Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
Corso di Laurea in Ingegneria Elettronica

Statistical analysis of Total Ionizing Dose response in 25-nm NAND Flash memories

Laureanda:
Federica Ferrarese

Relatore:
Prof. Alessandro Paccagnella

Correlatori:
Dott. Marta Bagatin
Dott. Simone Gerardin

Anno Accademico 2013/2014

Abstract

Flash memory is one of the most widely used non-volatile information-storage device today, as portable store media in cellphones, cameras, music players and other portable devices. The floating-gate transistor cell, which is the base element on Flash memory, is being aggressively scaled, especially for NAND devices, to increase the memory capacity.

In addition, Flash memories are also interesting for space applications. Commercial devices are attractive for space, due to low cost per bit, but they are sensitive to radiation effects. The reason is that commercial Flash memories are characterized by higher performance than their rad-hard analogues, however it is necessary to study the radiation response to improve them against radiation-induced malfunctions. Ionizing particles, impinging on devices operating in space environment, cause a wide range of effects, from the function wear out to temporary effects or permanent ones, leading to unacceptable conditions for complex system, which have to work for long times without maintenance.

The purpose of this work is to analyze the bit error variability in response to Total Ionizing Dose (TID) in 25-nm SLC NAND Flash memories. A large number number of devices were exposed to gamma rays from Co-60 source and radiation induced Floating Gate errors, i.e. bit flips, were collected. The amount of errors were large enough to allow a statistical analysis of the results. Furthermore, radiation-induced effects were analyzed taking into account sources of statistical variability, related to scaling issue, since the nanometric size of the devices introduces several reliability problems, due to granularity of charge and matter.

For this reason more than 1 Terabit of cells were irradiated, from two different lots, analyzing cell-to-cell and lot-to-lot variability. In this work for the first time FG errors, due to TID effects, are statistically analyzed in Flash NAND samples. There are no previous studies with such a large amount of memories irradiated. The results of this thesis intend to produce a contribution in the study of Flash memory reliability, since this device is becoming more and more attractive for space market, as solid state driver for data storage.

Sommario

Attualmente la memoria Flash è uno dei più diffusi dispositivi non volatili per l'immagazzinamento di informazioni, utilizzata in molte applicazioni di uso quotidiano come telefoni cellulari, fotocamere, lettori musicali ed alti dispositivi portatili. Il transistor a gate flottante, che costituisce l'elemento base della memoria Flash, ha subito un'aggressiva riduzione delle dimensioni, soprattutto nei dispositivi di tipo NAND, per aumentare la capacità della memoria.

Inoltre la memoria Flash è utilizzata anche in applicazioni spaziali, grazie al basso costo per bit, nonostante la sensibilità alle radiazioni. La ragione è che le memorie Flash commerciali sono caratterizzate da prestazioni superiori rispetto agli analoghi rad-hard, tuttavia è indispensabile studiarne la risposta alla radiazione, per prevenire eventuali malfunzionamenti. Le particelle ionizzanti, incidenti sui dispositivi che operano in ambiente spaziale, possono causare una vasta gamma di effetti, dall'usura ad effetti temporanei o permanenti, portando a una condizione inaccettabile in un sistema complesso, il cui funzionamento deve essere garantito per un lungo tempo, senza alcun intervento di manutenzione.

L'obiettivo di questo lavoro è di analizzare la variabilità dei bit errors dovuti alla dose totale ionizzante (TID) in memorie Flash SLC da 25 nm. Un gran numero di dispositivi è stato esposto a raggi gamma da Co-60 e sono stati misurati gli errori del Floating Gate, ovvero i bit flip, indotti dalla radiazione ionizzante. La quantità degli errori osservati è stata significativa, in modo da permettere un'analisi statistica dei risultati. Inoltre, gli effetti indotti dalla radiazione ionizzante sono stati considerati congiuntamente alle sorgenti di variabilità statistica relative allo scaling dei dispositivi, poiché le dimensioni nanometriche della cella introducono una serie di questioni affidabilistiche, dovute alla granularità della carica e della materia.

Per questo motivo è stato irraggiato più di un Terabit di celle, prelevate da due differenti lotti, ed è stata analizzata la variabilità tra lotti e tra dispositivi. Per la prima volta, irraggiando una così ampia popolazione di campioni, si sono analizzati statisticamente gli errori del Floating Gate dovuti a TID, in memorie Flash di tipo NAND. Con questa tesi si intende apportare un contributo nello

studio dell'affidabilità delle memorie Flash in applicazioni per uso spaziale, dato che questi dispositivi si stanno largamente diffondendo sul mercato come unità a stato solido per l'immagazzinamento di informazioni.

Contents

Abstract **I**

Sommario **III**

1 Introduction **1**

1.1 Non-volatile memory 1

1.2 Motivation 3

1.3 Thesis Organization 4

2 The Device **5**

2.1 Floating Gate Transistor 5

2.1.1 The Reading Operation 8

2.1.2 Charge Injection and Removal Mechanisms 9

2.1.3 Threshold Voltage Distribution 13

2.2 Array Organization 16

2.2.1 NOR Architecture 16

2.2.2 NAND Architecture 18

2.3 Reliability 20

2.3.1 Retention 20

2.3.2 Endurance 21

2.3.3 Statistical Effects 22

2.4 Scaling Issues 30

2.5 Conclusions 31

3 Radiation Effects **33**

3.1 Space Radiation Environment 33

3.2 Basic Concepts 37

3.2.1 Dosimetry 38

3.2.2 Basic Radiation Effects in Device 39

3.3 Total Ionizing Dose 39

3.3.1 General Overview 39

3.3.2 Charge Yield 42

3.3.3 Oxide Traps Neutralization 43

3.4 Radiation Induced Failure in Flash memory 44

3.4.1 Floating Gate Cell 44
3.4.2 Peripheral Circuitry 48
3.5 Conclusions 50

4 Experiment **51**
4.1 Estec Co-60 Facility 51
4.2 Experimental Procedure 53
4.2.1 Calibration 53
4.2.2 TID test setup 59
4.3 Conclusions 65

5 Variability Analysis **67**
5.1 Threshold Voltage Spread 67
5.1.1 Lot-to-lot variability 71
5.1.2 Chip-to-chip variability 75
5.2 Errors Spatial Distribution into Pages 82
5.2.1 Device Decap 88
5.3 Annealing 88
5.4 Current 92
5.5 Conclusions 94

6 Final Considerations **97**

A Program and Read Scripts **99**

B Errors distribution across pages **101**

Bibliography **106**

List of Figures

1.1 Non volatile memory market 2
1.2 Flash applications 3

2.1 Schematic of a charge-based cell 5
2.2 Drain current as a function of gate voltage 6
2.3 Schematic of capacitive coupling components in the FG cell. . . . 7
2.4 Energy band diagram 9
2.5 Energy band diagram for HEI 11
2.6 FN tunneling 12
2.7 FN tunnel current versus electric field 13
2.8 ISPP algorithm 15
2.9 Threshold voltage distribution SLC and MLC 15
2.10 NOR architecture 17
2.11 Program and erase in NOR architecture 17
2.12 NOR architecture 19
2.13 Program and erase in NAND architecture 20
2.14 Threshold voltage window closure 22
2.15 Over-erasing effect 23
2.16 Erratic bit 24
2.17 Effects of cycling on SILC 25
2.18 SILC 25
2.19 TAT 26
2.20 Anode Hole Injection 27
2.21 TAT and 2-TAT 27
2.22 EIS 28
2.23 Random Telegraph Signal 29

3.1 Van Allen belts 34
3.2 Van Allen belts, AP-8 and AE-8 35
3.3 Solar cycles 36
3.4 Coronal mass ejection 36
3.5 Interaction of cosmic galactic rays rays with the atmosphere . . . 37

3.6 TID damage 41
3.7 Mechanisms of TID induced V_T shift 45
3.8 Schematic representation of a FG memory cell 46
3.9 TID effects on NOR 47
3.10 Sketch of the of V_T distribution shift 48
3.11 Charge pump degradation 50

4.1 ESTEC Facility 52
4.2 Board and socket 54
4.3 Calibration setup 56
4.4 Dose rate distribution normalized to dosimeter readings 56
4.5 Calibration: errors build up 57
4.6 Calibration: errors percentage build up 58
4.7 Calibration: check program time 59
4.8 TID test setup 63
4.9 Errors vs total dose 65
4.10 Errors normalized vs total dose 66

5.1 Errors percentage versus total dose 68
5.2 V_T spread vs technology node 69
5.3 ΔV_T vs ΔW 70
5.4 ΔV_T vs ΔL 70
5.5 C_{PP} vs technology node 72
5.6 Electrons per bit (N) vs technology node 72
5.7 Distribution of pre-rad errors 74
5.8 Ratio of mean errors per chip 74
5.9 Histogram and QQ of V_T distribution with applied RDD and ITC. 77
5.10 Probability distribution 79
5.11 Probability distribution compared with Poisson distribution . . . 80
5.12 Sketch of threshold voltage distribution shift 81
5.13 Errors distribution after a total dose of 33 krad(Si). 81
5.14 Errors spatial distribution into pages 83
5.15 Errors spatial distribution into blocks 84
5.16 Even and odd blocks during calibration errors buid up 84
5.17 Retention errors distinguishing even and odd blocks 85
5.18 Errors spatial distribution into pages, fresh devices 85
5.19 Probability distribution, even and odd blocks, Lot A 86
5.20 Probability distribution, even and odd blocks, Lot B 87
5.21 Decapped memory 88

5.22 Annealing errors 89
5.23 Annealing, normalized errors, up to 24 hours 90
5.24 Annealing, normalized errors 90
5.25 Errors distribution over time 93
5.26 Current during E/P/R cycle 94
5.27 Program current 95
5.28 Program time 95

B.1 Errors spatial distribution into pages: A11-A20 102
B.2 Errors spatial distribution into pages: A21-A30 103
B.3 Errors spatial distribution into pages: B1-B10 104
B.4 Errors spatial distribution into pages: B11-B20 105

List of Tables

2.1 Differences between NOR and NAND Flash memories 17

3.1 Radiation effects in spacecraft electronics 40

4.1 Details of the SLC NAND memories 54
4.2 Gamma TID run on the calibration device 55
4.3 Gamma TID runs during week 47 and 48 61
4.4 Gamma TID runs during week 49 and 50 62
4.5 Mean of errors per lot 63

5.1 QQ plot 78

Chapter 1

Introduction

In this introductory chapter the main concepts and metrics of non volatile memory will be presented, just followed by a brief historical evolution of the previous technologic solutions which have lead to the Flash memory device, object of the thesis. Alternative solutions to the data storage mechanism used by floating gate devices and future technological trends will be presented.

1.1 Non-volatile memory

A non volatile memory is a memory able to retain digital information without any power supply. This feature makes it an essential element of most systems, since critical information must be retrieved even if power is lost to the system. Non volatile memory market has been growing in the last two decades, because of the broad diffusion of portable electronic devices, such as smartphones, laptops, digital cameras and multimedia players. The huge success and market increase of Flash memories is due to their flexibility against the previous technology, read-only memory (ROM) and Programmable-ROM, but especially due to the remarkable scaling of these devices, which allowed a 50× cost-per-bit reduction from 1.5 μm (1987) to 20 nm (2012) technology nodes [1]. Flash memories combine the electrical erase from EEPROMs (electrically erasable and programmable read-only memory), which have the disadvantage to use a large area, and the high density from EPROMs, which are electrically programmable but erasable via ultraviolet exposure.

The most important connotative parameters for non-volatile memories are retention and endurance. They give the measure of how much strong is the NVMs, because this type of memory is more subject to wear-out than standard digital circuits.

Retention is the ability of the cell to keep the information for a long amount of time and it is measured in years. At this moment a typical value is ten years.

Endurance is the number of erase/program operations that can be executed

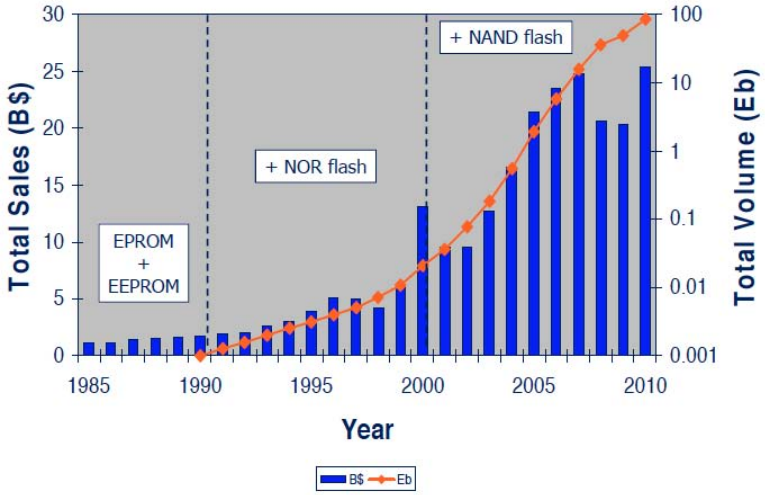


Figure 1.1: Non volatile memory market (Source: Micron [2])

without cell's properties degradation. It is measured in number of cycles (a typical value for floating gate memories is 10^5 cycles).

Other characteristics of interest for non-volatile memories are the following.

1. Density: the number of bits of information that may be stored on a single device.

2. Program or write time: the time necessary to write information into the memory.

3. Read time: the time necessary to read information from the memory.

4. Erase time: the time necessary to erase information from the memory. A peculiarity of non-volatile memories is that they must perform an erase operation before the program operation. This implies a complex peripheral circuitry and a speed penalty, because of the additional operation.

5. Operating power: the power necessary to perform the write, erase or read operation.

6. Standby power: the power consumed by the memory when no operations are being performed. In a non-volatile memory the power required to retain information is zero.

Various storage mechanisms differentiate various non volatile memories. The simplest one is the storage of a net amount of charge, whose presence or absence, different quantities or its sign represent the digital information. To this type of NVMs belong Floating Gate (FG), Charge Trap (CT), and Nanocrystal (nXTL)

memories. Phase change memories base the information storage on the phase (amorphous or crystalline structure) because of the very different behavior of these materials depending on the phase. Ferroelectrics materials use the direction of a remanent polarization (as a consequence of removal of an electric field) to make non-volatile memory bits. Instead ferromagnetic materials use the magnetoresistance, the property to change resistivity when immersed in a magnetic field. Other prototypes of storage mechanisms are under developing, but still immature to get commercialized, such as nanotube RAM, resistance RAM and Conductive-Bridging RAM [3].

1.2 Motivation

Flash memory is the most widely used non-volatile memory device today. As represented in Fig. 1.2, market application is very extended and the demand for increasing memory capacities, achieved by device scaling down, is in continuos increase. Nowadays in mainstream market are available 1 Gb NOR Flash memory manufactured in 45-nm technology, while NAND Flash memory has reached 32 Gb capacity in 25-nm, since NAND structure is characterized by high-density array.

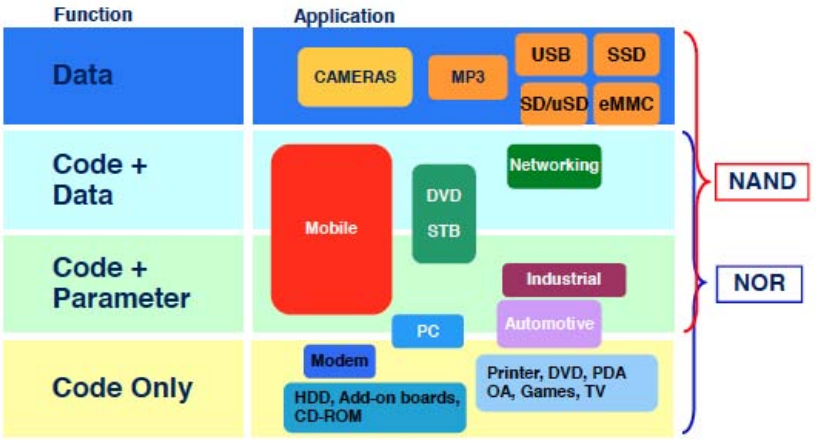


Figure 1.2: Applications of Flash memory device (source: Micron [2]).

Commercial NMVs are also used in space applications, since rad-hard market presently provides low capacity devices. In particular Flash memory devices are of great interest for space, because of the high-density storage and non-volatility. Obliviously, parts for space systems are subjected to reliability test, to study radiation induced effects on devices operation.

With the scaling down trends, many reliability issues have emerged, both intrinsic and radiation-related. Scaled devices have shown to be more prone to

intrinsic charge loss mechanisms, such as stress induced leakage currents, erratic bits, etc. Also variability increases as cell size decreases. Several studies analyzed the variability effects on nanoscale Flash memories [4] [5] [6]. On the other hand, investigations on ionizing radiation effects have showed that in advanced Flash memory devices errors appear at lower radiation doses than in older devices.

However, the combined effect of radiation induced effects and intrinsic sources of variability is quite unexplored. This research field is very interesting in order to allow precise prediction of radiation-induced failure during space missions. The purpose of the work is to investigate the impact of variability on TID response of NAND floating gate cells. For this reason, a large number of FG cell has been tested (more than 10^{12} cells), from two different lots. Statistical distribution of parameters has been analyzed, aiming at looking for variability response to TID, such as correlation between pre-rad and post-rad results.

1.3 Thesis Organization

The thesis is organized in 5 chapters, including the introductive chapter.

In chapter 2, a description of the floating gate memory cell and its operation principle are provided. The architecture types are presented and charge injection mechanisms are analyzed. Then, reliability issues, even related with the scaling down trend, are discussed.

Chapter 3 studies radiation effects on Flash memories. The first section introduces the basic concepts of the radiation environment. Subsequently, after an overview of the radiation effects on electronic devices, both Total Ionizing Dose and Single Event Effects on Flash memory device are explained.

In chapter 4 the experimental conditions of the TID test are provided, performed at ESTEC (Noordwijk, NL). The tested device characteristics and the test setup are described. The final section presents the results, which are analyzed in the following chapter.

Chapter 5 develops the statistical analysis of the results, providing explanation to both the cell-to-cell and lo-to-lot variability observed.

Chapter 2

The Device

This chapter is devoted to the description of Flash memory cell. Fist of all, the operating principle of the FG cell will be exposed, thus the physic structure and the electric characteristics. Moreover two possible array organizations will be presented. Then the main reliability issues related to this data storage solution will be analyzed and discussed.

2.1 Floating Gate Transistor

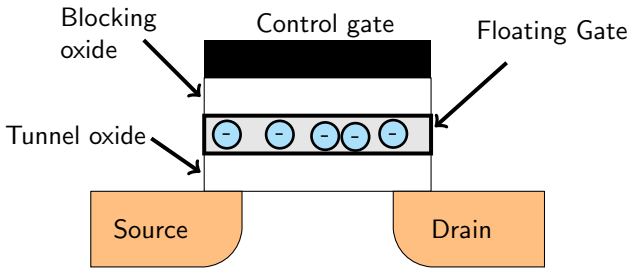


Figure 2.1: Schematic of a charge-based cell.

A Flash memory is a non-volatile memory that can be erased and reprogrammed. By the definition, a NMV cell doesn't need any power supply to retain digital values. The mechanism employed to retain the non-volatile digital information is the charge storage, for which it is necessary to create a potential well, where the charge can be confined. The basic element is the floating-gate transistor, represented in Fig. 2.1. It similar to a MOSFET, but there is an addition element, the charge storage element, placed between the silicon bulk and the gate, isolated from them by a tunnel oxide and a blocking oxide respectively. The charge storage element consists in a conductive polysilicon floating gate[1], in which the charge is collected at the oxide interface or in discrete trapping sites.

[1]For other charge-based memories, like Charge Trap (CT) and and Nanocrystal (nXTL) memories, the charge storage element is respectively a dielectric film, with high density of traps to capture charge carriers, and a layer of nanocrystals, each able to store charge.

The amount of charge in the storage layer is responsible for the change of the threshold voltage of the transistor. Fig. 2.2 represents the drain current versus the gate voltage for a floating gate transistor. The left curve is obtained in the condition of a positive or no charge in the floating gate, which is called *erased*. The right curve is proper to a net negative charge in the storage element, that is the *programmed* state.

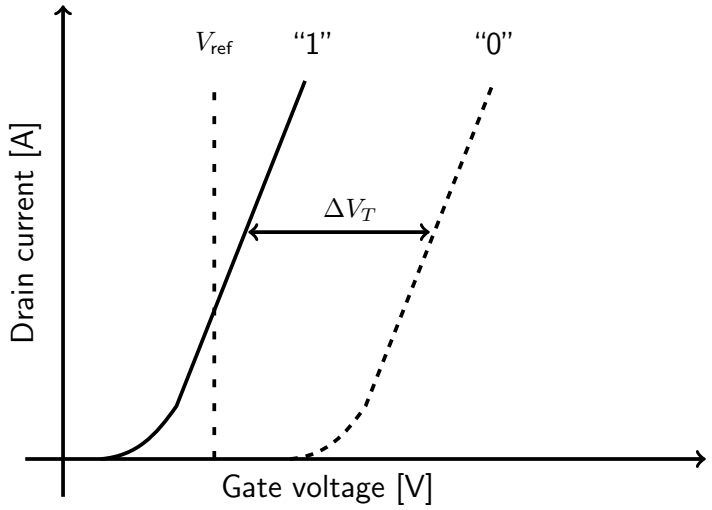


Figure 2.2: Drain current as a function of gate voltage for cell in *erase* ("1") and *program* ("0") state.

The model in Fig. 2.3 helps to understand the behavior of the FG device. C_{PP} is the capacitance between the FG and the control gate, while C_S, C_B and C_D between the floating gate and respectively the source, the substrate and the drain[2]. In the condition of no charge stored in the FG:

$$Q = 0 = C_{PP}(V_{FG} - V_{CG}) + C_S(V_{FG} - V_S) + C_B(V_{FG} - V_B) + C_D(V_{FG} - V_D) \quad (2.1)$$

where V_{FG} is the potential on the FG, V_{CG} is the potential on the control gate, V_S, V_B and V_D are the potential on the source, the bulk, and the drain respectively. Defining $C_{TOT} = C_{PP} + C_S + C_B + C_D$ as the sum of the capacitances and $\alpha_J = C_J/C_{TOT}$ as the ratio between the capacitance on one of the four terminals and the total capacitance, it is possible to write the equation 2.1, carrying out the FG potential:

$$V_{FG} = \alpha_G \cdot V_{CG} + \alpha_S \cdot V_S + \alpha_B \cdot V_B + \alpha_D \cdot V_D \quad (2.2)$$

The floating gate results to be capacitively coupled to other terminals, therefore its potential is set not only by the control gate, but also by the other terminals and by the neighbor cells.

[2]In this model it is not considered the capacitive coupling between neighbor cells C_P

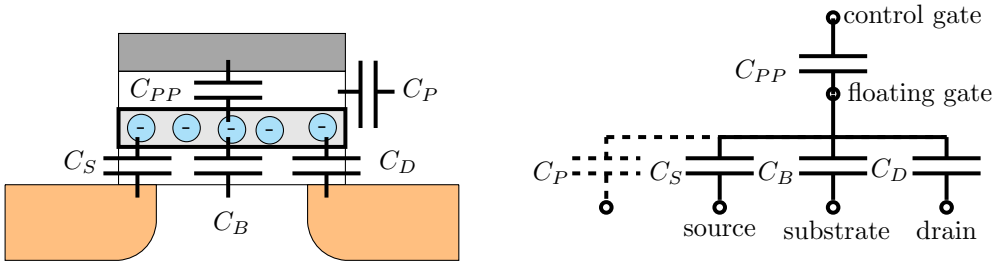


Figure 2.3: Schematic of capacitive coupling components in the FG cell.

If the source and the bulk are grounded, equation 2.2 can be rearranged:

$$V_{FG} = \alpha_G(V_{CG} + \frac{\alpha_D}{\alpha_G} \cdot V_D) = \alpha_G(V_{CG} + f \cdot V_D) \tag{2.3}$$

where

$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_{PP}} \tag{2.4}$$

The equations of the floating gate transistor can be obtained by substituting the MOS gate V_{GS} voltage with the V_{FG} potential and transforming the device parameters, threshold voltage V_T and conductivity factor β [7].

$$V_T^{FG} = \alpha_G \cdot V_T^{CG} \tag{2.5}$$

$$\beta^{FG} = \frac{1}{\alpha_G}\beta^{CG} \tag{2.6}$$

The I-V (current-voltage) characteristic equations for the floating gate transistor become:

- **Triode region:** $|V_{DS}| < \alpha \cdot V_{GS} + f \cdot V_{DS} - V_T$

$$I_{DS} = \beta\Big[(V_{GS} - V_T) \cdot \Big(f - \frac{1}{2 \cdot \alpha_G}\Big) \cdot V_{DS}^2\Big] \tag{2.7}$$

- **Saturation region:** $|V_{DS}| \geq \alpha \cdot V_{GS} + f \cdot V_{DS} - V_T$

$$I_{DS} = \frac{\beta}{2}\alpha_G(V_{GS} + f \cdot V_{DS} - V_T)^2 \tag{2.8}$$

where β and V_T are measured with respect to the control gate.

These equations show the floating gate transistor behavior, different from a standard MOS transistor.

i) Looking at the voltage equation for triode region, the floating gate transistor can conduct current even when $|V_{GS} < |V_T||$, because the channel can be turned on raising the drain voltage ($f \cdot V_{DS}$). This is called the "drain turn-on".

ii) While for a standard MOS in the saturation region, the I_{DS} is almost independent from the V_D, for the floating gate transistor the drain current continues to rise as drain voltage increases. The result is that no drain current saturation occurs for high drain voltage values.

iii) The boundary between the two operation ragion is

$$|V_{DS} = \alpha \cdot V_{GS} + f \cdot V_{DS} - V_T| \tag{2.9}$$

respect to

$$V_{DS} = V_{GS} - V_T \tag{2.10}$$

for the MOS transistor.

iv) The transconduttance in saturation region is

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS\,(V_{DS}=\text{costant})}} = \alpha_G \beta (V_{GS} + f V_{DS} - V_T) \tag{2.11}$$

and increases with V_{DS}, instead in conventional MOS transistors it is independent of the drain voltage.

v) The capacitive coupling ration f depends on C_D and C_{PP} only ($f = \alpha_D/\alpha_G = C_D/C_FG$), and its value in the saturation region is:

$$f = -\frac{\partial V_{GS}}{\partial V_{DS\,(I_{DS}=\text{constant})}} \tag{2.12}$$

Many techniques have been proposed to extract the capacitive coupling ratio from simple dc measurements [7]. These methods require the measurement of the electrical parameters 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.

2.1.1 The Reading Operation

Before explaining the mechanisms that induce charge variations in the floating gate, let us show how equations 2.3, 2.5 and 2.7 become, considering a net charge $Q \neq 0$ in the FG.

$$V_{FG} = \alpha_G V_{CG} + \alpha_D V_D + \frac{Q}{C_{TOT}} \tag{2.13}$$

$$V_T^{CG} = \frac{1}{\alpha_G} V_T^{FG} - \frac{Q}{C_{TOT}\alpha_G} = \frac{1}{\alpha_G} V_T^{FG} - \frac{Q}{C_{PP}} \tag{2.14}$$

$$I_{DS} = \beta\Big[\Big(V_{GS} - V_T - \Big(1 - \frac{1}{\alpha_G}\Big)\frac{Q}{C_{TOT}}\Big)V_{DS} + \Big(f - \frac{1}{2\alpha_G}\Big)V_{DS}^2\Big] \tag{2.15}$$

Equation 2.14 demonstrates that the threshold voltage depends on the charge in the FG. The shift ΔV_T is obtained subtracting to 2.14 the V_T value when $Q = 0$ (2.5):

$$\Delta V_T = V_T - V_{T0} = -Q/C_{PP} \tag{2.16}$$

represented in Fig. 2.2. The charge greatly affects the current level, which corresponds to the two different cell states, erased or programmed.

2.1.2 Charge Injection and Removal Mechanisms

Fig. 2.4 shows the energy band diagram of the cell for the two states, erased and programmed. It is clear that the two oxide layers play a role of potential barriers, while the floating gate is a potential well. Without any charge in the storage element, the cell is in flat band condition. If the charge is stored in the floating gate, it creates a electrical field, that bends the oxide layer bands. The higher is the amount of the stored charge, the higher is the bands bend; as consequence, they become thinner.

