Erratic Cell Behavior in Channel Hot Electron Programming of NOR Flash Memories

Marco Grossi, Massimo Lanzoni, and Bruno Riccò, *Fellow, IEEE*

***Abstract*—This paper describes for the first time an erratic behavior found in NOR array cells of Flash memories after cycling when programming is performed by channel hot electron injection. The effects of different program conditions (i.e., drain and bulk bias, as well as program speed) on such an erratic behavior are discussed and a possible explanation is given. Implications in terms of memory reliability are discussed, in particular for multilevel applications.**

***Index Terms*—Erratic, Flash, memories, programming.**

I. Introduction

MULTILEVEL (ML) Flash memories [1] allow for storage of more than 1 bit in a single cell, thus increasing memory density for the same technology and cell dimension. However, ML memories are more critical than their "1-bit per cell" counterparts in terms of reliability, read and write disturbs, and sensing accuracy.

In particular, tight constraints are required for programming, i.e., the placement of the right amount of charge into the memory floating gate (FG), since for the same threshold voltage (V_T) window (i.e., the difference between the highest and lowest V_T) with the increasing number (n) of bits stored in each cell, the V_T distribution width (ΔV_T) must decrease as 2^n (for example, in a 3-bit per cell memory ΔV_T of 300 mV is typically required, while for the 4-bit per cell case the maximum ΔV_T is only 150 mV).

Because of these tight constraints, any type of program disturbs or instability can easily cause a bit error.

This paper introduces a new effect, observed for the first time during channel hot electron (CHE) programming of NOR Flash memory arrays, that can degrade the device reliability. This new phenomenon, hereafter called erratic programming effect (EPE), presents some analogies with the erratic erase effect discussed in several papers [2]–[4], but differs substantially because it is connected with CHE programming instead of Fowler–Nordheim tunneling (FNT) through the cell oxide.

In virgin arrays (i.e., a memory array not yet cycled through a number of program-erase operations) EPE is not present, and all the cells show a reproducible behavior in successive program operations. On the contrary, after a few thousand program-erase cycles, a small number of cells exhibit a random variation in their program characteristic, leading to different V_T values for the same program conditions. Monitoring the entire array during several program operations, the EPE cells are found not always to be the same, but to vary randomly in the whole array, as is characteristic of typical "erratic" phenomena.

Much work has been done on the reliability problems related to program-erase cycling: in particular, many papers illustrate the oxide degradation due to cycling stress on programming and erasing times, in turn causing V_T window degradation [5], [6]. The EPE effect is, however, different from this phenomenon: in fact, while the programming time degradation discussed in [5], [6] presents a static behavior (i.e., the programmed V_T degradation is a deterministic function of the number of program-erase cycles), the EPE has an erratic behavior. In particular, in any cell it can appear and disappear from a cycle to the next one or, alternatively, if a whole array is studied, it can affect randomly a small number of cells in the array.

EPE is important not only from a physical and scientific point of view, but also for applications, as it can affect reliability, in particular for ML memories with three or four bits per cell. In this case, programming is normally obtained by program and verify (P&V) procedures, where, after each small program pulse, a read operation is carried out so as to decide whether or not a new program step should be made. As known, P&V procedures achieve narrow V_T distributions at the cost of long programming time (due to verify operations and the switching back and forth from program and read operations). From this point of view, EPE has a negative effect in that it results in a further slow down of P&V algorithms, thus reducing the program throughput (PT), i.e., number of bits programmed per second, since 1) P&V operations must start from a lower V_T value, because of induced broadening of the programmed V_T distribution, and b) the presence of a small number of cells randomly featuring enhanced programming imposes smaller programming steps (i.e., shorter program pulse durations during P&V). In the worst case, EPE can lead to cell overprogramming when it affects the last step of the P&V procedures.

EPE has never been reported until now since the produced variation in cell V_T is relatively small, so in practice it can be detected only when dealing with very narrow V_T distributions (in the order of 100 mV) and thousands of program-erase cycles. This is not the case with today's state-of-the-art two bits per cell Flash memories, whose V_T distribution width of about 500 mV are too large to be critically sensitive to EPE. Nevertheless, with the increasing number of bits per cell (and consequent reduction of ΔV_T), EPE will play a greater role in device reliability.

In order to introduce and describe EPE, this paper is organized as follows. Section II describes the Flash memories test chips and the setup used in this paper. Section III illustrates an

Manuscript received November 3, 2003; revised July 1, 2004. The review of this paper was arranged by Editor C.-Y. Lu.

The authors are with the Department of Electronics, University of Bologna, Bologna 4013, Italy (e-mail mgrossi@deis.unibo.it).

Digital Object Identifier 10.1109/TED.2004.834903

0018-9383/04$20.00 © 2004 IEEE

algorithm (based on ramped voltage programming) used to detect the presence of EPE cells and shows how this phenomenon is strongly related to program erase cycling. Section IV shows how EPE affects cycled devices also when programming is performed by box waveform programming or P&V procedures are used. In Section V the effects of program conditions on EPE are discussed and Section VI is dedicated to a simple physical explanation. Finally, conclusions are drawn in Section VII.

II. Devices and Experiments

The experiments of this paper have been carried out using test chips of Flash memories with NOR, common ground architecture, fabricated by STMicroelectronics, Milan, Italy with a production-quality 0.18-μm technology.

These test chips contain $4 \cdot 10^6$ cells (4 *Mcells*) divided into eight sectors, each with 512 Kcells, organized in 256 rows each of 2048 cells and are configurable by means of suitable latches allowing to select the operation to be performed as well as the number of cells to be simultaneously programmed.

The Flash cells are programmed by CHE injection and erased by FNT. Program operation is performed either by ramped voltage programming [7], [8], that provides good control on the charge injected into the FG, or by the conventional box waveform programming.

The cell array is realized in triple well technology, allowing the application of negative and positive substrate bias during programming and erasing, respectively [9].

The experimental setup used for this paper features two acquisition boards providing all the digital and analog signals used for memory address and control. The whole setup (acquisition boards and external instruments) is controlled via a GPIB bus, by means of application programs running under LabVIEW.

