tag{7}$$

it is possible to compare the current–voltage (I–V) equations of a conventional and an FG MOS transistor in the triode region (TR) and in the saturation region (SR) [5].

Conventional MOS transistor

$$\begin{aligned} \text{TR}\quad & |V_{\rm DS}| < |V_{\rm GS} - V_T| \\ & I_{\rm DS} = \beta\left[(V_{\rm GS}-V_T)V_{\rm DS} - \frac{1}{2}V_{\rm DS}^2\right] \\ \text{SR}\quad & |V_{\rm DS}| \geq |V_{\rm GS} - V_T| \\ & I_{\rm DS} = \frac{\beta}{2}(V_{\rm GS}-V_T)^2 \end{aligned}$$

FG MOS transistor

$$\text{TR}\quad |V_{\rm DS}| < \alpha_G|V_{\rm GS} + fV_{\rm DS} - V_T| \qquad I_{\rm DS} = \beta\left[(V_{\rm GS}-V_T)V_{\rm DS} - \left(f - \frac{1}{2\alpha_G}\right)V_{\rm DS}^2\right] \tag{8}$$

$$\text{SR}\quad |V_{\rm DS}| \geq \alpha_G|V_{\rm GS} + fV_{\rm DS} - V_T| \qquad I_{\rm DS} = \frac{\beta}{2}\alpha_G(V_{\rm GS}+fV_{\rm DS}-V_T)^2 \tag{9}$$

where β and V_T of (8) and (9) are measured with respect to the control gate rather than with respect to the FG of the stacked gate structure. They are to be read as $\beta(\text{control gate}) = \beta^{\rm CG}$ and $V_T(\text{control gate}) = V_T^{\rm CG}$.

Several effects can be observed from these equations, many of them due to the capacitive coupling between the drain and the FG, which modifies the I–V characteristics of FG MOS transistors with respect to conventional MOS transistors [5].

1) The FG transistor can go into depletion-mode operation and can conduct current even when $|V_{\rm GS}| < |V_T|$. This is because the channel can be turned on by the drain voltage through the $f \cdot V_{\rm DS}$ term in (8). This effect is usually referred to as "drain turn-on."
2) The saturation region for the conventional MOS transistor is where $I_{\rm DS}$ is essentially independent of the drain voltage. This is no longer true for the FG transistor, in which the drain current will continue to rise as the drain voltage increases and saturation will not occur.
3) The boundary between the triode and saturation regions for the FG transistor is expressed by

$$|V_{\rm DS}| = \alpha_G|V_{\rm GS} + f\cdot V_{\rm DS} - V_T| \tag{10}$$

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

4) The transconductance in SR is given by

$$g_m = \frac{\partial I_{\rm DS}}{\partial V_{\rm GS}}_{(V_{\rm DS}=\text{constant})} = \alpha_G\beta(V_{\rm GS} + fV_{\rm DS} - V_T) \tag{11}$$

where g_m increases with $V_{\rm DS}$ in the FG transistor in contrast to the conventional transistor, where g_m is relatively independent of the drain voltage in the saturation region.

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

$$f = -\frac{\partial V_{\rm GS}}{\partial V_{\rm DS}}_{(I_{\rm DS}=\text{constant})} \tag{12}$$

in the saturation region.

Many techniques have been presented to extract the capacitive coupling ratios from simple dc measurements [6]–[8]. The most widely used methods [9], [10] are 1) linear threshold voltage technique, 2) subthreshold slope method, and 3) transconductance technique. These methods require the measurement of the electrical parameter in both a memory cell and in a "dummy cell," i.e., a device identical to the memory cell, but with floating and control gates connected. By comparing the results, the coupling coefficient can be determined. Other methods have been proposed to extract coupling coefficients directly from the memory cell without using a "dummy" one, but they need a more complex extraction procedure [11]–[13].

C. The Reading Operation

Let us consider the case when charge is stored in the FG, i.e., $\bar{Q} \neq 0$. All the hypotheses made above hold true, and the following modifications need to be included.

Equations (4), (6), and (8), respectively, become

$$V_{\rm FG} = \alpha_G V_{\rm GS} + \alpha_D V_{\rm DS} + \frac{\bar{Q}}{C_T} \tag{13}$$

$$V_T^{\rm CG} = \frac{1}{\alpha_G}V_T^{\rm FG} - \frac{\bar{Q}}{C_T\alpha_G} = \frac{1}{\alpha_G}V_T^{\rm FG} - \frac{\bar{Q}}{C_{\rm FC}} \tag{14}$$

$$I_{\rm DS} = \beta\left[\left(V_{\rm GS} - V_T - \left(1 - \frac{1}{\alpha_G}\right)\frac{\bar{Q}}{C_T}\right)V_{\rm DS} + \left(f - \frac{1}{2\alpha_G}\right)V_{\rm DS}^2\right]. \tag{15}$$

Equation (14) shows the V_T dependence on $\bar{Q}$. In particular, the threshold voltage shift ΔV_T is derived as

$$\Delta V_T = V_T - V_{T0} = -\bar{Q}/C_{\rm FC} \tag{16}$$

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

Equation (15) shows that the role of injected charge is to shift the I–V curves of the cell. If the reading biases are fixed (usually $V_{\rm GS} \simeq 5$ V, $V_{\rm DS} \simeq 1$ V), the presence of

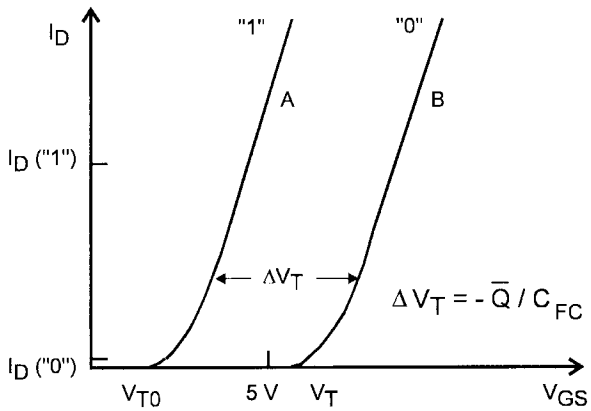

Fig. 4. I–V curves of an FG device when there is no charge stored in the FG (curve A) and when a negative charge $\bar{Q}$ is stored in the FG (curve B) [14].

charge greatly affects the current level used to sense the cell state. Fig. 4 [14] shows two curves: curve A represents the "1" state and curve B the same cell in the "0" state obtained with a 3-V threshold shift. In the defined reading condition I_D ("1") is approximately 100 μA and I_D ("0") $\simeq 0$.

III. Charge Injection Mechanisms

There are many solutions used to transfer electric charge from and into the FG. For both erase and program, the problem is making the charge pass through a layer of insulating material.

The hot-electron injection (HEI) mechanism generally is used in Flash memories, where a lateral electric field (between source and drain) "heats" the electrons and a transversal electric field (between channel and control gate) injects the carriers through the oxide.

The Fowler–Nordheim (FN) tunneling mechanism starts when there is a high electric field through a thin oxide. In these conditions, the energy band diagram of the oxide region is very steep; therefore, there is a high probability of electrons' passing through the energy barrier itself.

It is interesting to notice how these two mechanisms have been deeply investigated for MOS transistors in order to avoid their unwanted degradation effects. In Flash cells, they are exploited to become efficient program/erase mechanisms.

A. Hot Electron Injection

The physical mechanism of HEI is relatively simple to understand qualitatively. An electron traveling from the source to the drain gains energy from the lateral electric field and loses energy to the lattice vibrations (acoustic and optical phonons). At low fields, this is a dynamic equilibrium condition, which holds until the field strength reaches approximately 100 kV/cm [15]. For fields exceeding this value, electrons are no longer in equilibrium with the lattice, and their energy relative to the conduction band edge begins to increase. Electrons are "heated" by the high lateral electric field, and a small fraction of them have enough energy to surmount the barrier between oxide and

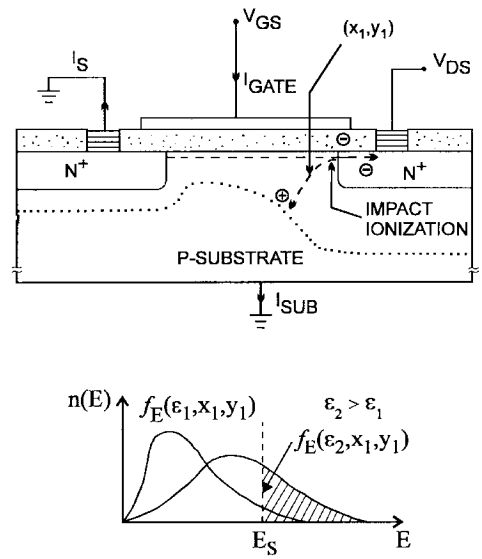

Fig. 5. Schematic cross section of a MOSFET. The energy-distribution function at point X_1, Y_1 is also shown [16], [19].

silicon conduction band edges. For an electron to overcome this potential barrier, three conditions must hold [16].

1) Its kinetic energy has to be higher than the potential barrier.
2) It must be directed toward the barrier.
3) The field in the oxide should be collecting it.

To evaluate how many electrons will actually cross the barrier, one should know the energy distribution $f_E(\mathcal{E}, x, y)$ as a function of lateral field $\mathcal{E}$, the momentum distribution $f_k(E, x, y)$ as a function of electron energy E (i.e., how many electrons are directed toward the oxide), the shape and height of the potential barrier, and the probability that an electron with energy E, wave vector k, and distance d from the Si/SiO_2 interface will cross the barrier. Each of these functions needs to be specified in each point of the channel (see Fig. 5). A quantitative model, therefore, is very heavy to handle. Moreover, when the energy gained by the electron reaches a threshold, impact ionization becomes a second important energy-loss mechanism [17], which needs to be included in models.

Nevertheless, a description of the injection conditions can be accomplished with two different approaches. The HEI current is often explained and simulated following the "lucky electron" model [18]. This model is based on the probability of an electron's being lucky enough to travel ballistically in the field $\mathcal{E}$ for a distance several times the mean free path without scattering, eventually acquiring enough energy to cross the potential barrier if a collision pushes it toward the Si/SiO_2 interface. Consequently, the probability of injection is the lumped probability of the following events [19], which are depicted in Fig. 6.

1) The carrier has to be "lucky" enough to acquire enough energy from the lateral electric field to overcome the

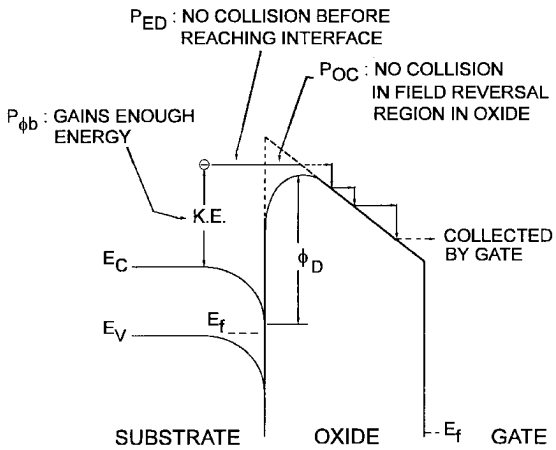

Fig. 6. A schematic energy band diagram describing the three processes involved in electron injection [19].

oxide barrier and to retain its energy after the collision that redirects the electron toward the interface (P_{Φ_b}).

2) The carrier follows a collision-free path from the redirection point to the interface (P_{ED}).
3) The carrier can surmount the repulsive oxide field at the injection point, due to the Schottky barrier lowering effect, without suffering an energy-robbing collision in the oxide (P_{OC}).

Although this simple model does not fit precisely with some experiments, it allows a straightforward and quite successful simulation of the gate current.

A more rigorous model is based on the quasi-thermal equilibrium approach [20], [21]. It assumes that the electron can be treated as a gas in quasi-thermal equilibrium with the electric field. This electron gas is characterized by an "effective temperature," which is different from the lattice temperature. The model establishes a nonlocal relation between the effective electron temperature and the drift-field [20]. Thus, the carrier probability to acquire certain energies depends on the complete profile of the electric field in the channel region [22].

Both models allow the prediction of the following relation between the substrate current I_{sub} and the injection current: I_G (Fig. 7)

$$I_G/I_{\mathrm{ch}} \sim I_{\mathrm{sub}}/I_{\mathrm{ch}} e^{-\Phi/\Phi_i} \tag{17}$$

where I_{ch} is the channel current, Φ_i is the impact-ionization energy, and Φ is the energy barrier seen by the carriers to be injected in the oxide [19]. This latter barrier has to be corrected for the image-force lowering [23] and tunneling components [24] of the gate current.

The substrate current is composed of holes generated by impact ionization in the drain region. Holes are always generated since the energy ionization threshold Φ_i ($\sim$1.6 V) is lower than the injection energy barrier Φ ($\sim$3.2 V). Some holes can acquire enough energy from the lateral electric field to be injected into the oxide, thus degrading it. The ionization process also generates a lot of carriers that can be injected in regions of the oxide, where they can be trapped

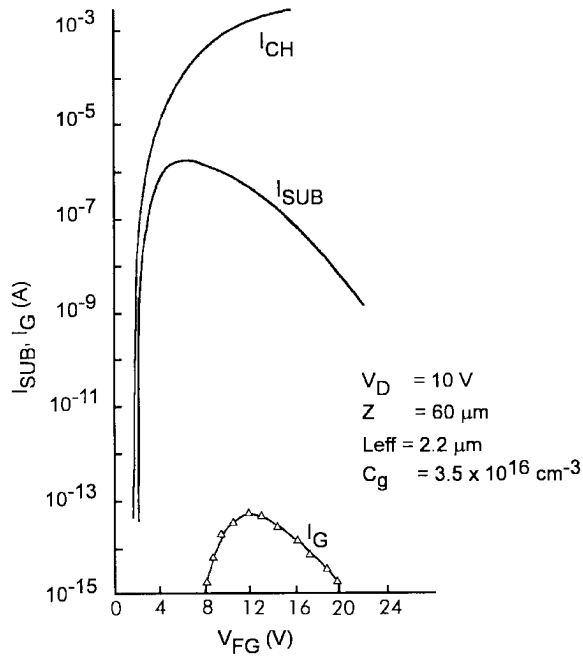

Fig. 7. Channel and substrate currents (continuous measurements) and gate current as a function of FG voltage (device geometries and experimental setup in legend) [16].

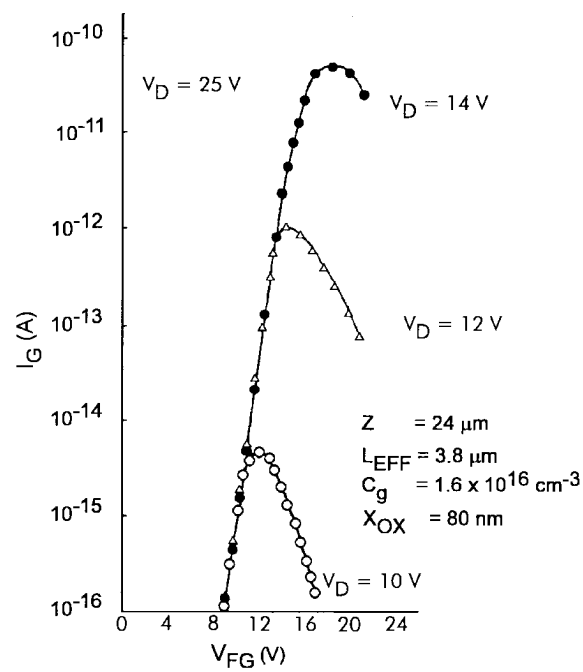

Fig. 8. Gate current as a function of FG voltage with drain voltage as a parameter [16].

near the interface or generate interface states [22], thus degrading device performances.