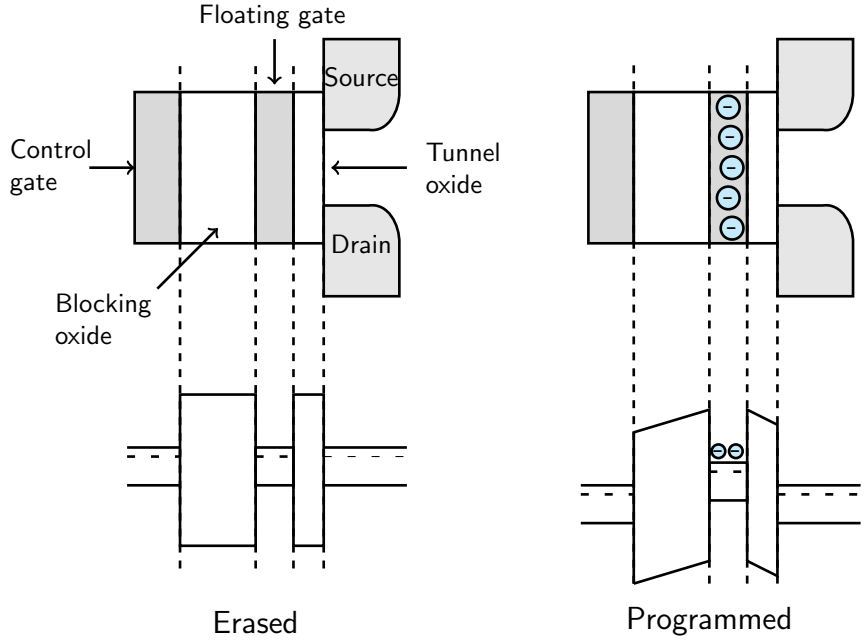


Figure 2.4: Energy band diagram for FG cell.

Two mechanisms perform injection (and/or removal) of the charge in the floating gate: the injection of electrons (or holes) over the potential barrier, called **Channel Hot Electron Injection** (HEI), and, through the barrier, **Fowler-**

Nordheim (FN) tunneling. The blocking oxide, thicker than the tunnel oxide, allow us to confine the carriers in the storage element, reducing the leakage currents. It is formed by three-layer Oxide Nitride Oxide (ONO) structure, where a thin film of Si_3N_4 is placed between two SiO_2 layers. On the other hand, the tunnel oxide is made of high quality SiO_2.

Channel Hot Electron

When a carrier is traveling from source to drain, it is powered by the electric field and it interacts with the lattice, losing its energy throughout lattice vibrations. If the electric field is high enough[3], electrons are not in dynamic equilibrium with the lattice as for low fields, and a fraction of them can gain enough energy to jump over the potential barrier and be injected into the floating gate. Three conditions are necessary for electrons to overcome the potential barrier:

1. the electron potential energy must be higher than the potential barrier;
2. it must be directed toward the barrier;
3. the field in the oxide must be directed to collect the charge.

To evaluate how many electrons could overcome the barrier, we must analyzed the energy distribution $f_E(\varepsilon, x, y)$ as a function of the lateral field ε, the momentum distribution $f_k(E, x, y)$ as a function on electron energy E, the shape and heigh of the barrier and the probability that an electron with energy E, wave vector k and distance from the Si/SiO_2 interface will overcome the barrier. Each function must be defined along the entire channel. The model becomes more complex considering the impact ionization, which is a second important energy loss mechanism.

The HEI current can be explained introducing the "lucky electron" model. With this approach, the injection mechanism is attributed to the probability of an electron "lucky" enough not to be scattered during its travel in the electric field ε, gaining enough energy to jump over the barrier. The total probability of the event is the sum of three probabilities of the following events [7].

1. The carrier has to be "lucky" enough to acquire enough energy to jump over the barrier and to retain its energy after the collision that directs it toward the interface (P_{Φ_b}).
2. The carrier traveling path from the direction point to the interface must be without collisions (P_{ED}).

[3]For fields exceeding the value of 100 kV/cm [8].

3. The carrier can surmount the repulsive oxide field at the injection point, due to the Schottky barrier lowering effect, without collision in the oxide (P_{OC}).

This is a simple model that is in quite good agreement with simulation results, even if not very accurate.

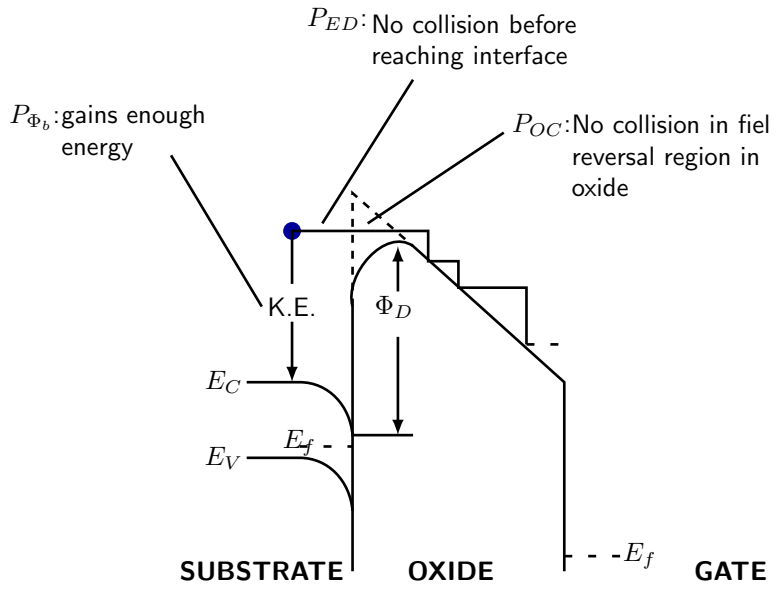


Figure 2.5: Energy band diagram describing the three processes in involved HEI.

Another model to explain HEI, more rigorous than the previous, is based on the quasi-thermal equilibrium [9].

In both models the relation between the substrate current I_{sub} and the injection current I_G is

$$I_G/I_{ch} \sim I_{sub}/I_{ch}e^{-\Phi/\Phi_i} \tag{2.17}$$

where I_{ch} is the channel current, Φ_i is the impact ionization energy and Φ is the energy barrier seen by electrons to be injected in the oxide.

The substrate current is composed of holes generated by impact ionization close to the drain voltage. 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 in the oxide, degrading it. The ionization process generates a lot of carriers that can be injected in the oxide and be trapped at the interface, producing interface states, thus degrading the device performance.

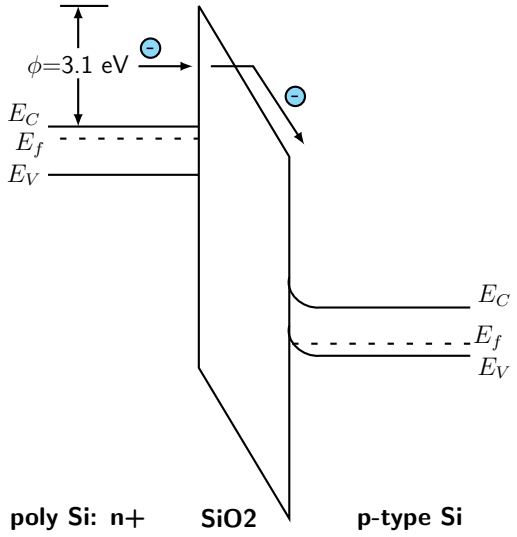


Figure 2.6: FN tunneling throught a potential barrier in a MOS structure.

Fowler-Nordheim Tunneling

Fowler-Nordheim tunneling can be explained through the quantum mechanic, according to an electron is able to penetrate a potential barrier, and the probability that this could happen depends on the distribution of the occupied states in the injecting material and on the height, width, and shape of the barrier. The tunneling process in an MOS structure is represented in Fig. 2.6. A negative bias is applied to the metal electrode with respect to the p-type silicon substrate. The following equation describes the tunneling current density:

$$J = A \cdot E^2 \cdot exp\Big(-\frac{B}{E}\Big) \tag{2.18}$$

where E is the oxide electric field, A and B are constants related to the Si/SiO_2 barrier heigh:

$$A = \frac{q^3 \cdot m_0}{16\pi^2 \cdot \hbar \cdot m_{ox}\phi} \qquad B = \frac{4 \cdot \sqrt{2 \cdot m_{ox}} \cdot \phi^{3/2}}{3 \cdot \hbar \cdot q} \tag{2.19}$$

where m_0 and m_{ox} are the electron effective mass in vacuum and in SiO_2 respectively, ϕ is the Si/SiO_2 barrier height, q the elementary charge, and $\hbar$ the reduced Planck's constant. It should be noted that the Fowler-Nordheim current is a function of the oxide field, and not of the applied gate voltage, but the mechanism can happen only if the oxide field is high enough, thus a gate voltage high enough is applied. In particular the following equation must be satisfied:

$$E \geq \frac{\phi}{q \cdot t_{ox}} \tag{2.20}$$

being t_{ox} the oxide thickness. An optimum tunnel oxide thickness of about 10 nm

is a tradeoff between performance requirements, such as programming speed and power consumption, and reliability constraints, which would require thick oxides, to avoid large tunnel current density. In fact, remembering that the oxide electric field is roughly equal to the applied voltage divided by t_{ox}, since the tunnel current is exponentially dependent on the oxide electric filed (Fig. 2.7), an increase of few MV/cm leads the current density to rise some orders of magnitude [7]. This could be a critical problem concerning process control, because a very small variation in the tunnel oxide thickness among the cells belonging to the same array produces a great difference in programming and erasing current. For this reason, a very good process control is required. The result is the threshold voltage distribution spread. This mechanism is more efficient than CHE, but slower.

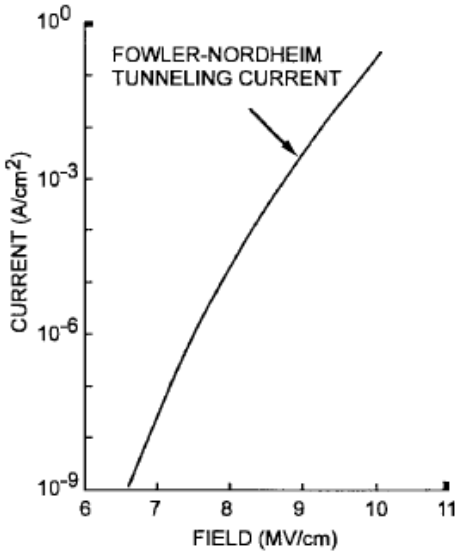


Figure 2.7: FN tunnel current as a function of the electric field [10].

2.1.3 Threshold Voltage Distribution

The large array organization of the cells leads to a wide distribution of the parameters. The threshold voltage distribution for each state, pictured in Fig. 2.9 should be as narrow as possible, but in general it has a Gaussian shape, even larger if no constraining algorithms are applied during program and erase operations. For this purpose, two reference levels are introduced, the program verify and the erase verify levels. The program/erase algorithms ensure the cell threshold voltage is beyond the erase verify level after an erase operation, and above the program verify level after programming.

Programming is obtained by applying pulses to the control gate and to the drain simultaneously, when the source is grounded. This operation can be per-

formed selectively by applying the pulse to the word line (connected to gates) and biasing the bit line (connected to drain). The threshold voltage of the cell shifts to higher value because of the negative charge injected in the floating gate. The width of the programming pulse influences the amount of carriers injected, thus the threshold voltage shift [7]. Both temperature and drain voltage have an effect on threshold voltage shift: higher temperature reduces the number of electrons available for injection in the FG, increasing the programming time, and V_T depends linearly on the drain voltage.

Accurate programming of Flash memories is usually obtained by the incremental step pulse programming algorithm (ISPP), consisting in the application to the cell control gate of short programming pulses of equal duration and increasing amplitude. The constant increase V_{step} of the control gate pulses drives to a very tight threshold voltage distribution, thus the V_T variation per step ($\Delta V_{T,\text{step}}$) rapidly converges to V_{step}. After each program pulse, a verify operation is inserted: V_T is sensed and compared to a verify level. If it exceeds the program verify (PV) level, the algorithm stops and the program operation is concluded; otherwise another control gate step is applied to the memory cell. The purpose of the method is to induce small variations of the cell V_T by transferring small charge packets from the substrate to the floating gate. However, due to cell size scaling, the number of electrons controlling the cell state has critically decreased, so that a very low number of electrons are transferred during each voltage step. This is one of the new issues related to device scaling, introducing a spread in the final V_T distribution. To overcome this problem a double-verify (DV) algorithm has been introduced [11], able to get thinner V_T distribution in presence of electron injection statistic (EIS), which will be further presented.

The erase operation requires a high voltage pulse to be applied to the source, while control gates (connected to WL) are grounded and drains (connected to BL) floating. Before the operation, all the cell are programmed with the same threshold voltage value. After the erase pulse, the threshold voltage shift to lower values depends on the applied source voltage value: each volt reduction corresponds a one-order-of-magnitude increase in erasing time. V_T depends on the oxide thickness too. If two FG cells with the same oxide thickness have different threshold voltage values while programmed, after the erase they will have the same threshold voltage. Since, in large array, FG transistors can have little oxide thickness differences, after erasing there is a read operation to check if all the cells are erased or not. If not, another erase pulse and check operation are applied. The algorithm will de repeated until all the cells' threshold voltages are below the erase verify level. The final erased V_T has a Gaussian shape (Fig. 2.9).

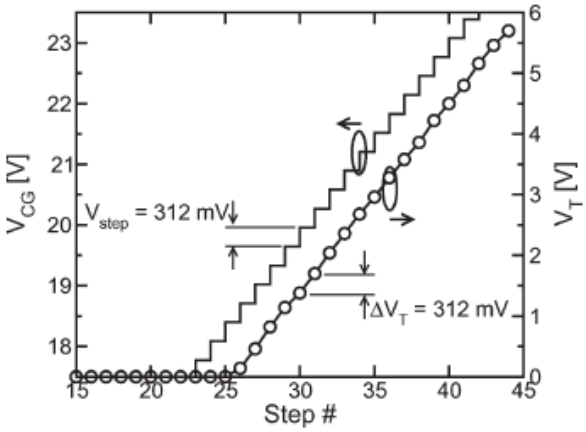


Figure 2.8: Control-gate voltage waveform used to program a NAND cell and resulting VT transient on a 60-nm device. Note that only positive V_T values can be sensed in our NAND array [12].

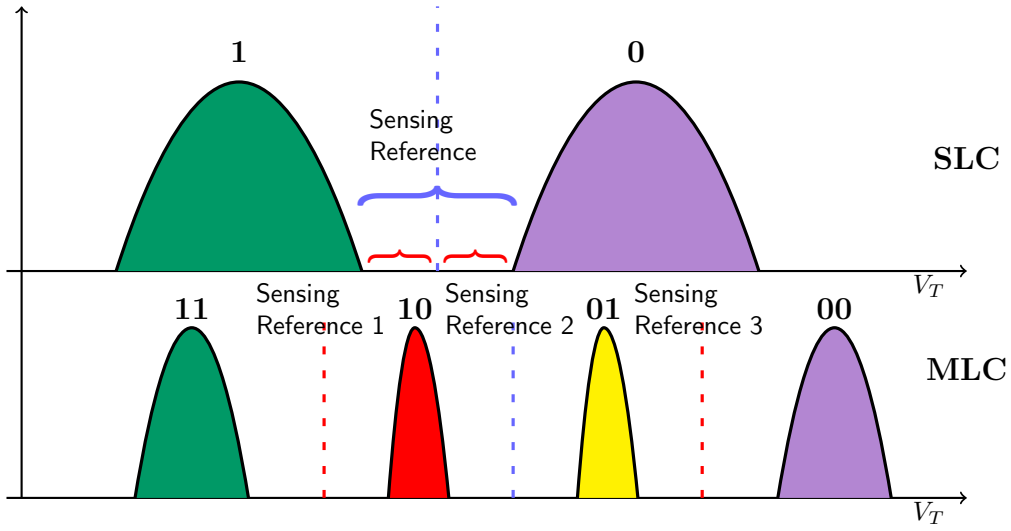


Figure 2.9: Threshold voltage distribution for SLC (above) and MLC (below) memory.

Depending on the number of bits stored for each cell there are two possible architectures.

- **Single Level Cell (SLC)**: is the simplest one, because each cell stores a single bit, "1" or "0".
- **Multi-Level Cell (MLC)**: modulating the amount of charge in the FG, each cell can store multiple bits. The reading/erasing/programming operations are complex because of the threshold voltage distributions are thinner than in SLC devices. For this reason MLC memories are slower and less robust, but it is possible to double the memory density. The output levels must not be necessarily a power of 2, in fact it is possible to store 1.5 bit per cell, resulting 3 output levels. In the present market, memories with up to 4 bits per cell are available.

2.2 Array Organization

Flash memories take advantage of the electrical erase from EEPROMs and the high density from EPROMs, since the erase operation is performed in blocks and there is not the selection device (necessary in EEPROMs), achieving high density arrays. Furthermore Flash memories are very fast because of the intrinsic parallelism.

Depending on data access and data write organizations, there are mainly two types of architecture for Flash memories.

- **NOR**: the first in time to be developed, used in permanent data storage or rarely modified data; this architecture is convenient for code storage.
- **NAND**: different respect to NOR because of the array organized in serial chain of cells, convenient for data storage.

Tab. 2.1 resumes the differences between NOR and NAND-type Flash memories.

In this thesis, devices under investigation are NAND Flash memories, which are experiencing a dramatic increase in demand thanks to their high density storage. However also a NOR architecture description will be provided.

2.2.1 NOR Architecture

The NOR array organization is shown in Fig. 2.10. Each cell has the source grounded through the source line and the drain shorted in the bit line, common to many cells.

NOR Flash Memory	NAND Flash Memory
Cell size = $10F^2$	Cell size = $4F^2$
Random access	Serial access
Slower Program/Erase	Faster Program/Erase
CHE programming	FN tunneling programming
Byte/Word Program, Block Erase	Page Program, Block Erase
Lower density	Higher densiry
Code storage applications	Data storage applications

Table 2.1: Main differences between NOR and NAND Flash memories.

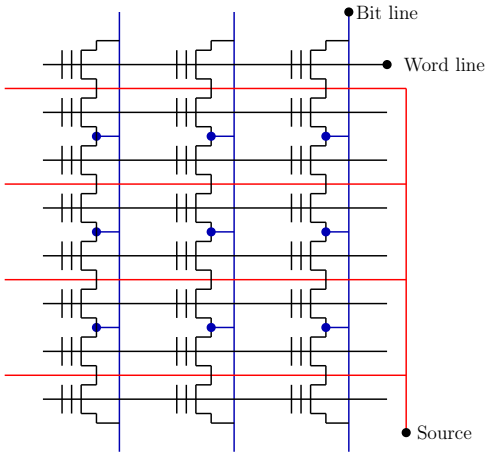
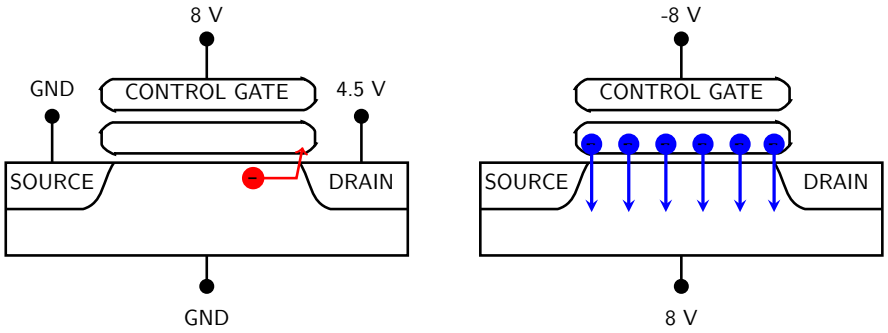


Figure 2.10: Flash NOR architecture



(a) Channel Hot Electron program **(b)** Fowler-Nordheim erase

Figure 2.11: Program (a) and erase (b) in NOR architecture

For the read operation, the cell address has to be provided to the row decoder and the column decoder. The row decoder selects the word line, raising its voltage. The bit line, selected by the column decoder, and its current, measured by a sense amplifier, will be significant if the cell is in the erased state, because the threshold voltage is under the read voltage. If no current is sensed, the cell is in the programmed state.

For programming and erasing operations, the word line selects the cell subject to the operation. Program is carried out by channel hot electron injection: the high current needed for this mechanism limits the parallelism of the operation. Supposing that "0" is the data to be written, the bit line will drive a high voltage, to allow the channel hot electron injection in the FG and consequently to increase the threshold voltage.

Erase is performed by FN tunneling. This operation is complex, first of all because it is applied on an entire section, secondly because the threshold voltage of some cells must be checked, to avoid too low values and, in case, to raise the threshold voltage to higher value. A high electric field must be applied between source and gate, to allow the carriers to flow out from the floating gate throughout FN tunneling.

The bit size is $\sim 10\,F^2$, where F is the feature size of the technology, therefore NOR cells can not be made smaller than $\sim 10\,F^2$. This limit is imposed by the difficulty of scaling the gate length, associate to the HEI mechanism for programming, and the array organization with a drain contact shared for each couple of cells. The consequence is a fast random access, about 100 nm, and a slow program operation, carried out at word level, about 5 μs. The erase operation too is slow, performed at block level, typically 100 ms. With these time performances, NOR memory is use as read-mostly memory, with rare necessity of programming or erasing.

In NOR architecture the manufacturer guarantees that each single bit is functional and complies retention and endurance requirement. No ECC (Error Correction Code) must be implemented by the user. Typically there are different buses for code and data.

2.2.2 NAND Architecture

The base element of the NAND architecture is not the single cell, but a serial chain of 16 or 32 floating gate transistor, each string connected to the Drain Selection line (DS) and to the Source Selection line (SS). Several strings are connected to the same bit line. The elimination of the shared drain contact between cells and the scaling of the channel length, because of the different injection mechanism (FN tunneling), allows to achieve a smaller area respect to NOR arrays. In

fact the bit size is $4\,F^2$.

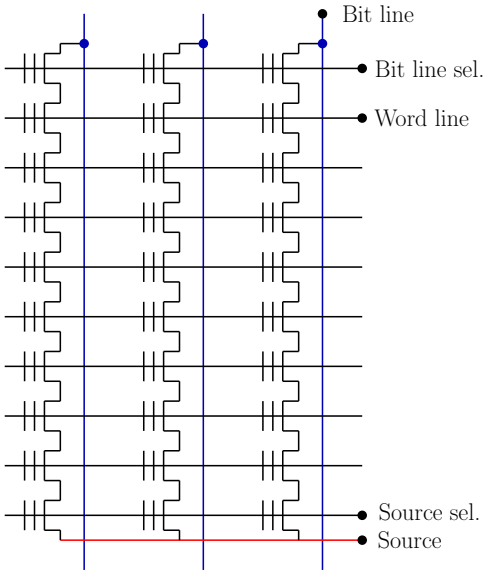


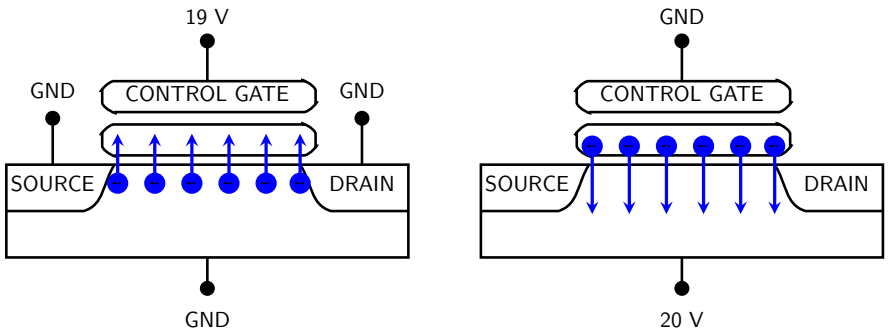
Figure 2.12: Flash NAND architecture

During reading operation, the selected cell has 0 V applied at the control gate, while other cells in series are biased at a voltage V_{PASS}, that is higher than the programming V_T[4], so that string current is determined only by the V_T of the selected cell. The sense amplifier, placed at the end of the bit line, will detect current only if the selected transistor has a negative threshold voltage, so it is in the erased state. An entire page is read at once in a single parallel operation.

Both program and erase operation use FN tunneling (with opposite polarity), that is more efficient than HEI, then currents are smaller and it is possible to reach a great level of parallelism. Program is performed at page level, applying positive voltage to the control gate, which capacitively couples to the FG and induces electrons to tunnel up from the channel. Erase operation is executed at block level, removing electrons by reversing the bias. Programming takes about 0.2 ms, erasing about 2 ms. This array organization makes NAND structure suitable for data storage, where random access is relatively important and latency is not critical.

An external ECC is indispensable because manufacturer does not guarantee each single bit and bad blocks are not infrequent in commercial devices. A block is defined bad block when it contains at least one page that has more bad bits that can be corrected by the minimum required ECC. The manufacturer specifies the minimum number of valid blocks (NVB) of the total available blocks, which will not fall below NVB during the endurance life of the memory. However, thanks

[4]In case of MLC, V_{PASS} is higher than the highest programming V_T level.



(a) Fowler-Nordheim program **(b)** Fowler-Nordheim erase

Figure 2.13: Program (a) and erase (b) in NAND architecture

to ECC, NAND Flash memory devices containing bad blocks, can be used quite reliably. ECC increase the latency, but allow to realize more scaled devices.

2.3 Reliability

Reliability of floating gate devices is a very complex problem, because of numerous factors that influence the two key parameters, endurance and retention, which give a measure of the quality of the device. Reliability aspects pose the most challenging questions for modeling, at both single-cell and array level.

There are two main categories of reliability mechanisms:

- **intrinsic mechanisms**, due to defects in the device structure, can affect all the cells in a uniform way and can be studied on single cell in test structures;
- **single-bit mechanisms**, due to extrinsic defects or to unfortunate configurations of intrinsic point defects, which occur in few cells.

In the following section will be summarized the most important factors responsible of reliability degradation in Flash memories, both intrinsic failure and statistical effect due to single-bit mechanisms.

2.3.1 Retention

Retention is critical in any NVM device, since the ability of retain information without power supplied is the definition of non-volatile memory. Manufactures define data retention time, for which the device is assured to retain the stored charge. A typical benchmark for Flash memories is 10 year, thus the device should not lose more than a small fraction of its carriers for 10 years.

The charge loss in the retention state is determined by tunneling leakage currents throughout dielectrics, which could be amplified if the dielectric contains defects. Fast program and erase operations require high voltages and current through thin oxides, which are easily degraded. Let illustrate three intrinsic mechanisms that lead to charge loss or charge gain.

The **fiel-assisted electron emission** consists of the move of one or more electrons from the floating gate to the oxide interface, and from there they can tunnel in the substrate, resulting a charge loss. If the cell is erased, the opposite injection can happen. The probability that the electron could tunnel from the floating gate to the substrate depends on the drop voltage between FG and substrate, and the FG potential depends on control gate potential, throughout the coupling coefficient between control gate and floating gate α_G. Thus the leakage current produced depends on α_G and on the stress level. The charge Q stored in the FG decreases as α_G decreases, because the electron injection during programming is less efficient, or as level stress increases, because this causes more negative charge trapped in the oxide. The leakage current depends exponentially on the electric field[5] around the FG, which is proportionally to the charge Q:

$$E = \frac{Q}{2\epsilon_{\text{ox}}\sigma} \tag{2.21}$$

where σ is the floating gate area.

The **thermodynamic emission** is a mechanism of carriers emission from the FG, above the potential barrier. This phenomenon is negligible at room temperature, while become relevant at high temperature.

The last is the **electron detrapping** in the gate oxide, producing charge loss, thus threshold voltage reduction.

2.3.2 Endurance

Flash memory is required to retain its properties on being subjected to repeated program/erase cycles. However when oxides are repeatedly stressed at high fields, interface and bulk traps develop in the dielectric. Program/erase cycles cause uniform degradation of the cell performance and limit the Flash memory endurance. Fig. 2.14 shows the result of an endurance test on a single cell. The variations of program and erase threshold voltage give a measure of the oxide aging.

The increase of erase V_T over cycles is due to the generation of negative traps, whole the reduction of the program V_T is attributed to oxide traps and interface

[5]Because in a tunneling mechanisms, the current density is $J = AE_{ox}^2 exp(\frac{-B}{E_{ox}})$

states generation. The variation of the threshold program/erase voltage level, thus the "window" closure is the result of three effects:

- variation of the transistor V_T, measured at the floating gate;
- oxide conduction variation;
- transconductance degradation, leading to higher V_G for the same I_D.

To prevent program/erase window closure is indispensable the optimized cell design. Flash memories are typically expected to least 100,000 program/erase cycles without wearing out. However, multi-level cells have lower endurance benchmarks, because of a stringent threshold voltage window.

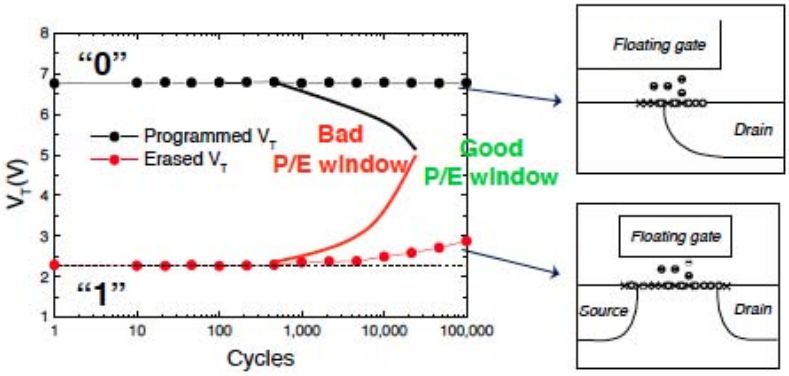


Figure 2.14: Threshold voltage window closure as a function of program/erase cycles on a single cell, showing endurance characteristic [2].