The bit-line voltage ($V_{\rm ds}$ as seen by the cell transistors) is provided by a programmable dc power supply whose floating ground terminal is connected to a shunt resistance current–voltage (*I–V*) converter, allowing measurement of current absorption during programming without affecting bias conditions.

The word-line voltage, i.e., the control gate (CG) bias of the cell transistor ($V_{\rm cg}$), is provided by a circuit integrated on the memory board. During programming, the initial digital state of $V_{\rm cg}$ is loaded into an 8-bit counter whose clock is generated by an external pulse generator, allowing signal period to be set so as to obtain the required ramp slope. The output of the counter is then used as the input for an 8-bit digital-to-analog (DAC) (dividing a 0–10 V interval in 256 steps, each of $\Delta V_{\rm step} =$ 39 mV) to provide the analog $V_{\rm cg}$ value to the cell CG.

During read operations or box programming, the word-line voltage is generated in the same way as in ramped voltage programming, but this time the clock signal is low: thus the DAC output assumes a fixed value.

Cell V_T determination is carried out by applying increasing value of $V_{\rm cg}$ until the cell begins to conduct (i.e., a 3 μA current flows between drain and source).

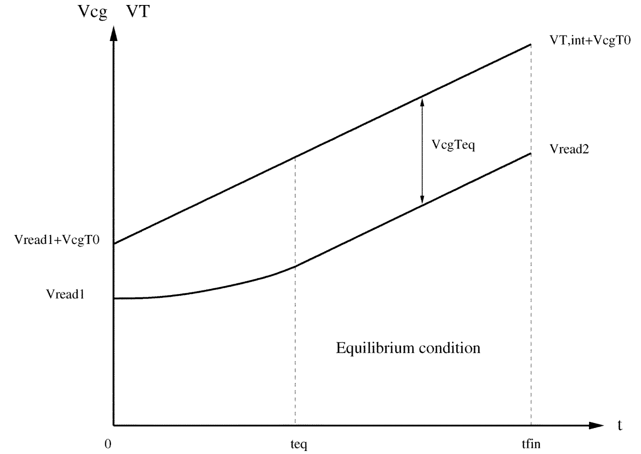

Fig. 1. Schematic representation of the evolution of $V_{\rm cg}$ and V_T with ramped voltage programming. $t_{\rm fin}$ denotes the total programming time while $t_{\rm eq}$ denotes the time required to reach the equilibrium condition where $V_{\rm cg} - V_T$ is constant.

III. EPE With Ramped Voltage Programming

In order to better understand our experiments, an important property of the ramped voltage programming method [7], [8] must be recalled. With such a method, after a transient period, an equilibrium condition is reached where the drain-source ($I_{\rm ds}$) current of the cell transistor is constant, in spite of a steadily increasing $V_{\rm cg}$. Under this condition the cell V_T increases linearly with the same slope as $V_{\rm cg}$, thus $V_{\rm cg} - V_T$ (namely the equilibrium overdrive $V_{\rm cgTeq}$) is constant (Fig. 1) at a value depending only on ramp slope ($V_{\rm cg,sl}$) and capacitative coupling between CG and FG, while the final value of V_T is determined by the time at which programming is stopped (i.e., the bit-line is grounded). Because of the dispersion in cell characteristics, $V_{\rm cgTeq}$ is different from cell to cell but for each cell it should be the same in successive program operations with the same operating conditions. However, this is not the case for cells affected by EPE, found only after substantial program-erase cycling.

To show this behavior, a whole sector of a virgin array has been program-erase cycled with the following conditions: ramped voltage programming is performed with a single voltage ramp applied to CG while $V_{\rm ds} = 3.5$ V, $V_{\rm sb} = 2$ V, $V_{\rm cg,sl} = 0.04$ V/μs, $V_{\rm T,fin} = 8$ V; instead erase is obtained with three pulses of 10 ms each with $V_{\rm cg} = V_{\rm sb} = -8.1$ V. After 0, 1500, 5000, 10000, and 20 000 program-erase cycles, a four-step programming algorithm (hereafter called program test) is carried out to look for the presence of erratic cells.

The program test (shown in detail in Fig. 2), applied to a whole sector of the array after this has been erased and then soft-programmed at $V_T = 1.6$ V, consists of the following steps.

1) The initial value ($V_{\rm read1}$) of the cell V_T is determined.
2) The cell is programmed from $V_{\rm read1}$ to an intermediate target value $V_{\rm T,int}$ using a ramped CG voltage with slope $V_{\rm cg,sl}$ and the same overdrive $V_{\rm cgT0}$ for all cells.
3) The obtained value ($V_{\rm read2}$) of V_T after this program step is determined.
4) The cell is programmed from $V_{\rm read2}$ to the final value $V_{\rm T,fin}$ with a CG voltage of slope $V_{\rm cg,sl}$ and overdrive $V_{\rm cgT} = V_{\rm cgT0} + \Delta$, where $\Delta = V_{\rm T,int} - V_{\rm read2}$.