Fig. 8 [16] shows the injected current I_G measured in an FG device with an indirect technique based on the relation $I_G = dQ_{\mathrm{FG}}/dt \approx C_{\mathrm{FC}}\Delta V_{\mathrm{FG}}/\Delta t$, where Q_{FG} is the FG charge, C_{FC} is the coupling capacitance between the control and floating gates, and V_{FG} is the FG potential. The shape of these curves is correlated to the injection-mechanism dependence on V_G and V_D. When $V_G < V_D$,

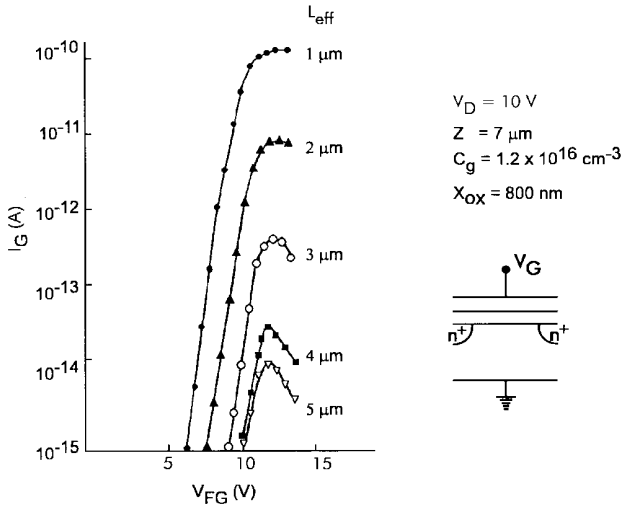

Fig. 9. Gate current as a function of FG voltage with effective channel length as a parameter [16].

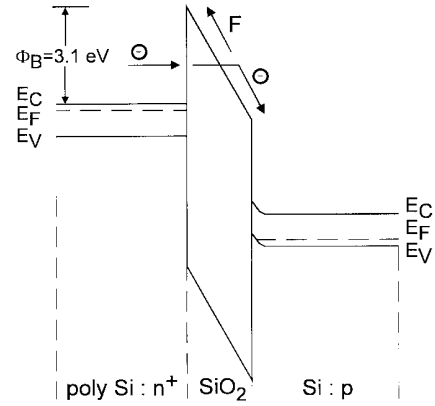

Fig. 10. FN tunneling through a potential barrier in a MOS structure.

there is a point along the channel where the transverse electric field in the oxide reverses its direction and rejects electrons rather than collecting them. The pinch-off point is closer to the source side than the inversion point [16]. Thus, the injection is dominated by hot electrons created in the relatively low lateral field region near the pinch-off point. As V_G increases, the average lateral field decreases [25] but the point of injection is shifted closer to the drain edge due to the change in the inversion point with V_G. As a result, part of the distribution of electrons available for injection is in a higher lateral electric field region, leading to a rapid increase in I_G. When V_G becomes greater than V_D, the hot electron distribution is subjected to the average lateral field near the drain, which decreases on increasing V_G, thus reducing the gate current.

I_G also depends on channel length, as shown in Fig. 9. A decrease in the channel length results in an increase of I_G due to the increased lateral electric field, even at lower V_G's. This is due to the coupling between the FG and drain, which is higher in shorter devices. Note that the gate current in the injection-limited region exhibits a decreasing dependence on $V_{\rm FG}$ as the channel length is decreased. This stems from the considerable increase in the lateral electric field $\mathcal{E}(x, \bar{y}(x))$ for shorter channel devices such that the reduction in this field due to the increase in $V_{\rm FG}$ will be noticeable only at higher values of gate voltage [16].

B. Fowler–Nordheim Tunneling

In the framework of quantum mechanics, the solutions of the Schrödinger equation represent a particle. The continuous nonzero nature of these solutions, even in classically forbidden regions of negative energy, implies an ability to penetrate these forbidden regions and a probability of tunneling from one classically allowed region to another [26]. The concept of tunneling through a potential barrier applies well to MOS structures with thin oxide. Fig. 10 shows the energy-band diagram of a MOS structure with negative bias applied to the metal electrode with respect to the p-doped silicon substrate. The probability of electron tunneling depends on either the distribution of occupied states in the injecting material or the shape, height, and width of the barrier.

Using a free-electron gas model for the metal and the Wentzel–Kramers–Brillouin (WKB) approximation for the tunneling probability [27], one obtains the following expression for current density [28]:

$$J = \frac{q^3 F^2}{16\pi^2 h^2 \Phi_B} \exp\left[-4(2m^*_{\rm ox})^{1/2} \Phi_B^{3/2} / 3\hbar q F\right] \quad (18)$$

where Φ_B is the barrier height, $m^*_{\rm ox}$ is the effective mass of the electron in the forbidden gap of the dielectric, h is the Planck's constant, q is the electronic charge, and F is the electric field through the oxide.

Fig. 11 [29] shows $\log J$ versus F. Since the field is roughly the applied voltage divided by the oxide thickness, a reduction of oxide thickness without a proportional reduction of applied voltage produces a rapid increase of the tunneling current. With a relatively thick oxide (20–30 nm) one must apply a high voltage (20–30 V) to have an appreciable tunnel current. With thin oxides, the same current can be obtained by applying a much lower voltage. An optimum thickness (about 10 nm) is chosen in present devices, which use the tunneling phenomenon to trade off between performance constraints (programming speed, power consumption, etc.), which would require thin oxides, and reliability concerns, which would require thick oxides. In fact, in Fig. 11, it is evident that with a field of 7 MV/cm, the current density is about 10^{-8} A/cm^2, while with a field of 10 MV/cm it is about 10^{-1} A/cm^2. There is a variation of about seven orders of magnitude in tunnel current. A slightly greater field range allows a difference of 12 orders of magnitude. The tunneling-injection mechanism is widely used in NVM, particularly in EEPROM. There are three main reasons for this choice: first, tunneling is a pure electrical mechanism; second, the involved current level is quite low and thus allows the internal generation of supply voltages needed for all operations; third, it allows one to obtain the time to program ($<$ 1 ms) 12 orders of magnitude shorter than retention time ($>$10 y), which is a fundamental request for all NVM technologies.

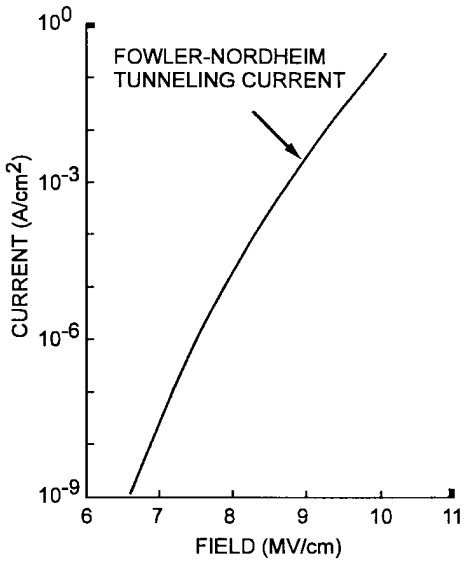

Fig. 11. FN tunneling current as a function of electric field [29].

On the other hand, the exponential dependence of tunnel current on the oxide-electric field causes some critical problems of process control because, for example, a very small variation of oxide thickness among the cells in a memory array produces a great difference in programming or erasing currents, thus spreading the threshold voltage distribution in both logical states. A very good process control is therefore required. Moreover, the tunneling currents may become important in device reliability at low fields either in the case of bad-quality tunnel oxides or when thin oxides are stressed many times at high voltages [30]. In fact, bad-quality oxides are rich of interface and bulk traps, and trap-assisted tunneling is made possible since the equivalent barrier height seen by electrons is reduced and tunneling requires a much lower oxide field than 10 MV/cm.

The oxide defects (whose density increases at decreasing oxide thickness) must be avoided to control program/erase characteristics and to have good reliability. In any case, frequent program and erase operations induce an increase of trapped charge in the oxide. This affects the barrier height, which is lower in the case of positive and higher in the case of negative trapping, respectively, thus increasing or decreasing the tunnel currents.

Although the simple and classic form of FN current density [(18)] is in quite good agreement with experimental data, many features have been still undervalued: the temperature dependence of the phenomenon, the quantum effects at the silicon interface, the influence of band bending at the Si/SiO_2 interface, and the voltage drop in silicon, the fact that the correct statistics for electrons are not Maxwellian but Fermi–Dirac, and the collision-broadening barrier lowering [31]. These features are of great importance in device simulation either to develop a quite general model of the tunneling injection to enable the full simulation of novel structures or to have a deep understanding of the influence of scaling in device performances.

First of all, the classic theory is based on the assumption that electrons, as well as holes, at the semiconductor surface can be treated as a three-dimensional gas of free-particles with a Boltzmann distribution of energy. But when the silicon surface is inverted or accumulated (which are the usual conditions during tunnel injection in MOS devices), these particles are confined into a narrow potential well, so the quantum-mechanics laws require their motion perpendicular to the interface to be quantized. Thus, the correct treatment is a two-dimensional quantum-mechanical gas [32]. Briefly, the results of this treatment are as follows [33].

1) The barrier height is voltage dependent and is lower than the classical one.
2) The oxide field is lower than the classical one due to a much greater voltage drop in the silicon substrate.

It is possible to rewrite (18) in the simplest form

$$J = A \cdot F^2 \exp[-B/F] \tag{19}$$

and to use A and B as functions of the electric field, which include quantum effects [34]. This approach is quite satisfactory in a lot of cases but leads to different values of A and B depending on the injecting electrode and on device polarization.

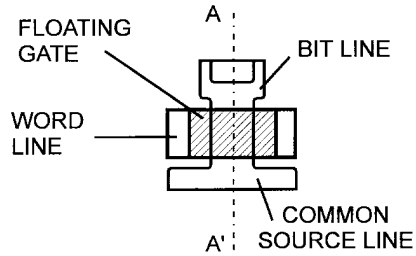

Fig. 12. Layout of a typical T-shaped double-polysilicon stacked gate cell.

IV. Industry-Standard Flash Cells

An EPROM memory cell is programmed via channel hot electron (CHE) injection and erased via ultraviolet light. It is composed of only one transistor. The T-shaped cell (Fig. 12) allows a very tight memory array and then high density (larger than 16 Mb). On the other hand, the advantages of this array organization are paid in terms of versatility: program is by single bit but erase is on the whole array.

EEPROM memory cells are programmed and erased via FN tunneling and are composed of two transistors: storage and select. In this way, they allow byte alterability and good endurance performances with more than 100 000 cycles. On the other hand, two transistors per cell require a larger area and consequently reduce the achievable density (less than 1 Mb).

A Flash memory cell represents the synthesis of EPROM and EEPROM, since it is programmed and erased electrically but composed by a single transistor.

The first cell based on this concept was presented in 1979 [35]; the first commercial product, a 256-K memory chip,

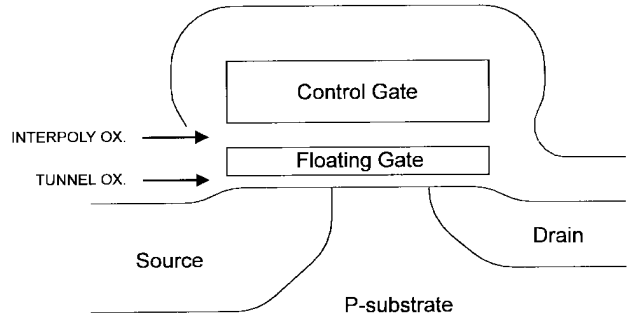

Fig. 13. Schematic cross section of a Flash cell along line A-A′ in Fig. 12.

was presented by Toshiba in 1984 [36]. The market did not take off until this technology was proven to be reliable and manufacturable [37].

The first Flash prototypes needed an external supply voltage for programming and external management of the erasing algorithm. They featured only a bulk-erase capability and their endurance was very poor (less than 10 000 cycles). As an advantage versus EPROM's, they offered just an electrical-erase capability. Modern Flash memories have an embedded microcontroller to manage the erasing algorithm and offer sector erase capability and single power supply.

The growing demand of high-density NVM for the portable computing and telecommunications market has encouraged serious interest in Flash memory with the capability of multilevel storage [38], [39] and low-voltage operation [40]–[42]. Multilevel storage implies the capability of storing two bits in a single cell. To do so, four different threshold voltages need to be correctly identified in the FG transistor. The program-verify procedure trades off between the accuracy of V_T and programming speed and imposes serious limitations to V_T levels possibly stored in cells. It is therefore mandatory to accept a reduction of margin among V_T levels [43].

A. Basic Operations

Fig. 13 shows the cross section of an industry-standard Flash cell. This cell structure was presented for the first time by INTEL in 1988 and named ETOX[1] (EPROM tunnel oxide) [44]. Though it is derived from an EPROM cell, there are a few meaningful differences.

First, the oxide between the substrate and FG is very thin ($\approx$ 10 nm). Therefore, if a high voltage is applied at the source when the control gate is grounded, a high electric field exists in the oxide, enabling tunneling effects from the FG to the source. This bias condition is dangerously close to the breakdown of the source-substrate junction. Therefore, the source diffusion is realized differently from the drain diffusion, which does not undergo such bias conditions. To do so, a new mask is added to the technological process to discriminate source and drain implants. The cell is not symmetrical, but this is the only difference with the standard EPROM process. It is a great advantage, since all the accumulated experience in process development can be used to produce these devices.

[1] ETOX is a trademark of INTEL.

Table 1 Source, Control Gate, and Drain Biases During Operations of a Typical Flash Cell. Typical Reference Values Can Be $V_{cc} = 5$ V, $V_{pp} = 12$ V, $V_{dd} = 5 - 7$ V, and $V_{\rm read} = 1$ V

	SOURCE	CONTROL GATE	DRAIN
READ	GND	V_{cc}	V_{read}
PROGRAM	GND	V_{pp}	V_{dd}
ERASE	V_{pp}	GND	FLOAT

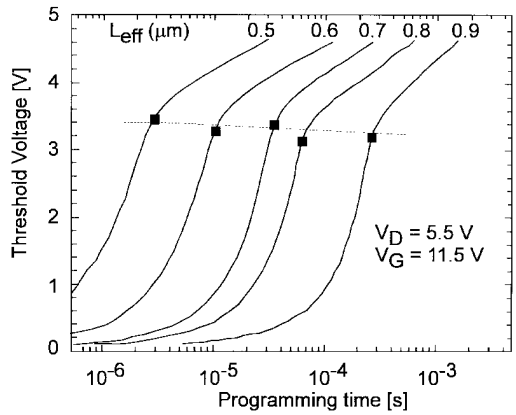

Fig. 14. Programming curves of Flash cells with different channel lengths; $t_{\rm ox}$ is 12 nm [46].