2.3.3 Statistical Effects

Multiple sources of statistical variability affect the FG cell threshold voltage, thus the array reliability during program, erase and data retention, which are associated with the discreteness of the charge. These variability sources increase their impact on V_T variability as technology scaling proceeds.

First of all, the statistical distribution of defects in the cell tunnel oxide leads to RTS, SILC and erratic bits. In addition, the cell-to-cell parameter variation, due to a more critical process control, introduces a great variability. Finally, the miniaturization of the area leads to a reduction of the dopants and of the electrons transferred to/from the floating gate to change the cell state, making the discreteness of the matter a fundamental source of statistical variability.

This paragraph is devoted to the brief description of sources of statistical variability.

Over-erasing

In Flash memory each erasing is actually formed by a series of elementary erase operation. After the first erase pulse, an algorithm checks if the operation is completed, controlling the threshold voltage. The verification of the complete erasure of all the cell belonging to the same block, which can be very large, is one of the biggest issues in Flash technology. The threshold voltage distribution of the erased state spreads around an average voltage, with a shape like Gaussian, but it is not symmetrical toward lower values (Fig. 2.15). A very small percentage of the cells has a very large threshold voltage variation, and these few cells have an high relevance. The erasing speed of these cells is too high respect to the other cells. Since all the cells are erased simultaneously, the time required to erase the slowest may be long enough to over-erase the fastest cell. If the over-erased cells exhibit a negative or zero threshold voltage, all the cells connected to the same bit line would be read as "1" independently of they actual content. The population

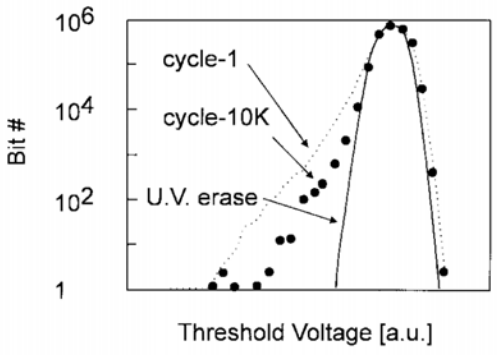


Figure 2.15: Threshold voltage distribution after different erase procedures: UV erase, after the first cycle and after 10 K cycles [7].

related to faster cells is too large to be attributed to extrinsic defects, and it is supposed to be due to statistical fluctuations of oxide charge and to the structure of the injecting electrode [13]. Positive charges in the tunnel oxide and irregular polycrystal grains may induce a local increase of the electric field, making some cells to be erased faster than average. This physical explanation is consistent with the tail behavior after programming/erasing cycles.

Erratic Bits

A relevant mechanism of single bit failure during programming/erasing cycles is the occurrence of an "erratic bit". Such bits show an unstable and unpredictable behavior in erasing, since its erase threshold voltage changes randomly from cycle to cycle, from the center of the Gaussian shape distribution to the lower part of

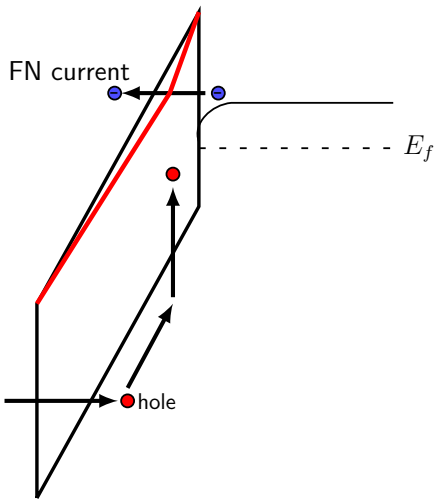


Figure 2.16: Erratic bit due to hole trapping.

the tail. This behavior is ascribed to an hole trapping/detrapping in the tunnel oxide. It has been demonstrated [14] that the statistical distribution of hole traps give a low but finite probability of having clusters of two or more positive charges whose combined electric field effect induces a huge increase in the tunnel current. Trapping/detrapping of a single positive charge cause a detectable change in the erase speed, leading to over-erase failures. Since this behavior is due to statistical fluctuations of intrinsic oxide defects, erratic bit events can be reduced by process optimization, but not completely eliminated. They are menaged by internal algorithm and ECC.

Stress Induced Leakage Current

One of the effect of trap generation in the oxide is the Stress Induced Leakage Current (SILC), that appears as a gradual and continuos increase of the leakage current. It is characterized by an uniform conduction across the whole oxide area, due to neutral trap generation inside the insulating layer caused by electrical stress. Different components contribute to SILC, that were classified by K. Sakakibara *et al.* [15] [16] [17]. The *DC* component is the predominant for 10 nm-oxide, such as tunnel oxide in FG memories. This component is modeled by a trap-assisted tunneling (TAT), pictured in Fig. 2.19. The electrical stress generates neutral defects uniformly distributed across the whole oxide area. These defects allow the trap assisted tunneling of electrons. SILC is proportional to the defects concentration and the trap-assisted tunneling probability. The strong dependence of SILC on the number of program/erase cycles is pictured in Fig. 2.17. The reduction of the oxide thickness reduces the distance between the neutral defects and the oxide interface, raising the probability that traps could capture

or emit electrons. In Fig. 2.18 it is show the threshold voltage distribution of two samples of the same lot, but with different oxide thickness. During retention the V_T shift is higher for the sample with the thinner oxide.

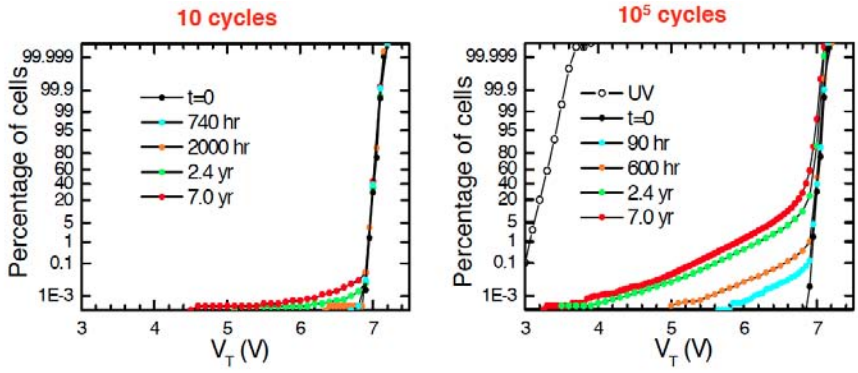


Figure 2.17: Effects of cycling on SILC. Data retention test at room temperature [2].

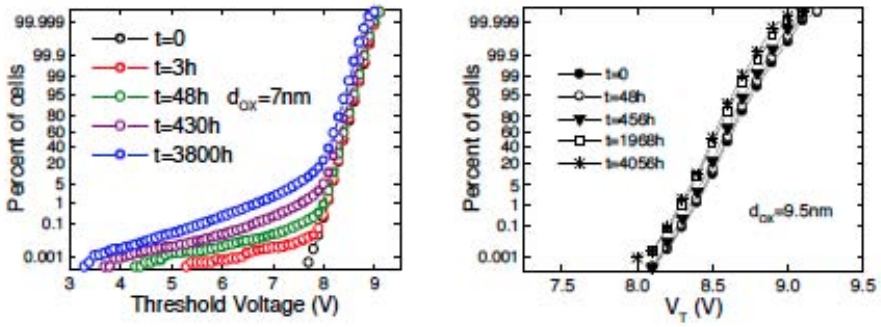


Figure 2.18: Effects of tunnel oxide thickness on SILC. V_T distribution of two sales from the same lot, but with different tunnel oxide thickness [2].

There are various theories on the origin of the neutral traps responsible of the SILC.

- **Hydrogen-induced defects**: an hydrogen atom is captured in an oxygen vacancy, this compound forms a hydrogen bridge[6] for electrons.
- **Anode Hole Injection (AHI)**: electrons with high kinetic energy injected into the floating gate cause anode hole generation in the floating gate (Fig. 2.20), which leads to degradation of the cell, such as threshold voltage shift and, ultimately, the breakdown [18].
- The third model [19] includes both holes and hydrogenous species. Holes, generated and injected into the oxide, can be trapped or react with Si−H

[6]The hydrogen bridge is a complex of a hydrogen atom and an oxygen vacancy: the hydrogen replaces an oxygen atom and the structure can be positively, negatively or neutrally charged depending on the electrons trapped (none, two or one respectively).

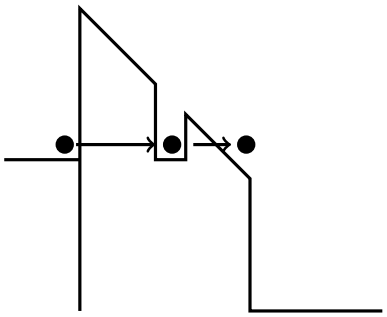


Figure 2.19: Trapped-assisted tunneling.

bonds in the dioxide layer, producing hydrogen release, according to the following reaction:

$$\mathrm{Si{-}H + h^+ \rightarrow Si + H^+}$$
$$\mathrm{Si{-}H + h^+ \rightarrow Si^+ + H}$$

The first forces will lead to the release of positive hydrogen ions, while the second the release of neutral hydrogen atoms. The mobile hydrogen can then drift or diffuse towards the interface, where it can break a Si−H bond and create an interface trap, according to one of the following reactions:

$$\mathrm{Si{-}H + H^+ + e^- \rightarrow Si + H_2}$$
$$\mathrm{Si{-}H + H \rightarrow Si + H_2}$$

Hydrogen and hydrogen-related compounds are present in the silicon dioxide, because the oxide is grown in H_2O ambient and because H is used to passivate the silicon dangling bonds at the interfaces. All these mechanisms well explain the defect generation and they may all contribute simultaneously to produce traps involved in SILC.

Anomalous SILC than can not be explained with the trap assisted tunneling (TAT) mechanism have been attributed to two trap assisted tunneling (2TAT) (Fig. 2.21).

Electron Injection Spread

The discrete nature of the electric flow charging the floating gate determines the accuracy of the program algorithm in nanoscale cell. As illustrated before, in order to obtain tight V_T distribution after program, the control gate voltage is not kept constant during the programming transient, but s increased as a staircase

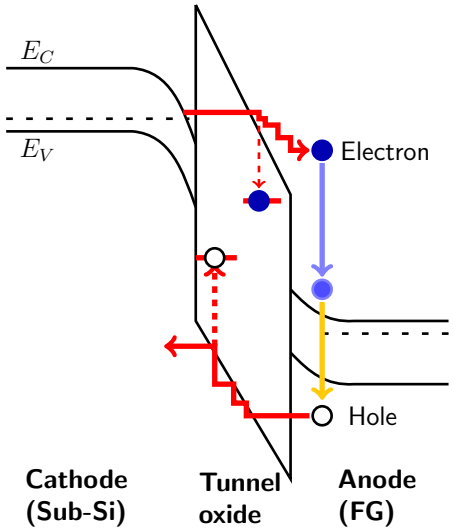


Figure 2.20: Diagram of anode hole injection in Flash memory cell.

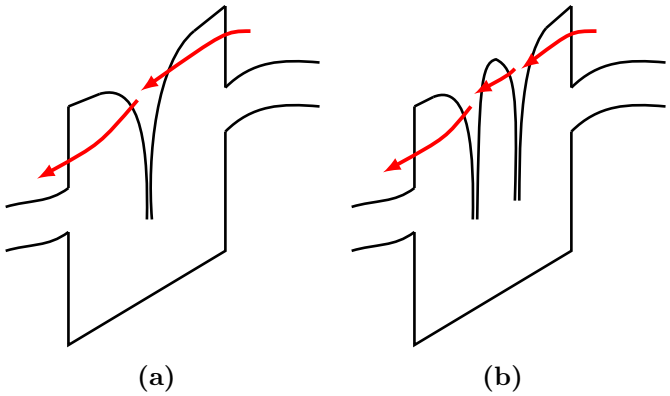


Figure 2.21: Trap-assisted tunneling (a) and tunneling assisted by 2 traps (b).

with fixed step amplitude. After each step, V_T is sensed and compared to a verify level. Small variation of V_T are obtained transferring small charge packets from the substrate to the floating gate. Due to memory cell miniaturization, the number of electrons for each packet is very low. The statistic of the electron injection introduces a spread in the final V_T distribution.

Compagnoni *et all.* [12] have presented a model for the EIS. The spread of ΔV_T is related to the spread of the number of injected electron n by the following equation:

$$\sigma_{\Delta V_T} = \frac{q}{C_{PP}}\sqrt{\sigma_n^2} \tag{2.22}$$

where q is the electronic charge. By assuming that n is ruled by a Poisson statistic, its variance σ_n^2 is equal to its mean $\overline{n}$ and equation 2.22 become:

$$\sigma_{\Delta V_T} = \frac{q}{C_{PP}}\sqrt{\overline{n}} = \sqrt{\frac{q}{C_{PP}\overline{\Delta V_T}}} \tag{2.23}$$

since $\Delta V_T = qn/C_{PP}$.

The effect of the injection spread on the V_T distribution respect to the program-verify level V_{pv} is shown in Fig. 2.22. While neglecting the injection spread, the maximum value reached by V_T is $V_{\text{pv}} + V_{\text{step}}$, considering the EIS, larger values can be obtained, allowing the cell to move farther away from the verify level. The

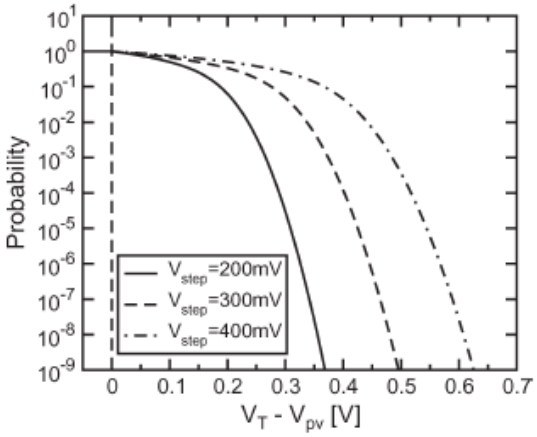


Figure 2.22: Cumulative probability for $V_T - V_{\text{pv}}$ assuming constant-current NAND programming with different V_{step} values (60nm technology)[12].

Poissonian behavior should be carefully considered for evaluating the impact of EIS on the programming algorithm accuracy. Furthermore the miniaturization of cell leads to a reduction of C_{PP}, which according to equation 2.23, determine an increase of $\sigma_{\Delta V_T}$.

Random Telegraph Signal

Random Telegraph Signal (RTS) is a stochastic fluctuation between two levels of the device threshold voltage, induced by the capture or emission of a single electron in a gate oxide trap. Fig. 2.23 is an example of RTS fluctuation detected in a Flash memory cell. Random variations of cell V_T may give rise to erroneous data reads or erroneous program operation, since also program algorithm includes verification steps. This is particularly critical for MLC application, whose threshold voltage cell control is more critical.

Adopting a simple electrostatic approach and assuming trap located at the substrate/tunnel oxide interface, the following equation estimates the threshold voltage shift caused by the capture/release of a single electron [20]:

$$\Delta V_T^{st} = \frac{q t_{\text{ox}}}{LW \epsilon_{\text{ox}} \alpha_G} \tag{2.24}$$

where q is the electron charge, t_{ox} the tunnel oxide thickness, L and W the channel length and width respectively, and α_G the capacitive coupling coefficient between control gate and floating gate. Equation 2.24 is a simple electrostatic approach: in the case of $W = L$ =45 nm, t_{ox} =7 nm and α_G =0,65, ΔV_T^{st} result 25 mV, a quite low value.

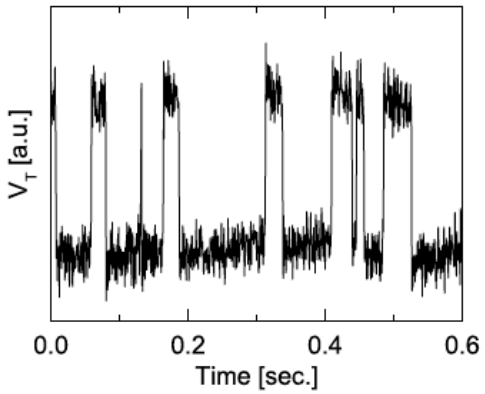


Figure 2.23: Example of RTS measured in the time domain for a single Flash memory[20].

However it has been reported that NVMs feature RTS with magnitude much larger than the value predictable by equation 2.24, calling this phenomenon "Giant RTS" [20].

The amplitude of RTS V_T fluctuations has been shown increase upon channel size downscaling. This size dependence is explained by percolation effects,

through non-uniform V_T landscape in the random doped channel and by confinement effects of the current at the channel edges, due to local field enhancement.

Other Variability Sources

The first and the most relevant source of statistical variability in nanoscale devices is the **random discrete dopants (RDD)**. Statistical variability simulations [5] demonstrate that RDD causes the reduction of the average threshold voltage. The reason is that RDD induces current percolation path in the channel, which leads to early turn on of the FG cell.

The second most important source of variability in Flash memory cell is **interface trapped charge (ITC)**. The average threshold voltage increase because of charge trapped at the surface of the device. As for RDD, the distribution of the trapped charges is governed by Poisson distribution.

Line edge roughness (LGR) of gate induce a smaller variability respect to RDD and it is due to the short channel effects around the nominal channel length. Line edge roughness affects also shallow trench isolation (STI), resulting in random STI roughness (**LSR**). In this case the induced variability (half the value of LGR) is due to the channel width dependence of the threshold voltage, which is weaker compared to the channel length dependance.

Oxide thickness fluctuation (OTF) has a small impact on statistical variability considering the relative thick (8-nm) oxide in Flash memory cells.

Finally **polysilicon granularity (PSG)** causes an increase of the average threshold voltage because of the increased potential at grain boundaries, but he variation is small due to the thick tunnel oxide.

2.4 Scaling Issues

The effort of semiconductor memory technology is constantly focused on accommodate more cells per unit of wafer area. The main purpose is the cost-per-bit reduction, obtained by cell size scaling. However downscaling must be supported by redesign in new technology, considering that high density require high performance dielectrics and high voltage architectures.

While the cell basic structure has not changed throughout different generation, scaling down has been carried out on the cell area, through active area and passive elements scaling. The reduction of the channel length not only allows increasing of density, but also speeds up the program operation, because carrier injection into the FG is more efficient. However decreasing L, the capacitive coupling between FG and drain increases, enhancing the probability of punch-through and drain turn on. The fair channel length comes from a tradeoff between performance and

disturb.

Another relevant issue in Flash memory is the relatively high voltage needed to programming and erasing, operations based on physical mechanisms (FN and HEI) whose parameters do not scale. Concerning the tunnel oxide thickness, it is limited at 7-8 nm, in order to comply data retention requirement of at least 10 years[7]. In fact, thinner oxide facilitates trap assisted tunneling process caused by oxide aging. As consequence, amplitude of RTN V_T fluctuations increases. The interpolydielectric too is limited to about 12-13 nm by the coupling coefficient α_G. Furthermore, reducing the cell-to-cell distance causes parasitic coupling cross talk, which contribute to V_T distribution width. This issue is especially relevant for NAND Flash memory, because of its high density: since cells are very close to each other, scaling of NAND Flash memory is limited by parasitic interferences between adjacent cells. This is especially critic for multi-level cells.

Finally scaling down implies less electron per bit, making the stochastic nature of quantum-mechanical tunneling be more evident. The control of the amount of charge injected is degraded and the precision of program and erase operation is affected. Statistical effects, such as Electron Injection Spread, are more effective. Therefore scaling of planar Flash is getting to the end, because of structural limit and device physic limits [2].

Future alternative will follow mainly two paths [21]: one aimed at overgonig the scaling limit of the planar Flash by 3D structures, such as FinFET and stacked/vertical Flash, the second implementing alternative charge storage mechanisms, such as phase-charge memory (PCM) and resistive switching memory (RRAM).

2.5 Conclusions

Basic concepts of Flash memory have been reviewed: the physics mechanisms used to store and remove charge from an FG, thus enabling information storage. The main reliability aspects have been analyzed, since reliability is becoming a critical issue with downscaling. Problems related with oxide defects and its wear out, and especially issues coming from the discrete nature of the electrons flow have been analyze, since the small (25 nm) size of the devices object of this work. These concept will be recalled during the analysis of the experimental results.

[7] In Chapter 3 the maximum number of electrons lost from the FG will be quantitative evaluate.

Chapter 3
Radiation Effects

Failure on satellite electronic components from the natural space radiation environment has become to acquire interest side the early 1960's, when high altitude nuclear tests[1] increased the radiation levels in the Van Allen belts. The complexity of failure in space system electronics has increased as the devices trend to miniaturization. Furthermore, the development of many device types, levels of integration and technologies, has widened the failure mechanisms.

Nowadays we are almost totally dependent on successful space systems, whether they be military, research or commercial missions. It is essential to assure a high level of confidence that space system electronics will complete their missions without radiation-induced failures. This purpose requires not just accurate ground-based test data for the device and a accurate model for prediction the device performance in space, but also requires precise modeling of the space radiation environment. Over-specification, thus over-hardening, leads to unnecessary expense, whole under-hardening leads to time consuming anomalies and premature failure.

In this chapter a brief introduction will focus on the space radiation environment and its effect on electronics. Then, it will focus on radiation induced effects on Flash memory device.

3.1 Space Radiation Environment

The natural space radiation environment is composed by two major elements: the transient environment, made up by all of the elements of the periodic table, and that trapped, consisting of particles confined by the magnetic fields of most planet.

[1]The Starfish test of 1962 contaminated the inner belt with electrons form many years.

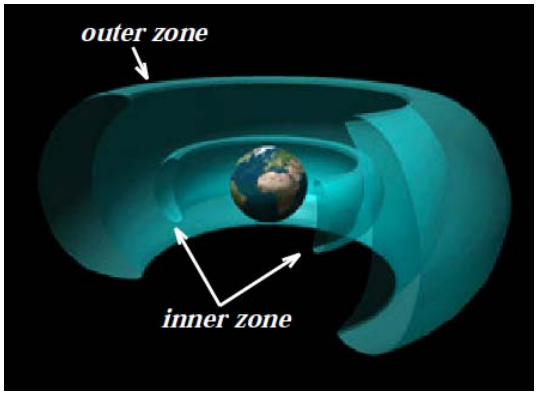


Figure 3.1: Diagram of the Earth's Van Allen radiation belts.

Van Allen Belts

Particles with defined charge, mass, energies and trajectories can be captured by Earth's magnetic field, forming the Van Allen belts (Fig. 3.1). The presence of the these trapped particles is attributed to several physic mechanisms: the acceleration of low energy particles by magnetic storms, the decay of energetic neutrons produced by the interaction of cosmic galactic rays with the atmosphere[2], and solar radiation.

Van Allen belts are divided in two, an inner belt, extending to 7 earth radii and consisting of energetic protons up to 600 MeV and electrons, up to several MeV, and an outer belt, trapping mainly electrons, extending to 10 earth radii, with energies up to 7 MeV. The Earth's atmosphere is the lower bound for Van Allen belts, because trapped particles, interacting with terrestrial atmosphere, lose their energy. The standard models of the trapped protons and electrons are AP-8 and AE-8 respectively[3]. Fig. 3.2 is a sketch of the Van Allen belts as predicted by AP-8 and AE-8 trapped particle models. On the left a cross section of the trapped protons is depicted, on the right the one for trapped electrons.

Transient Environment

Between many types of radiation making up the the transient environment, the two most important component for radiation effect in spacecraft are Galactic Cosmic Rays (GCRs) and particles emitted during solar events.

GCRs are a flux of energetic charged particles, formed by 83% protons (hydrogen nuclei), 13% alpha particles (or helium nuclei), 3% electrons and 1% heavier

[2]This mechanism is called cosmic ray neutron albedo decay (CRAD) and it is the most significant source of energetic particles in the inner zone.

[3]Ions with Z >1 can also be trapped by Earth's magnetic field, although the intensities for these ions are lower than those for protons and electrons and their effects on microelectronic systems are second order in most case.

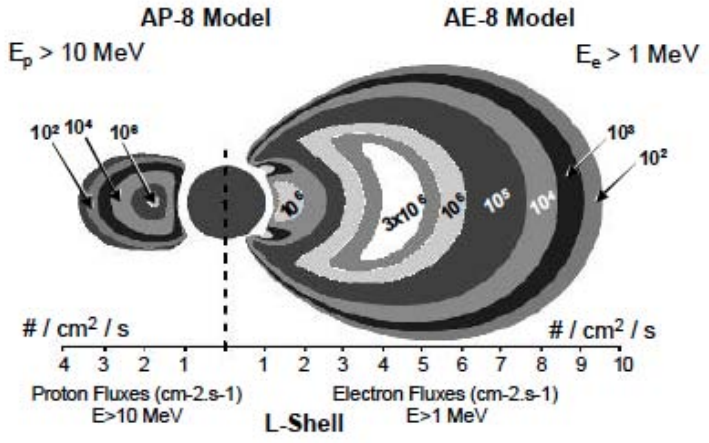


Figure 3.2: Artistic representation of the Van Allen belts as predicted by AP-8 and AE-8 models.

ions (Z>2), covering the full range of elements. The ion energies range from tens of MeV/amu to hundreds go GeV/amu. The source of CGRs is far from our solar system, for this reason the flux is omnidirectional. They are partly kept out by the Earth's magnetic field and have easier access at the poles compared with the equator. If a cosmic ray enters in the Earth atmosphere, it interact with air nuclei, generating a cascade of secondary particles, called secondary GCRs.

Not all of radiation impinging on Earth originate from distant sources. The sun is responsible for changes in the interplanetary and near-Earth radiation levels. The magnetic solar field is highly variable: it has a long term variation, that occurs in a 22-year cycle, and a short term variation, in the form of intense and short storm phenomena (solar flares and coronal mass ejection) [22]. Actually, the solar cycle is 11 years long: at the end the solar magnetic polarity reverses, and another 11-years cycle follows (Fig. 3.3[4]) . However the polarity change has not effects on trapped particles in earth magnetic field, but it influences only the CGRs flux. In the 7 years during solar maximum, the sun emits particles, comprising both protons and heavier ions (96.4% protons, 3.5% alpha particles, 0.1% heavier ions), accelerated during solar flares and coronal mass ejection. Typical energies range up to several hundred MeV, occasionally several GeV. During 4 years of solar minimum the activity levels are low. The solar wind, composed by plasma (i.e. ions and free electrons) and ionized gas, continually emitted by the sun, tends to modify the external magnetic earth field (up to 5-6 earth radii), which is in order subjected to alteration depending on the solar cycle. The magnetosphere results compressed at the sun side, while the opposite

[4]The F10.7 index is a measure of the solar radio flux per unit frequency at a wavelength of 10.7 cm, near the peak of the observed solar radio emission. F10.7 is often expressed in SFU or solar flux units (1 SFU = 10^{-22} W/m^{2}Hz).

side is more extended (Fig. 3.4). Solar particles are less penetrating than CGRs and only a few events in each cycle can reach aircraft altitudes or ground level. However the sum of GCRs and high energy solar particles (SEP=*Solar Energetic Particles*), emitted during solar storm, could considerably increase the incident flux of particles on spacecrafts and penetrate, according to their energies, in the magnetic earth field, being significant for satellites.

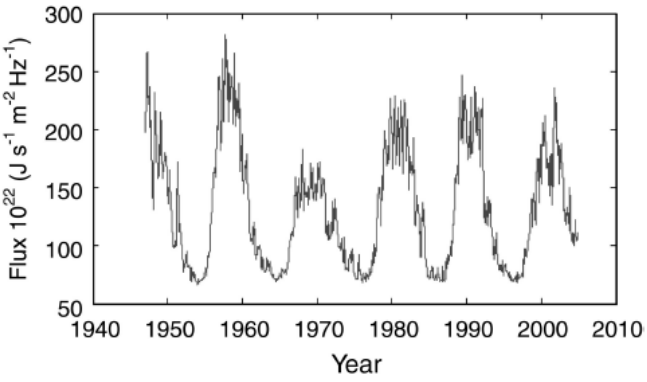


Figure 3.3: Measured values of solar 10.7 cm radio flux.

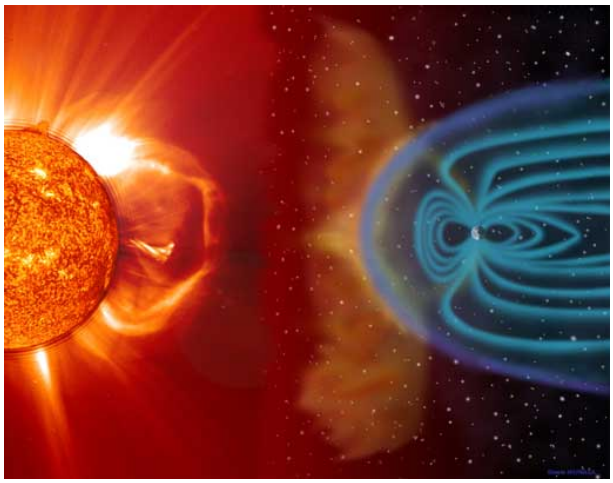


Figure 3.4: Coronal mass ejection and the solar wind effect on the magnetic field

Terrestrial Environment

GCRs and solar particles which penetrate the magnetosphere, is attenuated by the interaction with the earth's atmosphere, because of the interaction with oxygen and nitrogen atoms. The result is showers of secondary particles, such as protons, electrons, neutrons, muons and pions (Fig. 3.5). However, neutrons are the most important component concerning radiation effects on electronic devices.