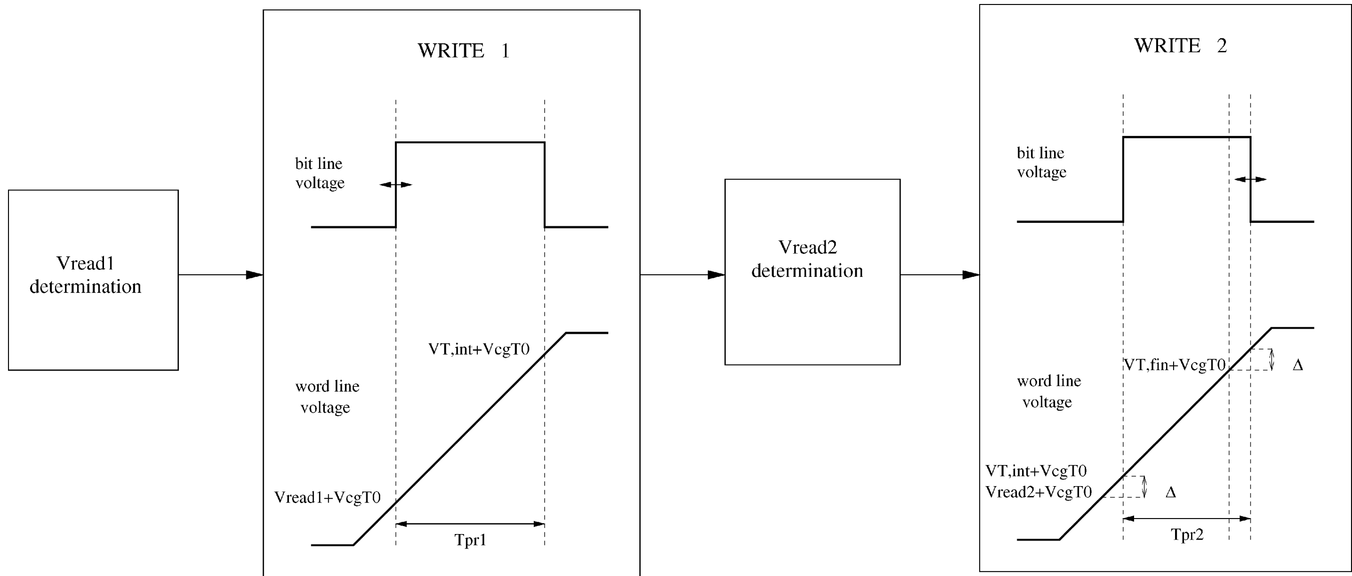

Fig. 2. Schematic representation of the four-step algorithm used to detect EPE cells. Word line and bit line voltages ($V_{\rm cg}$ and $V_{\rm ds}$, respectively) during the two program steps are shown. $V_{\rm read1}$ modulates the initial CG value of the ramp during WRITE 1, while $V_{\rm read2}$ modulates the final CG value of the ramp during WRITE 2. $\Delta = V_{\rm T,int} - V_{\rm read2}$ represents an overdrive compensation and is the difference between the target threshold value $V_{\rm T,int}$ and the value $V_{\rm read2}$ reached at the end of the first write step.

As a result of this operation, cells not exhibiting EPE have their programmed V_T value in a range of 160 mV centered on the target V_T value ($V_{\rm T,fin}$). On the contrary, cells affected by EPE (in one of the two program steps) end up outside the target voltage range.

The program test described above is derived from a method recently proposed [10] to program high-throughput ML Flash memories and is particularly interesting in the context of this paper because V_T (and so $V_{\rm cgTeq}$) is determined for each cell at the end of the two program steps separated only by a read operation. Thus, a variation in $V_{\rm cgTeq}$ between the two write operations can be only due to programming, thus making it possible to separate the effects of CHE programming from those of FNT erasing.

The results of the dependence of EPE on the number of program erase cycles in terms of maximum $V_{\rm cgTeq}$ shift and number of erratic cells are shown in Fig. 3(a) and (b), respectively. In particular, Fig. 3(a) represents the maximum deviation from $V_{\rm T,fin}$ achieved by the programming operation on a whole sector of the array, and the (represented) experimental points belong to different cells. Of course, when such a deviation exceeds the target ΔV_T, an error is produced as an effect of EPE. In this picture, the "normal" cell would produce points contained in a small region centered on the x axis. As can be seen, until 2 K program-erase cycles, no EPE is observed. After 2 K program-erase cycles, the number of cells presenting EPE and the maximum $V_{\rm cgTeq}$ shift increase with the number of program-erase cycles and tends to saturate after 10 K cycles.

Monitoring the addresses of the erratic cells during different program tests, this effect is found to be random-like, i.e., cells that are erratic during a program test, exhibit a typical behavior during others and vice versa.

For a further proof of the randomness of EPE, an entire sector of 20 K cycled cells have been programmed four times using ramped voltage programming ($V_{\rm ds} = 4$ V, $V_{\rm sb} = 1.5$ V, and

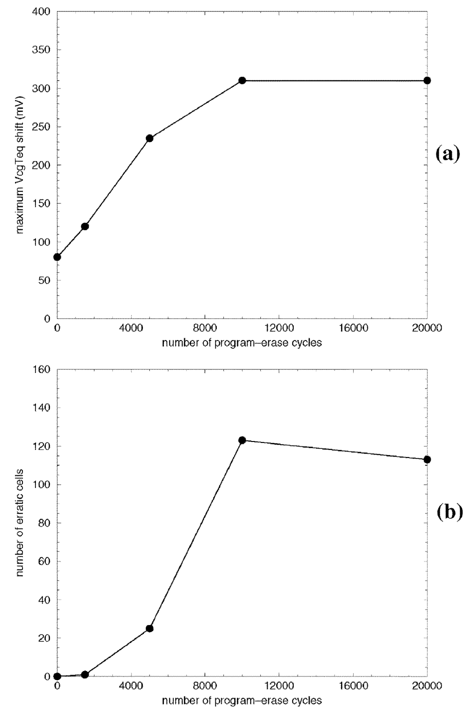

Fig. 3. Effects of the number of program-erase cycles on EPE. (a) The maximum $V_{\rm cgTeq}$ shift and (b) the number of erratic cells are represented as a function of the number of program-erase cycles. Until the array is subjected to about 2 K program erase cycles, the maximum $V_{\rm cgTeq}$ shift is less than 160 mV (the accuracy achieved by the proposed program test), so the number of erratic cells is considered to be 0. Program test is carried out under the following conditions: $V_{\rm ds} = 3.5$ V $V_{\rm sb} = 2$ V $V_{\rm cg,sl} = 0.01$ V/μs.