Oxide/nitride/oxide (ONO) interpoly dielectrics are used in Flash memories. Interpoly dielectric thickness heavily influences program/erase speed and the magnitude of read current for an industry-standard Flash cell [45]. Low defect density and long mean time to failure, together with charge retention capability, are important reliability issues.

Flash-cell read, program, and erase bias configurations are summarized in Table 1. Special attention needs to be given to erase, which is the most critical operation.

1) Program: HEI is used to move charge in the FG, thus changing the threshold voltage of the FG transistor. Programming is obtained by applying pulses to the control gate and to the drain simultaneously when the source is grounded. This operation can be performed selectively by applying the pulse to the word line (WL), which connects the control gates, and biasing the bit line (BL), which connects the drains. Hot electrons are injected in the FG, and the threshold voltage of the selected transistor becomes high. The change in threshold voltage depends upon the width of the programming pulse. To have a voltage shift of around 3, 3.5 V, a pulse width with typical values in the 1–10 μs range must be applied. See the curve corresponding to $L_{\rm eff} = 0.6$ μm in Fig. 14 [16]. A rapid change in cell V_T occurs initially. Then, as the FG potential drops below the drain potential, V_T saturates. At this point, we can define an intrinsic threshold. The electric field in the tunnel oxide close to the drain reverses and electron injection into the FG is much less favorable [47]. Intrinsic threshold voltage shift roughly does not depend on the channel length

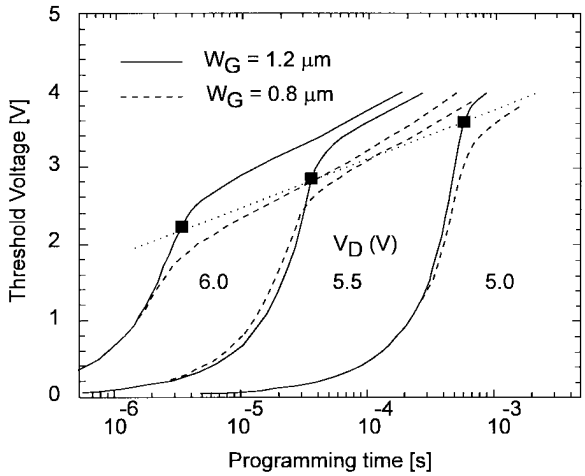

Fig. 15. Programming curves of Flash cells with different coupling ratios and at different V_D; $L_{\rm eff} = 0.7$ μm, $t_{\rm ox} = 12$ nm [46].

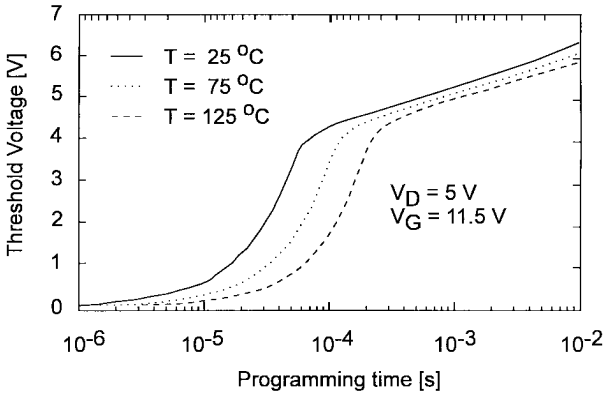

Fig. 16. Programming curve of a Flash cell at different temperatures; $L_{\rm eff} = 0.7$ μm, $t_{\rm ox} = 12$ nm.

of the cell (Fig. 14) but depends on the coupling ratios, i.e., the overlap between the FG and control gate on field oxide (Fig. 15, [46]). From the same figure, one can see that intrinsic threshold voltage shift also depends linearly on drain voltage. Temperature also has an influence on programming speed (Fig. 16): a higher temperature reduces the number of hot electrons available for injection into the FG, hence retarding the programming characteristics [46].

2) Erase: The erase operation requires a high voltage pulse to be applied to the source (common to all the transistors in the array/block) when control gates (WL) are grounded and drains (BL) floating. Before applying the erase pulse, all the cells in the array/block are programmed to start with all the thresholds approximately at the same value. After that, an erase pulse having a controlled width is applied. The threshold shift depends on source voltage (Fig. 17) and, as a rule of thumb, a one-order-of-magnitude increase in erasing time occurs for each volt reduction in source voltage [47]. Threshold voltage depends on oxide thickness (Fig. 18) [48], as explained in Section III-B. From the same figure, one can infer that after electrical erase, cells with the same oxide thickness but different initial values of threshold voltage will reach the same

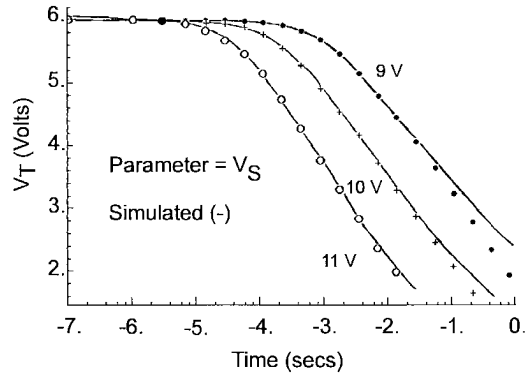

Fig. 17. Erase curves of a cell with $T_{\rm ox} = 12$ nm, when different source voltages are applied [47].

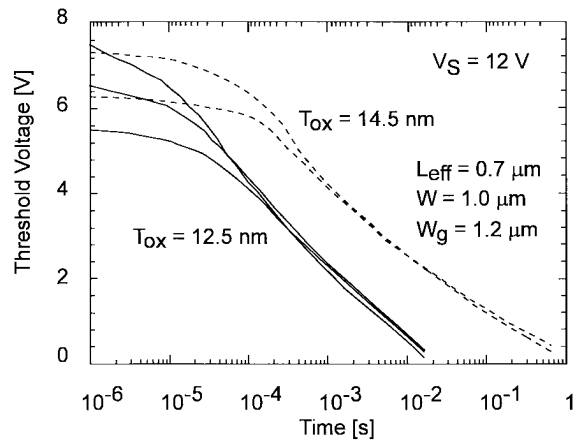

Fig. 18. Erase curves of two Flash cells having different oxide thicknesses and the same $L_{\rm eff}$ [48].

threshold voltage at the end of the erase operation. Since FG transistors in the array may have slightly different gate oxide thicknesses, and the erase mechanism is not self-limiting, after an erase pulse we may have "typical bits" and "fast erasing bits" (Fig. 19) [49]. Subsequently, the whole array/block is read to check whether all the cells are erased or not. If not, another erase pulse is applied and another read operation follows. This algorithm is applied until all the cells have threshold voltages lower than the "erase verify level." At the end of this procedure, cells will not have all the same threshold voltages but their thresholds will be on a Gaussian distribution, except for a tail of bits that erase faster, which will be analyzed in Section IV-B-4. Typical erasing times are in the range 100 ms–1 s.

Electrical erase is achieved via FN tunneling of charge from the FG to the source. To have a junction that can sustain the high applied voltages without breaking down, the source junction needs to be carefully designed. A new mask is added to the process to allow for a lighter and deeper junction. Details of the source junction are sketched in Fig. 20. A high electric field through the tunnel oxide means that even the electric field at the surface of the silicon is very high, and this can give rise to a leakage current due to band-to-band tunneling (BBT) or breakdown of the source/substrate junction.

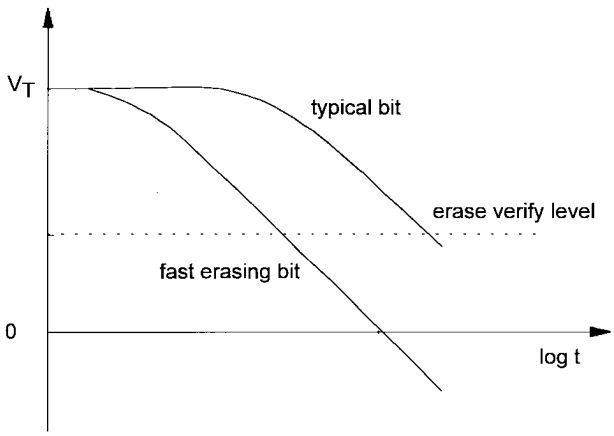

Fig. 19. Erase curves of a "typical bit" and of a "fast erasing bit" in a Flash array [49].

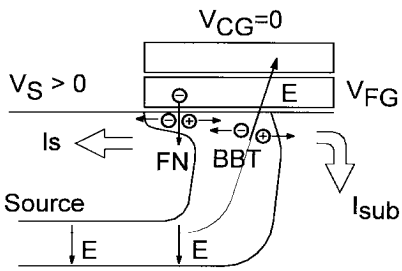

Fig. 20. Detail of the n^+ source junction in a Flash cell showing band-to-band tunneling (BBT) and FN currents $V_S > 0$ and $V_{\rm CG} = 0$ V [47].

If band bending is higher than the energy gap of the semiconductor, and the surface electric field is higher than 1 MV/cm, tunneling of electrons from the valence band to the conduction band becomes significant, and holes are left in the valence band. Electrons are collected at the source terminal; holes are injected into the substrate, thus generating a substrate leakage current. This substrate current depends only on the vertical electric field in the oxide, i.e., on the voltage drop between source and gate. The lateral electric field does not allow the inversion layer to be generated at the n^+-Si/SiO_2 interface and leads the space charge region in deep depletion, sweeping all the free carriers. When source voltage is high enough, impact ionization becomes significant and contributes to the leakage current, thus starting the breakdown mechanism. The minimum voltage to start BBT decreases on decreasing the oxide thickness, and this is one of the major scaling limits. Generated holes can gain enough energy to be injected in the oxide where they are trapped at the Si/SiO_2 interface.

Source breakdown is one of the major limiting factors to erase time reduction, since the higher the voltage applied to the source, the shorter the erasing time. A solution to the problem is achieved by optimizing the source junction profile to a more gradual one in order to reduce the electric field at the junction and, consequently, the substrate current of some order of magnitude.

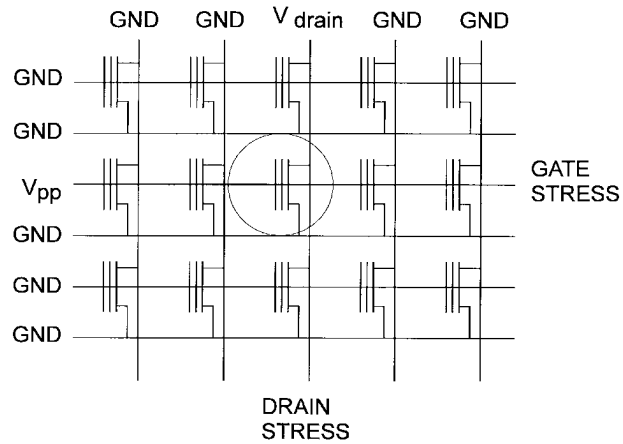

Fig. 21. Programming disturbs.

In a conventional dual-voltage Flash, where besides V_{cc} (3 or 5 V) a high-voltage V_{pp} (about 12 V) is available, erasing is obtained by applying a high positive voltage to the source region (V_S) while the WL terminal (control gate of the memory cell) is grounded. In a single-voltage Flash, the lack of the high voltage V_{pp} implies the on-chip generation of a negative voltage by means of charge pumps. In fact, in this case, the necessary voltage drop between the source and the control gate is obtained by applying V_{cc} to the source and a negative voltage $V_{\rm GN}$ to the control gate [50]–[52]. No matter how the voltage drop is obtained, in both cases the high electric field in the oxide between the FG and source gives rise to a gate current due to FN tunneling. Simultaneously, the high electric field in the silicon is responsible for the source/substrate current due to BBT tunneling, which is a function of source voltage.

B. Reliability

The reliability issues for EPROM and EEPROM memory cells are both present in Flash memories. The confidence in Flash memory reliability has grown together with the understanding of memory-cell failure mechanisms. Cycling and retention experiments are performed to investigate Flash-cell reliability.

The high degree of testability allows the detection at wafer level of latent defects, which may cause single bit failures related to programming disturbs, data retention, and premature oxide breakdown, thus making Flash memories more reliable than full-featured EEPROM's at equivalent density [53]. Flash arrays are verified analyzing array disturbs and erase-threshold distribution. New architecture solutions, however, may open new issues on Flash array reliability.

1) Programming Disturbs: Consider an array as in Fig. 21. If we want to program the highlighted transistor, a high voltage ($V_{pp} = 12$ V) is applied to the WL and a sufficiently high voltage ($V_{\rm drain} = 5$–7 V) is applied to the BL to generate hot electrons to program the cell. In this bias condition, though, there are two major disturbs, one due to the high voltage applied to the WL and to the transistors on the same line, the second to the medium-

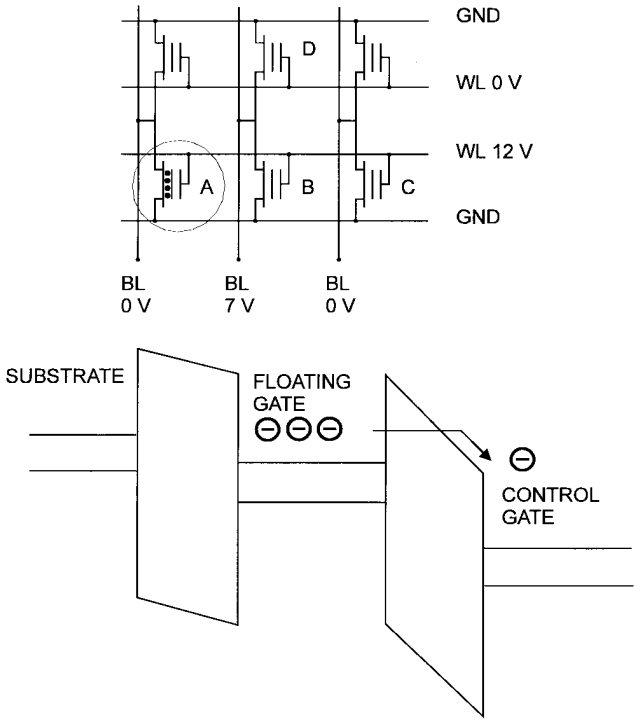

Fig. 22. Programming disturbs—dc erasing of a programmed cell. Cell A is programmed, cell C is nonprogrammed.