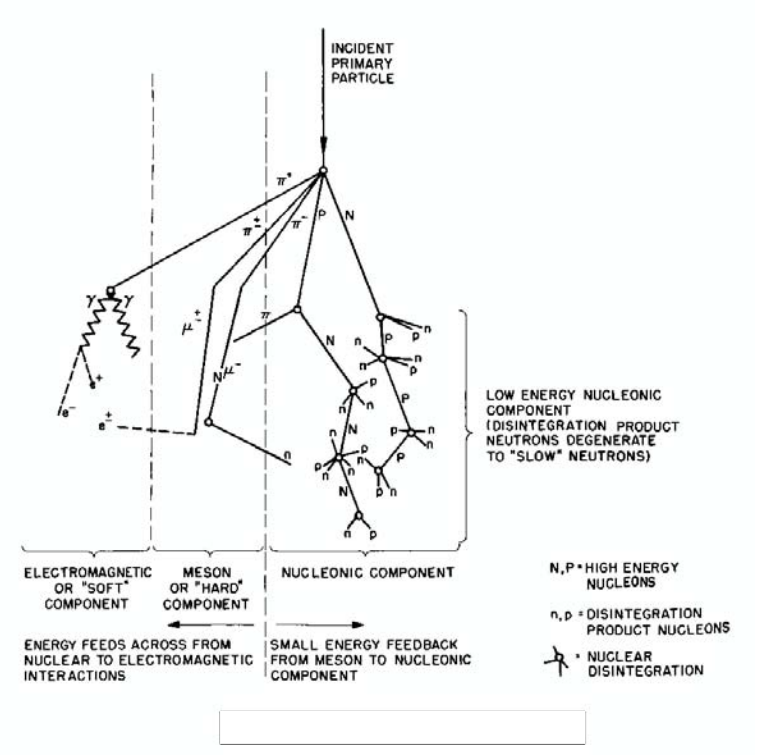


Figure 3.5: Interaction of cosmic galactic rays rays with the atmosphere.

The intensity is maximum at an altitude of 20 km and drops off at sea level where the fluence of neutrons with energy >20 MeV is 10^5 neutrons/cm^2 a year. For avionic applications, neutrons are the dominant factor in producing Single Event Upsets (SEUs).

In addition, on earth are present both natural and man-made radiation, but dose level are below 1 rad per year [23], suggesting that is not very relevant for ionizing or non-ionizing radiation damage in electrical components.

3.2 Basic Concepts

An atom results ionized when it looses one or more electrons, becoming an ion. The ionizing radiation is the radiation able to induce the emission of electrons from the atom with which it interacts. It is caused by the interaction of protons, electrons, energetic heavy ions and photons with the matter. From the standpoint of radiation interaction with matter, different types of incident particles produce different effects. The first important distinction is between ionizing radiation, able to ionize atoms and molecules of the target matter, and non-ionizing radiation, which interacts with the material by transferring thermic energy.

3.2.1 Dosimetry

Dosimetry is the study of how dose is deposited in matter. To define same basic concept rigorously will help to understand radiation interaction with matter.

- **Flux**: the number of impinging particles for area and time units:

$$\phi = \frac{Particles}{Area \times Time} \qquad (\mathrm{cm}^{-2} \cdot \mathrm{s}) \qquad (3.1)$$

- **Fluence**: the number of impinging particles for area unit, thus it is the time integral of the flux:

$$\Phi = \frac{Particles}{Area} \qquad (\mathrm{cm}^{-2}) \qquad (3.2)$$

- **Dose**: the mean energy adsorbed per unit mass of irradiated material at the pout of interest, P. If ΔE_D is the mean energy imparted by ionizing radiation to matter of mass Δm, then:

$$D = \frac{\Delta E_D}{\Delta m}. \qquad (3.3)$$

If ΔE_E is the ionizing radiation energy entering the mass element Δm and ΔE_L is the energy leaving it, then $\Delta E_D = \Delta E_E - \Delta E_L$. The SI unit of dose is the gray (Gy):

$$1\,\mathrm{Gy}{=}1\,\mathrm{J/kg}$$

The customary unit of dose is the rad (radiation absorbed dose):

$$1\,\mathrm{rad}{=}0.01\,\mathrm{Gy}.$$

Since the adsorbed dose depends on the material of interest, the specific material is always referenced in parentheses right after the name of the unit (e.g. rad(Si), Gy(GaAs), etc.).

- **Dose Rate**: the time ratio of charge of absorbed dose. If dD is the increment of absorbed dose in the time interval dt, then:

$$\dot{D} = dD/dt. \qquad (3.4)$$

Radioisotope sources, such as ^{60}Co and ^{137}Cs irradiators, have such a slowly varying dose rate that it can be assumed constant for the duration oh the test[5].

3.2.2 Basic Radiation Effects in Device

Three important effects occur when a component is exposed to radiation:

1. Effects due to **Total Ionizing Dose** (TID)
2. **Single Event Effects** (SEEs)
3. Effects due to **Displacement Damage Dose** (DDD)

Tab. 3.1 resumes radiation effects on spacecraft electronics, categorizing particles by their origin. Several mitigation techniques have been developed against radiation effects, at different levels: circuit level, using specific technologies and process, design level, by *ad hoc* logic structures, system level, modifying software/hardware.

3.3 Total Ionizing Dose

Ionizing radiation is radiation that has enough energy to break atomic bonds and create electron-hole pairs in the material of interest. Ionization damage of a target material is caused by protons, electrons, energetic heavy ions and photons.

The amount of ionization is described by the total dose absorbed. The main consequence of this energy deposition is the trapping of either or both electrons and holes created in dielectric materials, and the subsequent alteration of the proprieties of the devices. Degradation of a single transistor making up a device typically comes in the form of increased leakage current or as a result of threshold voltage shift.

3.3.1 General Overview

Analyzing in detail, the physical processes that leads from the initial deposition of energy by ionizing radiation to the creation of ionization defects are:

1. generation of electron-hole pairs;
2. recombination of a fraction of pairs;
3. transport of free carriers remaining in the oxide;

[5]The absorbed dose rate produced by pulsed radiation source such as flash x-ray generators and LINACs changes very rapidly with time. In this case the dose rate is taken to mean the absorbed dose rate at peak of the pulse.

Particle Origin	Particle	Typical effect
Trapped	Protons	TID SEEs DD Solar cell degradation
	Electrons at L<2.8	TID Solar cell degradation
	Electrons at L>2.8	TID Solar cell degradation Electrostatic discharging
	Heavy Ions	Possible SEE Dose exposure for Humans
Transient	Solar Protons	TID SEE DD Solar cell degradation
	Solar Heavy Ions	SEEs
	Galactic Cosmic Rays	SEEs Dose exposure for Humans
	Plasma Electrons	Deep Dielectric Charging
Secondary	Neutrons-Atmospheric	SEUs in Avionics
	Neutrons-Spacecraft Shielding	DD

Table 3.1: Radiation effects in spacecraft electronics.

4. formation of trapped charge via hole trapping in defect sites or the formation of interface traps.

A fraction of the incident particle's kinetic energy is lost for the creation of electron-hole pairs. The mean energy E_p needed to ionize the material is strongly dependent on the band gap of the material. Thus the number of pairs generated for a given dose depend on E_p as well as the material density[6].

Fig. 3.6 is a plot of a MOS band diagram for a p-substrate and a positive applied gate bias. Once generated, a fraction of the pairs are annihilated through either columnar or geminate recombination, depending on the radiation LET: low LETs cause geminate recombination, while high LETs cause columnar recombination. Furthermore electron-hole pairs surviving geminate recombination is greater than that surviving columnar. The recombination rate is also a function of the electric field within the material. Surviving pairs increase as the local electric field increase. The fraction of electron-hole pairs which escape recombination

[6] In SiO_2 E_p is 17 eV and the pairs density per rad is 8.1×10^{12} pairs/cm^3.

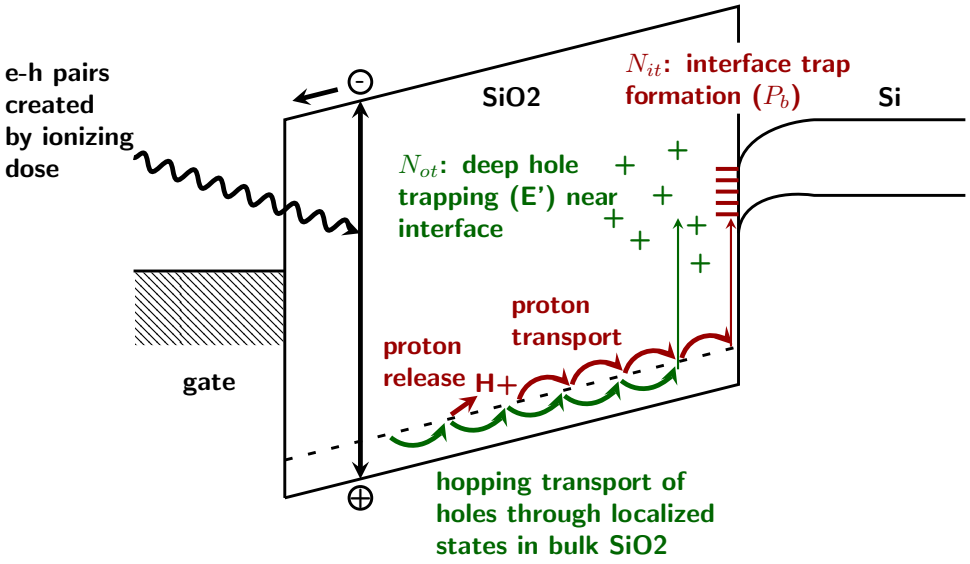


Figure 3.6: Illustration of the main processes in TID damage

is called electron-hole yield. Electrons which escape recombination will rapidly drift toward the gate and holes will drift toward Si/SiO_2 interface. Those holes which escape initial recombination are transported through the SiO_2 toward the Si/SiO_2 interface by hopping through localized states in the oxide. As the holes approach to the interface, a fraction of them will be trapped in oxygen vacancy and subsequently an E' center is formed. The net positive oxide trapped charge, Q_{ot}, depends on the electric field in the dielectric material. Hydrogen ions (H^+) are released as holes "hop" through the oxide or as they are trapped near the Si/SiO_2 interface. The hydrogen ions can drift to the Si/SiO_2 where they may react to form interface traps (N_{it}). In fact, Si/SiO_2 interface is characterized by the presence of silicon dangling bonds, called P_b centers, in which the central silicon is back bonded to three other silicons. Generally dangling bonds are passivated by hydrogen, but proton (H^+) diffusing or driven by the electric field to the Si/SiO_2 interface can remove the hydrogen atoms from the H-passivation.

In addition to oxide-trapped charge and interface-trapped charge in the gate oxide, charge build up also occurs in field oxide and silicon-on-insulator (SOI) oxide. The radiation-induced charge buildup in gate, field and SOI oxides can cause device degradation and circuit failure. For example, positive charge trapping in the gate oxide can invert the channel interface causing leakage current to flow in the off state condition (when the transistor is turned off, V_{GS}=0). This will result in an increase in the static power supply current of the device and may also lead to failure. The same for positive charge buildup in field and SOI oxides can cause rage increase in the static power supply leakage current (caused by parasitic leakage paths in the transistor). For advanced devices,with very thin

gate oxide, radiation-induced charge buildup in field oxide and SOI oxides dominates the radiation-induced degradation. Large concentrations of interface traps can decrease the carrier mobility and increase the threshold voltage of n-channel transistors. Transistor parameters result degraded.

Considering a MOS device, the threshold voltage shift caused by interface traps, ΔV_{it}, is:

$$\Delta V_{\text{it}} = -\frac{q\Delta N_{\text{it}}}{C_{\text{ox}}} \tag{3.5}$$

where ΔN_{it} is the interface trapped charge. ΔQ_{it} can be positive, negative, or neutral, depending on the Fermi level at the interface and on the substrate type. Interface traps are called amphoteric states: the upper half above the intrinsic level are acceptors, that means that they "accept" an electron, becoming negative charged, while the lower half are donors, thus they "donate" an electron to the silicon, becoming positive charged. For a p-channel transistor, the lower part of the band gap (above the Fermi level) is characterized by positive interface traps, on the other hand, for n-channel transistor, the upper part of the band gap (between Fermi level and intrinsic level) has negative charged interface traps.

The threshold voltage shift caused by oxide traps, ΔV_{ot}, is:

$$\Delta V_{\text{ot}} = -\frac{q\Delta N_{\text{ot}}}{C_{\text{ox}}} = -\frac{qt_{\text{ox}}}{\epsilon_{\text{ox}}}\Delta N_{\text{ot}} \tag{3.6}$$

where ΔN_{ot} is the equivalent charge density, considered as all trapped placed at the interface. For previous consideration on the oxide thickness, N_{ot} is relevant for oxide thickness >6 nm.

The final threshold voltage shift considering both interface traps and trapped-oxide charge is $\Delta V_T = \Delta V_{\text{it}} + \Delta V_{\text{ot}}$.

3.3.2 Charge Yield

As previously described, if an electric field exists across the oxide of the transistor, once electron-hole pairs produced, electron and holes are transported in opposite directions. Since electrons have a high mobility in silicon dioxide, they are swept out in picoseconds. However, some of them recombine with holes in the oxide valence band. The amount of pairs which recombine is highly dependent on the electric field in the oxide and the energy and type of the incident particles. Strongly ionizing particles generate a dense electron-hole pairs column, and the recombination rate is relatively high. On the other hand, weakly ionizing particles form quite isolate charge pairs, and recombination rate is lower.

As the electric field increases, the recombination probability decreases. The total amount of holes, N_h which escape this first recombination is given by:

$$N_h = f(E_{\text{ox}}) g_0 D t_{\text{ox}} \tag{3.7}$$

where $f(E_{\text{ox}})$ is the hole yield as a function of the oxide field, D is the dose, t_{ox} the oxide thickness and g_0 is a material parameter which gives the initial pair density per rad of dose [24]. In the case of silicon dioxide g_0=8.1 $\times$ 10^{12} pairs/cm^3.

3.3.3 Oxide Traps Neutralization

The trapped charge in the oxide will be neutralized, depending on the temperature and electric field. Thus ΔV_{ot} decreases following a logarithmic time dependance [24]. Oxide-trap charge decreases as temperature increases, because detrapping is a thermally activated process. Furthermore trapped-charge neutralization is bias dependent: as the bias voltage rises, ΔV_{ot} decreases.

Two mechanisms allow the neutralization of the oxide-trap charge:

- the tunneling of electrons from the silicon into the oxide traps;
- the thermal emission of electrons from the oxide valence band into oxide traps.

The probability of an electron tunneling from the silicon into the oxide is given by [24]:

$$p_{tn} = \alpha e^{-\beta x} \tag{3.8}$$

where α in the attempt to escape frequency, x the distance of the trap from the Si/SiO_2 interface, and β is a tunnel parameter related to the electron barrier height. The tunneling probability is independent of temperature, but exponentially related with the distance of the trap from the Si/SiO_2 interface. However the medium free path for an electron into the oxide is about 2 nm, thus a trap more than 2 nm from the interface is inaccessible to the electron. The rate of oxide traps neutralized by electron tunneling is highly dependent on the spatial distribution of traps into the oxide.

The second neutralization mechanism, the thermal emission, is ruled by the probability p_{em} on an electron being emitted from the oxide valence band into a trap:

$$p_{em} = AT^2 e^{-\phi_t q/kT} \tag{3.9}$$

where ϕ_t is the energy difference between the oxide valence and and the trap and A is a constant, which depends on the capture cross-section of the trap. The phenomenon is strongly dependent on the temperature, in fact it is thermally activated, and independent on the position of the trap. In this case, the neutralization rate depends on the energy distribution of traps.

Temperature and bias dependence of oxide-traps neutralization is affected by the traps spatial and energy distribution, which both depend on the device fabrication conditions.

In addition to neutralization by electron tunneling and thermal emission, oxide-traps can be compensated by radiation-induced electrons trapped at electron trap sites associated with trapped holes.

Concerning interface-traps, they do not anneal at room temperature.

3.4 Radiation Induced Failure in Flash memory

Both Total Ionizing Dose and Single Event Effects impact the reliability of Flash memories. Several studies about TID and SEE on FG memory demonstrated that the control circuitry is the most vulnerable part of commercial devices. It includes the charge pump, used to obtain the high voltages needed for read, programming and erasing, the output buffers and the state machine, used to control the circuit. However information in FG cell is based on the presence or absence of charge on an electrical conductor and the charge generated by the radiation may lead to the corruption of the stored data, through the degradation of the threshold voltage. Galactic cosmic rays and protons from solar flares can upset the in internal circuitry as well s the FG arrays. Upsets such as incorrect read/write operation can occur, or functional interrupt until the internal circuitry will reinitialize.

As Flash memory devices are going towards higher density structure, throughout scaling down, the radiation sensitivity is becoming more and more critical.

The following analysis of radiation effects will start from the FG cell, then will examine the peripheral circuitry.

3.4.1 Floating Gate Cell

In general, an error occurs in a FG cell when the ionizing radiation induces a threshold voltage shift large enough to bring the V_T of the cell below the read voltage. This happens both for total ionizing dose and single event effects, but since this thesis deals with TID effect on FG cell, the description of only this type of radiation effects on Flash memories will be provided.

Total Dose Effects

Since the threshold voltage is not visible to the user, but only the digital value, the error is detected if the output is “1” instead of “0”. As well know, this is due to the radiation-induced threshold voltage shift.

Three main mechanisms, depicted in Fig. 3.7, have been identified as responsible for the threshold voltage shift [25].

1. Injection into the FG of holes generated by the impinging radiation in the oxides surrounding the FG: these holes recombine with the stored electrons, reducing the negative stored charge.

2. Charge trapping in the tunnel oxide. However, this component is typically small because of the small oxide thickness, but it may give rise to visible effects over time after exposure, causing error reduction.

3. Electrons stored in the FG can gain enough energy from the ionizing radiation to be emitted from the FG over SiO_2 barrier towards the substrate or the control gate.

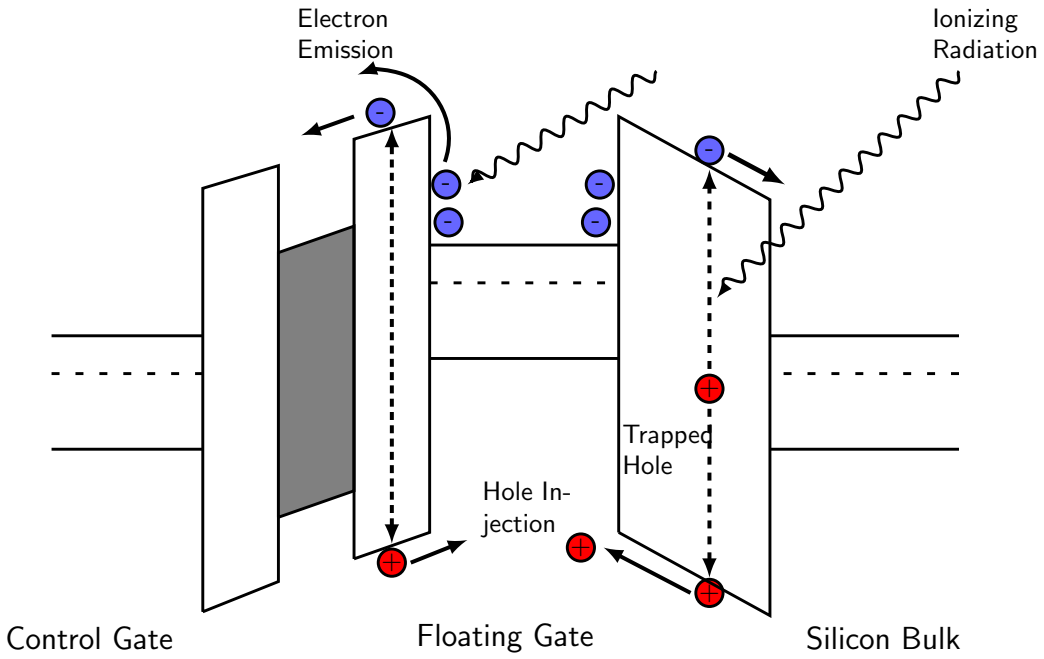


Figure 3.7: Three mechanisms responsible for the TID induced shift of the threshold voltage in floating gate cells.

Let analyze in detail the process. Ionizing radiation imparts energy to oxide regions surrounding the FG and electron-hole pairs are generate (an electron-hole pair is generated for every 17 eV lost by radiation in the SiO_2). The number of carriers generated depends on the material and the volume available. A fraction of electron-hole pairs will recombine in very short time. This fraction depends on the electric field across the oxide E: the greater the electric field, the greater the fraction that escape recombination. The remaining carriers are subject to thermalization. Electrons, which are more mobile than holes, are swept away from

the oxide, under the influence of the oxide electric field. Really this description of different phenomena in sequential order is not realistic for 10 nm (or less) oxide thickness, but recombination, thermalization, electron transit times happen at the same time [26].

Hole Injection and Trapping

The remaining holes drift under the influence of E toward the FG. The positive charge injected in the FG reduces the net amount of electron charge stored because of the recombination, thus decreases the threshold voltage. Positive or negative charge trapping in defects generated by these holes can be neglected in thin (< 10 nm) oxides.

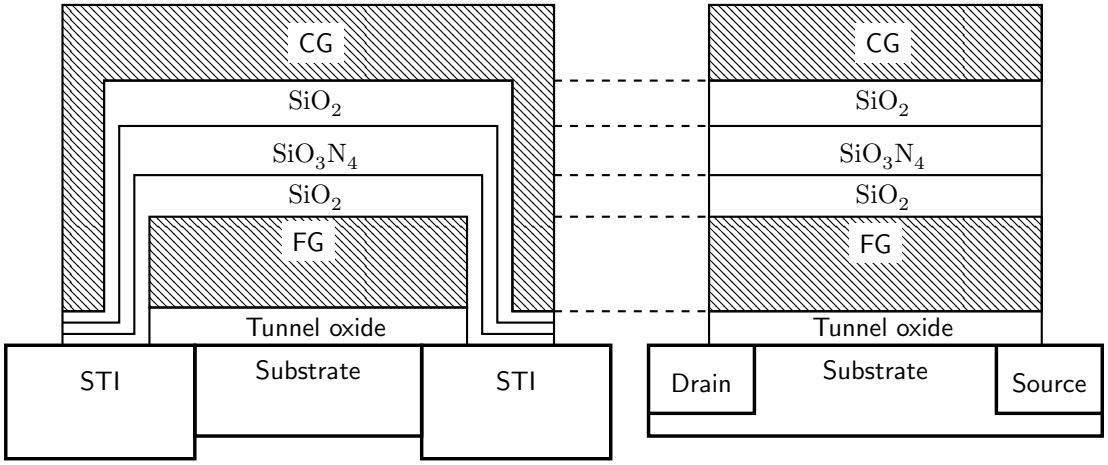


Figure 3.8: Schematic representation of a FG memory cell (out of scale).

Instead, carriers behavior in the oxide-nitride-oxide (ONO) dielectric over and around the FG is more complex. In fact the nitride layer has a very high recombination rate and the charge generated into it is almost zero. First considering the SiO_2 layer in contact with the FG (Fig. 3.8), all the holes generated there that survive recombination are injected in the FG. There they recombine with electrons, so that the amount of charge stored in the FG decreases. In the second SiO_2 layer, over the nitride one, holes surviving recombination are injected in the Si_3N_4, as well as electrons surviving recombination in the SiO_2 layer in contact with the FG. Because of the reduced barrier height of the nitride in comparison with SiO_2, carriers injected in it will not drift to the SiO_2. Modeling the charge in the nitride layer as a charge sheet at the interface between bottom SiO_2 and Si_3N_4, whose amount is given by the difference between positive and negative charge [26]. This charge is almost zero if the two SiO_2 layers have the same thickness, otherwise this charge sheet distribution must be considered. However its

effects can be neglected, being less than 1% the charge generated in the tunnel oxide and in the bottom SiO_2 layer.

Electron Photoemission

Photoemission is the third mechanism present during irradiation. It is due to the interaction between the radiation and the charge in the FG. Electrons hit by the impinging radiation gain enough energy (4.3 eV or more) to surmount the potential barrier. Once the electrons are in the oxide, they are quickly swept away by the electric field. It should be noted that this mechanism does not decrease as the FG thickness or oxide thickness is reduced, partially because there is no more scaling of oxide thickness from one generation to another due to reliability issues [26].

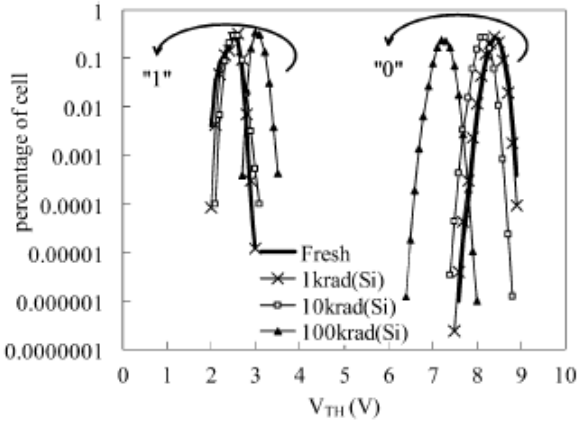


Figure 3.9: Total ionizing dose effect of threshold voltage distribution in NOR arrays after irradiation with 1 krad(Si), 10 krad(Si), 100 krad(Si), with 10-keV X-rays [27].

Fig. 3.9 shows the changes induced by total dose irradiation on the threshold voltage distribution of a NOR architecture. The programmed distribution moves towards lower voltage values, while the erased distribution shifts to higher values. Both go towards the intrinsic distribution, which is the state of no charge net charge in the floating gate and is typically between the two distributions in NOR devices. Depending on the position of neutral threshold voltage, total-dose-induced errors occur prevalently in the programmed or in the erased cells. Fig. 3.10 illustrates the concept. If the neutral threshold voltage is placed below 0 V, a complete discharge of the floating gate will cause the programmed cell to be read as erased. On the contrary, if the neutral threshold voltage distribution is placed above 0 V, the erased cell will be read as programmed. If the neutral threshold voltage distribution stays on both sides of 0 V, both the types of errors can be

detected. In the devices object of the experiment, only errors in programmed cells have been detected, as will be presented in the following chapter.

In case of multi-level cells, the output interpretation may be more complex, because of the multiples threshold voltage distributions related to multiple program levels, however, as for single level cells, it has been observed [28] a globally shift toward the neutral distribution.

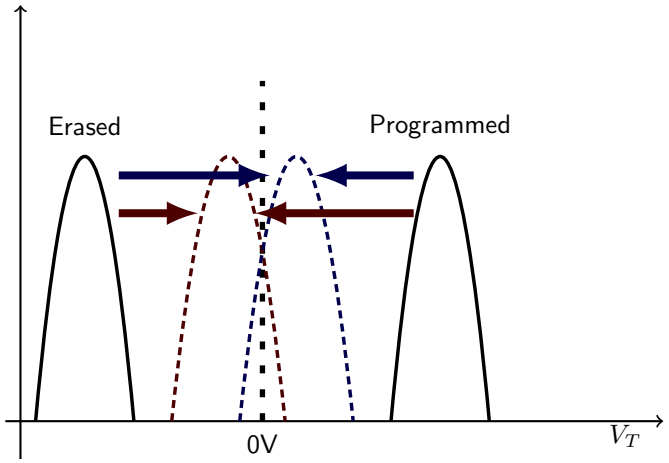


Figure 3.10: Sketch of the of V_T distribution shift due to TID.

3.4.2 Peripheral Circuitry

The corruption of the peripheral circuitry leads to a type of errors called **dynamic errors**, opposite to **static errors** caused by floating gate failures. Dynamic errors are present in one read cycle and disappear in the next one, because of an incorrect reading operation on a whole block. Such failures are called Single Event Functionally Interrupt (SEFI) and are delated to a strike in the embedded microcontroller [3]. Analyzing the NAND architecture, to perform a read operation, the row decoded selects the desired page in the FG array. The selected page is transferred to the page buffer (PB), from where the data are output byte by byte to the device pins. For the writing, operation sequence is the opposite. While in the PB, data are sensitive to radiation: if a ion strike the PB, a dynamic error occurs. Since each time the page is accessed in the FG array the page buffers are corrected, the SEFI is visible only in one cycle. SEFI can be restored by repeating the operation failed, resetting or by power cycle. This is one of the Flash memory advantage, since no data loss occurs after a power cycle.