$V_{\rm cg,sl} = 0.08$ V/μs) without any erase operation between each programming operation. The cells initially have a V_T value of 1.5 V and are programmed four times with a final CG voltage of 3, 4.5, 6, 7.5 V, respectively: after each programming operation, V_T is measured to evalutate $V_{\rm cgTeq}$ as the difference between the final CG voltage and the measured V_T. Fig. 4 shows the value

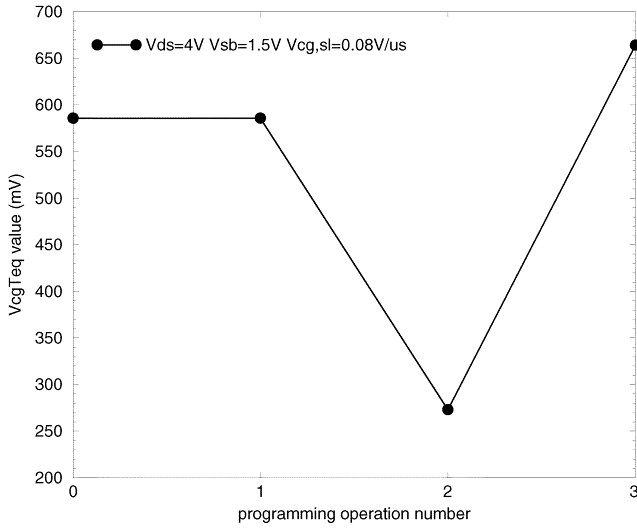

Fig. 4. $V_{\rm cgTeq}$ value during different programming operations for a cell experiencing EPE. $V_{\rm cgTeq}$ should be normally constant but, as can be seen, it decreases when EPE occurs (thus, giving rise to cell over-programming).

of $V_{\rm cgTeq}$ during the four programming operations for a cell experiencing EPE: as can be seen, $V_{\rm cgTeq}$, normally constant for all write operations on the same cell, decreases when EPE occurs, thus producing enhanced programming for one or two program operations before returning to its typical behavior.

IV. EPE With Box Waveform Programming

This section shows that EPE is found also when Box Waveform Programming (instead of the Ramped Voltage method) is used for write operations.

Different cells of the 20 K program-erase cycled array are programmed using the following conditions: $V_{\rm cg} = 7$ V, $V_{\rm ds} = 3.5$ V, $V_{\rm sb} = 2$ V and a program time of 40 μs. A large number of program operations are carried out for each cell (about 7 K) and between two consecutive ones an erase operation is performed using four pulses of 10 ms each and $V_{\rm cg} = V_{\rm sb} = -8.3$ V. Fig. 5 shows the measured variation in the programmed V_T (represented as the number of DAC states shift, each of 39 mV) with respect to the previous programming as a function of the number of operations. Fig. 5(a) represents the behavior for a normal cell, where the variation in the programmed V_T is never higher than 3 DAC steps (120 mV). Fig. 5(b), instead, shows the case of a cell subjected to EPE after about 3500 program operations: as can be seen the variation in programmed V_T is 9 DAC steps (360 mV), i.e., three times higher than the normal behavior.

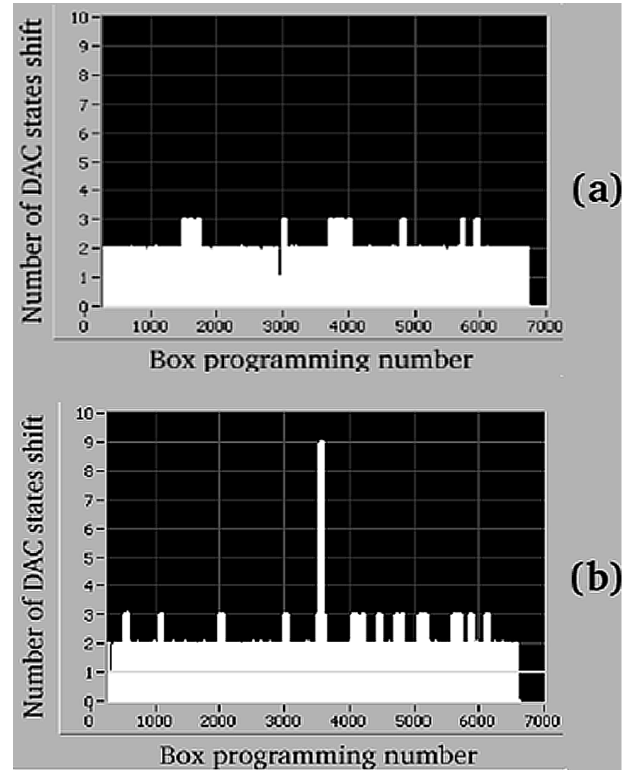

Fig. 5. Variation in programmed V_T for a 20 K cycled device between two consecutive write operations (expressed as number of DAC steps, each of 39 mV) as a function of the number of successive operations. About 7 K write operations are carried out using box waveform programming with $V_{\rm cg} = 7$ V, $V_{\rm ds} = 3.5$ V, $V_{\rm sb} = 2$ V and a program time of 40 μs. Between two program operations, the cells are erased with four pulses of 10 ms each and $V_{\rm cg} = V_{\rm sb} = -8.3$ V. The x axis represents the number of programming operations, while the y axis represents the difference in the programmed V_T with respect to the previous program operation. (a) Normal cell behavior where the maximum V_T shift is 120 mV. (b) Anomalus cell behavior due to EPE (the maximum V_T shift is now 360 mV).

As mentioned in the introduction, EPE can endanger the correct operation of ML memories even if P&V algorithms are used. In order to illustrate this point, we have used the P&V algorithm schematically represented in Fig. 6: program pulses of the same width (T_W) but with voltage amplitude increasing of $\Delta V_{\rm pp}$ from one step to the other are applied to the CG, while $V_{\rm ds}$ is constant during programming; between two write pulses, a read operation is carried out to determine if further programming is needed. The first write pulse has a sufficiently low voltage amplitude ($V_{\rm cg,min}$) to guarantee that fast cells are not overprogrammed, then the pulse amplitude increases to minimize the number of verify steps for slow cells. With the right choice for T_W, V_T distribution width as narrow as $\Delta V_{\rm pp}$ can be achieved [11].