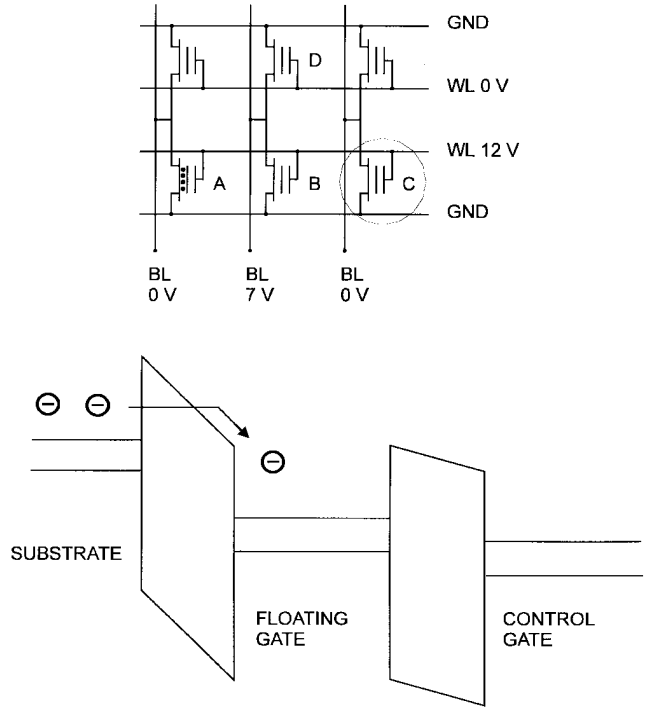

Fig. 23. Programming disturbs—dc programming of a nonprogrammed cell. Cell A is programmed, cell C is nonprogrammed.

high voltage applied to the BL and to the transistors on the same column.

High voltages applied to the WL can stress the gate of transistors that have their gate connected to the WL but are not selected. There might be tunneling of electrons from the FG to the control gate through the interpoly oxide in all the programmed cells, i.e., in those cells where the FG is filled with electrons, since they have 12 V applied to the gate and 0 V on both source and drain. This is the "dc-erasing" disturb (Fig. 22), which induces charge loss and therefore reduces the margin for the high level of threshold voltage.

There might be also tunneling of electrons from the substrate to the FG in all the nonprogrammed cells, i.e., in those cells where the FG is "empty." This is the "dc-programming" disturb (Fig. 23), which induces charge-gain and reduces the margin for the low level of threshold voltage.

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

A relatively high voltage ($V_{\text{drain}} = 5$–7 V) applied to the BL can stress the drains of all FG transistors in the same column. Namely, in cells which share the BL with cells which are to be programmed, electrons tunnel from the FG through the gate oxide to the drain [54]; moreover, holes can be generated via impact-ionization in the substrate and then injected in the FG. This disturb, called "drain disturb" (Fig. 24), causes charge loss and, consequently, a decrease in the high value of the threshold voltage. The same disturb can result from extensive reading cycles and can be used as a gate oxide quality monitor.

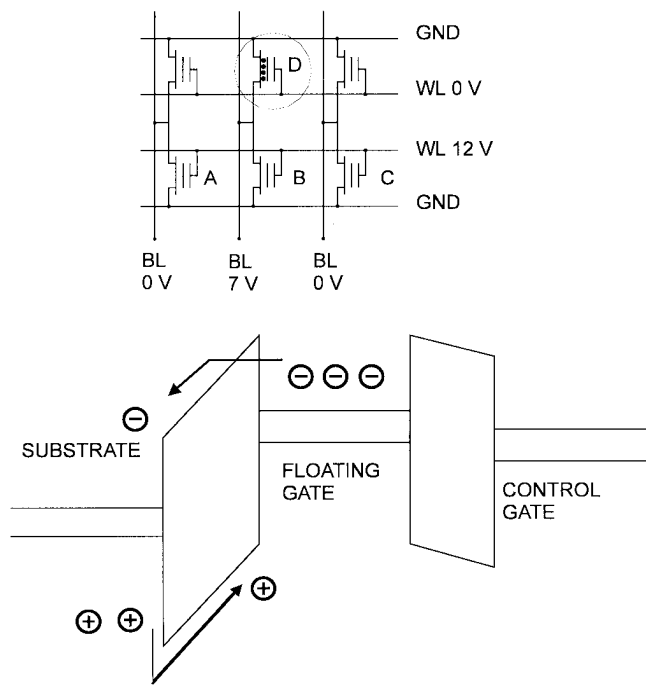

Fig. 24. Programming disturbs—drain disturb. Cell D is programmed.

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

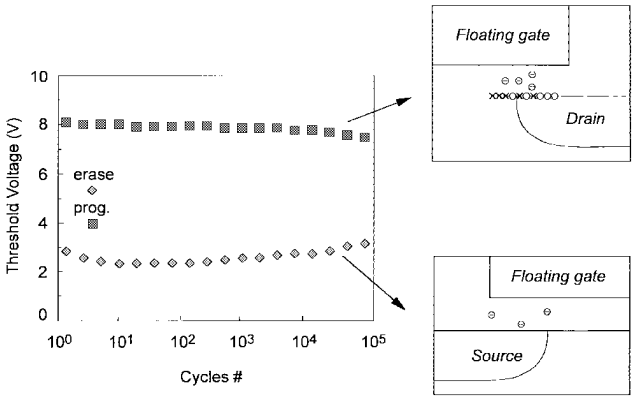

Fig. 25. Threshold voltage window closure as a function of program/erase cycles on a single cell [53].

2) Retention: Fast program and erase operations require high voltages and currents through thin oxides, which in turn are easily degraded. In modern Flash cells, FG capacitance is approximately 1 fF. A loss of only 1 fC can cause a 1-V threshold voltage shift. If we consider the constraints on data retention in ten years, this means that a loss of less than five electrons per day can be tolerated.

Mechanisms that lead to charge loss or charge gain can be divided into two categories: extrinsic and intrinsic. The former are due to defects in the device structure; the latter are due to the physics mechanisms that are used for program and erase operations.

a) Intrinsic mechanisms: Intrinsic mechanisms that contribute to charge variations are field-assisted electron emission, thermionic emission, and electron detrapping.

The first mechanism, field-assisted electron emission [55] consists of the motion of electrons stored in the FG of a programmed cell, which can migrate to the interface with the oxide and from there tunnel into the substrate, thus causing charge loss. If the cell is erased, i.e., has a low threshold voltage, the opposite injection can happen. Experiments have demonstrated that the leakage current due to these mechanisms depends on the floating-to-control-gate coupling coefficient α_G and on the stress level [55]. The probability of an electron's passing through the oxide barrier due to tunneling depends upon the voltage drop between the FG and substrate; FG potential depends on control-gate potential through α_G. Therefore, the charge induced on the FG during the program operation depends on α_G.

The charge Q stored on the FG decreases either on decreasing α_G (since electron injection during the program operation is less efficient) or on increasing the stress level, i.e., the negative charge trapped in the oxide. The leakage current depends exponentially on the electric field (it is a tunneling phenomenon), and the electric field around the FG is

$$E = \frac{Q}{2\epsilon_{\text{ox}}\sigma} \tag{20}$$

where ϵ_{ox} is the silicon dioxide dielectric constant and σ is the FG area.

The second mechanism of charge loss, thermionic emission, is a mechanism of emission of carriers above the potential barrier. At room temperature, the phenomenon is negligible, but it becomes relevant at high temperatures [56].

Last, detrapping of electrons in the gate oxide is a charge-loss mechanism that reduces the program threshold voltage.

b) Extrinsic causes: Extrinsic causes that can influence the charge storage are oxide defects (which increase on decreasing oxide thickness) and ionic contamination. Oxide defects can cause charge loss or gain [56], [57]. In fact, if the cell is programmed, its FG has a negative potential due to the stored charge. This potential induces an electric field in the oxide surrounding the FG itself; in thin oxides, these electric fields can be as high as some MV/cm. Therefore, defects can induce conductive paths, which tend to program the cell. If the cell is overerased and stores a positive charge, the electric field will induce charge gain.

Ionic contamination is a big issue in every nonvolatile memory technology [54], [56]–[58]. Ions, usually positive ones, are attracted to the FG which is negatively charged, thus shielding its effects and inducing charge loss. Memory chips can be affected by contaminations, which, during passivation deposition, can penetrate through defects in passivation glasses or from chip edges. The quality of passivation layers has to be increased in order to reduce this effect.

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

3) Endurance: Cycling is known to cause a fairly uniform wear-out of cell performance [59], which eventually limits Flash memory endurance. A typical result of an endurance test on a single cell is shown in Fig. 25 [53]. As the experiment was performed applying constant pulses, the variations of program and erase threshold levels give a measure of oxide aging. In real devices, this corresponds to longer program/erase times.

The evolution of erase threshold voltage is similar to that typically observed in EEPROM cells. It reflects the dynamics of net fixed charge in the tunnel oxide as a function of the injected charge [60]: the initial lowering of the erase V_T is due to a pileup of positive charge, which enhances tunneling efficiency, while the long-term increase of the erase V_T is due to the generation of negative traps. The reduction of program threshold voltage at high cycling numbers has been explained in [61]; it is attributed to oxide traps and interface state generation at different locations depending on bias conditions. At the beginning of programming, there is no charge in the FG, and the FG transistor is in the linear region. Interface states are created over the drain region, and their influence is negligible. At the end of the programming cycle, there is charge stored in the FG, which is now at a lower voltage than the drain; the FG transistor is in saturation, and interface states are created at the drain side of the channel, producing a detectable degradation of device performance.

4) Erase Distribution: In Flash memory integrated circuits, the complete erase operation is indeed a sequence of elementary erase operations. A first erase pulse is generated

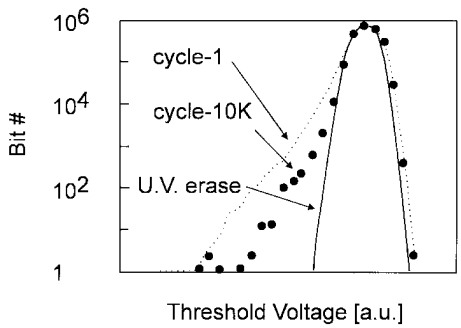

Fig. 26. Threshold voltage distribution after different erase procedures: UV erase (solid curve) after the first cycle (dotted line) and after 10 K cycles (dots) [53].

internally and sent to the logic circuitry, which controls the erase operation and is integrated in the same chip with the memory itself. An algorithm controls whether or not the erase operation is completed. The verification of the complete erasure of all the cells in a block, which can be very large, is one of the biggest issues in Flash technology. The parameter that is checked after erase is the value of the threshold voltages of the erased cells; these values have a distribution that is specific to the process. This distribution spreads around an average value and needs to be easily controlled for each process and to have a small variance. Fig. 26 shows the threshold voltage distribution for a 1-Mb Flash device [53]. The distribution seems to be Gaussian, but it is not symmetrical toward lower values. Another way to plot this kind of data is shown in Fig. 27 [29]. In this case, the cumulative percentage of erased cells is reported versus the threshold voltage of the erased cell (data refer to different devices). As can be seen, a high percentage (99.9%) of the cells have a very small variation of the threshold voltage (0.5 V), and only 0.01% show very large threshold variations. This very small percentage has a great relevance. It is used as a process monitor since it is the limiting factor for the whole Flash technology. In fact, if we have a negative or zero threshold voltage, every single erased cell would be normally on in the reading condition since there is not a selection transistor; if we want to read a single cell, we read all the erased cells in the same column, causing logic errors.

The exponential tail in threshold voltage distribution (see Fig. 26) represents a large population of cells that erase faster than typical bits. This population is too large to be attributed to extrinsic defects, and it is believed to be related to statistical fluctuations of oxide charge and to the structure of the injecting electrode [53]. Positive charges in the tunnel oxide and irregular polycrystal grains may induce a local increase of the electric field, thus enhancing current injection locally and making individual cells erase faster than average. This explanation is consistent with the observation of the narrowing of the tail on increasing the number of program/erase cycles performed on the array (Fig. 26). In fact, a larger current density corresponds to a higher negative trap generation rate, resulting in faster aging. The generated negative charge partially neutralizes electric field peaks, making the current injection more uniform and the erasing speed of tail bits closer to that of typical bits [53].

A relevant mechanism of single bit failure during program/erase cycling is the occurrence of an "erratic bit" [53], [62], [63]. An erratic bit shows an unstable and unpredictable behavior in erasing since its erase threshold voltage changes randomly from cycle to cycle, from the bulk Gaussian distribution to the lower part of the tail. This behavior is expected to be due to hole trapping in the tunnel oxide. WKB calculations [62], [63] have shown that the statistical distribution of hole traps gives a low but finite probability of having clusters of three or more positive charges whose combined electric field effect induces a huge local increase in the tunnel current. In this condition, trapping/detrapping of an individual positive charge causes a detectable change in the erasing speed and threshold level. Since this behavior is due to statistical fluctuations of intrinsic oxide defects, erratic bit occurrence can be reduced by process optimization but cannot be completely eliminated. Therefore, design solutions have been developed to cope with this problem at the circuit level.

Other failures are related to the erase mechanism. Since FN tunneling is not self-limiting, it can lead to overerasing of the cells, i.e., more electrons than those which have been trapped are removed from the FG. The device has less negative charges than in the nonprogrammed case and a net positive charge is now present, thus transforming the device from an enhancement to a depletion device. The window that includes the distribution of erased cells is between the lower limit set by the overerasing (around 0.5 V) and an upper limit given by the peak of the Gaussian.

The aforementioned erase algorithm is used to compensate variations and to avoid overerasing. The algorithm is based on an erase procedure followed by a read procedure in the whole array. Initially, an erase pulse shorter than one-tenth of the typical erase time is applied, and all the bits in the array are checked to verify whether there are threshold voltages higher than a prefixed value. If there are cells that are not erased enough, another pulse is applied. The algorithm is applied until there are no cells with a threshold voltage higher than the prefixed value. There is no check on the lower threshold, since it is known that it is going to be 1.5–2 V lower than the higher one. By applying many erasing pulses while running the algorithm, we do not change significantly the threshold of the "already erased" cells. In fact, erase V_T changes following the logarithm of the erasing time (Fig. 28). Therefore, all of the pulses that are applied after the one that erases the cell do not significantly change the erased threshold voltage distribution.

The trend of decreasing supply voltages for new Flash generation imposes a tight erase V_T distribution. Besides the already mentioned method of iterating the erase algorithm, two self-convergent methods have been proposed, both relying on reprogramming techniques after erasure: hot-carrier injection (HCI) reprogramming [64] (i.e., ap-

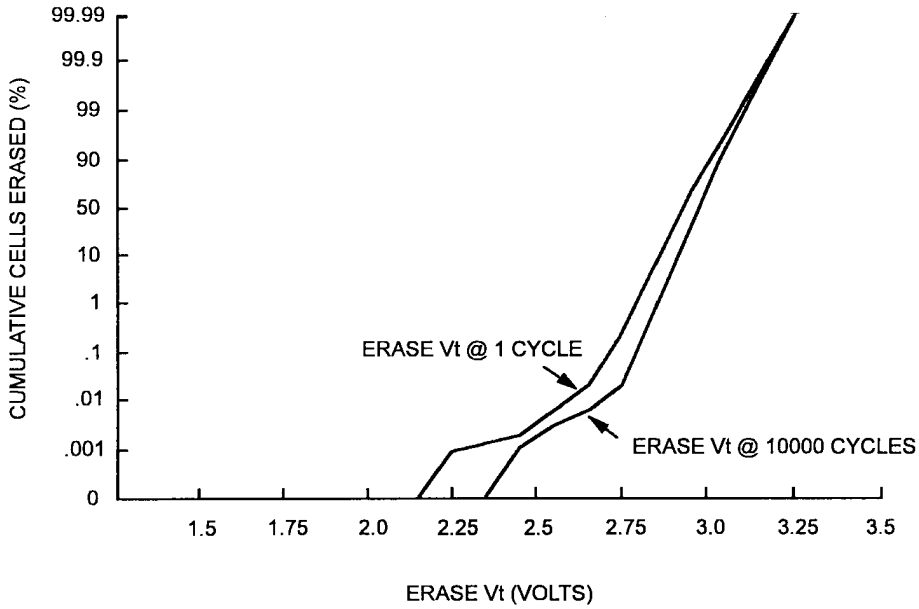

Fig. 27. Threshold voltage distribution after electrical erase after different program/erase cyclings [29].