Concerning total dose, the most sensitive part is the pump circuitry. This block have the function of providing the high voltage needed for programming

and erasing, and also for reading in modern devices. Generally three different charge pumps are present.

1. The *read pump* generates the high voltage needed to read the array and verify the correctness of program/erase operations.

2. The *program pump* provides the high voltage needed to injects charge through Fowler-Nordheim tunneling in the floating gate. It is used during program and erase operations by applying voltage at different terminals, in order to inject or removal carriers.

3. The *pass pump* is used during programming and erasing, to avoid injection of charge in unwanted cells belonging to the same word line of the selected ones.

Charge pumps work by charging some capacitors and connecting them in series to reach the high voltage. This block is the most critical for ionizing radiation tolerance. The voltage output must be well calibrated for make the memories work correctly and this low tolerance, combined with high voltage values, leads to failures. In fact pump circuity is one of the firs cause of failure during total dose tests. Fig. 3.11 is an example of charge pump voltage degradation. Charge pumps are also sensitive to heavy ions: single event gate rupture have been reported.

The voltage generated can be degraded because of V_T shift in the transistor and leakage increase, due to charge trapping in the STI, both after TID exposure and heavy-ion irradiation. For example, in NAND architecture the read pump must provide a voltage (V_{READ}) high enough to guarantee that both erased and programmed cells are turned on. If this voltage is lower that designed, because of total dose, and it is not able to turn on even a single transistor of the string, all the cell belonging to this string will be read as programmed (no current sensed), whatever their actual state. Degradation in the program voltage provided by the program pump (V_{PROG}) has been reported [29]. The failure dose is lower in this case because of the higher generated voltage respect to V_{READ}. Furthermore variation in program pumps are more critical than in read pump.

Still referencing to NAND architecture, radiation effects in the row decoder exhibit a behavior similar to the charge pumps degradation, but at higher doses. All the cells are read as programmed, regardless they programming condition before irradiation. The failure happens because some transistor, that should be turned on, are actually off, or because leakage discharges some nodes that were pulled high, blocking the current flowing in the string. In particular one of the following condition may occur [29]:

i) SSL or/and DSL are not turned on;

ii) the gate of the cell to be read is not grounded, due to failure in some pass transistor;

iii) a voltage lower than V_{READ} is provided to the transistor belonging to the same string of the cell to be read.

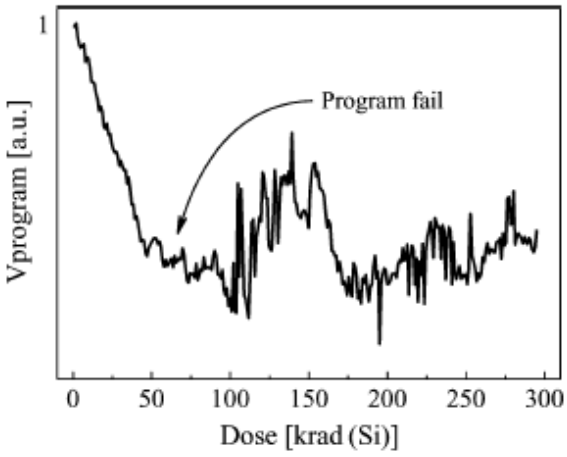


Figure 3.11: Degradation of the output voltage of a Flash charge pump as a function of total dose.

Heavy ions with high LET could lead to destructive events (DE) [30], consequence of a current spike, without the possibility of any recovery action. It has been demonstrate that charge pumps play a crucial role in DE.

3.5 Conclusions

In this chapter have been presented the main issues related with radiation effects on electronic devices. An initial general overview has been useful to roughly understand the ionizing radiation sources and their effects on semiconductor devices.

It has been paid particular attention to floating gate cell response to TID, since Flash memories capacity and non-valatility makes them attractive for space applications, due to the lack of corresponding rad-hard parts with comparable size. Total dose affects primarily the control circuitry, because of the degradation of the charge pump circuitry. Failure dosed vary widely between manufacturers and different generations, since the ratio between deposited charge by the impinging radiation and stored charge increases.

This analysis of ionizing radiation induced effects of Flash memory will allow to interpret experimental result presented in the following chapters.

Chapter 4

Experiment

In this chapter the experimental procedure developed for the TID test will be exposed. Features of the Flash memories under test and of the tools used to realize measurements will be described.

4.1 Estec Co-60 Facility

TID measurements were performed at the ESA-ESTEC Co-60 Facility, Noordwijk, The Netherlands. The new radiation facility has been reloaded in October 2011 with a 2000 Ci[1] Co-60 gamma source. It is useful to present basic characteristics of this element.

Cobalt-60 is a radioactive isotope of cobalt, with a half-life of 5.2714 years. It decays by beta decay to the stable isotope nichel-60, emitting two gamma rays with energies of 1.17 and 1.33 MeV. The equation of the nuclear reaction is:

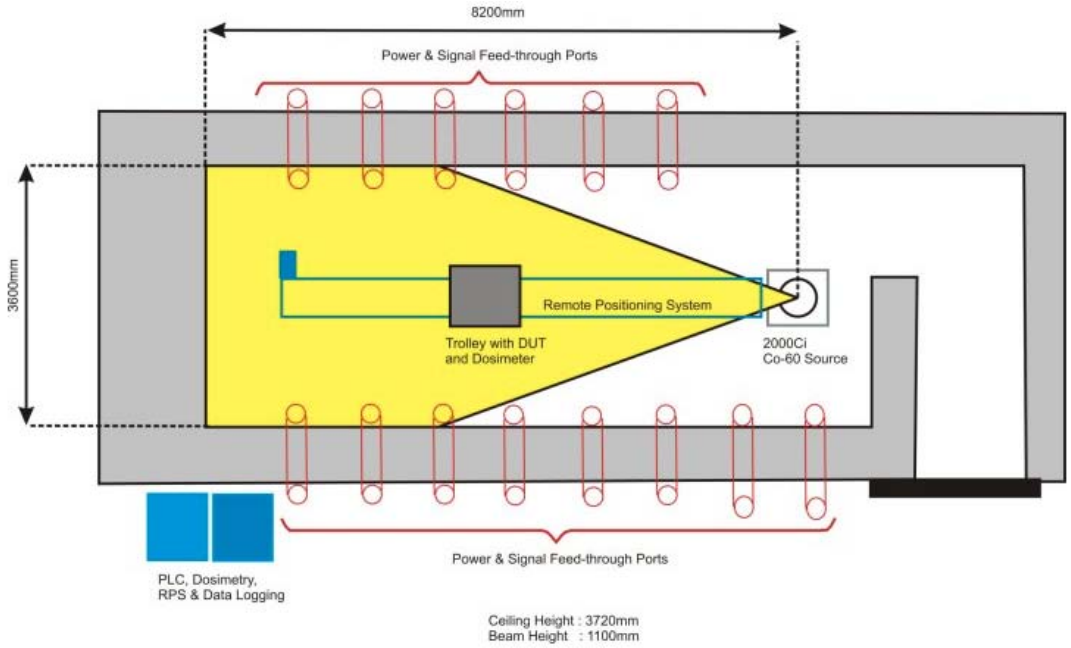
$^{59}_{27}\mathrm{Co} + \mathrm{n} \rightarrow {}^{60}_{28}\mathrm{Ni} + \mathrm{e}^- + \overline{v}_e + \text{gamma rays}$

One of the Co-60 application is as gamma source for tests and studies on ionizing radiation effects on electronic components.

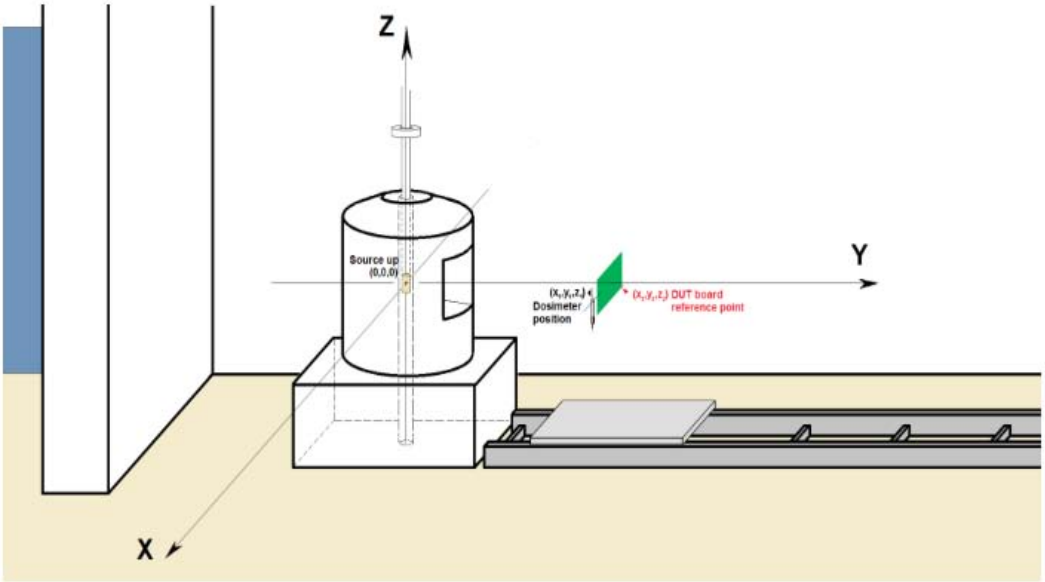
The ESTEC Co-60 source consists of multiple small rods, about 50 mm long, held around the periphery of a 30mm diameter container. The container is of double-welded steel construction, made to the internationally approved standards. The source is stored in its own special housing, built of steel with integral lead shielding. When the source is raised to the irradiation position, the gamma beam, produced by the Co-60 decay, exits the irradiator unit through a collimator window into the radiation cell [31].

The source is placed in radiation cell, depicted in Fig. 4.1 There are 14 cable feed-throughs enabling monitoring and control of experiments under test from the control room, just close to the radiation cell. Total dose and dose rate are measured using a dosimeter, placed near the DUT, and monitored from a PC in

[1] 1 Curie= 3.71×10^{10} decays per second.



(a) Radiation cell



(b) Co60 irradiator head and board positioning

Figure 4.1: ESA-ESTEC Co-60 Facility [31].

the control room, from which it is possible also to change the experiment position, therefore the dose rate, through a rail system installed in the radiation cell.

The device under test can be placed at minimum distance of 40 cm and at maximum distance of about 800 cm, which corresponds to a dose rate of 161.64 rad/min(Si) and 0.495 rad/min(Si) respectively.

4.2 Experimental Procedure

The subject of the study is 25-nm SLC NAND Flash memory by Micron. Part number is MT29F32G08ABAAA. The device size is 32Gb, organized in 4096 blocks, each of 128 pages. The page size is 8640 bytes. 40 devices were tested, from two different lots. The two lots were called A and B, to distinguish between them. In addition, one Flash memory from the first lot (A) has been used for initial calibration.

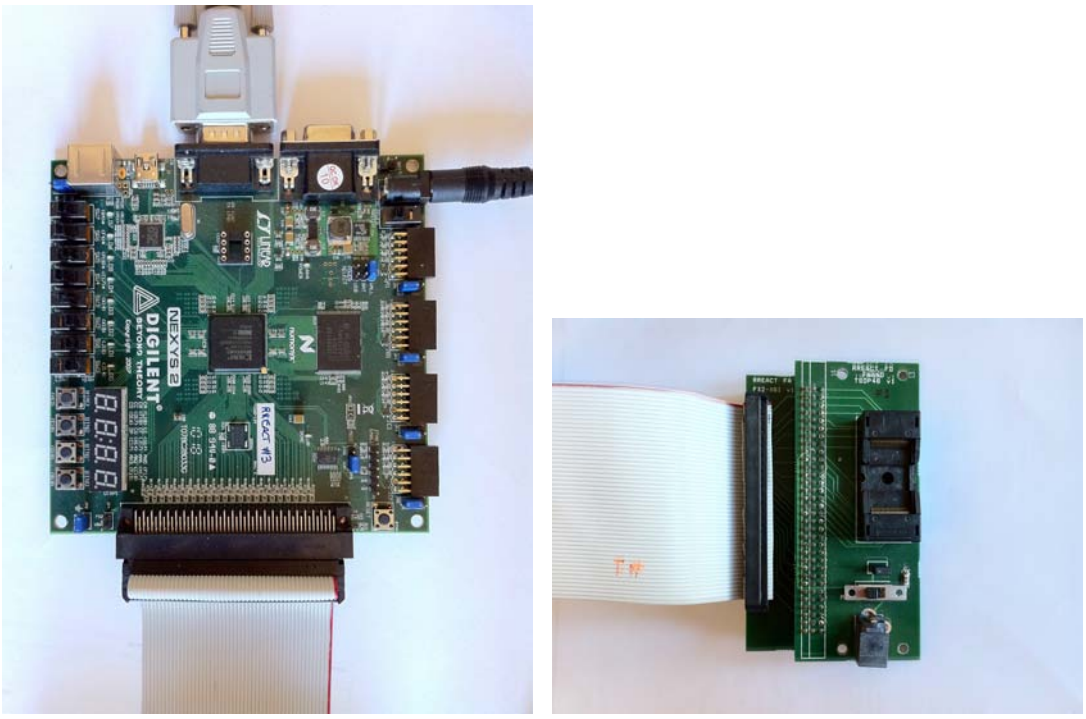
Two boards were used to perform all the operations on the devices (Fig. 4.2a). They are implemented by an FPGA and are connected by an IDE ribbon cable to the socket (Fig. 4.2b), where the memory is placed. Boards were connected to the power supply and controlled from the PC throughout a USB-serial controller. The RREACT Forgetful program, developed by the RREACT Group, allows to menage all the operations on the device from a clear interface, thus selecting the device configuration, checking the SR (Status Register) and the ID (Identification number), and performing the fundamental operations on the Flash memory: erasing, programing and reading. If it is necessary to implement complex operation, such erase/program/read loops, it possible to execute scripts, setting list of operations to perform.

Tab. 4.1 resumes the test conditions and the partition of the 41 devices during five weeks.

4.2.1 Calibration

The first week has been employed for the calibration, which helped to choose a correct total dose for the test. In fact the goal was to detect a significant amount of errors due to charge loss in the floating gate cells and to avoid failures of the peripheral circuitry, since we know that this is the most critical part. So the choice of the total dose has been an essential step to lead the experiment to consistent results with its purpose.

The dose should be low enough to allow to detect the errors build up and to avoid device failure. At the same time, the amount of errors caused by TID had to be large enough to statistically analyze the results. In fact, during calibration, the DUT was continuously measured, allowing to collect several measurements



(a) (b)

Figure 4.2: The board (a) and the socket (b).

Internal reference number	A11-A20	B1-B10	A21-A30	B11-B20	A10
Supply voltage	unbiased	unbiased	unbiased	unbiased	3.3V
Relative Humidity [%]	37.2-38	37.2-38	37.2-38	37.2-38	37.2-38
Operating temperature [°C]	22.4-22.6	22.4-22.6	22.4-22.6	22.4-22.6	22.4-22.6
Pressure [mbar]	1014.6-1031.8	1014.6-1031.8	1014.6-1031.8	1014.6-1031.8	1014.6-1031.8
Test week	47	48	49	50	46

Table 4.1: Details of the SLC NAND memories.

during total dose rise. We decided to irradiate the calibration device with a total dose of about 60 krad(Si), because previous studies on the same devices [32] showed devices failure at total dose of about 50 krad(Si).

During calibration, both the Flash memory, positioned on the socket (Fig. 4.4a), and the board were placed in the irradiation room. The distance from the source was 120 cm, in order to obtain the desired dose rate (0.2951 rad(Si)/s). Obliviously the board was shielded with lead bricks, to minimize as possible the radiation exposure. The setup is pictured in Fig. 4.3. From the controller room, a power supply of 3.3 V were connected to the memory, instead 5 V were provided to the board. The current was limited at 100 mV and 1 V respectively for the memory and the board. A multimeter allowed to measure the current during test. Details of the calibration are provided in Tab. 4.2. The dose rate was 0.2951 rad(Si)/s, i.e. 17.7 rad(Si)/min, thus 56 hours and 30 minutes were expected to reach the chosen total dose.

Run	Start	Stop	Dose rate(Si) [rad/s]	Dose(Si) [krad]	DUTs	Operating condition
1	13/11/13 11:50	15/11/13 14:43	0.2951	54.06	A10	3.3V

Table 4.2: Gamma TID run on the calibration device.

The script developed for the calibration managed the operations on the device. 2 sets of blocks, each one of 30 blocks, were firstly erased, then programmed with checkerboard pattern (i.e. “01” pattern). The first set of 30 blocks of the memory was continuously erased and programmed during radiation exposure, to check the erase and program errors, in order to detect failures in the charge pumps. The other 30 blocks, once programmed, were read every 17 minutes during the radiation exposure, to check the retention errors.

After each E/P/R cycle on the exercised blocks and reading of the blocks kept in retention, the memory was powered off. In fact, to simulate the TID test on unbiased devices, the calibration memory had to be mostly powered off. In this condition, the peripheral circuitry is less vulnerable to wear out. With this power off interval between operations, the device was unbiased for about 95% of the total test time. Because of the long duration of the test, the script provided a reset mechanism in case of system block. The log files, including the current measurement, were automatically saved.

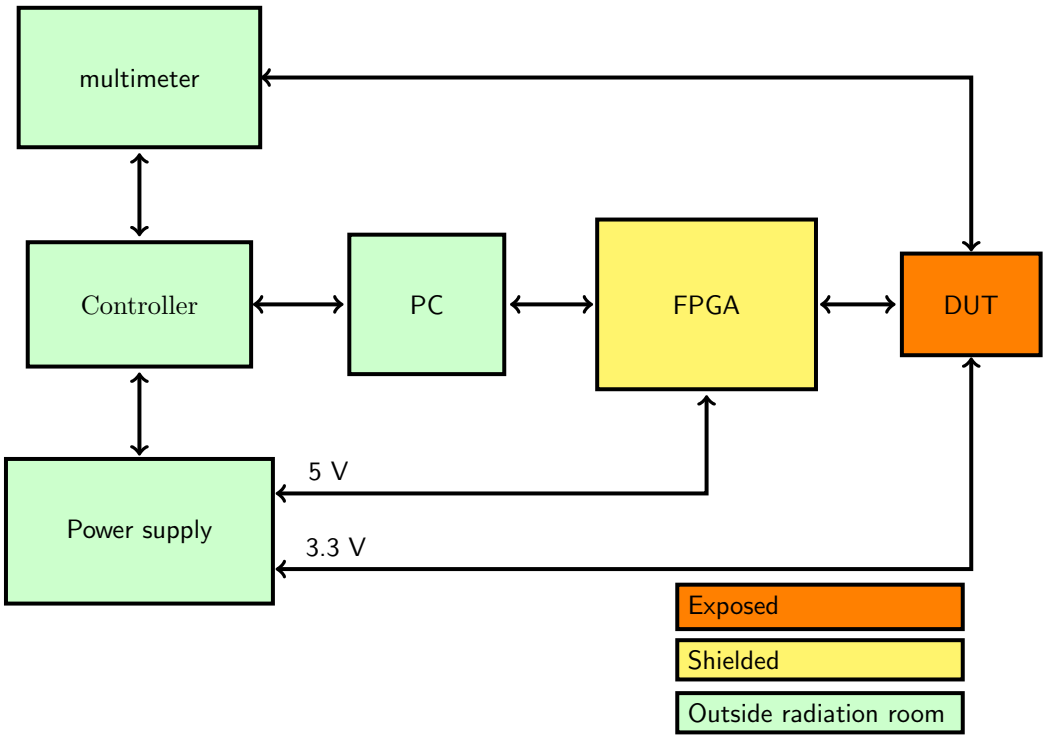
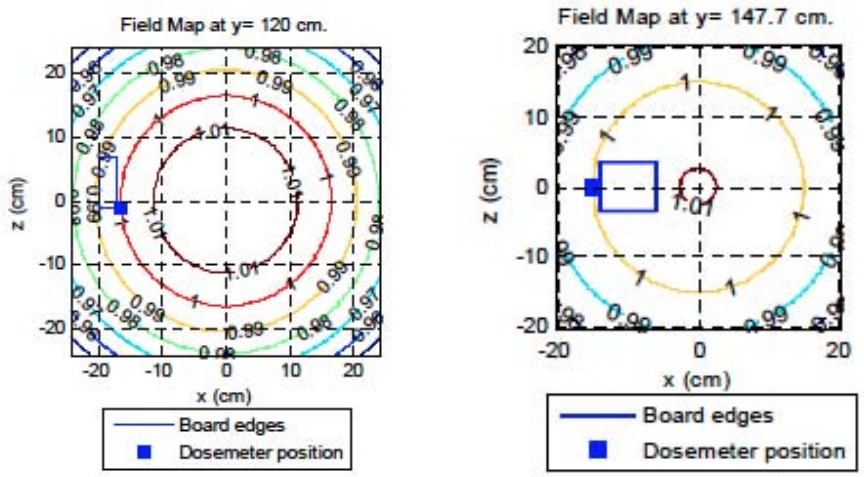
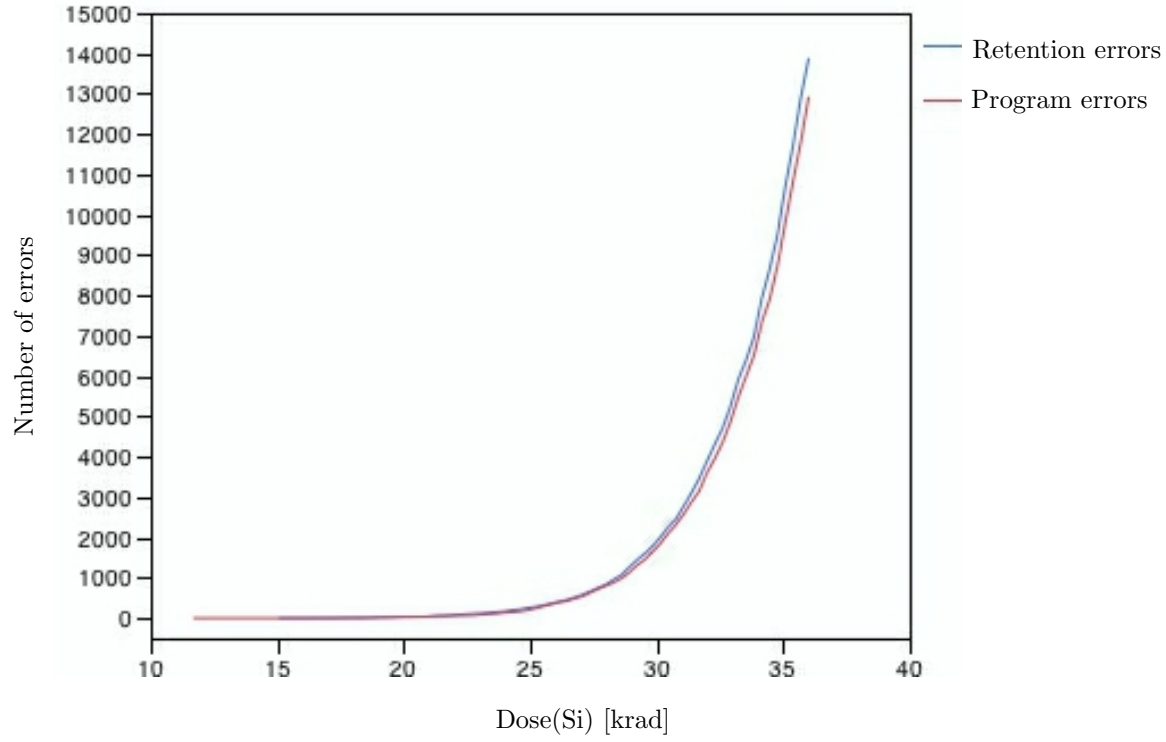


Figure 4.3: Calibration setup

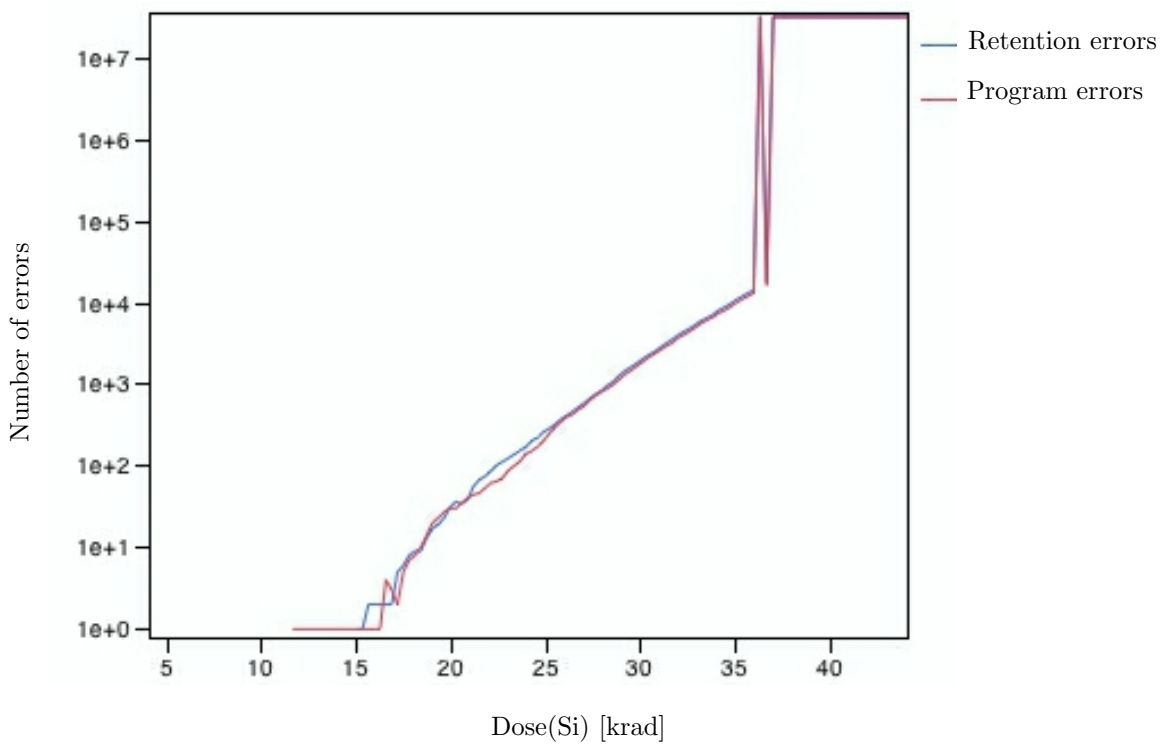


(a) Calibration **(b)** TID test

Figure 4.4: Dose rate distribution normalized to dosimeter readings.



(a)



(b)

Figure 4.5: Retention errors and program errors, as a function of total dose during calibration.

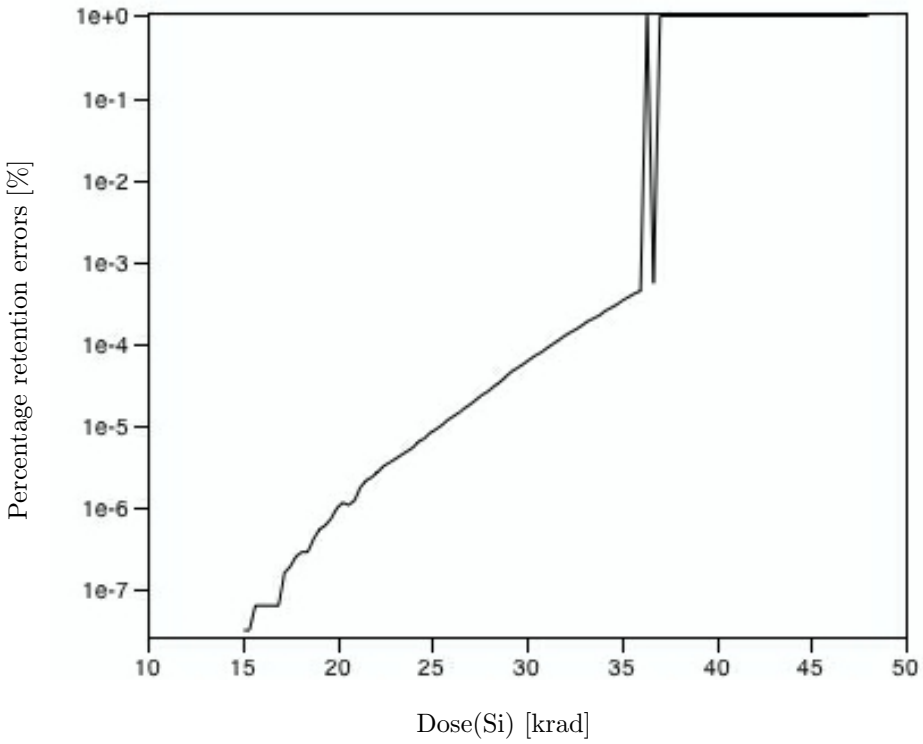


Figure 4.6: Percentage of bit error in programmed cells versus total dose.