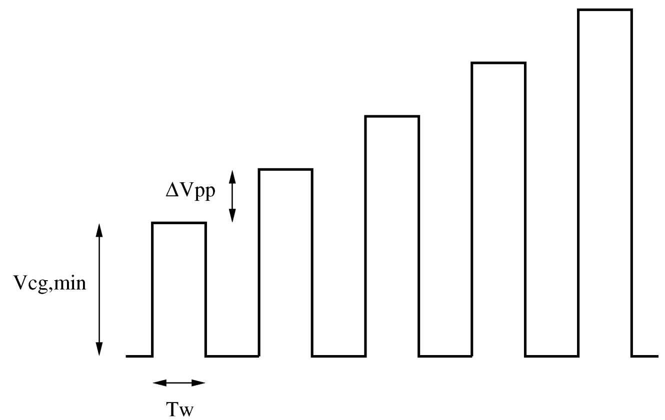

Fig. 6. Schematical representation of the algorithm of P&V used in this work. Write pulses have all the same width T_W, while the voltage amplitude increases of $\Delta V_{\rm pp}$ at each step. Between two write pulses a read operation is carried out.

This P&V algorithm has been applied to a whole sector of both virgin and cells that have experienced 20 K program-erase

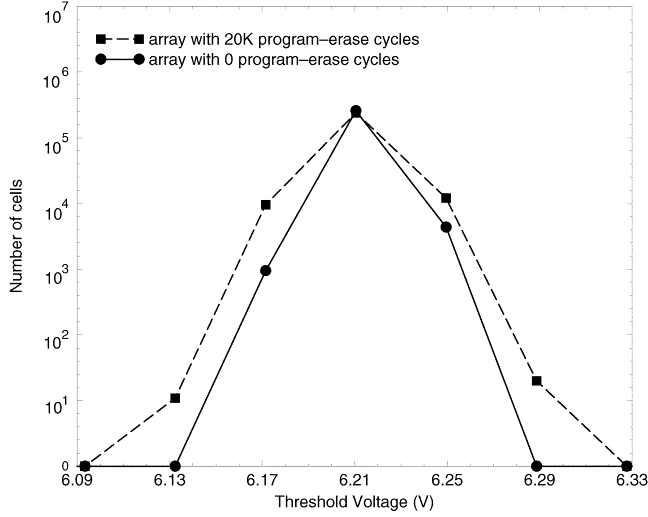

Fig. 7. V_T distributions obtained with a P&V procedure using $V_{\rm ds} = 3$ V, $V_{\rm sb} = 2.5$ V, $V_{\rm cg,min} = 6$ V, $\Delta V_{\rm pp} = 40$ mV, $T_W = 3\ \mu$s, $V_{T,\rm TARGET} = 6.21$ V for a virgin and a cycled array rispectively. In the cycled distribution the few cells overprogrammed to $V_{T,\rm TARGET} + 2\Delta V_{\rm pp}$ (6.29 V) are due to EPE, while the cells with the lowest V_T (6.13 V) are due to charge loss caused by drain stress. V_T determination has an accuracy of 40 mV.

cycles. The initial V_T distribution (centered on 5.5 V) has been programmed with the following conditions: $V_{\rm ds} = 3$ V, $V_{\rm sb} = 2.5$ V, $V_{\rm cg,min} = 6$ V, $\Delta V_{\rm pp} = 40$ mV, $T_W = 3\ \mu$s, $V_{T,\rm TARGET} = 6.21$ V. The obtained V_T distributions (represented in Fig. 7) show how in the case of the cycled sector the V_T distribution is twice as large as that for virgin cells. In fact, due to EPE in the cycled sector a small number of cells are overprogrammed to $V_{T,\rm TARGET} + 2\Delta V_{\rm pp}$ (6.29 V), contrary to the case of virgin cells, where the highest V_T is $V_{T,\rm TARGET} + \Delta V_{\rm pp}$, (namely 6.25 V). In the cycled sector the distribution broadens also on the low V_T side because of charge loss caused by drain stress and thus, of course, broadens the dispersion on the final distribution [12].

V. Effects of Program Conditions With Ramped Voltage Programming

In this section the effects on EPE of the electrical conditions used for both program-erase cycling and Program Test are investigated.

To this purpose, two virgin sectors of the memory array have been programmed and erased 20 K times using two different erase methods: sector A is erased with three pulses of 10 ms each with $V_{\rm cg} = V_{\rm sb} = -8.1$ V, while sector B is erased with 10 pulses of 10 ms each with increasing value for $-V_{\rm sb} = -V_{\rm cg} = 4.7\ {\rm V} + 0.039\,06 \cdot k$ (where $k = 0, 1, 2, \ldots, 9$ is the pulse number). The method used for sector B is much less degrading for tunnel oxide than that applied to sector A [13]. Nevertheless, the maximum $V_{\rm cgTeq}$ shift as a function of program-erase cycle number is found to be the same for the two cases, thus clearly indicating that the oxide degradation related to EPE is essentially due to CHE programming.

A further proof of the difference between EPE and erratic erase is provided by the different behavior with respect to program-erase cycling: as described in the literature [5], erratic erase is stronger on fresh devices and tends to decrease with the number of program-erase cycles, while EPE increases with the number of cycles (as pointed out in Section III) and tends to saturate after 10 K cycles.

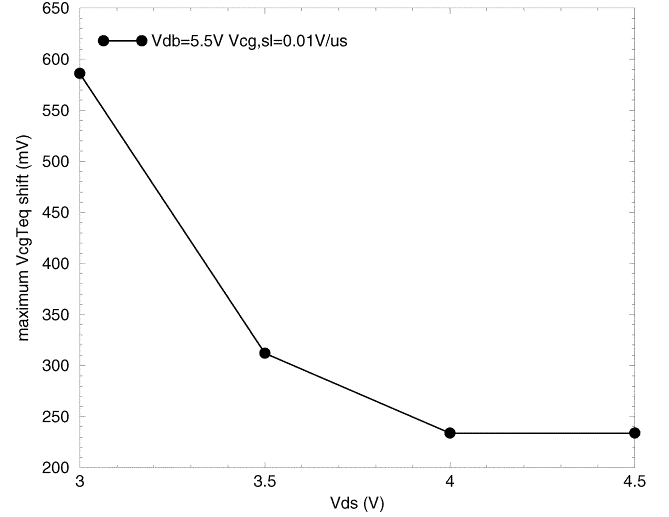

Fig. 8. Maximum $V_{\rm cgTeq}$ shift due to EPE as a function of $V_{\rm ds}$ with $V_{db} = 5.5$ V.