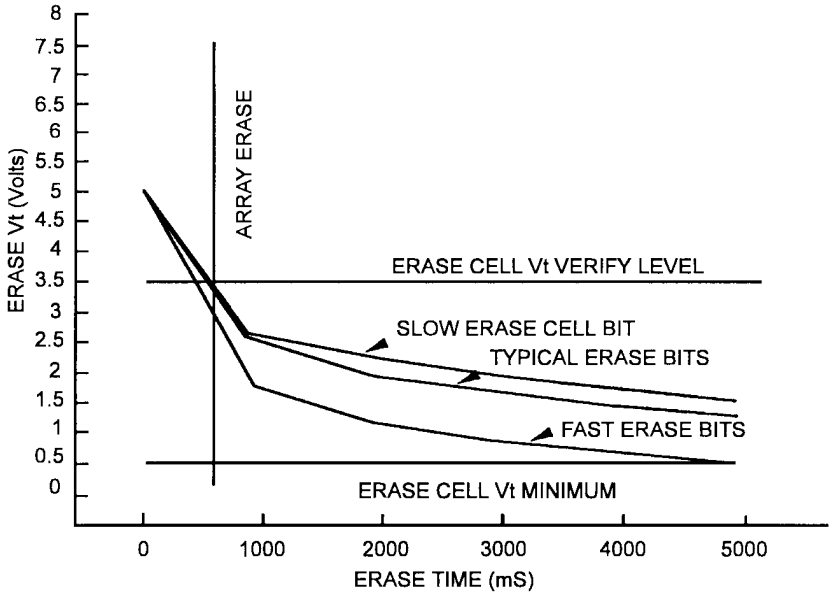

Fig. 28. Erase threshold voltage profile versus erase time [29].

plying 5 V on drain and grounding word-lines for 0.5 s) or FN channel reprogramming [65] (i.e., applying 16 V at the word-lines). The erase scheme proposed in [64] uses channel-electron-induced avalanche HCI to bring the erase V_T distribution to convergence after FN tunneling erase. This scheme does not require either programming all bytes before erasure by FN tunneling or iterating the erase/verify sequence. After hot-carrier stress, the self-convergence reprogramming becomes faster [66]; hot-carrier stress would cause interface degradation and larger gate-induced drain-leakage current [67], [68]. The reliability issues of using this technique are a concern.

The erase scheme proposed in [65] is a two-step procedure that starts by applying a negative high voltage to the control gate to erase the cell, followed by the application of a positive high voltage to the control gate to inject a few electrons from the substrate back to the FG to reprogram the cell to decrease the erase V_T distribution. This technique is not truly self-convergent, however, since erase V_T distribution spreads after reaching a minimum. To reach the minimum spread, several iterations of reprogramming and reerase may be used [66].

Therefore, for low-voltage high-density Flash memories, FN channel reprogramming can be used when high relia-

bility and cycling are required. When these specifications are not an issue, the HCI technique has the advantage of self-convergence and low operating voltage.

C. Scaling Issues

The architecture of an industry-standard Flash-cell array is typical of a NOR gate array, where every single cell is addressed by two signals, one for the BL and one for the WL; the source line and body are common to the whole array. Moreover, in standard arrays, a contact is shared between two cells, thus consuming a lot of cell area. The common issue among the different solutions and applications is the cost-per-bit reduction, which will be provided mainly by technology scaling. No consolidated theory has been developed for Flash-cell scaling [1].

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

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

Another relevant issue in Flash memories is the need for high voltages for program and erase. While CMOS technology scaling requires the reduction of the operating voltages, the actual program/erase operations are based on physical mechanisms whose major parameters do not scale (3.2-eV energy barrier for CHE's and at least 8–9 MV/cm for FN data alteration in 0.1–1 s). Moreover, the trend toward increasing the programming throughput will even force the internal voltage to rise. Double voltage supply simplifies memory design and minimizes the area, since there is no need to generate the high voltage internally. Therefore, they are preferably used when present in the circuit, even though internal generation is sometimes preferred for the correct operation of the memories in the chip. If internal generation is to be done, many issues need to be discussed. For example, hot-electron programming is not efficient if only a 5-V voltage supply is used, and programming times can become unacceptably slow. On the other hand, internal charge pumping circuits can be used only when small currents flow in the channel. Erasing opens similar issues.

Nonvolatility implies at least ten years of charge retention. Nonvolatility issues affect the scalability of thin active dielectrics (tunnel and interpoly). A direct tunneling mechanism fixes the tunnel oxide limit to 6 nm, which needs to be increased more realistically up to 7–8 nm due to trap-assisted electron tunneling caused by oxide aging

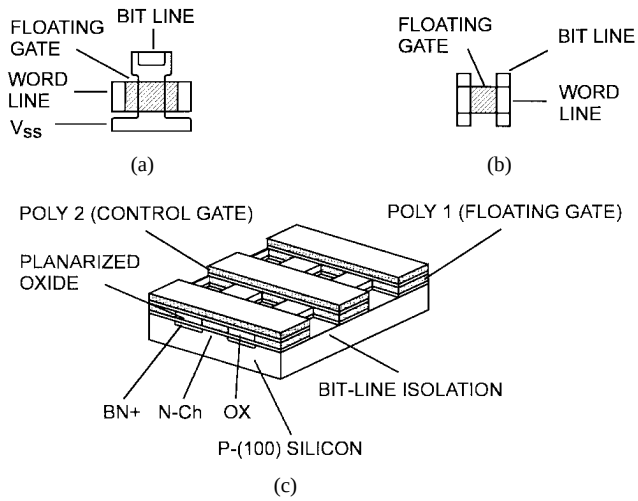

Fig. 29. (a) Layout of a T-shaped staked-gate Flash cell. A contactless (virtual ground) cell is also shown in (b) and (c) [72].

[69]. The scalability limit of the interpoly dielectric (ONO) has been reported to be around 12–13 nm [70]. These thicknesses can be combined to give an equivalent memory cell oxide (defined as tunnel oxide thickness divided by the coupling coefficient α_G), which sets the limit for the memory-cell poly length. Other constraints limit the minimum poly length.

- CHE injection requires some minimum drain-gate overlap and abrupt junction to maximize injection efficiency.
- FN tunneling to the source requires an overlap with the n^+ region below the gate.
- FN tunneling to the channel requires small gate/diffusion overlaps.

Moreover, when charge is injected from the polysilicon FG, the number of poly grains in the tunneling area plays an important role in determining the V_T distribution width [71].

In this scenario, the search for higher integration goes toward new architectural solutions, the reduction of the number of contacts, and the reduction of alignment tolerances. Contactless (virtual ground) configurations have been proposed and used [72]. Fig. 29 shows the layouts of a T-shaped staked-gate Flash cell and of a contactless one. In the second case, a 50% area reduction can be achieved only by self-aligning every single device, but this induces a higher complexity.

V. Flash Array Architectures

Other structures alternative to industry-standard Flash cells have been considered for Flash memories. Differences mainly are due to the array organization, program/erase mechanisms, and approaches to overerasing (which is solved algorithmically in standard structures). Many new cells and new arrays have been presented in the last five years and have reached different levels of maturity. This variety can be related to three main concepts.

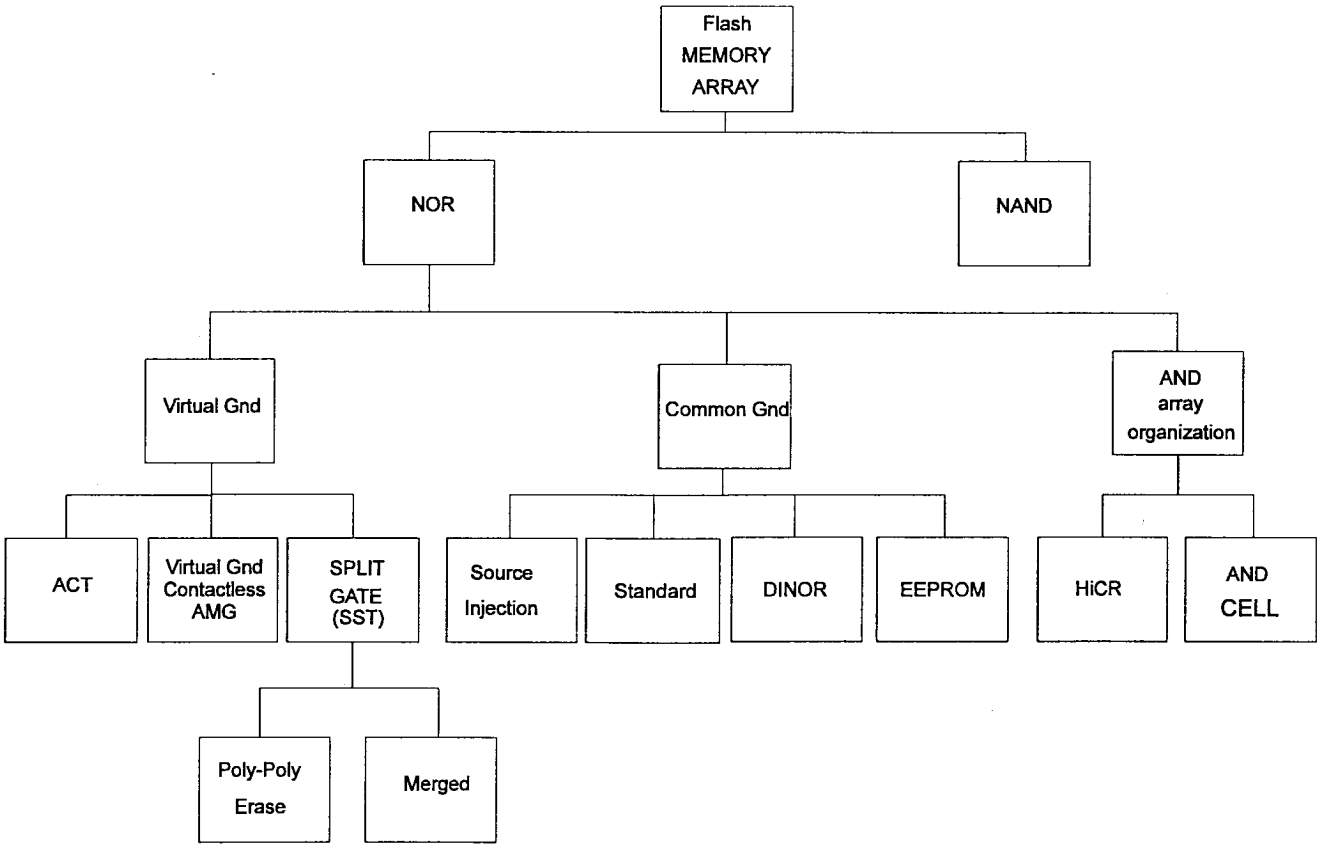

Fig. 30. The tree of Flash architectures, split according to array organizations [73].

1) Flash physics is not yet fully understood: dielectric scaling, program, and erase mechanisms leave open issues that can be resolved in different ways.
2) A company's past experience will direct it toward different attitudes on Flash-cell design and array organization.
3) The variety of applications (embedded memory, mass storage, or new concepts) imposes specific constraints.

New architectures are derived from a combination of these three different needs.

The actual scenario can be summarized as in the diagram in Fig. 30 [73]. Another way to catalogue the Flash-array scenario is to divide the different architectures according to data access and data write organization. Arrays that have random access and random program (parallel) are consistent with embedded applications, while page read and page program (serial) are consistent with mass storage applications. The different architectures can be catalogued by combining these topics with the program and erase mechanisms, thus obtaining the tree of Flash architectures shown in Fig. 31 [1].

In the following, we will describe some of the Flash memories proposed in the recent literature. Cells which use the CHE program and FN erase can be grouped into two main categories: 1) one-transistor cells and their array architectures (NOR common ground and alternate metal virtual ground [AMG]) and 2) merged cells (split-gate triple poly, split-gate source injection).

Cells that use FN programming and erase can be grouped as:

1) NOR arrays (divided bit-line NOR (DINOR), asymmetrical contactless transistor (ACT), and EEPROM-like cells);
2) AND arrays (AND, high capacitive coupling ratio (HiCR) cells);
3) NAND cells.

The motivation for using FN tunneling to the channel for both programming and erase comes from the need to change the programming mechanism to simplify the supply scaling and to reduce cell sizes. Moreover, it comes from the experience already accumulated in EEPROM memory development.

A. Alternate Metal Ground

AMG cells [74] come as an extension of a similar AMG EPROM cell, and the cell concept is the same as the industry standard. AMG cell-size scaling is accomplished by sharing one metal line per two diffusion bit lines; a new segmentation scheme and fieldless array allow the achievement of the minimum design rule of the process, which typically is the pitch of polysilicon. Cross sections of AMG cells are shown in Fig. 32. The layout is simple

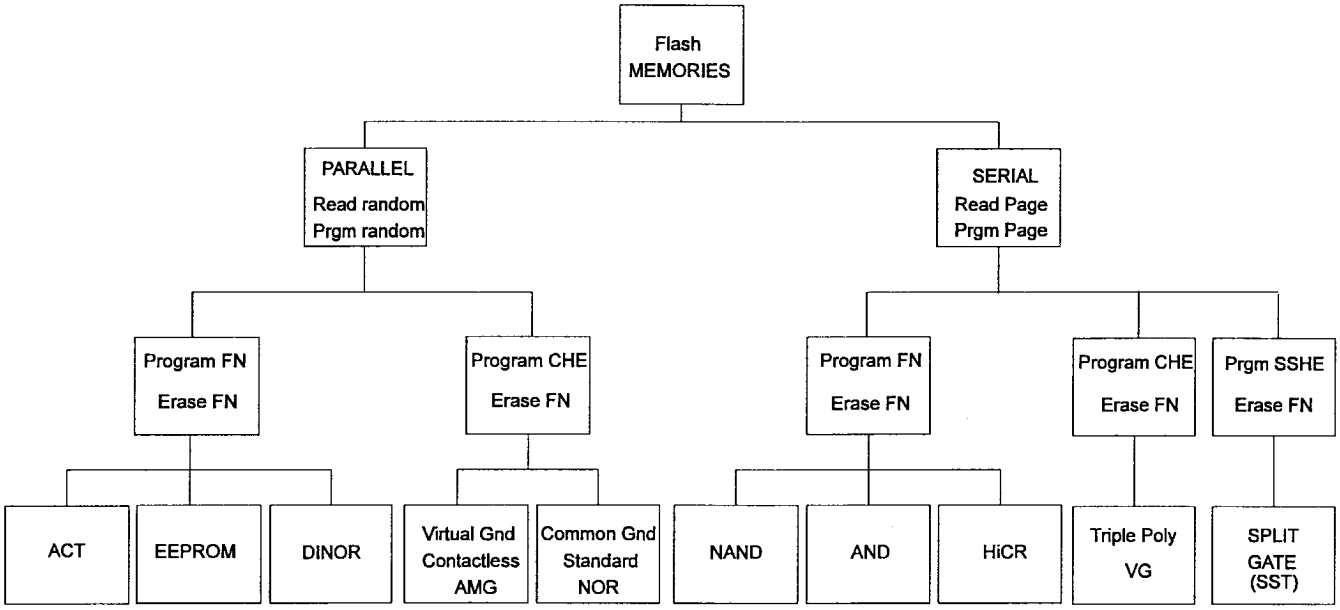

Fig. 31. The tree of Flash architectures, split according to the parallel/serial access to the array [1].