Calibration results

Fig. 4.5 shows the errors build up as a function of the dose. Retention errors appeared at 15 krad(Si) and increased with the dose, following an exponential trend. All the retention errors were observed in program cells (i.e. storing electrons), thus bit flipping from "0" to "1". In fact, in NAND architecture, erased cells store positive charge (i.e. holes) in the FG and the neutral threshold voltage, that is the V_T with no charge in the FG, is placed below 0 V. With this arrangement a complete discharge of the FG will cause the cells to be read as erased. As presented in the second chapter, mechanisms that cause the FG discharge are photoemission, charge recombination, and positive charge trapping in the oxides surrounding the floating gate.

At 36.35 krad(Si), after about 34 hours from the beginning of gamma run, an abrupt increase of the number of errors occurred. This was due to control circuitry failure. So the irradiation test was stopped. The assumption of the peripheral circuitry failure is confirmed by the fact that the second reading, after the first failure, detected a number of errors comparable with the previous values, before the charge pump failure. Probably, a temporary functional recovery occurred. The same happened for program errors: the first error appeared at 11.68 krad

and the abrupt increase occurred at the same dose of that for retention errors.

Comparing with previous samples [33] (34-nm SLC NAND Flash memories by Micron), errors build up started at lower dose and it is more rapid. Furthermore, the device failure, due to the control circuitry sensibility, happened at lower total dose values than less scaled devices. Devices with different size show different TID sensitivity. However past studies demonstrated that TID effect on FG errors does not depend much on scaling if only the cell width and length (W and L) are scaled [27]. On the contrary, small changes in the tunnel oxide thickness could increase the charge loss rate, because of the formation of discharge paths, thus the number of errors rises more rapidly.

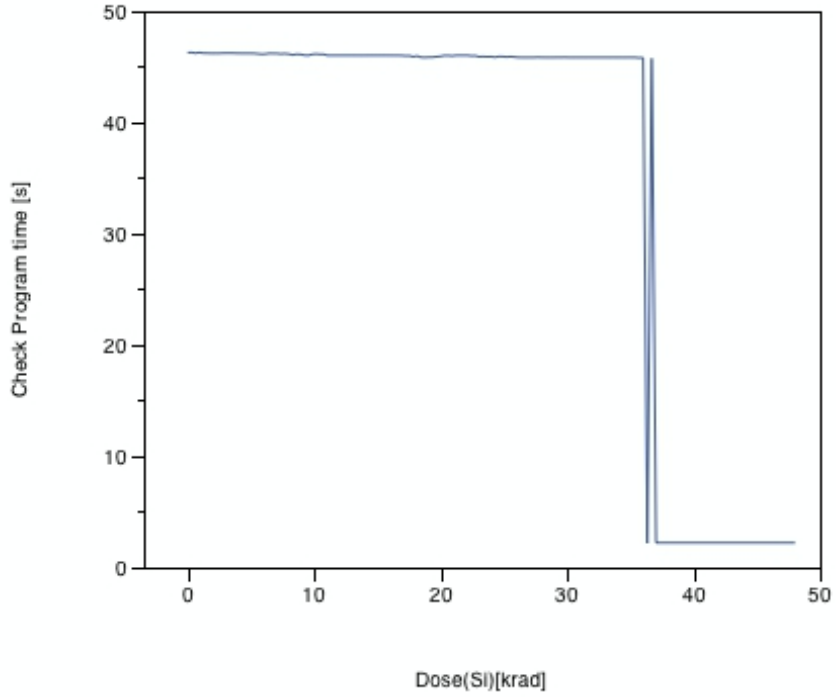


Figure 4.7: Check program time as a function of total dose during the calibration

Fig 4.7 illustrates the degradation of the reading time as the dose increases. The time dropped down at the constraint value of 2 s after the control circuitry failure. As for errors, a functional recovery was observed just after the first failure appearance.

4.2.2 TID test setup

Based on calibration results, the total dose provided for the test should be absolutely lower than 36 krad(Si). It has been decided to irradiate the devices with a total dose of 33 krad(Si), to avoid failures in the control circuitry. At this dose level thousands of errors were expected. This dotal dose was divided

in two steps, to allow us to perform an intermediate measure. In fact during irradiation the devices were unbiased, placed on a support perpendicular to the plane (Fig. 4.4b). After 20 krad(Si), the gamma run was stopped to extract the DUTs from the irradiation room. At this dose, hundreds of errors were expected, based on the calibration. Thus the second dose step was of 13 krad(Si). The dose rate was 0.1928 rad(Si)/s. Small dose rate changes are supposed to not affect the TID response.

10 memories were measured per week. Memories were tested, in couples: each pair was exposed to radiation one hour after the previous pair. The device were extracted from the irradiation room after a dotal dose of 20 krad(Si), in the same chronological order. One hour is the maximum supposed time to read a memory in the worst case[2] of about 10^{-2} % bit errors. Tab. 4.3 and 4.4 report the details of the gamma runs.

Before irradiation, all the blocks of the memories were erased, bad blocks were checked, and then devices were programmed with checkerboard pattern, except for 10 block, programmed with random pattern, to check the proper functionality of the memory peripheral circuitry. In the Appendix A are reported the script for programming and reading. Obliviously, bad blocks were not considered (the program and read scripts acted in order to skip bad blocks[3]). A read operation was performed just before the exposure, to detected possible program errors and errors appeared after programming (devices were programmed about three days before the radiation test). Once programmed, memories have never been erased or reprogrammed. Only on 30 blocks in each device were performed erase/program/read cycles, to allow measurements of the erase/program/read currents. Devices in pair were read as simultaneously as possible, to avoid different times between the run stop and the read operation, thus to minimize variations due to different annealing times. The samples have been measured again 24 hours, one and more weeks after the radiation exposure, to control and compare annealing effects. All the tests and measurements were performed at room temperature.

Operations performed on each Flash memory device are summarized in the following list:

i) program operation (10 blocks with random pattern, 4086 blocks with checkerboard pattern) about three days before irradiation;

ii) read operation, just before the gamma run start;

iii) irradiation with a total dose of 20 krad(Si);

[2]From previous studies [32] .

[3]NAND Flash memory is specified to have at least 4016 valid blocks of the total available blocks, while the total available blocks are 4096.

(a)

Run	Start	Stop	Dose rate(Si) [rad/s]	Dose(Si) [krad]	DUTs	Operating condition
1	18/11/13 10:57	19/11/13 16:00	0.1928	19.998	A11-A12	unbiased
2	18/11/13 11:57	19/11/13 17:00	0.1928	20	A13-A14	unbiased
3	18/11/13 12:58	19/11/13 18:02	0.1928	20.045	A15-A16	unbiased
4	18/11/13 13:57	19/11/13 19:00	0.1928	20.009	A17-A18	unbiased
5	18/11/13 14:56	19/11/13 19:58	0.1928	20.007	A19-A20	unbiased
6	20/11/13 14:39	21/11/13 09:33	0.193	12.999	A11-A12	unbiased
7	20/11/13 15:38	21/11/13 10:31	0.193	12.9994	A13-A14	unbiased
8	20/11/13 16:36	21/11/13 11:30	0.193	12.9994	A15-A16	unbiased
9	20/11/13 17:34	21/11/13 12:28	0.193	12.9996	A17-A18	unbiased
10	20/11/13 18:32	21/11/13 13:26	0.193	12.9998	A19-A20	unbiased

(b)

Run	Start	Stop	Dose rate(Si) [rad/s]	Dose(Si) [krad]	DUTs	Operating condition
1	25/11/13 11:32	26/11/13 16:40	0.1928	19.07	B1-B2	unbiased
2	25/11/13 12:05	26/11/13 17:09	0.1928	19.996	B3-B4	unbiased
3	25/11/13 13:07	26/11/13 18:10	0.1928	20	B5-B5	unbiased
4	25/11/13 14:07	26/11/13 19:10	0.1928	19.996	B7-B8	unbiased
5	25/11/13 14:40	26/11/13 19:43	0.1928	19.996	B9-B10	unbiased
6	27/11/13 15:20	28/11/13 10:14	0.1929	13.003	B1-B2	unbiased
7	27/11/13 15:45	28/11/13 10:39	0.1929	13.003	B3-B4	unbiased
8	27/11/13 16:45	28/11/13 11:39	0.1929	13	B5-B6	unbiased
9	27/11/13 17:45	28/11/13 12:38	0.1929	12.997	B7-B8	unbiased
10	27/11/13 18:19	28/11/13 13:11	0.1929	13	B9-B10	unbiased

Table 4.3: Gamma runs during week 47 (a) and 48 (b).

(a)

Run	Start	Stop	Dose rate(Si) [rad/s]	Dose(Si) [krad]	DUTs	Operating condition
1	02/12/13 11:04	03/12/13 16:02	0.1955	19.995	A11-A12	unbiased
2	02/12/13 12:04	03/12/13 16:55	0.1955	19.9955	A13-A14	unbiased
3	02/12/13 12:57	03/12/13 17:48	0.1955	19.9958	A15-A16	unbiased
4	02/12/13 13:39	03/12/13 18:30	0.1955	19.996	A17-A18	unbiased
5	02/12/13 14:20	03/12/13 19:12	0.1955	20.0027	A19-A20	unbiased
6	04/12/13 14:38	05/12/13 09:14	0.1968	13.0008	A11-A12	unbiased
7	04/12/13 15:33	05/12/13 10:10	0.1968	13.0022	A13-A14	unbiased
8	04/12/13 16:25	05/12/13 11:04	0.1968	13.0027	A15-A16	unbiased
9	04/12/13 17:05	05/12/13 11:41	0.1968	13.0025	A17-A18	unbiased
10	04/12/13 17:51	05/12/13 12:27	0.1968	13.0001	A19-A20	unbiased

(b)

Run	Start	Stop	Dose rate(Si) [rad/s]	Dose(Si) [krad]	DUTs	Operating condition
1	09/12/13 11:00	10/12/13 15:58	0.196	20.001	B11-B12	unbiased
2	09/12/13 12:00	10/12/13 16:57	0.196	20.001	B13-B14	unbiased
3	09/12/13 12:59	10/12/13 17:57	0.196	20.0007	B15-B16	unbiased
4	09/12/13 13:39	10/12/13 18:38	0.196	20.0017	B17-B18	unbiased
5	09/12/13 14:20	10/12/13 19:18	0.196	20.0018	B19-B20	unbiased
6	11/12/13 14:30	12/12/13 09:13	0.1957	13.004	B11-B12	unbiased
7	11/12/13 15:35	12/12/13 10:11	0.1957	13.003	B13-B14	unbiased
8	11/12/13 16:32	12/12/13 11:11	0.1957	13.004	B15-B16	unbiased
9	11/12/13 17:14	12/12/13 11:51	0.1957	12.98	B17-B18	unbiased
10	11/12/13 17:56	12/12/13 12:32	0.1957	12.99	B19-B20	unbiased

Table 4.4: Gamma runs during week 49 (a) and 50 (b).

iv) read operation, just after the gamma run stop;

v) annealing of 22 hours and 30 minutes;

vi) irradiation with a total dose of 13 krad(Si);

vii) read operation, just after the gamma run stop.

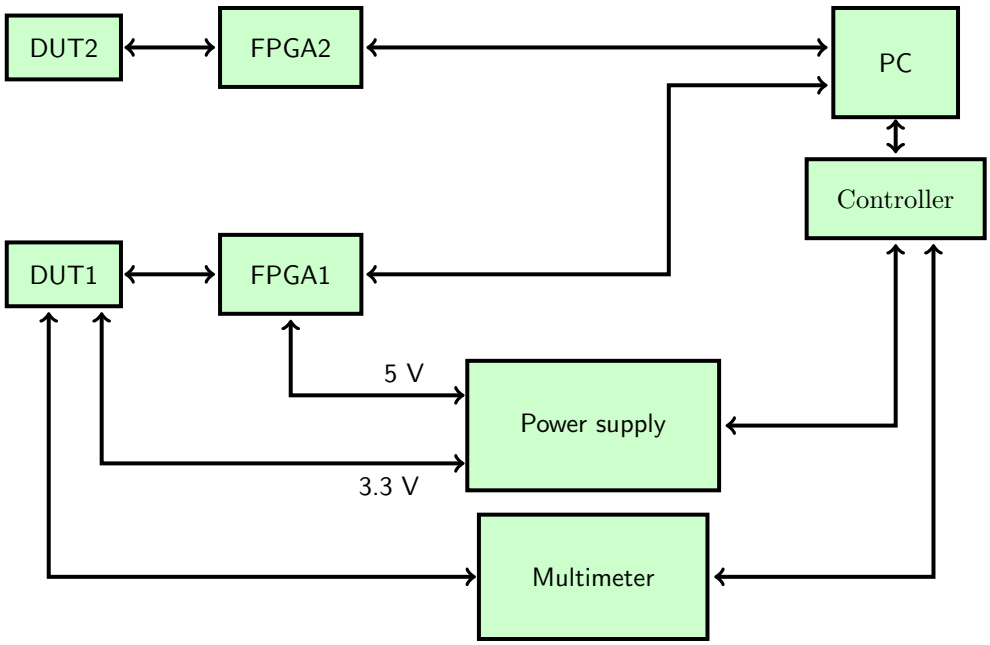


Figure 4.8: TID test setup.

First Results

As for calibration, all the errors have been detected in programmed cells, thus bit flipping from "0" to "1". The mean value of detected errors for the two dose levels (Tab. 4.5) were consistent with the values predicted, based on calibration.

	20 krad(Si)	**33 krad(Si)**
Lot A	158.95	46170.95
Lot B	497.9	141276.35
Total	323.425	93723.65

Table 4.5: Comparison between the mean of errors detected at 20 krad(Si) and 33 krad(Si) for each lot (A and B) and for all the samples without distinction of the lot (tot).

Fig. 4.9 shows the first evident result: Lot B has and average errors value higher than Lot A, both before and after the radiation exposure. Thus Lot B should be supposed to be weaker than the other lot. The graphic also provides the box plots for each lots and total dose level. This is a useful way to compare the two different groups, displaying statistical differences without making any assumption of the underlying statistical distribution. The variability (represented by lines outside the upper and the lower quartile) of the Lot B is related with the dose: the higher the dose, the higher the spread around the median value. On the other hand, Lot A did not show any dependence on the dose. The whisker below the mean value in Lot A before irradiation (0 krad(Si)) is due to no errors detected in 2 of the 20 samples. Furthermore, no outliers are observed in Lot A at 33 krad(Si), while Lot B has one or more upper outliers in each measurements.

From these few observations is already possible to extract interesting informations on the errors distribution. First of all, since the median (the band inside the box) is not placed in the middle of the box, especially in the post-radiation measurement, it is possible to foresee that normal distribution would not fit well the data. In addition, outliers display the presence of some devices which a high enough number of errors to be distant from other observations. However, those extremely high results must be taken into account, since they do not come from measurement errors, but they indicate a long-tailed errors distribution.

Fig. 4.10 represents the number of errors normalized on the errors detected during the pre-rad read operation[4]. Lot A has a higher mean value respect to Lot B. The reason is the small number of pre-rad errors, and not a stronger radiation-induced influence on this lot. As in Fig. 4.9, devices belonging to Lot B show a higher variability, independently on the pre-rad errors detected. On the contrary, in this chart Lot B has no outlier, while Lot A has. From this result one could suppose that the high number of errors detected in some of the devices belonging to Lot B is due to an intrinsic error tendency, being present already before the TID exposure.

In fact, all the results must be taken into account considering that, at these dose values, the energy deposition variability is negligible and the variation across the population is caused by parameter variability from sample to sample. Actually it must be considered that, to avoid control circuitry failures, the total dose was low enough to allow to detects errors only of the cells belonging to the lowest part of the program V_T distribution. Considering all the devices irradiated, excluding from each memory 30 blocks erased/programmed/read during each check operation, more than 10^{12} cells were exposed to TID and the floating gate errors

[4]Two samples belonging to the A lot, with zero pre-rad errors, have not been considered.

detected after the exposure were less than 1% of the amount of FG in the array. This relatively small amount of errors correspond to the integral of the left tail of the programmed V_T distribution, which cross the read voltage. TID effect on the devices is the V_T distribution shift towards the neutral threshold voltage. Thus as dose increases, the integral value of the left tail will increase.

The relevant difference between the two lots is ascribed to different threshold voltage distributions. Probably Lot B has a wider V_T distribution than Lot A. Furthermore radiation induced effects increased the width of V_T distribution in Lot B, as it is visible both in Fig. 4.9 and 4.10, where the variability, represented by the whisker, increases with total dose.

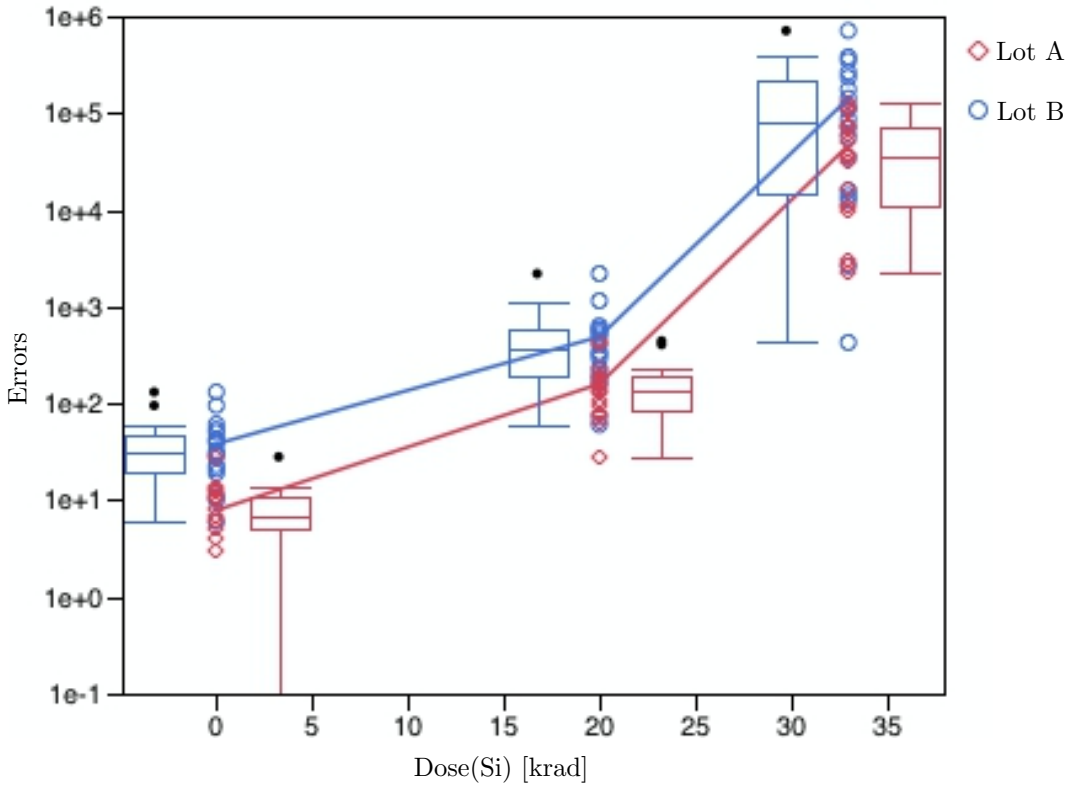


Figure 4.9: Errors detected before and after the radiation exposure, at 20 krad(Si) and 33 krad(Si).

4.3 Conclusions

Details of the experiment setup has been described, to have a clear vision of the experimental conditions of the test. The calibration demonstrated that in 25-nm devices errors occur at lower total dose than in older devices. From the fist evident results of the irradiated 40 Flash memories, just looking at the mean value of retention errors detected in each device, differences between lots

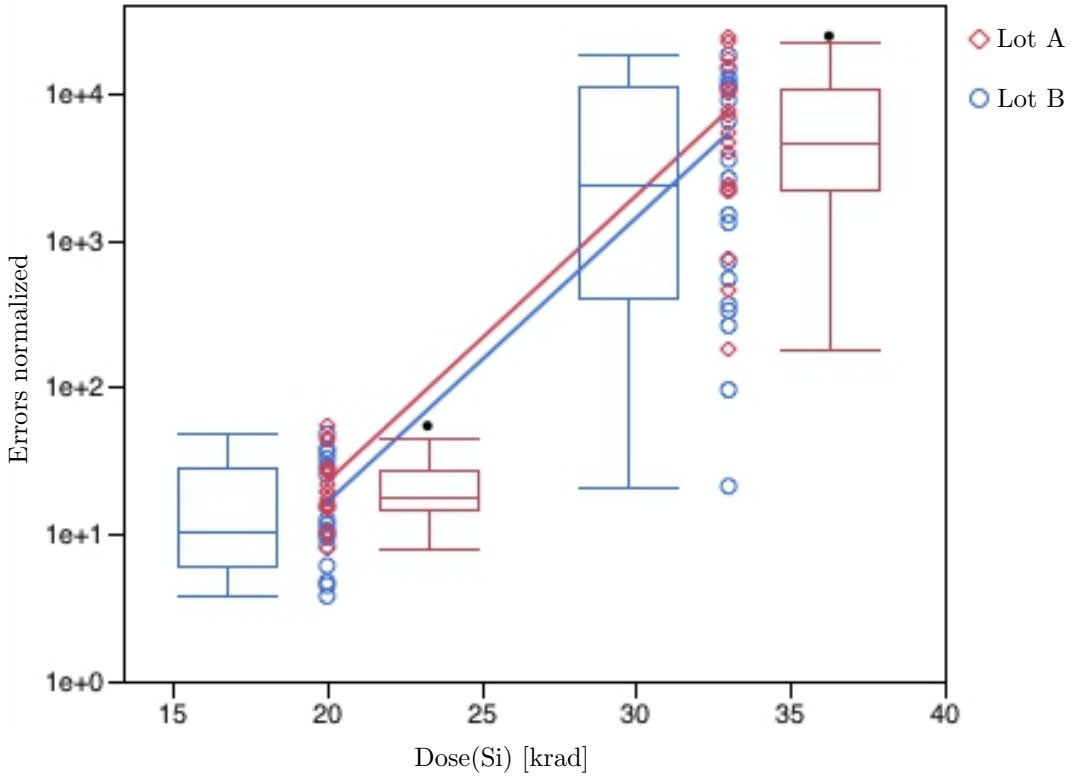


Figure 4.10: Errors normalized on the number of errors detected before irradiation, as a function of total dose.

are unequivocal. In the following analysis, sources of lot-to-lot and cell-to-cell variability will be investigated, shooting for explain the experimental results.

Chapter 5
Variability Analysis

From previous observations, the variability in the detected errors are ascribed to chip-to-chip variability, since at these dose level and for this radiation source, the variability of energy deposition among samples is negligible.

First of all, a preliminary consideration came from noticing the small amount of cells identified with errors, respect to the large number of irradiated ones. Fig. 5.1 represents the errors percentage versus total dose. Even at the maximum dose level, i.e. 33 krad(Si), the maximum percentage of detected FG errors is less than 1 % of more than 1 Terabit of irradiated cells. Therefore, the analysis of the results deals with the left tail of the threshold voltage distribution, since FG errors occur only in programmed cells.

Variability sources will be investigated. In addition, the sharp difference between lots is a further issue, for which possible explanations will be examined. Analysis will start from the lot-to-lot variability, aiming to find correlation between pre-rad and post-irradiation results. Then the variability among samples will be investigated.

5.1 Threshold Voltage Spread

Statistical sources of V_T spread affect the array operation during programming, erasing and data retention.

There are three main variability sources, following listed:

- cell-to-cell parameters variability, due to process control issues;
- the discreteness of matter and charge, that means reduction of the total amount of dopants in the cell active area and of the electrons transferred to/from the floating gate, because of the small size of the devices;
- statistical distribution of defects in the cell's tunnel oxide.

From experimental results of the radiation test, bit errors occur because of the program threshold voltage spread during programming and data retention,

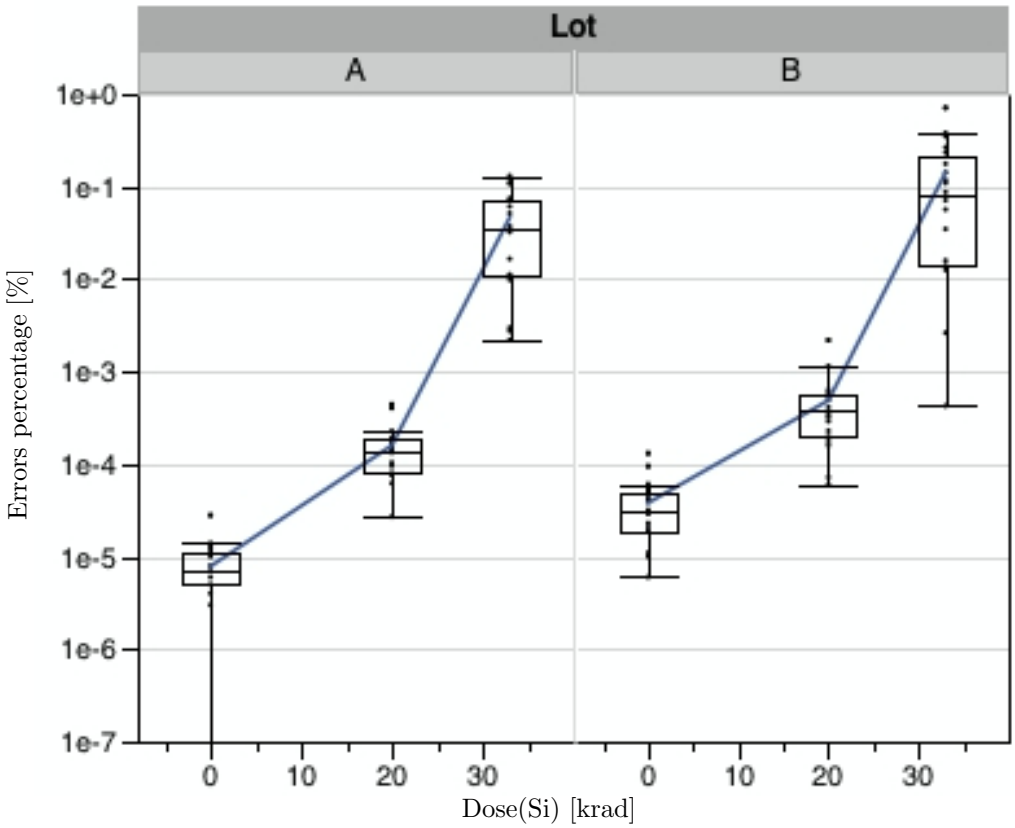


Figure 5.1: Errors percentage versus total dose, for each lot.

which brings the some cells' V_T above the V_{READ}, since no bit flipping from "1" to "0" occurred.

However, first of all, it must be considered that also the neutral threshold voltage distribution is supposed to be wider than previous technology nodes. As predicted in Fig. 5.2, the spread of the V_T distribution increases as scaling proceeds. The considered variability sources in this model [4] are: fluctuation in cell W and L, Random Dopant Fluctuations (RDFs) in the substrate, Oxide Traps Fluctuations (OTFs), referring to Si/SiO_2 interface traps, variability of α_G and variability of the Oxide Recess Height (EFH).

It is possible to estimate how some of these sources of statistical variability affects the devices under study.

Starting from the geometric parameters, Fig. 5.3 and 5.4 represent a rough evaluation of the threshold voltage shift induced by variation of cell W and L respectively, from the nominal value (25 nm), applying the equation $\Delta V_T = N \cdot q/C_{PP}$.

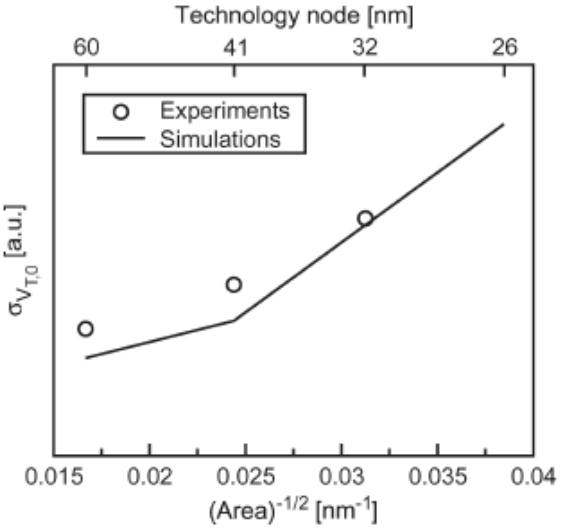


Figure 5.2: Spread of NAND Flash cell V_T as a function of the technology node [4].

C_{PP} can be calculate according to the following:

$$C_{PP} = \frac{\epsilon_{\text{ox}} \cdot A}{t_{ONO}}. \tag{5.1}$$

where A is the FG area covered by the control gate, given by:

$$A = \Big[W - 2t_{FG}\cot\theta + 2\big(\frac{t_{FG} - EFH}{\sin\theta}\big) + 2t_{ONO}\Big]L, \tag{5.2}$$

assuming a trapezoidal shape for the floating gate [4]. For 25-nm technology, $W = L = F$, with F being the technological node, the FG height $t_{FG} \simeq$45 nm, the oxide thickness (t_{ox}) is about 8 nm and the equivalent ONO layer thickness (t_{ONO}) is 14.5 nm [2].