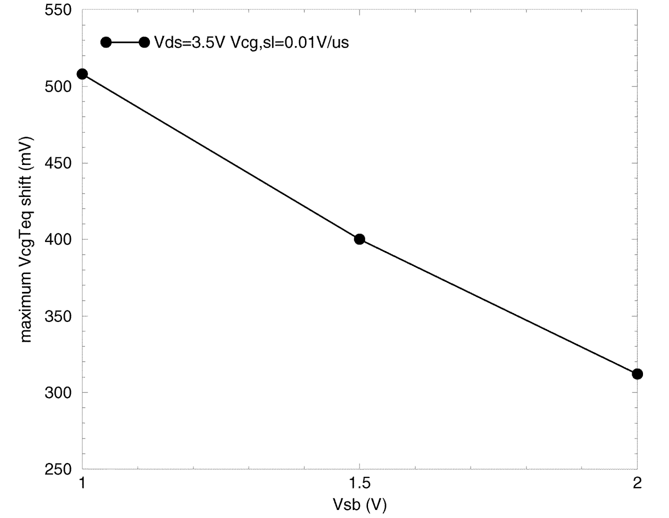

Fig. 9. Maximum $V_{\rm cgTeq}$ shift due to EPE as a function of $V_{\rm sb}$ with $V_{\rm ds} = 3.5$ V.

The effect on EPE of $V_{\rm ds}$, $V_{\rm sb}$, and $V_{\rm cg,sl}$ during the program test for a sector subjected to 20 K program-erase cycles are shown in Figs. 8–10.

In Fig. 8 the maximum $V_{\rm cgTeq}$ shift due to EPE is represented as a function of $V_{\rm ds}$, with constant $V_{\rm db} = V_{\rm ds} + V_{\rm sb} = 5.5$ V. As can be seen EPE is stronger for low values of $V_{\rm ds}$.

In Fig. 9, instead, the maximum $V_{\rm cgTeq}$ shift is represented as a function of $V_{\rm sb}$ with $V_{\rm ds} = 3.5$ V and EPE is found to decrease with the increase of $V_{\rm sb}$.

Finally, Fig. 10 illustrates the dependence on $V_{\rm cg,sl}$, for $V_{\rm ds} = 3.5$ V $V_{\rm sb} = 2$ V and, as can be seen, EPE is found to be stronger for faster voltage ramps, (i.e., higher $V_{\rm cg,sl}$).

In all the cases EPE appears to be more prominent for operating conditions with higher value of $V_{\rm cgTeq}$.

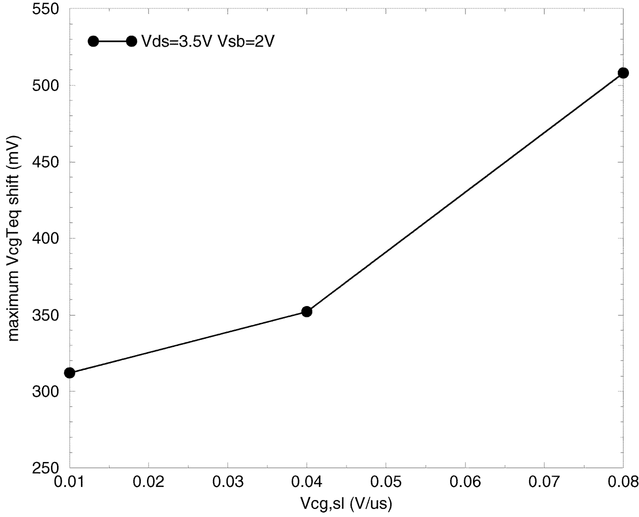

Fig. 10. Maximum V_{cgTeq} shift due to EPE as a function of the ramp slope $V_{\mathrm{cg,sl}}$ with $V_{\mathrm{ds}} = 3.5$ V $V_{\mathrm{sb}} = 2$ V.

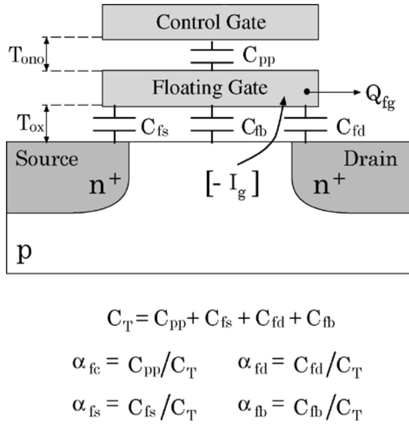

Fig. 11. Capacitative model for a flash memory cell.

VI. Physical Interpretation

This section presents a simple and preliminary physical interpretation for the observed phenomenon at the microscopical level that, although needs to be confirmed with further experiments, is consistent with our results. The proposed explanation for EPE is based on the presence of traps created at the channel-oxide interface near the drain region of the transistor cell, i.e., that where essential electron injection into the FG during CHE programming takes place.

If we consider the capacitative model for a Flash memory cell [7] (Fig. 11), FG voltage (V_{fg}) is expressed as

$$V_{\mathrm{fg}} = V_{T-\mathrm{fg}} + \alpha_{\mathrm{fc}}(V_{\mathrm{cg}} - V_T) + \alpha_{\mathrm{fd}}(V_{\mathrm{ds}} - V_{\mathrm{dT}}) \quad (1)$$

where $V_{T-\mathrm{fg}}$ is the FG threshold voltage (i.e., the V_{fg} value at which the cell turns on), V_{dT} is the drain voltage used for V_T measurements, while α_{fc} and α_{fd} are coupling ratios defined in Fig. 11.

Assuming ramped voltage programming with $V_{\mathrm{ds}} = 3.5$ V, $V_{\mathrm{sb}} = 2$ V, $V_{\mathrm{cg,sl}} = 0.08$ V/μs (thus $V_{\mathrm{cgTeq}} = 1.2$ V) and $V_{\mathrm{T-fg}} = 0.8$ V, $\alpha_{\mathrm{fc}} = 0.8$, $\alpha_{\mathrm{fd}} = 0.1$, $V_{\mathrm{dT}} = 1.9$ V, it is $V_{\mathrm{fg}} = 1.92$ V.