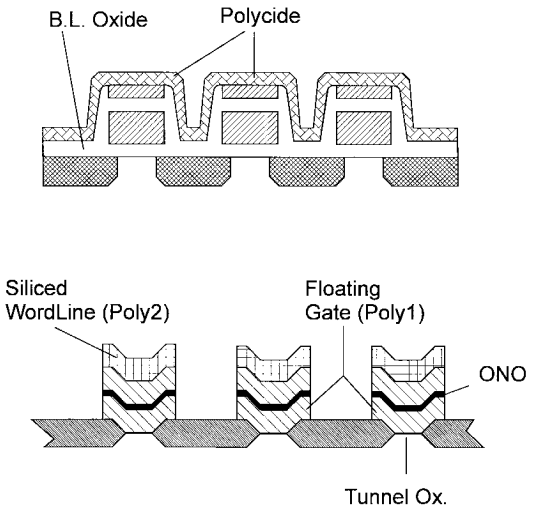

Fig. 32. AMG cross sections along X and Y directions [74].

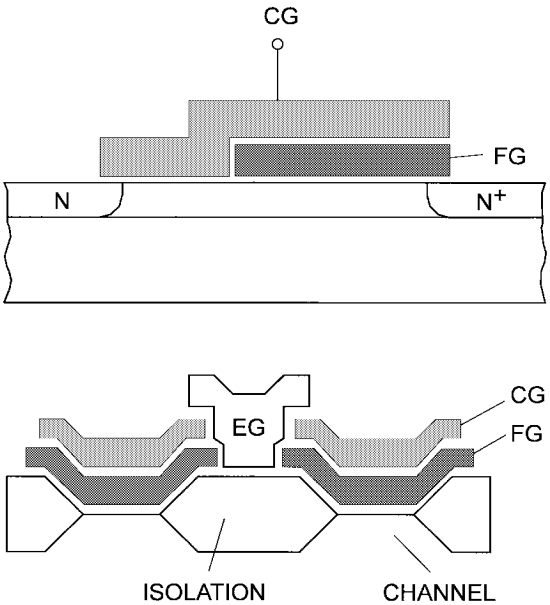

Fig. 33. Split-gate triple-poly cell cross section. FG = floating gate (poly 1), CG = control gate (poly 2), and EG = erase gate (poly 3) [75].

and overcomes most of the critical issues of standard cells; nevertheless, it has some critical features, which can be summarized in double-poly width control and BL side diffusion. The array architecture adopts the virtual ground schemes, using one metal line every two BL's. It is a general-purpose product, which can have many applications. It is possible to use it in low-voltage applications, and it can reach high speed. Moreover, multilevel programming is possible.

B. Split Gate

Merged cells add new features to the Flash array and simplify the design since monitoring the erase distribution is not critical. They improve CHE program by using a very low program current. Split-gate triple-poly cells [75] (Fig. 33) use a different approach: there is a MOS transistor and an FG device fabricated on the same cell. They come as an extension of self-aligned split-gate EPROM products and use a third poly level along the WL as an erase gate. This enables positive voltage low power erase, eliminating the overerase issue but introducing a poly-poly trapping issue. Since the read current is inherently low, a smaller diffusion area is used.

The layout has some critical features given by the BL side diffusion, alignment of the three poly layers, distance between metal BL's, and alignment of poly layers to diffusions. It has the advantage of eliminating the contact design rule limitations.

The process uses self-aligned drain and poly-poly dielectrics. It is a single metal process, and cell-size limitations are given by poly layer alignments to diffusions. Some problems can be introduced by poly etching.

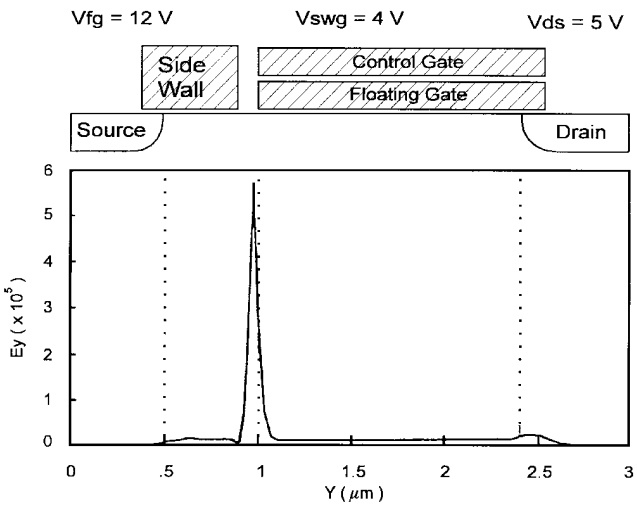

Fig. 34. Split-gate source-injection cell cross section. V_{fg} is floating gate potential and V_{swg} is sidewall gate bias. Electric field profile along the channel during programming is also shown [76].

The choice of a poly-poly erase split gate gives many advantages since the erase operation requires only positive voltages, and the erase V_T distribution is inherently good. In fact, the FG device can even go into depletion, but the total transistor will maintain the threshold voltage of the n-MOS transistor.

Due to the excellent natural segmentation of the array, this kind of cell can find many applications as dedicated mass storage. It is possible to go to low voltages, but the low read current excludes high speeds. Multilevel programming is possible.

C. Source Injection

The concept behind source injection cells [76] is the separation of the regions of acceleration and injection of carriers, thus enabling CHE injection at low current and voltages (Fig. 34). The cell is composed of a stacked-gate MOS transistor with a sidewall select gate on its source side. The array architecture uses a common ground and has the advantages of low-current CHE, no overerase, and positive voltages only; it has the disadvantages of increased cell size and of poly-poly erase cycling.

By using merged cells, dual-bit split-gate (DSG) [77] area savings can be achieved. The cell consists of three channels directly connected through a shared select gate (Fig. 35). A transfer gate in poly-3 is added. The cell has two FG's, one control gate, one transfer gate, and one common select gate. The cell contains two bits, which share one pair of drain and source. Again, there is no overerase problem, but the erase V_T distribution becomes an issue and excludes multilevel operation. Moreover, complex decoding, use of negative voltages in erase, and high capacitances add to the disadvantages of this architecture.

D. Divided Bit-Line NOR

DINOR cells [78] (Fig. 36) are common ground cells fabricated with a triple-well, triple-level polysilicon, a tungsten plug, and two layers of metal. They allow scaling of

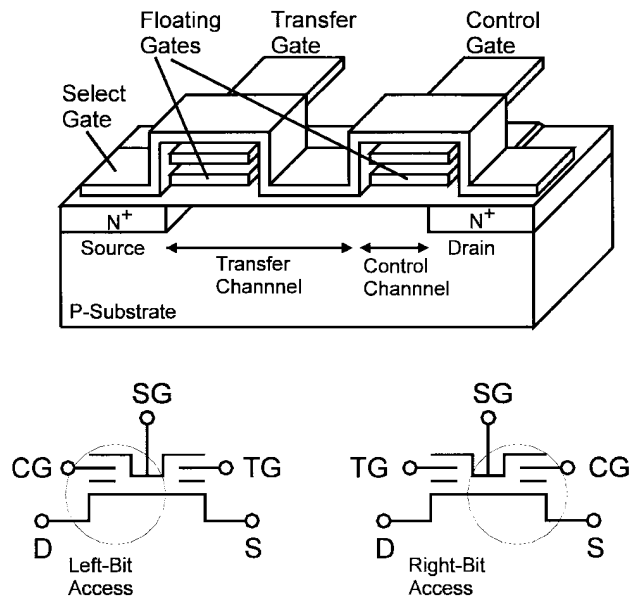

Fig. 35. Merged split-gate source injection cross section. Transfer gate and control gate are equal [77].

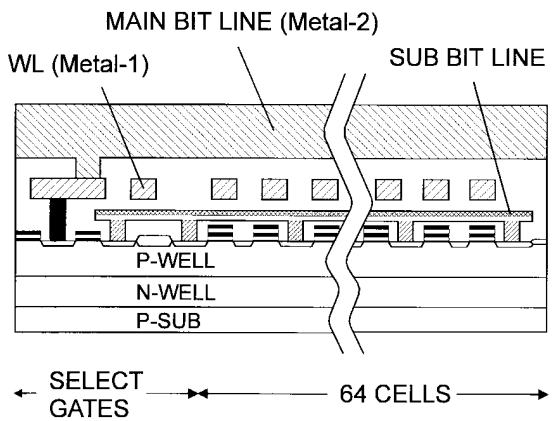

Fig. 36. DINOR array cross section [78].

cells and a better segmentation, which is related to the application. This array has blocks of 32 or 64 b and shares one metal-2 main BL every two sub-BL's and a metal-1 WL strapping. The cell size is reduced by forming a local BL in poly-3 and by using poly-plug drain contacts. There are no double diffused drains.

In DINOR cells, program and erase operations are opposite to the conventional NOR. Program means set the V_T of the selected cell to be low state and erase means set the V_T of the cells of the selected sector to be high state. Erase is obtained via channel FN tunneling, which requires 200-μs programming time and 1-MV/cm or higher electric field. Program is achieved via FN tunneling through the gate/drain overlapped area (like in the erase operation in industry-standard cells).

Disturbs are reduced by the high V_T erase, but there are some issues in the program/erase characteristics given by the high programming current from the charge pump and from the overhead in the circuitry.

DINOR is a general-purpose architecture that makes possible the low-voltage/high-speed operation, while multilevel

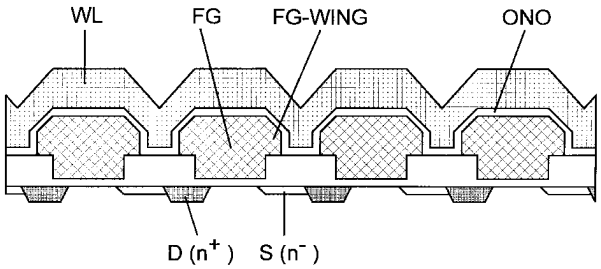

Fig. 37. ACT cell cross section parallel to a word line [79].

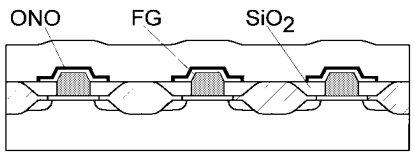

Fig. 38. AND cell cross section [80].

programming is very difficult. The small array is shadowed by the large periphery needed to reduce the programming time.

E. Asymmetrical Contactless Transistor

The ACT is a new cell structure for subquarter micrometer Flash memory (Fig. 37) introduced in 1995 [79]. It has been proposed for high-density data storage, which requires low-voltage low-power consumption and fast program/erase. The ACT cell is fabricated with a lightly doped source and heavily doped drain. It realizes a simple virtual ground array using the FN tunneling mechanism for both program and erase operations. FN tunneling on the drain side is used for program, while channel FN tunneling is used for erase. To achieve a high gate coupling ratio in the WL direction, a self-aligned FG wing technology has been used, which does not sacrifice cell area. Therefore, small cells with a high gate coupling ratio can be obtained. The low programming current of the ACT cell enables the use of multiple programming, making it possible to achieve fast programming times with a low single supply voltage (< 3 V). A good disturb immunity in program, erase, and read modes is also obtained.

F. AND

The main idea behind AND cells is to increase gate coupling in the WL direction, thus saving area since there is no need for coupling wings over the field (Fig. 38) [80]. Contacts between WL's are eliminated, and a very good segmentation can be achieved. The architecture shows local segments with diffused BL's and source lines and connected source select devices. Every block is separated from other blocks. The cell operates like a DINOR, and the key issue is the small programming margin. It enables small sector program and erase, thus improving the V_T distribution, with an excellent row redundancy, but the overhead of the sector over the cell is as high as 30%. This, in general, would have been called a virtual ground architecture. It is all separated, however, thus resulting in a common ground.

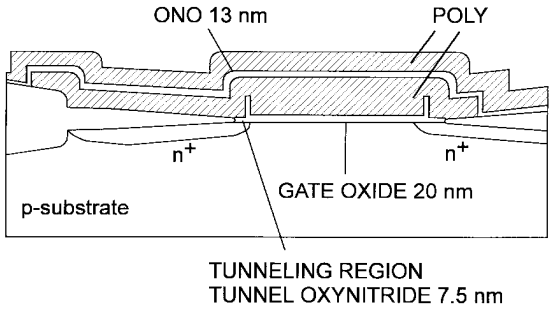

Fig. 39. HiCR cell cross section [81].

AND cells can be used as mass storage and in a low-power high-speed application. Multilevel programming is very difficult due to the small program window and to tunneling programming.

This concept emphasizes mass storage applications without significant speed degradation.

G. High Capacitive Coupling Ratio

HiCR cells were first presented in 1993 for 3-V-only 64-Mb Flash memories [81]. They are contactless cells with a high capacitive coupling ratio, programmed and erased by FN tunneling.

They use the same concept as AND cells and are realized defining a small self-aligned tunneling region underneath the FG side wall by means of an advanced rapid thermal process for 7.5-nm-thick tunnel oxinitride (Fig. 39).

The cell is programmed via FN tunneling across the tunnel oxinitride from the FG to the drain junction; it is erased via FN tunneling from source/drain junctions to the FG. The program/erase scheme requires positive voltages only, and can find applications in the same area as AND cells. The process is rather complicated, and a tunnel area has to be grown over an n^+. But the array architecture, with its segmentation of both BL and WL, enables effective block size, better speed, and good redundancy, paying it in terms of major overhead.

H. NAND

A completely different approach in array organization can be followed by using a NAND architecture, which greatly improves the results [82]. The elementary unit is not composed of the single three-terminal cell, which stores one single bit, but by more FG transistors connected in a series (eight or 16), which constitutes a chain connected to the bit line and ground through two selection transistors (Fig. 40). This organization allows the elimination of all contacts between WL's, which can be separated by their minimum design rule, thus reducing the occupied area by 40%. Moreover, a kind of memory organization with a unit element with a dimension of one byte (or one word) is closer to the ideal memory with parallel access. It allows even page (256-byte) programming, resulting in a greatly improved versatility.