Evaluation of C_{PP} is useful also for analyze its impact of the variability of α_G on the threshold voltage spread. The capacitive coupling ratio between control gate and floating gate α_G is one of the most important parameters for Flash memories, because it determines the control exerted on the floating-gate potential by the control gate, which allows the external access to the memory cell. A fluctuation of α_G results into a spread of V_T. Contributions to fluctuation of α_G come from W, L and t_{ox} variability, impacting the floating-gate to substrate capacitance, and from fluctuation of C_{PP}.

The mechanisms that cause charge loss from the floating gate (photoemission, charge recombination and positive charge trapping in the oxides) strongly depend

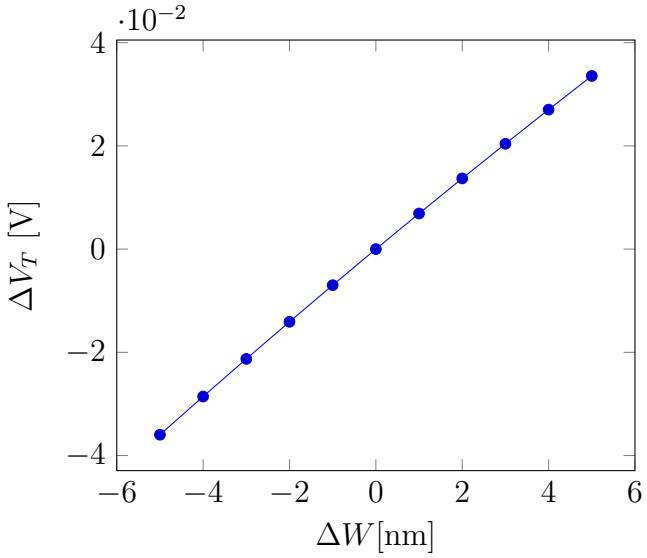


Figure 5.3: Rough evaluation of the threshold voltage displacement as a function of the displacement of cell W from the nominal value of the 25-nm technology.

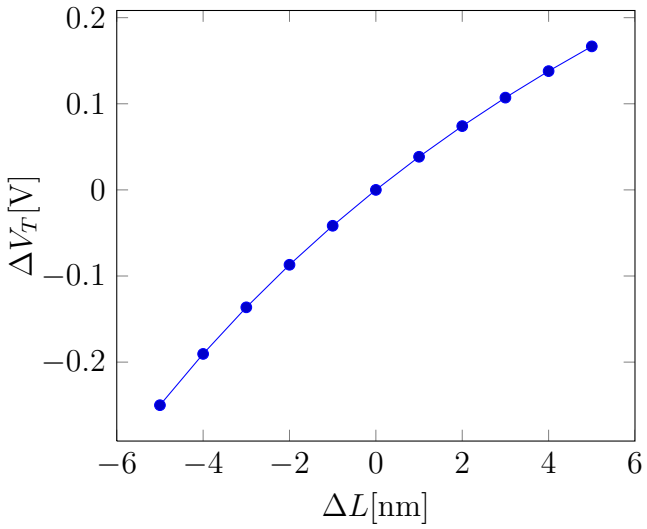


Figure 5.4: The same as Fig. 5.3, but changing L.

on the electric field in the oxide, which is in turn related to the threshold voltage:

$$E_{\text{ox}} = \frac{\alpha_G}{t_{\text{ox}}}(V_T - V_{T,0} + V_{FB}). \tag{5.3}$$

where $V_{T,0}$ in the neutral threshold voltage, V_{FB} is the flat band voltage, t_{ox} the tunnel oxide thickness and α_G the coupling coefficient between control gate and floating gate.

Since the threshold voltage of a programmed cell is:

$$V_T = V_{T,0} - \frac{Q}{C_{PP}} = V_{T,0} + \frac{N \cdot q}{C_{PP}}, \tag{5.4}$$

the threshold voltage shift depends on the number of stored electrons, N, and on the coupling capacitance C_{PP} between control gate and floating gate.

With the previous geometrical parameters:

$$C_{PP} = \frac{\epsilon_{ox} \cdot A}{t_{ONO}} = 20\,\text{aF}. \tag{5.5}$$

For the technological node under study, the number of electron per bit N is relatively small. It can be calculated as the number of electron necessary to induce a threshold voltage shift of 1 V:

$$Q = -C_{PP}\Delta V_T = -20\,\text{aF} \cdot 1\,\text{V} = -20\,\text{aC} = \sim 125e^{-}. \tag{5.6}$$

Thus charge loss of just about one hundred of electrons causes a FG error, in other words, the cell reverses its information from “0” to “1”. Such a small number of electrons per bit leads to consider the charging and discharging of the floating gate not a continuos phenomenon, but a sum of discrete stochastic events. Fig. 5.5 and 5.6 represent a prediction of C_{PP} and N as the feature size F of the FG cell scales down. The variability induced by small amount of electrons per bit, which are injected/removed during programming/erasing, becomes relevant for reliability issues.

The discrete nature of the electron flow can explain, first of all, the difference between lots of pre-rad errors, after programming, since the small amount of electrons injected during programming leads to an uncertainty of the number of stored charge after writing.

5.1.1 Lot-to-lot variability

As introduced in the FG cell overview, the program operation is performed by the ISPP algorithm, which applies short programming pulses (of V_{step} amplitude) to the control gate. Each program pulse is followed by read operation (verify

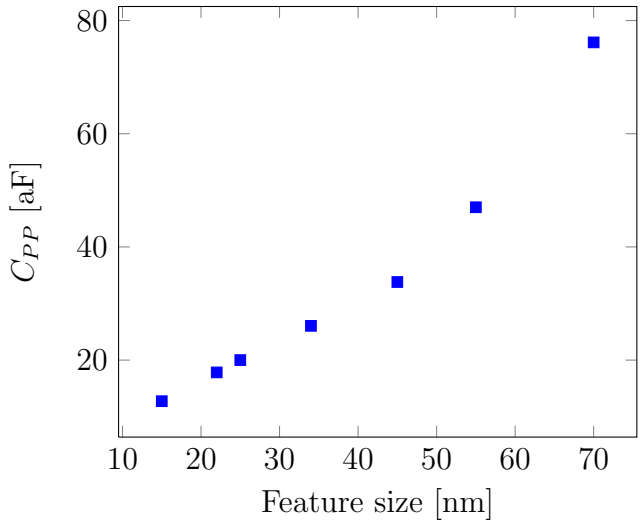


Figure 5.5: Prediction of capacitive coupling C_{PP} as a function of technology scaling.

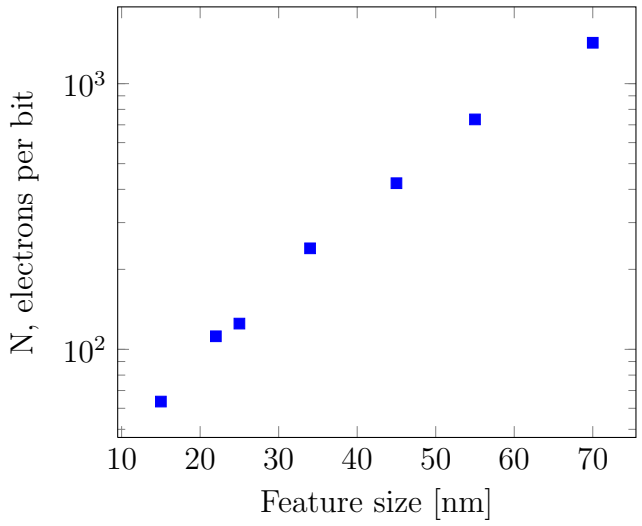


Figure 5.6: Prediction of electrons per bit (N) as a function of technology scaling.

phase), to measure V_T and compare it to the program verify level (PV). For each step, ΔV_T equals ΔV_{CG}. The efficiency of the ISPP algorithm, which would theoretically allow to obtain tight threshold voltage distribution, is compromise by the electron injection statistics (EIS). The final programmed threshold voltage, obtained with the program algorithm, results displaced from the PV level more than the constant increase V_{step} given to the control-gate pulses. This effect increases as the number of electrons transferred into the FG at each program pulse decreases, which is determined by the control-gate to floating-gate capacitance C_{PP}. In fact in every programming pulse, a fixed amount of charge equal to $-V_{\text{step}} \cdot C_{PP}$ is injected. The average number $\overline{n}$ of electrons is injected into the FG is given by:

$$\overline{n} = \frac{JA}{q} t_{\text{step}} \tag{5.7}$$

Assuming V_{step} of about 200 mV, at each program pulse a package of 25 electrons is injected.

The relation between V_T spread resulting from EIS is given by:

$$\sigma_{\Delta V_T} = \frac{q}{C_{PP}} \sqrt{\overline{n}} = 40\,\text{mV} \tag{5.8}$$

assuming the number of injected electrons, n, ruled by Poisson statistic (equation 2.23).

Furthermore, in addition to EIS, contribution in programming spread comes from small variations of t_{ox}, which directly impacts the program field in the Fowler-Nordheim equation. Because of the exponential dependence of the tunnel current on the oxide electric field, just a small variation of the tunnel oxide thickness among the cells of the array could lead to great difference in programming currents, spreading the program threshold voltage distribution.

In Fig. 5.7 the errors distribution before the radiation exposure, thus some days after the program operation, shows that devices belonging to Lot B are clearly affected by higher errors values, up to a factor of 4, since more cells for each device are below the read voltage level.

The graph in Fig. 5.8, plotting the ratio between the two lots of the average number of errors detected in each chip, shows that total dose does not increase the lot-to-lot difference. Instead, the pre-rad ratio is higher than after the irradiation, and the value is quiet constant from 20 krad(Si) to 33 krad(Si).

In conclusion, lot-to-lot variability may be ascribe to two main causes:

- stronger effect of electron injection spread (EIS) in Lot B;

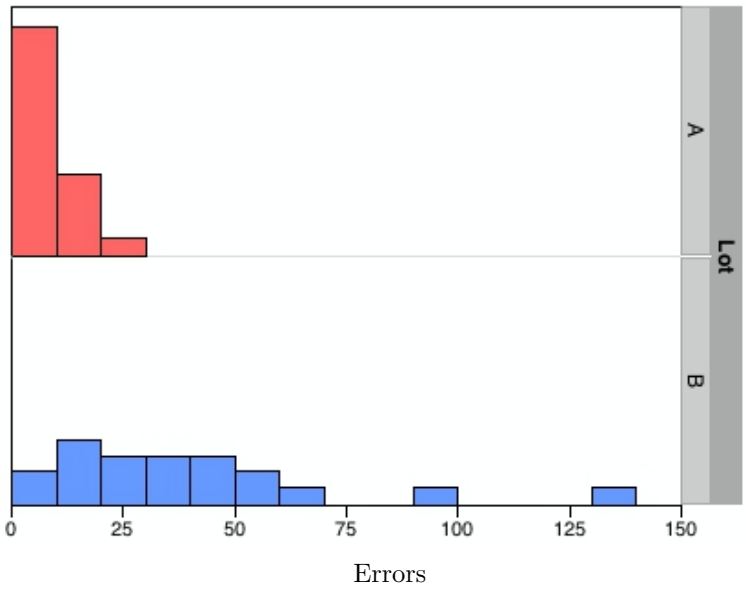


Figure 5.7: Distribution of errors detected in each device, before irradiation.

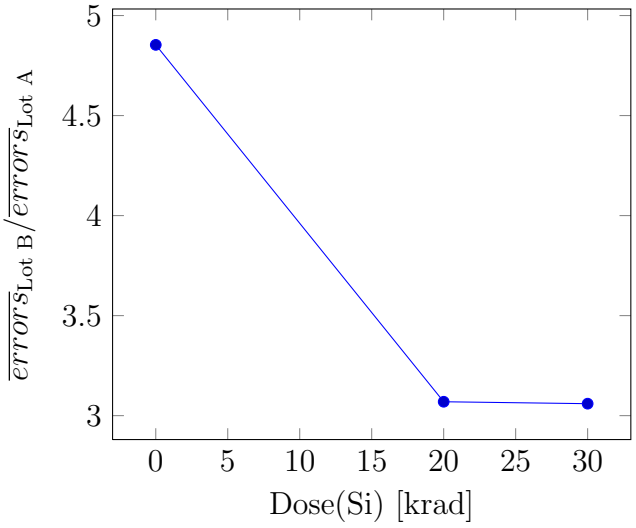


Figure 5.8: Ratio of the mean errors value per chip between Lot B and A, versus total dose.

- small variation of t_{ox} among arrays belonging to Lot B, probably related with process control issues.

The result is that Lot B has a more stretched program threshold voltage distribution. Furthermore TID effects make the threshold voltage distribution shift, explaining the larger amount of errors in Lot B both before and after irradiation.

5.1.2 Chip-to-chip variability

Lot B, in addition to a larger mean value of errors in each situation (pre-rad, at 20 krad(Si) and 33 krad(Si)), shows larger chip-to-chip variability. This means that the left tail of the programmed threshold voltage distribution, which crosses the read voltage level, not only has a wider area than the Lot A one, but stretches over more left values.

As introduced in chapter 2, several sources of statistical variability in floating gate cells have been identified. Variability sources of primary importance are: Random Discrete Dopants (RDD), Line Edge Roughness of Gate (LGR) and STI (LSR), Oxide Thickness Fluctuation (OTF) and Interface Trapped Charge (ITC). In this section will be investigate which among these variability sources has more impact on the 25-nm floating gate cells under study.

Random Discrete Dopant

The discreetness of substrate doping results into a V_T spread, calculated for MOSFET as [4]:

$$\sigma_{RDD} = 3.19 \times 10^{-8} \frac{t_{ox} N_A^{0.4}}{\sqrt{WL}}. \tag{5.9}$$

The V_T spread increases as feature size of devices, thus W and L, decreases, since t_{ox} can not be scaled as quickly as L, due to retention data constraint.

The discrete nature of dopants has also be invoked as one of the major causes of Random Telegraph Signal (RTS) instabilities. It is one of the most conditioning cell-to-cell variability source. The phenomenon consists in the variation of channel carrier density and mobility, due to trap-detrap phenomenon near Si/SiO_2 interface.

As for EIS, this problem arises from the granularity of the matter, because of the small size of the device, which becomes comparable to the distance among dopants. The amount or position of the dopants determines changes in V_T, affecting ΔV_T statistic. The total channel current is the sum of each current path, forced to flow in few narrow channels between negatively charged boron potential spikes, from source to drain [20]. If an electron is trapped near one of these current path, it can completely turn it off, causing large I_D or V_T variation. Fluctuations

between two threshold voltage or current levels, due to trapping/detrapping of a single electron in a gate oxide trap, may give rise to erroneous data reads or erroneous program operations, since the program algorithm includes verify steps.

From equation 2.24, threshold voltage variation ΔV_T induced by a single electron captured or released in 25-nm Flash cell can be estimate as:

$$\Delta V_T^{st} = \frac{q}{C_{OX} W L \alpha_G} = 85\,\mathrm{mV} \tag{5.10}$$

considering α_G of about 0.7. Actually this value may be grater, because of percolation effect [20]. As in charge/discharge of the floating gate, the Poisson distribution governs the number of dopants.

Interface Trapped Charge

The second most important variability source for devices under investigation is Interface Trapped Charge (ITC). As presented in chapter 2, defects are inevitably present in oxide bulk and interfaces. In particular the transition region between SiO_2 layer (amorphous) and the Si lattice (mono-crystalline) is characterized by *dangling bonds*, i.e. non-saturated bonds, which may be passivated with hydrogen. The oxide degradation is very sensitive to the fabrication process and device parameters, and the magnitude of the effects strongly depends on the device fabrication conditions. In particular, the growth temperature and pressure are the most important parameters that thermodynamically influence the quality of the insulator [34]. Hydrogen related defects, such as Si-H or Si-OH groups, are common extrinsic oxide defects. During TID exposure, ionizing radiation can break these bonds, releasing H atoms or H^+ ions, which easily move in the SiO_2 lattice structure, interacting with other dangling bonds, as explained in chapter 3. Assuming a Poissonian fluctuation of the trapped charge, its contribution in the resulting spread in cell V_T is given by:

$$\sigma_{ITC} = K_{\mathrm{ox}} \cdot t_{\mathrm{ox}} \cdot \frac{\sqrt{Q_{\mathrm{ox}}}}{\sqrt{WL}} \tag{5.11}$$

where Q_{ox} is the surface density of traps. The resulting voltage spread increases as the factor $\sqrt{WL}$ decreases.

This type of source of variability is destined to be more and more affecting as the lifetime increases, even more as TID increases. Program/erase cycles, performed by Fowler-Noordheim tunneling, induce oxide degradation (Fig. 2.18), because of the high electric field, causing drift in the threshold voltage, according to equation 3.5.

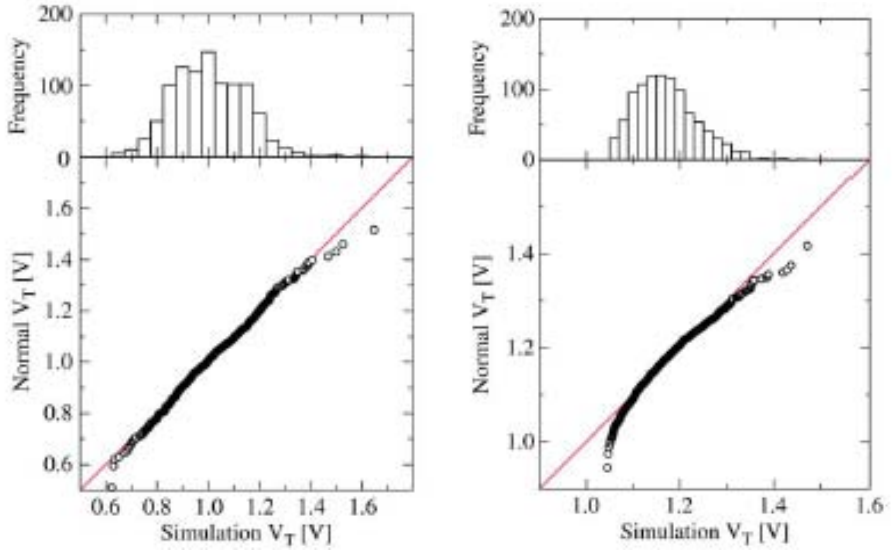


Figure 5.9: QQ plot and histogram of the distribution of V_T from 1000 microscopically different devices with applied RDD (a) and ITC (b) [5].

Histograms of the V_T distribution of 1000 32-nm devices due to RDD and ITC are shown in Fig. 5.9. Each histogram is associated with the quantile-quantile (QQ) plot. The QQ plot is used to compare the similarity of two distribution: the measured data are plotted versus the normal distribution. If the measured errors distributions is an ideal normal distribution, the data point will follow the red line. Any deviation from this line points out the departure from the normal distribution. The *skew* is an indication of the asymmetry of the distribution. If the skew is a negative value, it means that the left tail of the distribution is longer, that is the mean has a higher value than a normal distribution; on the other hand a positive skew is a measure of a lower mean value from a normal distribution. The *kurtosis* indicates how peaked the distribution is. A large positive values of kurtosis correspond to a sharply peaked distribution, with the majority of the errors close to the mean value, with fewer outliers; whereas large negative values mean that the distribution is flattened.

Both histograms associated to RDD and ITC show a skewed distribution, because of the Poisson statistic that governs the distribution of dopant and trapped charge respectively. The longest right tail indicates that the threshold voltage shift is small that that associated to a normal distribution.

It is not straightforward to connect these observations on the RDD and ITC effects on chip-to-chip variability with the TID test results, since we are able just to analyze User Mode (UM) results, in other words the digital output, and

the threshold voltage distribution is unknown. However it is possible to extract informations from the errors distribution.

Probability distribution of the floating gate errors detected in each memory after a total dose of 33 krad is plotted in Fig. 5.10, distinguishing the lots. Tab. 5.1 resumes mean, variance, skew and kurtosis values, comparing the lots.

	μ	s^2	**Skew**	**Kurtosis**
Lot A	46170,95	$1.501\,02 \times 10^9$	0,7539839	-0,312616
Lot B	141277,45	$3.115\,47 \times 10^{10}$	2,0409944	4,8093699
total	46993,625	$1.121\,03 \times 10^{10}$	4,0767298	20,572795

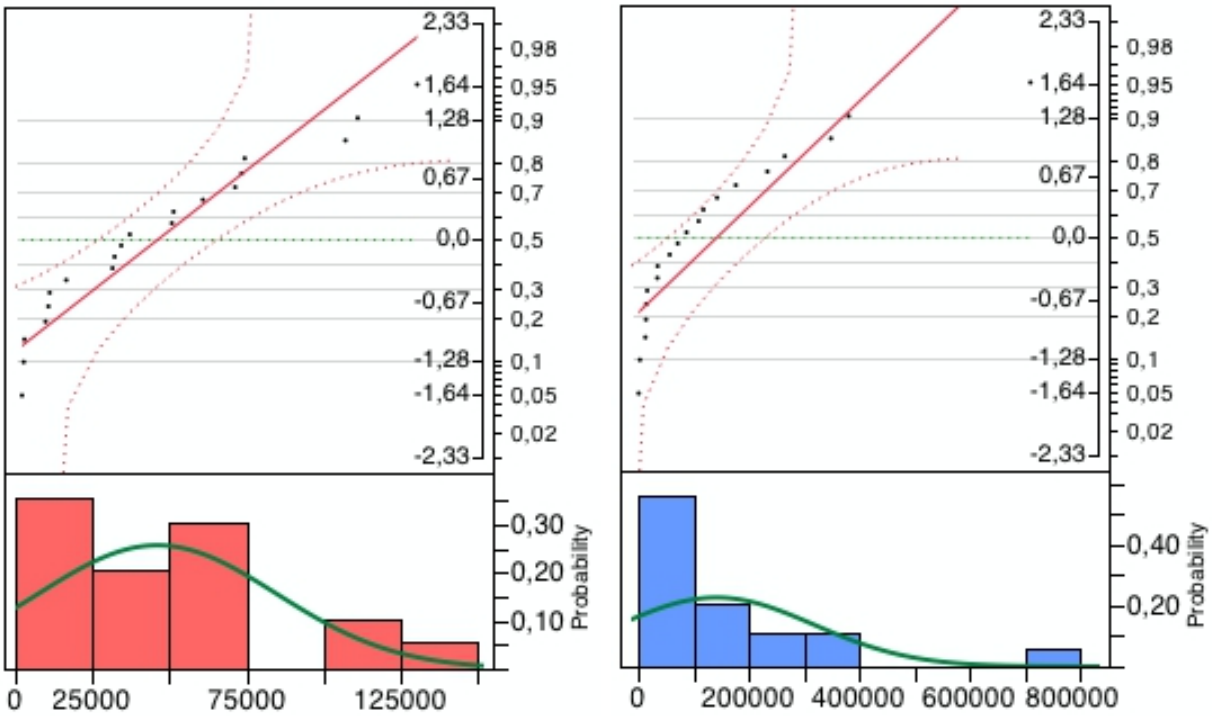
Table 5.1: Comparison between mean (μ), variance (s^2), skew and kurtosis of the errors distribution for each lot (A and B) and for all the samples without distinction of the lot (total), after a total dose of 33 krad(Si).

For both lots, experimental errors distribution is not very close to normal distribution, but it seems to fit quite well with Poisson distribution, specially in Lot B. Large positive skew, especially for Lot B, shows a strong departure from the normal distribution. Actually the right tail is longer than the left tail. While Lot B distribution is more peaked, Lot A distribution is quite flat.

Fig. 5.11 represents the errors distributions pre-rad and after both the dose steps, compared to Poisson distribution. Irradiation does not seem to influence the probability distribution of FG errors in both lots. Probability distribution is in good agreement with Poisson statistic, notably at the end of the radiation exposure.

Some considerations on the two main variability source come from these results. The large threshold voltage spread induced by RDD is consistent with the large variability of errors detected for each memory. In other words, cells whose experience large number of floating gate errors, are subjected to a larger threshold voltage spread, ascribed to RDD-induced statistical effects. Now considering ITC, since hydrogen-related defects in oxides are unavoidable, this statistical variability source may be related with the higher variability of floating gate errors in devices belonging to Lot B. Different oxide quality, because of small variations of the process conditions, may be an explanation of such clear different behavior between Lot A and Lot B.

Reminding that each post-rad error corresponds to a cell, whose threshold voltage shifted left to the V_{READ} voltage level, and that histograms in Fig. 5.11 represent the probability distribution of errors across devices, it is possible to imagine an hypothetical average threshold voltage distribution for each lot, con-



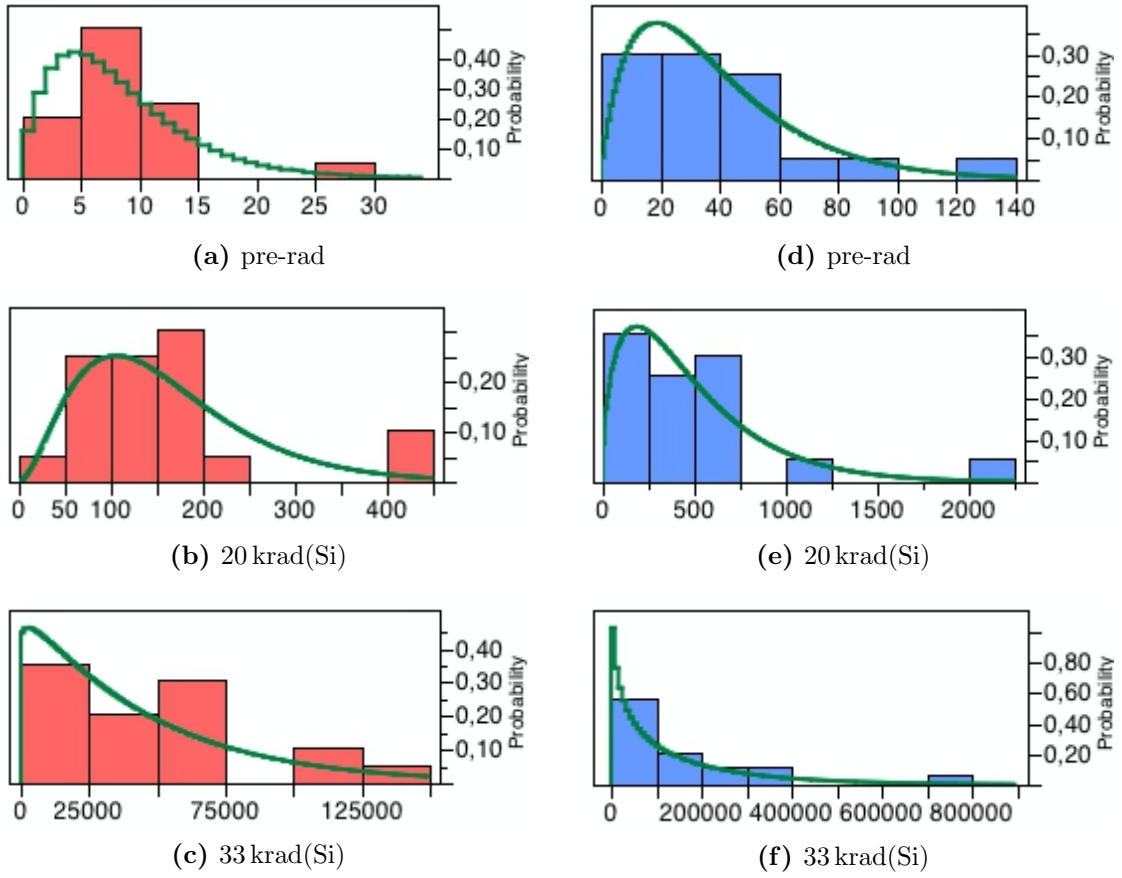
(a) Lot A **(b)** Lot B

Figure 5.10: Probability distribution of retention errors detected for each Flash memory

sidering differences previous discussed between lots. Fig. 5.12 is a qualitative representation of the programmed threshold voltage, comparing lots A and B. After the program operation (blue line), threshold voltage distribution for both lots is narrower than erase distribution, thanks to the ISPP algorithm. However, Lot B could be characterized by an average pre-rad threshold voltage distribution wider and flatter that Lot A (the total integral area must be theoretically the same for both lots, because the amount of irradiated cells is the same), because of EIS during ISPP algorithm and the oxide thickness fluctuation among cells belonging to the same array.

Furthermore, total dose acts on the threshold voltage distribution (red arrow), making it shift towards the neutral voltage. After the radiation exposure (black line), the radiation-induced shift makes the V_T distribution to cross the V_{READ} level, below which cells are read as erased ("1"). The number of bit errors corresponds to the integral of the colored area. As total dose raises, the area below the V_{READ} voltage level increases, since more and more cells lose the negative charge stored in the floating gate.

In addition, phenomena such as Random Telegraph Signal, mainly caused by the discrete nature of dopants in devices so scaled, and ITC, which affects



(a) pre-rad **(d)** pre-rad

(b) 20 krad(Si) **(e)** 20 krad(Si)

(c) 33 krad(Si) **(f)** 33 krad(Si)

Figure 5.11: Distributions of FG errors in Lot A (a, b and c), and Lot B (d, e and f) samples, before irradiation, after 20 krad(Si), and after 33 krad(Si). Experimental data are compared with Poisson distribution.