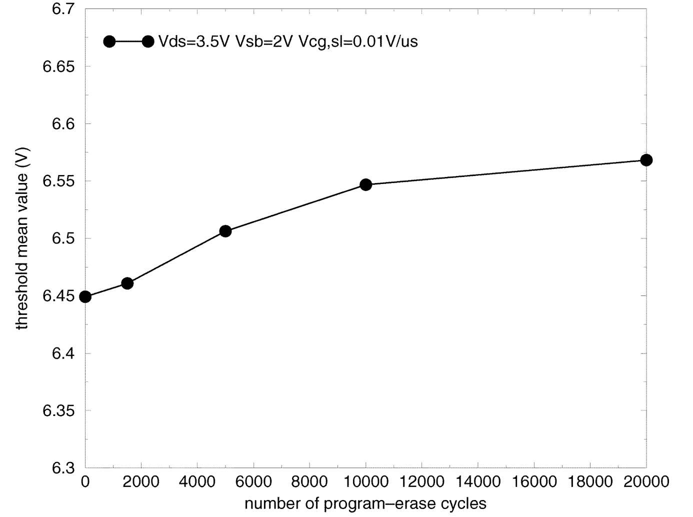

Fig. 12. V_T distribution mean value as a function of the number of program-erase cycles. Writing is performed by ramped voltage programming under the conditions $V_{\mathrm{ds}} = 3.5$ V, $V_{\mathrm{sb}} = 2$ V, $V_{\mathrm{cg,sl}} = 0.01$ V/μs.

Since $V_{\mathrm{fg}} < V_{\mathrm{ds}}$ the transverse electric field at the drain region of the cell transistor is repulsive for electrons and actractive for holes. Thus, holes tend to accumulate at the channel-oxide interface near drain region and, if they can be captured (in the traps mentioned before), the transverse electric field becomes less repulsive for electrons, thus producing enhanced CHE programming. The erratic nature of this phenomenon is then inherently due to carrier trapping de-trapping phenomena.

Fig. 12 represents the mean value of the programmed V_T distribution during the 20 K program-erase cycling (writing is performed by ramped voltage programming under the conditions $V_{\mathrm{ds}} = 3.5$ V, $V_{\mathrm{sb}} = 2$ V, $V_{\mathrm{cg,sl}} = 0.01$ V/μs), and shows a slight increase with the number of program-erase cycles. This is consistent with the results of [6]: in fact, with our ramped voltage programming, V_{fg} is fixed during the whole operation [7] at a value around $V_{\mathrm{ds}}/2$; thus, interface states are created near the drain region and almost no programming V_T degradation is expected. Instead, the slight increase in the V_T distribution mean value is due to the CHE programming enhancement of a small number of the array cells in the tail of the distribution due to the effect of EPE.

With this model for EPE, the results of Section V can be easily explained. In fact, as pointed out in [14], CHE degrades the channel-oxide interface near the drain region, creating interface traps whose number increases with stress time (thus with the number of program erase cycles) with a considerable saturation of the induced trap density under the condition of constant I_{sub} stress. Our test chips are cycled using the ramped voltage programming, essentially leading to constant I_{ds} during the programming operation. Under this condition, presumably also I_{sub} is constant, although in our experiments it cannot be measured since the substrate is common to all cells in the test chip. More precisely, program-erase cycling produces oxide degradation, inducing small I_{ds} variations (in particular, during the 20 K cycles I_{ds} varies less than 10%); thus I_{sub} too can be considered to exhibit a similar behavior with program cycling.

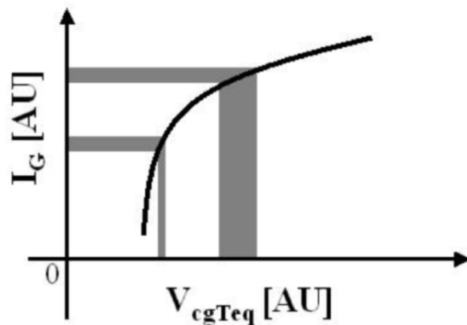

Fig. 13. Qualitative representation of the injection current into the FG as a function of V_{cgTeq} for different bias conditions. As can be seen, the same dispersion on I_g corresponds to larger dispersion in V_{cgTeq} for higher values of V_{cgTeq}.

The behavior of EPE is strictly related to the trapping and de-trapping dynamics of the interface states, hence with physical mechanisms similar to those responsible for "random telegraph signals" in devices current [15], [16]. Therefore, the fact that both the number of erratic cells and the maximum V_{cgTeq} shift increase with the number of program erase cycles and tend to saturate after 10 K cycles is then explained. Furthermore, decreases in V_{ds} and V_{sb} as well as increases in $V_{\mathrm{cg,sl}}$ all lead to higher values of V_{cgTeq}, and, as can be seen in Fig. 13, the same variations of I_g produce larger variation in V_{cgTeq} (thus on the programmed V_T) for higher values of V_{cgTeq} (for a detailed description of the $I_g - V_{\mathrm{cgTeq}}$ characteristic and numerical values of the parameters, see [8]). Consequently, for the same number of trapped holes, the variation of I_g is (about) the same for all bias conditions but the programmed V_T shift is larger for operation conditions featuring higher V_{cgTeq}.

VII. Conclusion

In this paper, a new effect that can degrade the reliability of ML Flash memories was described. This effect, called EPE consists of the fact that when cycled through some thousands program-erase operations, a few cells in a large array are programmed faster than they usually do. This effect is erratic in that the cells exibiting anomalous behavior are not the same, but change from one program operation to the next one. EPE is attributed to the presence of traps in the channel-oxide region at the drain side, produced by CHE writing.

It has been shown how, due to EPE, cycled devices present larger V_T distribution width also when P&V algorithms are used and, although the V_T distribution broadening is still not a problem for today technology, this could represent a severe limitation for future generations of Flash memories featuring three or four bits per cell.

Acknowledgment

The authors would like to thank Dr. A. Modelli, STMicroelectronics, Agrate, Italy, for helpful discussions and for providing the devices used in this work.

References

[1] B. Riccò, G. Torelli, M. Lanzoni, A. Manstretta, H. Maes, D. Montanari, and A. Modelli, "Nonvolatile multilevel memories for digital applications," *Proc. IEEE*, pp. 2399–2421, Dec. 1998.