Fig. 41 shows the cross section of an 8-b elementary block for a 4-Mb Flash memory organized as a NAND

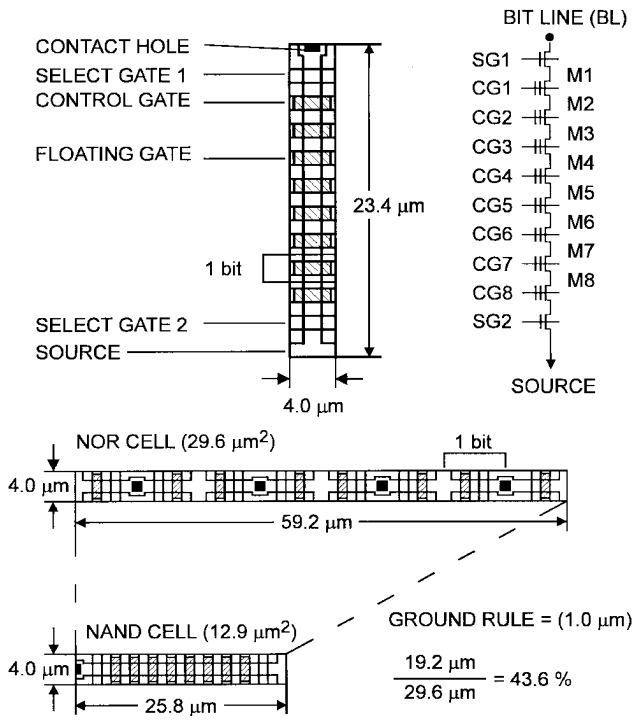

Fig. 40. NAND architecture. Dimensions of a NAND array are compared to those of a NOR array [82].

array with peripheral circuits [82]. Erase voltages are 20 V to the n-substrate, the p-well2, the drain, and the source, and 0 V to the control gate of the selected location. This biasing induces electron tunneling toward p-well2, resulting in a negative low threshold. There is no voltage drop between the drain and p-well, so that there is no breakdown of the junction. Programming voltages are 18 V to the selected control gate and 10 V to the other gates; p-well2 is grounded. Selection transistors are biased to connect the chain to the bit line and isolate it from the ground. If a "0" is to be stored, the bit line is grounded; sources and drains are grounded, and only the selected transistor has such an electric field in the oxide to induce electron injection from the substrate in the FG, increasing the threshold voltage to the high level. If a "1" is to be stored, the bit line is biased at 10 V; there is no tunneling and the threshold is negative. The reading operation is performed by applying 5 V to all the control gates except the selected one, which is grounded. Selection transistors are now conductive, and they connect the chain to the ground and bit line. The bit line is precharged, therefore, if the stored data is "1" (negative threshold voltage) and the selected transistor is conductive and discharges the bit line; if the stored data is "0," the transistor is off and the bit line holds the charge.

If the memory is organized as a NAND array, threshold voltage checking becomes critical; only very small variations on the nominal value are allowed. In NAND arrays, both program and erase mechanisms are electron tunneling. Since tunneling is more efficient than HEI, currents are smaller and different supply voltages can be internally generated by charge-pumping circuits implemented in the same die. NAND arrays are preferred for high-density Flash memories. Page access time is around 80 ns, and even programming can be by the page, thus reducing programming time to 300 μs per page. Erase is fast and takes 6 ms per block and around 10 ms per chip. Therefore, it is possible to use these memories instead of floppy disks or hard disks. Moreover, they are more reliable, since the programming mechanism, i.e., electron tunneling from the bulk, is uniform through the oxide if compared to HEI in the drain region or tunneling in the source region to erase the cell. Therefore, oxide damage is reduced and breakdown is less probable.

VI. Conclusion

This paper offers an overview of mainstream Flash memory cells and the basic principles for the fast development that Flash memory cells had in the last few years.

We reviewed the physics mechanisms used to store and remove charge from an FG, thus enabling information storage. The main reliability aspects have been analyzed, and their impact on the development of new structures has been stressed. A proper organization of the cell architecture can reduce the impact of intrinsic degradation mechanisms, which are responsible for the wear-out of device performances in program/erase cycling.

If we consider the number of bits per memory device as a function of time, the density of memory circuits doubled every 1.8 years. Extrapolating to the future with the same trend needs some breakthrough. In fact, if we consider the cell size, it has been scaling down at about a factor of two every four years; the die size almost doubled every two years [83]. If we extrapolate to the future, in the year 2025, density will be a little bit higher than 10^{12}, feature size will be 18 nm, cell size will be 50 $\times$ 50 nm, and die size will be 64 cm^2, equivalent to 18 dices on an 8″ wafer. This will not be a realistic scenario. Industry is already showing some slowing down. The large economic impact of research and development has increased both time to production and product lifetime. Flash technology scaling, as far as both cell size and voltage are concerned, can proceed through an evolutionary path until the 1-Gb generation, even if some major issues must be solved [1], while further scaling will require a real breakthrough and innovative concepts [84]. In the near future, the main market is going to be the embedded application market. In the mass storage market, cost is the driving factor for increasing density. The use of the same solutions for the embedded application market is welcome, but multilevel concepts will be introduced.

Flash memories have been demonstrated to be a reliable and flexible integrated circuit to be used in many new applications that could be covered neither by EEPROM's, because of their low density and high cost, nor by EPROM's, which do not support in-system reprogramming.

Acknowledgment

The authors wish to thank Prof. C. Canali of the University of Modena for his constant support and encouragement, Dr. N. Bovolon and Prof. A. Paccagnella of the University

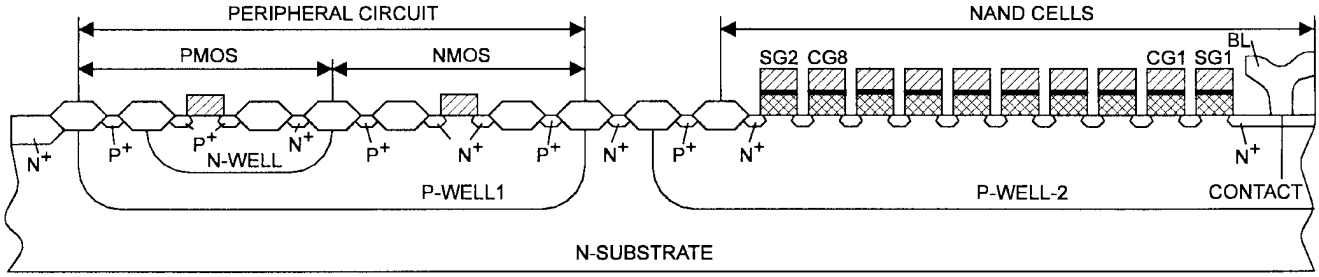

Fig. 41. NAND cells and peripheral circuit cross section [82].

of Padova for many fruitful discussions, and R. Formentini and S. Gheduzzi of the University of Modena for their help in figure editing. They are grateful to Dr. B. Eitan (W.S.I. Israel, Ltd.) for allowing them to use data and forecasts from his many tutorials on this subject.

REFERENCES

[1] G. Crisenza, R. Annunziata, E. Camerlenghi, and P. Cappelletti, "Non volatile memories: Issues, challenges and trends for the 2000's scenario," in *Proc. ESSDERC'96*, G. Baccarani and M. Rudan, Eds. Bologna, Italy: Editions Frontieres, 1996, pp. 121–130.

[2] S. Wells and D. Clay, "Flash solid-state drive with 6 MB/s read/write channel ABS data compression," in *Proc. ISSCC Conf.*, San Francisco, CA, no. WP 3.6, 1993, p. 52.

[3] E. Suzuki, H. Hirashi, K. Ishii, and Y. Hayashi, "A low-voltage alterable EEPROM with metal-oxide-nitride-oxide-semiconductor (MONOS) structure," *IEEE Trans. Electron Devices*, vol. ED-30, no. 2, pp. 122–127, 1983.

[4] E. Suzuki, K. Miura, Y. Hayashi, R.-P. Tsay, and D. Schroder, "Hole and electron current transport in metal-oxide-nitride-oxide-silicon memory structures," *IEEE Trans. Electron Devices*, vol. 36, no. 6, pp. 1145–1149, 1989.

[5] S. T. Wang, "On the I–V characteristics of floating-gate MOS transistors," *IEEE Trans. Electron Devices*, vol. ED-26, no. 9, pp. 1292–1294, 1979.

[6] M. Wada, S. Mimura, H. Nihira, and H. Iizuka, "Limiting factors for programming EPROM of reduced dimensions," *IEDM Tech. Dig.*, 1980, pp. 38–41.

[7] A. Kolodny, S. T. K. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of floating gate EEPROM cells," *IEEE Trans. Electron Devices*, vol. ED-33, no. 6, pp. 835–844, 1986.

[8] K. Prall, W. I. Kinney, and J. Marco, "Characterization and suppression of drain coupling in submicrometer EPROM cells," *IEEE Trans. Electron Devices*, vol. ED-34, no. 12, p. 2463, 1987.

[9] M. Wong, D. K.-Y. Liu, and S. S.-W. Huang, "Analysis of the subthreshold slope and the linear transconductance techniques for the extraction of the capacitive coupling coefficients of floating-gate devices," *IEEE Electron Device Lett.*, vol. 13, no. 11, pp. 566–568, 1992.

[10] W. L. Choi and D. M. Kim, "A new technique for measuring coupling coefficients and 3-D capacitance characterization of floating gate devices," *IEEE Trans. Electron Devices*, vol. 41, no. 12, pp. 2337–2342, 1994.

[11] R. Bez, E. Camerlenghi, D. Cantarelli, L. Ravazzi, and G. Crisenza, "A novel method for the experimental determination of the coupling ratios in submicron EPROM and Flash EEPROM cells," *IEDM Tech. Dig.*, 1990, pp. 99–102.

[12] K. T. San, C. Kaya, D. K. Y. Liu, T. P. Ma, and P. Shah, "A new technique for determining the capacitive coupling coefficients in FLASH EPROM's," *IEEE Electron Device Lett.*, vol. 13, no. 6, pp. 328–331, 1992.

[13] B. Moison, C. Papadas, G. Ghibaudo, P. Mortini, and G. Pananakakis, "New method for the extraction of the coupling ratios in FLOTOX EPROM cells," *IEEE Trans. Electron Devices*, vol. 40, no. 10, pp. 1870–1872, 1993.

[14] M. Woods, "An E-PROM's integrity starts with its cell structure," in *Nonvolatile Semiconductor Memories: Technologies, Design, and Application*, C. Hu, Ed. New York: IEEE Press, 1991, ch. 3, pp. 59–62.

[15] P. E. Cottrell, R. R. Troutman, and T. H. Ning, "Hot-electron emission in n-channel IGFET's," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 520–532, 1979.

[16] B. Eitan and D. Froham-Bentchkowsky, "Hot-electron injection into the oxide in n-channel MOS devices," *IEEE Trans. Electron Devices*, vol. ED-28, no. 3, pp. 328–340, 1981.

[17] G. A. Baraff, "Distribution functions and ionization rates for hot-electrons in semiconductors," *Phys. Rev.*, vol. 128, no. 6, pp. 2507–2517, 1962.

[18] C. Hu, "Lucky-electron model for channel hot-electron emission," *IEDM Tech. Dig.*, 1979, p. 22.

[19] S. Tam, P. K. Ko, C. Hu, and R. Muller, "Correlation between substrate and gate currents in MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-29, no. 11, pp. 1740–1744, 1982.

[20] E. Takeda, H. Kune, T. Toyabe, and S. Asai, "Submicrometer MOSFET structure for minimizing hot-carrier generation," *IEEE Trans. Electron Devices*, vol. ED-29, no. 4, pp. 611–618, 1982.

[21] K. Hess and C. T. Sah, "Hot carriers in Silicon surface inversion layers," *J. Appl. Phys.*, vol. 45, p. 1254, 1974.

[22] K. R. Hofmann, C. Werner, W. Weber, and G. Dorda, "Hot-electrons and hole-emission effects in short n-channel MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-32, no. 3, pp. 691–699, 1985.

[23] C. N. Berglund and R. S. Powell, "Photoinjection into SiO_2: Electron scattering in the image-force potential well," *J. Appl. Phys.*, vol. 42, p. 573, 1971.

[24] C. Fiegna, E. Sangiorgi, and L. Selmi, "Oxide field dependence of electron injection from silicon into silicon dioxide," *IEEE Trans. Electron Devices*, vol. 40, no. 11, pp. 2018–2022, 1993.

[25] Y. A. El-Mansy and D. M Coughey, "Characterization of silicon-on-sapphire IGFET transistors," *IEEE Trans. Electron Devices*, vol. ED-24, no. 9, pp. 1148–1153, 1977.

[26] L. Esaki, "Long journey into tunneling," *Proc. IEEE*, vol. 62, pp. 825–831, June 1974.

[27] J. Moll, *Physics of Semiconductors*. New York: McGraw-Hill, 1964.

[28] M. Lezlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, 1969.

[29] "IEDM 90 short course: Non-volatile memory," IEEE, San Francisco, CA, 1990.

[30] P. Olivo, T. Nguyen, and B. Riccò, "High-field-induced degradation in ultra thin SiO_2 films," *IEEE Trans. Electron Devices*, vol. 35, no. 12, pp. 2259–2267, 1988.

[31] J. Tang and K. Hess, "Theory of hot electron emission from silicon into silicon dioxide," *J. Appl. Phys.*, vol. 54, pp. 5145–5151, 1983.

[32] T. Ando, A. B. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Rev. Mod. Phys.*, vol. 58, pp. 437–672, 1982.

[33] J. Suné, P. Olivo, and B. Riccò, "Quantum-mechanical modeling of accumulation layers in MOS structures," *IEEE Trans. Electron Devices*, vol. 39, no. 7, pp. 1732–1739, 1992.

[34] M. Lanzoni, J. Suné, and B. Riccò, "Advanced electrical-level modeling of EEPROM," *IEEE Trans. Electron Devices*, vol. 40, no. 5, pp. 951–957, 1993.

[35] D. C. Guterman, I. H. Rimawi, T. L. Chiu, R. D. Halvorson, and D. J. McElroy, "An electrically alterable nonvolatile memory cell using a floating-gate structure," *IEEE Trans. Electron Devices*, vol. ED-26, no. 4, pp. 576–586, 1979.

[36] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new Flash E^2PROM cell using triple polysilicon technology," *IEDM Tech. Dig.*, pp. 464–467, 1984.
[37] G. Verma and N. Mielke, "Reliability performance of ETOX based Flash memories," in *Proc. IRPS*, 1988, p. 158.
[38] M. Bauer *et al.*, "A multilevel 32 Mb Flash memory," in *Proc. ISSCC Conf.*, San Francisco, CA, 1995, p. 132.
[39] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome, and R. Shirota, "Fast and accurate programming method for multilevel NAND EEPROM's," in *Dig. VLSI Symp. VLSI Technology*, 1995, no. 10B-4, p. 129.
[40] N. Rodjy, "0.85 μm double metal CMOS technology for 5 V Flash EPROM memories with sector erase," presented at the 12th Nonvolatile Semiconductor Memory Workshop, Monterey, CA, Aug. 1992.
[41] A. Bergemont, H. Haggag, L. Anderson, E. Shacham, and G. Woltsenholme, "NOR virtual ground (NVG)—A new scaling concept for very high density FLASH EEPROM and its implementation in a 0.5 μm process," *IEDM Tech. Dig.*, 1993, pp. 15–18.
[42] A. Bergemont, M. H. Chi, and H. Haggag, "Low voltage NVG: A new high performance 3 V/5 V Flash technology for portable computing and telecommunication applications," in *Proc. ESSDERC'95*, H. C. de Graaf and H. v. Kranemburg, Eds. The Hague, The Netherlands: Editions Frontieres, 1995, pp. 543–547.
[43] B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti, and A. Modelli, "Multilevel Flash cells and their trade-offs," *IEDM Tech. Dig.*, 1996, pp. 169–172.
[44] V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, and S. Wells, "An in-system reprogrammable 256 K CMOS Flash memory," in *Proc. ISSCC Conf.*, San Francisco, CA, 1988, p. 132.
[45] S. Mori, Y. Yamaguchi, M. Sato, H. Meguro, H. Tsunoda, E. Kamiya, K. Yoshikawa, and N. Arai, "Thickness scaling limitation factors of ONO interpoly dielectric for nonvolatile memory devices," *IEEE Trans. Electron Devices*, vol. 43, no. 1, pp. 47–53, 1996.
[46] R. Bez, D. Cantarelli, and S. Serra, "The channel hot electron programming of a floating gate MOSFET: An analytical study," presented at the 12th Nonvolatile Semiconductor Memory Workshop, Monterey, CA, Aug. 1992.
[47] S. Keenney, R. Bez, D. Cantarelli, F. Piccinini, A. Mathewson, and C. Lombardi, "Complete transient simulation of Flash EEPROM devices," *IEEE Trans. Electron Devices*, vol. 39, no. 12, pp. 2750–2757, 1992.
[48] S. Keenney, F. Piccinini, M. Morelli, A. Mathewson, C. Lombardi, R. Bez, L. Ravazzi, and D. Cantarelli, "Complete transient simulation of Flash EEPROM devices," *IEDM Tech. Dig.*, no. 8.7.1, 1990, pp. 201–204.
[49] P. Cappelletti, "Non volatile memories," presented at the 8th Workshop on Dielectrics in Microelectronics, Venice, Italy, Nov. 1996, to be published.
[50] S. Cagnina, C. Chang, S. Haddad, J. Lien, N. Radjy, Y. Sun, Y. Tang, M. Van Buskirk, and A. Wang, "A 0.85 μm double metal CMOS technology for 5 V Flash memories with sector erase," presented at the 12th Nonvolatile Semiconductor Memory Workshop, Monterey, CA, Aug. 1992.
[51] B. Venkatesh, M. Chung, S. Govindachar, V. Santurkar, C. Bill, R. Gutala, D. Zhou, J. Yu, M. Van Buskirk, S. Kawamura, K. Kurihara, H. Kawashima, and H. Watanabe, "A 55 ns 0.35 μm 3 V-only 16 M Flash memory with deep power-down," in *Proc. ISSCC Conf.*, San Francisco, CA, no. TP 2.7, 1996, p. 44.
[52] 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," *IEDM Tech. Dig.*, 1992, pp. 595–598.
[53] P. Cappelletti, R. Bez, D. Cantarelli, and L. Fratin, "Failure mechanisms of Flash cell in program/erase cycling," *IEDM Tech. Dig.*, 1994, pp. 291–294.
[54] S. Aritome, R. Shirota, G. Hemnik, T. Endoh, and F. Masuoka, "Reliability issues of Flash memory cells," *Proc. IEEE*, vol. 81, pp. 776–788, May 1993.
[55] C. Papadas, G. Ghibaudo, G. Pananakakis, C. Riva, P. Ghezzi, C. Gounelle, and P. Mortini, "Retention characteristics of single-poly EEPROM cells," in *Proc. European Symp. Reliability of Electronic Devices, Failure Physics and Analysis (ESREF)*, Bordeaux, France, Oct. 7–10, 1991, pp. 517–522.
[56] A. Watts, "Built-in reliability for 10 FITS performance on EPROM and Flash memory," SGS-Thomson Microelectronic, Agrate Brianza, Italy, Tech. Art. TA 109, Nov. 1991.
[57] N. R. Mielke, "New EPROM data loss mechanisms," in *Proc. IRPS*, 1983, pp. 106–113.
[58] P. L. Hefley and J. W. McPherson, in *Proc. IRPS*, 1988, p. 176.
[59] S. Haddad, C. Chang, B. Swaminathan, and J. Lien, "Degradation due to hole trapping in Flash memory cells," *IEEE Electron Device Lett.*, vol. 10, no. 3, pp. 117–119, 1989.
[60] P. Olivo, B. Riccò, and E. Sangiorgi, "High field induced voltage dependent oxide charge," *Appl. Phys. Lett.*, vol. 48, pp. 1135–1137, 1986.
[61] S. Yamada, Y. Hiura, T. Yamane, K. Amemiya, Y. Oshima, and K. Yoshikawa, "Degradation mechanism of Flash EEPROM programming after program/erase cycles," *IEDM Tech. Dig.*, 1993, pp. 23–26.
[62] T. C. Ong, A. Fazio, N. Mielke, S. Pan, N. Righos, G. Atwood, and S. Lai, "Erratic erase in ETOX Flash memory array," in *VLSI Symp. Technology*, 1993, pp. 82–83.
[63] C. Dunn, C. Kaya, T. Lewis, T. Strauss, J. Schreck, P. Hefley, M. Middendorf, and T. San, "Flash EPROM disturb mechanism," in *Proc. IRPS*, 1994, pp. 299–308.
[64] S. Yamada, T. Suzuki, E. Obi, M. Oshikiri, K. Naruke, and M. Wada, "A self-convergence erasing scheme for a simple stacked gate Flash EEPROM," *IEDM Tech. Dig.*, 1991, pp. 307–310.
[65] K. Oyama, H. Shirai, N. Kodama, K. Kanamori, K. Saitoh, Y. S. Hisamune, and T. Okazawa, "A novel erasing technology for 3.3 V Flash memory with 64 Mb capacity and beyond," *IEDM Tech. Dig.*, 1992, pp. 607–610.
[66] A. Bergemont, M. Chi, and H. Haggag, "Low voltage NVG: A new high performance 3 V/5 V Flash technology for portable computing and telecommunications applications," *IEEE Trans. Electron Devices*, vol. 43, no. 9, pp. 1510–1517, 1996.
[67] I. C. Chen, C. Tencg, D. Coleman, and A. Nishimura, "Interface-trap enhanced gate-induced leakage current in MOSFET," *IEEE Electron Device Lett.*, vol. 10, no. 5, pp. 216–218, 1989.
[68] T. Chang, C. Huang, and T. Wang, "Mechanisms of interface trap-induced drain leakage current in off-state n-MOSFET," *IEEE Trans. Electron Devices*, vol. 42, no. 4, pp. 738–743, 1995.
[69] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," *IEDM Tech. Dig.*, 1992, pp. 139–142.
[70] S. Mori, Y. Yamaguchi, M. Sato, H. Meguro, H. Tsunoda, E. Kamiya, K. Yoshikawa, N. Arai, and E. Sakagami, "Thickness scaling limitation factors of ONO interpoly dielectric for nonvolatile memory devices," *IEEE Trans. Electron Devices*, vol. 43, no. 1, pp. 47–53, 1996.
[71] 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 principles for Flash memories," *IEDM Tech. Dig.*, 1994, pp. 847–850.
[72] A. T. Mitchell, C. Huffman, and A. L. Esquivel, "A new self-aligned planar array cell for ultra high density EPROM's," *IEDM Tech. Dig.*, 1987, pp. 548–553.
[73] B. Eitan, "Cell concepts and array architectures," presented at the Flash Memory Tutorial, 14th Nonvolatile Semiconductor Memory Workshop, Monterey, CA, Aug. 1995.
[74] R. Kazerounian, A. Bergemont, A. Roy, G. Wolsteholme, R. Irani, M. Shamay, H. Gaffur, G. A. Rezvani, L. Anderson, H. Haggag, E. Shacham, P. Kauk, P. Nielson, A. Kablanian, K. Chhor, J. Perry, R. Sethi, and B. Eitan, "Alternate Metal virtual Ground EPROM array implemented in a 0.8 μm process for very high density applications," *IEDM Tech. Dig.*, 1991, pp. 311–314.
[75] J. VanHoudt, L. Haspeslagh, D. Wellekens, L. Deferm, G. Groeseneken, and H. E. Maes, "HIMOS—A high efficiency Flash EEPROM cell for embedded memory applications," *IEEE Trans. Electron Devices*, vol. 40, no. 12, pp. 2255–2263, 1993.
[76] K. Naruke, S. Yamada, E. Obi, S. Taguchi, and M. Wada, "A new Flash-erase EEPROM cell with a sidewall select-gate on its source side," in *Nonvolatile Semiconductor Memories: Technologies, Design, and Applications*, C. Hu, Ed. New York: IEEE Press, 1991, ch. 5, pp. 183–186.
[77] Y. Ma, C. S. Pang, K. T. Chang, S. C. Tsao, J. E. Frayer, T. Kim, K. Jo, J. Kim, I. Choi, and H. Park, "A dual-bit split-gate EEPROM (DSG) cell in contactless array for single-Vcc high

density Flash memories," *IEDM Tech. Dig.*, 1994, pp. 57–60.
[78] H. Onoda, Y. Kunori, S. Kobayashi, M. Ohi, A. Fukumoto, N. Ajika, and H. Miyoshi, "A novel cell structure suitable for a 3 V operation, sector erase FLASH memory," *IEDM Tech. Dig.*, 1992, pp. 599–602.
[79] Y. Yamauchi, M. Yoshimi, S. Sato, H. Tabuchi, N. Takenada, and K. Sakiyam, "A new cell structure for sub-quarter micron high density Flash memory," *IEDM Tech. Dig.*, 1995, pp. 267–270.
[80] H. Kume, M. Kato, T. Adachi, T. Tanaka, T. Sasaki, T. Okazaki, N. Miyamoto, S. Saeki, Y. Ohji, M. Ushiyama, J. Yugami, T. Morimoto, and T. Nishida, "A 1.28 μm^2 contactless memory cell technology for a 3 V-only 64 Mbit EEPROM," *IEDM Tech. Dig.*, 1992, pp. 991–993.
[81] Y. S. Hisamune, K. Kanamori, T. Kubota, Y. Suzuki, M. Tsukiji, E. Hasegawa, A. Ishitani, and T. Okazawa, "A high capacitive-coupling ratio (HiCR) cell for 3 V-only 64 Mbit and future Flash memories," *IEDM Tech. Dig.*, 1993, pp. 19–22.
[82] F. Masuoka, M. Momodami, Y. Iwata, and R. Shirota, "New ultra high density EPROM and Flash EEPROM cell with NAND structure cell," *IEDM Tech. Dig.*, 1987, pp. 552–555.
[83] B. Eitan, "Flash cells future trends," presented at the 8th Workshop on Dielectrics in Microelectronics, Venice, Italy, Nov. 1996.
[84] K. Yano, I. Ishii, T. Sano, T. Mine, F. Murai, and K. Seki, "Single-electron-memory integrated circuits for Giga-to-Tera bit storage," in *Proc. ISSCC Conf.*, San Francisco, CA, 1996, p. 266.

Paolo Pavan (Member, IEEE) was born in Mirano, Venezia, Italy in 1964. He graduated in electronics engineering and received the Ph.D. degree from the University of Padova, Italy, in 1990 and 1994, respectively. From 1992 to 1993, he was a graduate student at the University of California, Berkeley.

From 1993 to 1994, he was with the University of California, Berkeley, as a Visiting Research Engineer. In 1994, he joined the University of Modena, Italy, where he is now a Senior Researcher. His research interests are in the fields of solid-state devices. He has studied breakdown phenomena. In particular, he has studied and modeled latch-up, hot-electron degradation, radiation effects in MOS devices, and impact ionization phenomena in bipolar transistors. His interests currently are also in Flash-cell reliability, were he studies dielectric properties, and in the development of new memory cells.

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

Since 1986, he has been working in the VLSI Process Development Group of SGS-Thomson Microelectronics, Agrate Brianza, Italy, in particular researching nonvolatile memories process architectures. Until 1989, he was engaged in the electrical characterization and modeling of EEPROM cells, contributing to the development of an original SPICE model of this device. From 1989 to 1993, his work focused on the development of very-high-density Flash memory devices, following the basic device architecture and studying the transient performances by the characterization of the programming and erasing physical mechanism. Since 1994, he has been Project Leader of the development of a single power supply Flash memory device. Currently, he is working in the Non-Volatile Memory Process Development Group of the Central Research and Development. His research interests have covered many aspects of microelectronics, from silicon planar technology through process architecture to the physics of electron devices, with particular attention to device modeling and reliability. He was a Lecturer in electron device physics at the University of Milan and in nonvolatile memory devices at the University of Padova and Polytechnic of Milan.

Piero Olivo was born in Bologna, Italy, in 1956. He received the Ph.D. degree from the University of Bologna in 1987.

In 1983, he joined the Department of Electronics and Computer Systems, University of Bologna. In 1991, he became Associate Professor of electronic instrumentation and measurements. In 1993, he became Full Professor of electronics at the University of Catania, Italy. In 1995, he joined the University of Ferrara, Italy. In 1986–87 and autumn 1989, he was a Visiting Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. His scientific interests are in the areas of solid-state devices and integrated circuit (IC) design and testing. In the field of solid-state devices, he has worked on SiO_2 physics, quantum effects, charge transport through thin SiO_2 structures, charge trapping in SiO_2, oxide breakdown and reliability, MOS measurement techniques, thin oxide properties, and nonvolatile memories characterization. In the field of IC design and testing, he has worked on signature analysis testing, design for testability techniques, fault modeling and fault simulation, IDDQ testing, self-checking circuits, and nonvolatile memory testing.

Enrico Zanoni (Senior Member, IEEE) was born in Legnago, Verona, Italy in 1956. He received the doctoral degree in physics from the University of Modena, Italy, in 1982.

He has worked on the characterization of electronic components for telecommunication systems and automotive electronics systems. His research interests include the study of the reliability of Si electron devices and of nonvolatile memories, the characterization of hot-electron phenomena in III–V devices. On these (and related) subjects, he has coauthored approximately 150 papers published in international journals and conference proceedings, ten invited papers, and five book chapters. He has collaborated with several electronics companies, including Alcatel, AT&T Bell Labs (now Lucent Technologies), CNET-France Telecom, Hughes Research Laboratories, IBM T. J. Watson Research Center, Marelli Autronica, NECSY, Siemens, and TRW. Since 1993, he has been a full Professor of Electronics at the Univeristy of Padova, Italy. He has participated in the past in the establishment of a laboratory devoted to the evaluation of quality and reliability of electronic components in Bari, Italy. At the University of Padova, he has been and is currently responsible for of the following European projects: EUREKA project PROMETHEUS-PROCHIP (Prometheus subprogram on Custom Hardware for Intelligent Processing) for reliability testing activities, ESPRIT III project MANPOWER, "Manufacturable Power MMIC's for Microwave Systems Applications," and Human Capital and Mobility projects. He is the Project Leader of the Standards, Measurements and Testing program "'PROPHECY' Procedures for the Early Phase Evaluation of Reliability of Electronic Components by Development of CECC Rules," a research project involving Alcatel, Siemens, ST Microelectronics, and several research laboratories.