[2] T. Ong, A. Fazio, and N. Mielke, "Erratic erase in *ETOX* flash memory array," in *Symp. VLSI Tech. Dig.*, 1993, pp. 83–84.

[3] A. Chimenton and P. Olivo, "Erratic erase in flash memories Part I—Basic experimental and statistical characterization," *IEEE Trans. Electron Devices*, vol. 50, pp. 1009–1014, Apr. 2003.

[4] ——, "Erratic erase in flash memories Part II—Dependence on operating conditions," *IEEE Trans. Electron Devices*, vol. 50, pp. 1015–1021, Apr. 2003.

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

[6] S. Yamada, Y. Hiura, and T. Yamane, "Degradation mechanism of flash EEPROM programming after program-erase cycles," in *IEDM Tech. Dig.*, 1993, pp. 23–26.

[7] D. Esseni, A. D. Strada, P. Cappelletti, and B. Riccò, "A new and flexible scheme for hot-electron programming of nonvolatile memory cells," *IEEE Trans. Electron Devices*, vol. 47, pp. 125–133, Jan. 1999.

[8] R. Versari, D. Esseni, G. Falavigna, M. Lanzoni, and B. Riccò, "Optimized programming of multilevel flash EEPROMs," *IEEE Trans. Electron Devices*, vol. 49, pp. 1641–1646, Aug. 2001.

[9] C. Auricchio, R. Bez, A. Losavio, A. Maurelli, C. Sala, and P. Zabberoni, "A triple well architecture for low-voltage operation in submicron CMOS devices," in *Proc. Eur. Solid State Device Res. Conf.*, 1996, p. 613.

[10] M. Grossi, M. Lanzoni, and B. Riccò, "A novel algorithm for high throughput programming of multi-level flash memories," *IEEE Trans. Electron Devices*, vol. 50, pp. 1290–1296, May 2003.

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

[12] A. Chimenton, A. S. Spinelli, D. Ielmini, A. L. Lacaita, A. Visconti, and P. Olivo, "Drain-accelerated degradation of tunnel oxides in flash memories," in *IEDM Tech. Dig.*, 2002, pp. 167–170.

[13] A. Chimenton, P. Pellati, and P. Olivo, "Constant charge erasing scheme for flash memories," *IEEE Trans. Electron Devices*, vol. 49, pp. 613–618, Apr. 2002.

[14] C. Duvvury, D. J. Redwine, and H. J. Stiegler, "Leakage current degradation in N-MOSFET's due to hot-electron stress," *IEEE Electron Device Letters*, vol. 38, pp. 579–581, Nov. 1988.

[15] Z. Celik-Butler, "Low-frequency noise in deep submicron metal-oxide-semiconductor field-effect transistors," *Proc. Circuits Devices Syst.*, pp. 23–31, Feb. 2002.

[16] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "Random telegraph noise of deep-submicrometer mosfet's," *IEEE Electron Device Lett.*, vol. 11, pp. 90–92, Feb. 1990.

Marco Grossi was born in Bologna, Italy, on May 31, 1973. He received the Laurea degree in electronic engineering from the University of Bologna, in 2000. In 2001, he joined the Department of Engineering, University of Bologna as a Ph.D. student, where he currently works.

His research interests are focused in characterization of nonvolatile memories. He is currently working in the field of Flash memories and the multilevel programming of these memories using the ramped-gate-technique.

Massimo Lanzoni was born in Bologna, Italy, on August 9, 1961. He received the Laurea degree in Ingegneria Elettronica from the University of Bologna in 1987.

He has been with the Microelectronics Research Group, Department of Electronics, University of Bologna, working on research projects in the fields of the experimental characterization and simulation of EEPROM memory cells and MOS devices and the automatic test of VLSI devices. His scientific interests cover the characterization of thin-dielectrics reliability, nonvolatile memory cell characteristics and reliability, MOS transistor experimental characterization, and new techniques for IC testing such as nonvolatile memories endurance testing and CMOS IC latch-up testing. He is now involved in projects concerning analog applications of nonvolatile memories and multilevel programming.

Bruno Riccò (SM'91–F'03) was born in Parma, Italy, on February 8, 1947. In 1971, he received the B.S. degree in electrical engineering from the University of Bologna, Bologna, Italy, and in 1975, he received the Ph.D. degree from the University of Cambridge, Cambridge, UK.

He has worked at the Cavendish Laboratory, Cambridge, UK. In 1980, he became Full Professor of applied electronics at the University of Padova, Padova, Italy, and in 1983, he joined the Department of Engineering, University of Bologna. Since 1978, he has been holding courses on electron devices, digital integrated electronics, semiconductor technology, and IC reliability and testing. He has been a Visiting Professor with the University of Stanford, Stanford, CT; the IBM Thomas J. Watson Research Center, Yorktown Heights, NY; and the University of Washington, Seattle. He has been consulting for major companies interested in IC fabrication and evaluation and for the Commission of the European Union in the definition, evaluation, and review of research projects in microelectronics. He has worked in the field of solid-state devices and ICs, making many contributions to the understanding and modeling of electron transport, tunneling in heterostructures and thin-insulating films, silicon dioxide physics, MOSFETs physics, latch-up in CMOS structures, device modeling, and simulation. He is currently working in the field of IC design, evaluation, and testing. He has authored or co-authored over 300 publications, more than half of which have been published in major international journals. He has also written three books and holds six patents in the field of nonvolatile memories.

Dr. Riccó received the G. Marconi Award from the Italian Association of Electrical and Electronics Engineers (AEI) in 1996 for his research in electronics. In 1996, he became President of the Group of Electron Devices, Technologies, and Circuits of AEI and, from 1998 to 2001, became President of the Italian Group of Electronics Engineers. He was European co-chair at the International Electron Device Meeting (IEDM) from 1992 to 2001, Vice-Chairman of the North Italy Section of IEEE from 1999 to 2001, Vice President of the North Italy Section of IEEE in 2000, and Chairman of the IEEE–North Italy since 2002. He was the European Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES from 1986 to 1996.