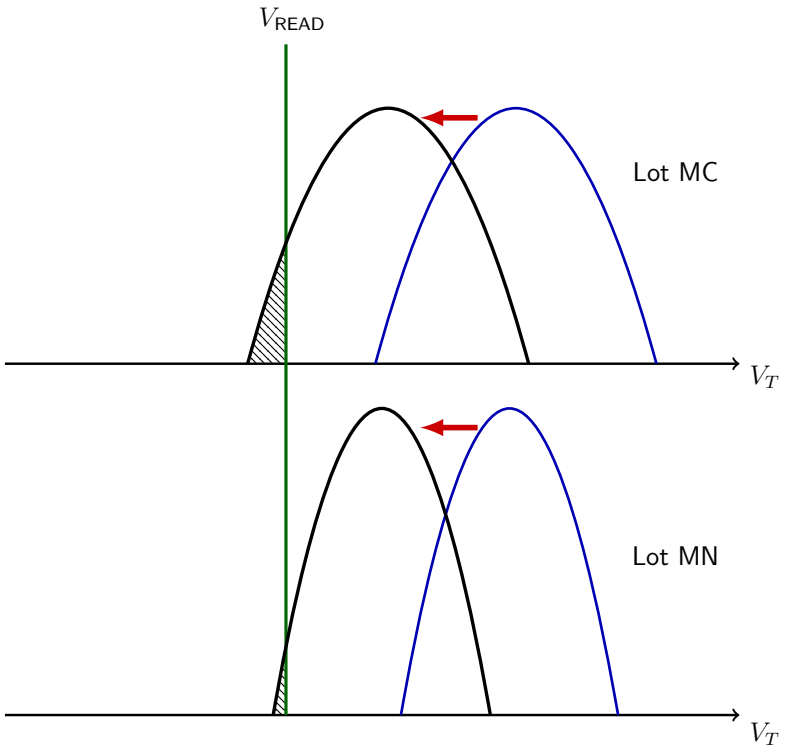


Figure 5.12: Sketch of the TID effects on the programmed threshold voltage.

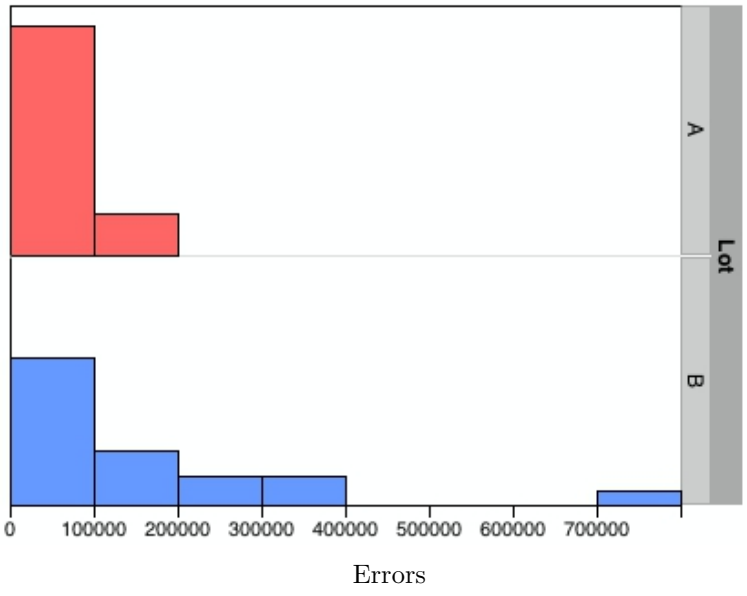


Figure 5.13: Errors distribution after a total dose of 33 krad(Si).

probably more Lot B, contribute to a further spread of the threshold voltage distribution. Adopting an uniform scaling of the x axis, histograms of the errors distribution in Fig. 5.13 show a clear higher spread for Lot B.

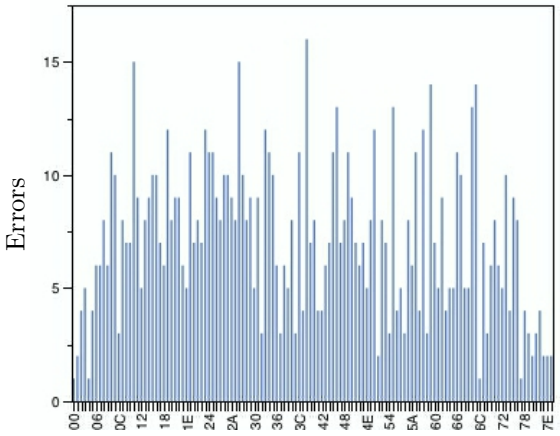
5.2 Errors Spatial Distribution into Pages

It is interesting to notice the errors spatial distribution into pages (Appendix B) at the end of the irradiation exposure. Plotting the number of errors in each page, it has been observed that errors are clearly more in odd pages than in even ones. The shape of the distribution of the FG errors in each page is a bell-shape, which means that errors are more concentrate on central pages. This characteristic is the same for both lots. Another particular behavior deduced by these graphics is the high errors concentration on the last pages, i.e. 0×7D, 0×7E and 0×7F, especially on this last one. At firs sight, this peculiar disposition could be attributed to control circuitry issue, related with different distances between pages and read pump. This hypothesis is reinforced by the fact that the same distribution is observed also for the blocks (Fig. 5.15): more errors appear in odd blocks than in even ones, but in this case this peculiar disposition is visible only on devices with a high amount of errors.

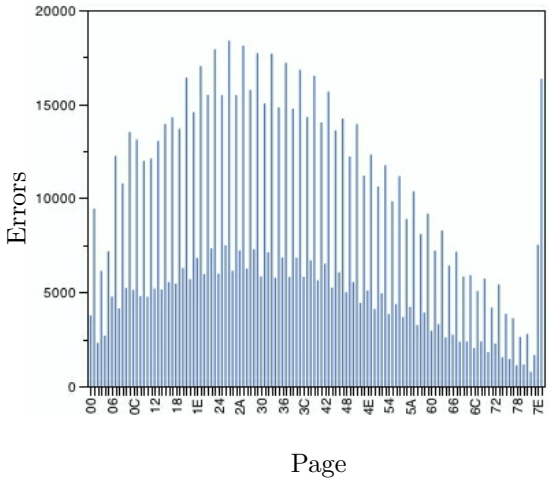
To investigate if even and odd blocks show different TID responses, calibration results have been taken into account, since, during the calibration, errors are sampled about every 300 rad(Si). Fig. 5.16 shows the error build up during calibration, distinguishing even and odd blocks. Retention errors appear first in odd blocks, then, after about 2 krad(Si), in even blocks. Furthermore, the total amount of retention errors is larger in odd blocks, confirming the post-rad distribution observed.

Finally, retention errors versus total dose in Fig. 5.17 shows that more errors appear in odd blocks only after the radiation exposure. The pre-rad error distribution does not follow this peculiar post-rad partition into even and odd blocks. Fig. 5.14 represents the errors spatial distribution into pages, including all the devices, before the TID test (a), after 20 krad(Si) (b) and 33 krad(Si) (c). The bell-shaped distribution is almost the same in both the total dose values, on the contrary pre-rad distribution does not seem to be ruled by any particular distribution.

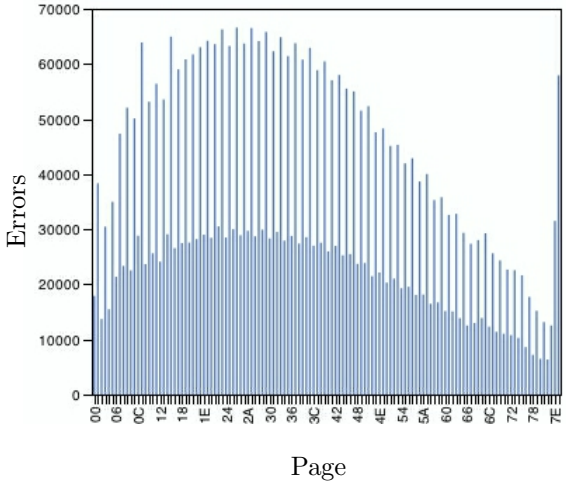
To verify if this peculiar disposition of errors in pages and blocks is caused by radiation-induced effects, 4 fresh memories, two from each lot, have been programmed as the irradiated ones. After one week (the same time gone by pre-rad programming and post-radiation reading) devices have been read to check retention errors. Fig. 5.18 shows the result. Errors are homogeneously distributed



(a) pre-rad



(b) 20 krad(Si)



(c) 33 krad(Si)

Figure 5.14: Errors spatial distribution into pages, considering 40 Flash memories, pre-rad (a) and at two dose level (b and c).

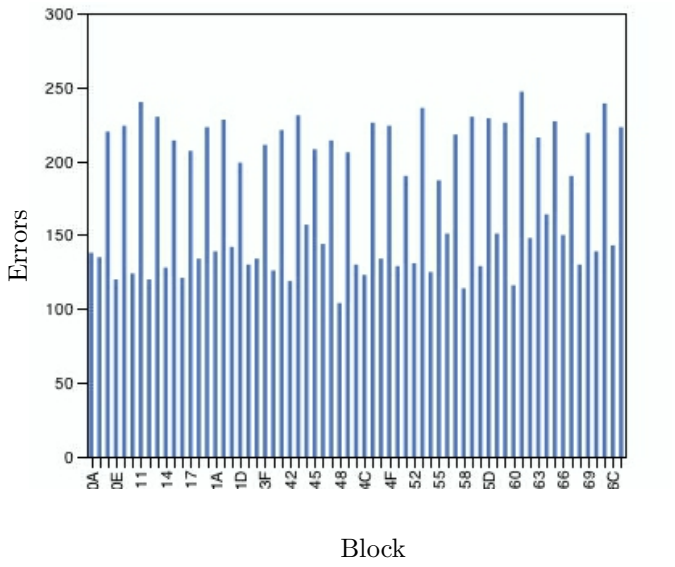


Figure 5.15: Errors spatial distribution into about 180 (0×A to 0×BC) blocks in device B1, after 33 krad(Si).

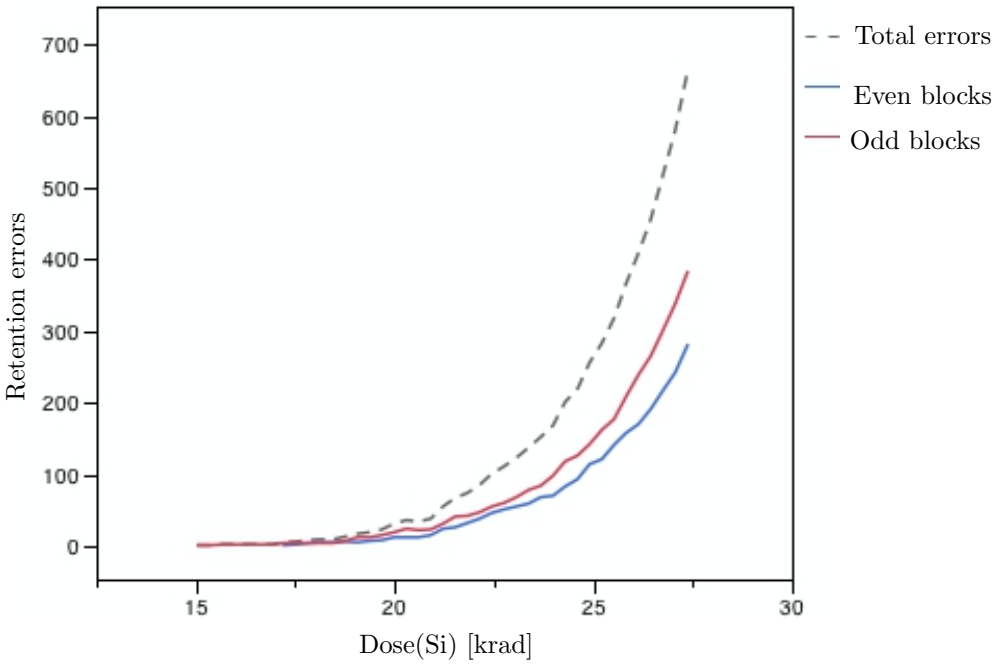


Figure 5.16: Errors build up during calibration, distinguishing even and odd blocks.

in even and odd pages. The same has been observed concerning even and odd blocks. As for previous tests, the devices belonging to Lot B have more errors. Therefore, this additional check confirms that the bell-shape distribution into pages and the difference in even and odd pages and blocks are radiation-induced effects.

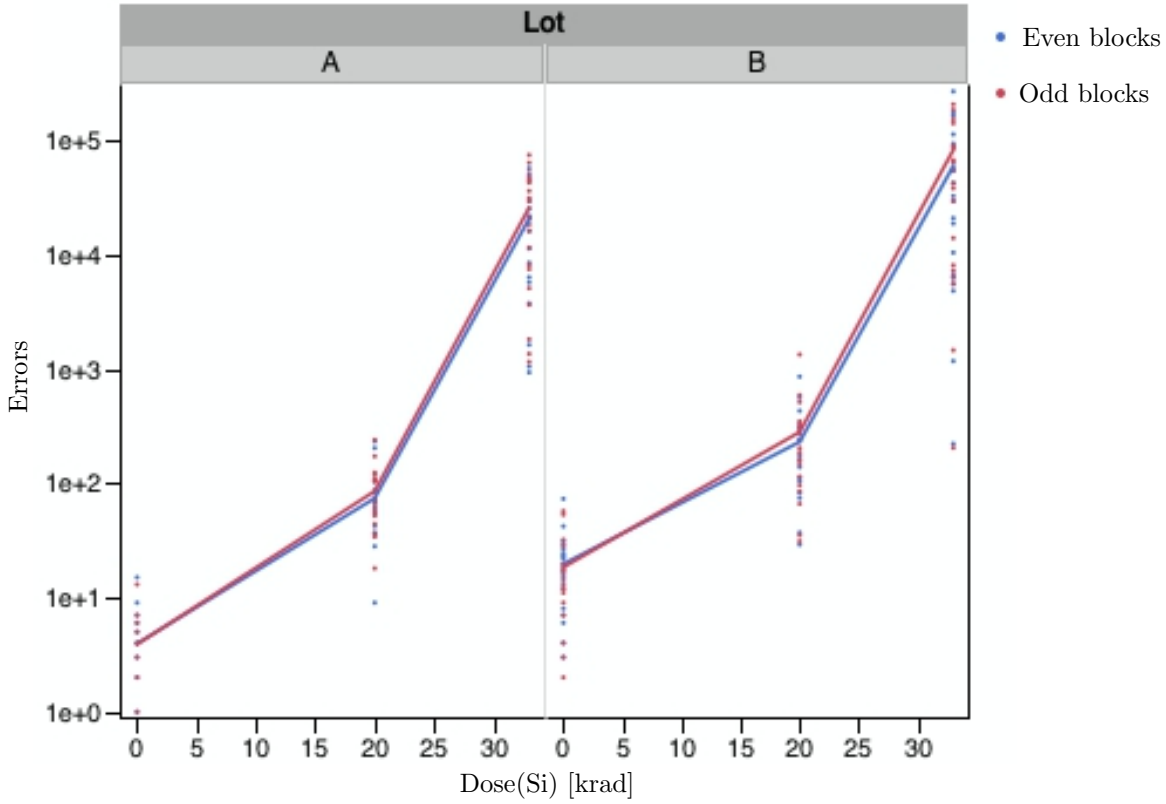
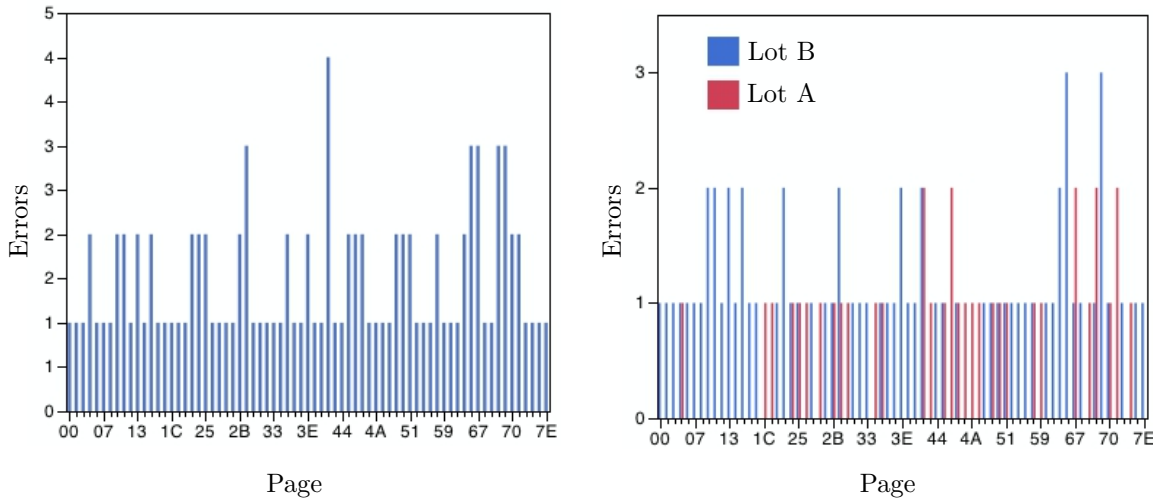


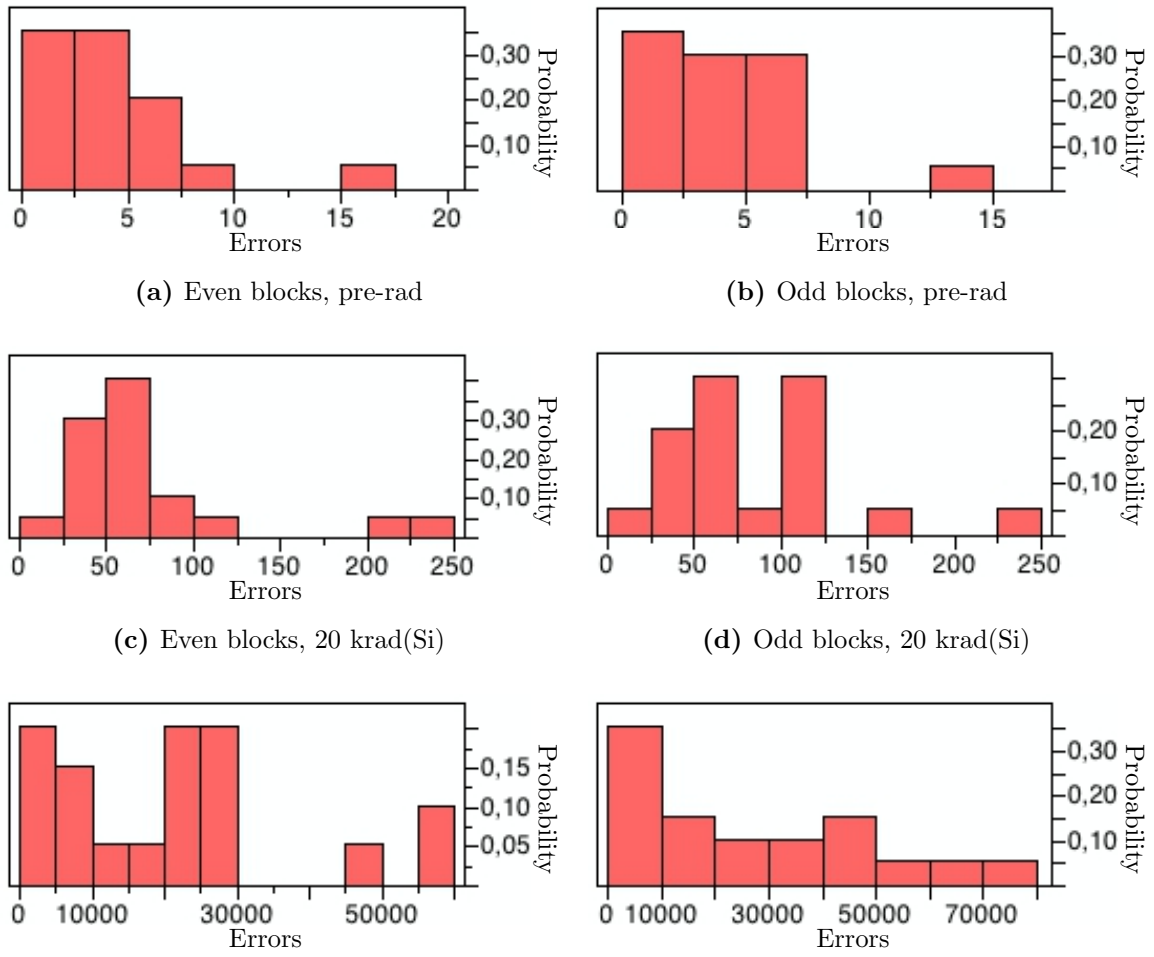
Figure 5.17: Errors versus total dose, distinguishing even (blue line) and odd (red line) blocks, for Lot A (left) and Lot B (right).



(a) **(b)**

Figure 5.18: Errors distribution into pages considering 4 fresh devices (a), and distinguishing between lots (b)

However, this issue does not affect the errors statistic. In fact, plotting the corresponding histograms of Fig. 5.10, but distinguishing between even and odd blocks, the retention errors distribution trend does not change (Fig. 5.19 and 5.20).



(a) Even blocks, pre-rad

(b) Odd blocks, pre-rad

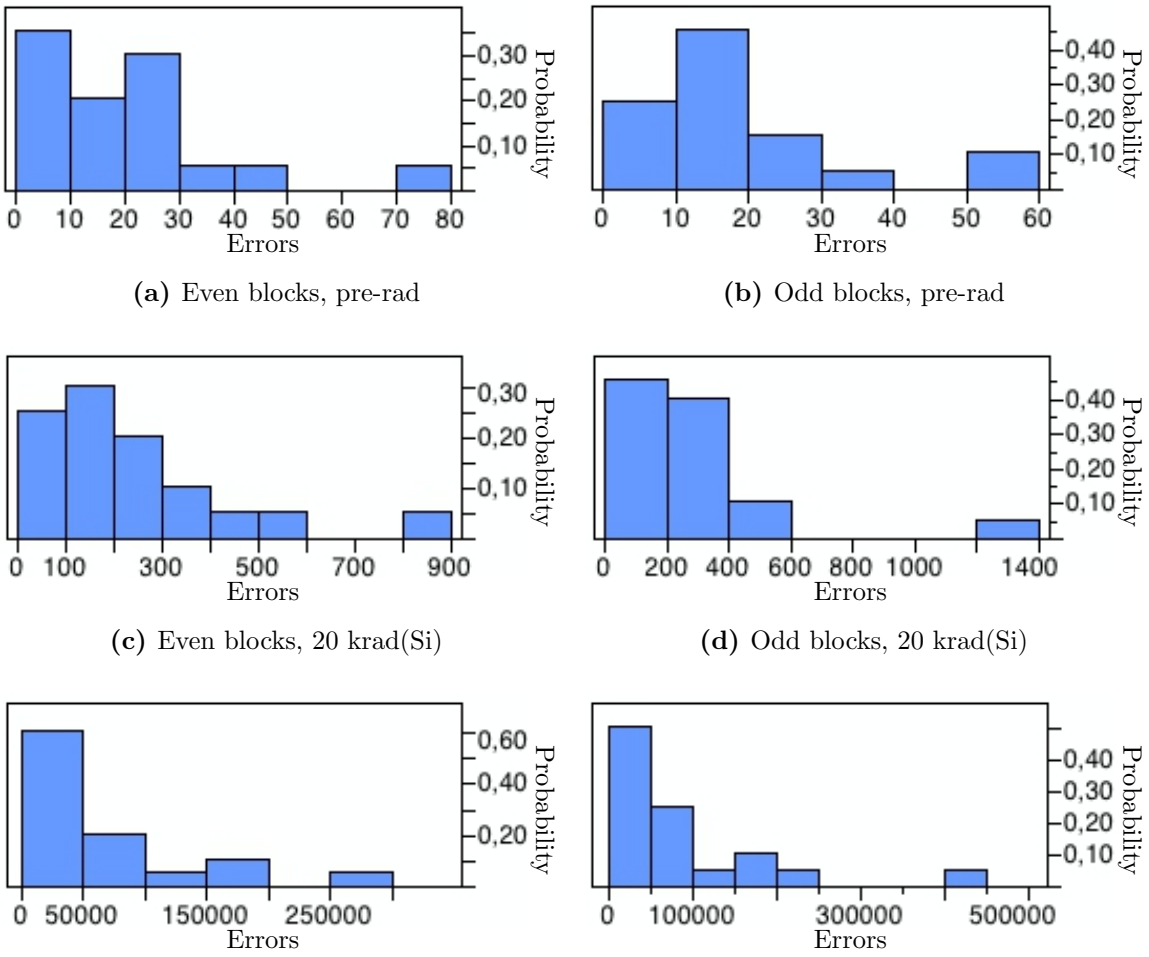
(c) Even blocks, 20 krad(Si)

(d) Odd blocks, 20 krad(Si)

(e) Even blocks, 33 krad(Si)

(f) Odd blocks, 33 krad(Si)

Figure 5.19: Distributions of FG errors in even and odd blocks, for Lot A.



(a) Even blocks, pre-rad

(b) Odd blocks, pre-rad

(c) Even blocks, 20 krad(Si)

(d) Odd blocks, 20 krad(Si)

(e) Even blocks, 33 krad(Si)

(f) Odd blocks, 33 krad(Si)

Figure 5.20: Distributions of FG errors in even and odd blocks, for Lot B.

5.2.1 Device Decap

In ESTEC laboratories, some Flash memories have been decapped. The operation required two steps: first, mechanical etching of a thin layer of plastic (about 0.10 mm), then chemical etching by nitric acid at high temperature (80° C).

Through this method it is possible to open the device, avoiding damaging it, and to observe the whole die (Fig. 5.21). The control circuitry is placed below (i.e. the small rectangles at the bottom of the die), while the memory array is divided in half planes. The gold bond wires are well visible.

Typically odd and even pages/blocks are placed on different half planes of the memory array. The left one probably contains even blocks, while the right one odd blocks . Each half plane is partitioned in two other section, that could be dedicated in turn to even and odd pages. The asymmetric position of the control circuitry between the two half planes, in addition to a slight radiation-induced degradation of the read pump, could be the cause of this particular disposition, characterizing all the tested devices. A lower read voltage for even blocks/pages leads to detect less errors, since erased cells are read as programmed.

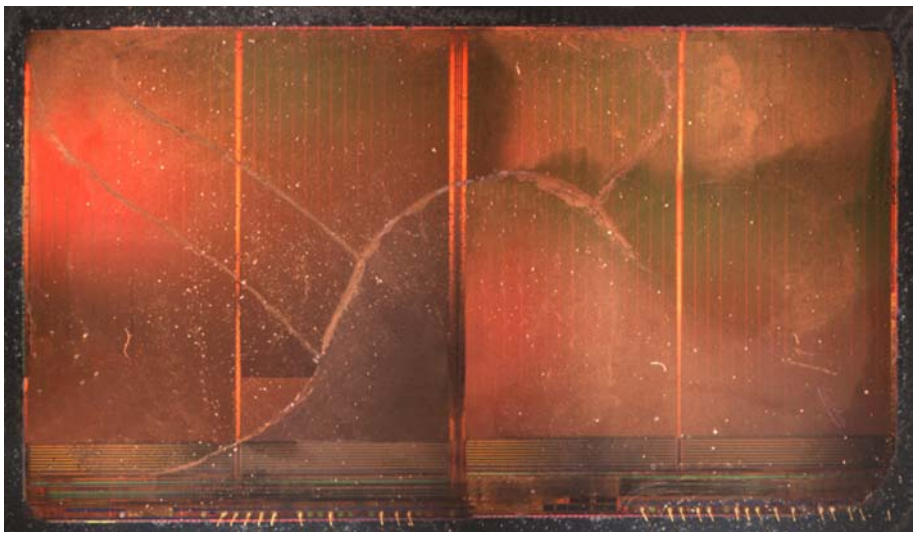


Figure 5.21: The die of a decamped memory.

5.3 Annealing

As suggested in the ESCC Basic Specification No. 2290 for total dose test, several read operations have been performed after the irradiation. Considering t_0=0 the gamma source stop time, the following read operations have been performed:

1. t_1=5 minutes;

2. t_2=24 hours;

3. t_3=196 hours;

4. t_4=744 hours;

It must be specify that not all these read operations have been performed on the same conditions. In fact the last reading, t_4, about one month after the radiation test, and the third reading of 10 memories of the Lot B (B11-B20) have been performed in Padua. The devices have been subjected to travel unknown conditions, such as warmer temperatures, that could have caused detrapping of the positive charge trapped in the oxide surrounding the floating gate. In fact it has been observed a reduction of the retention errors in the third read operation for the 10 memories read in Padua, while in other samples, retention errors increased. Performing several read operations in succession, retention errors decrease, probably because of the same reason: the device warming, due to the applied bias, could cause charge detrapping.

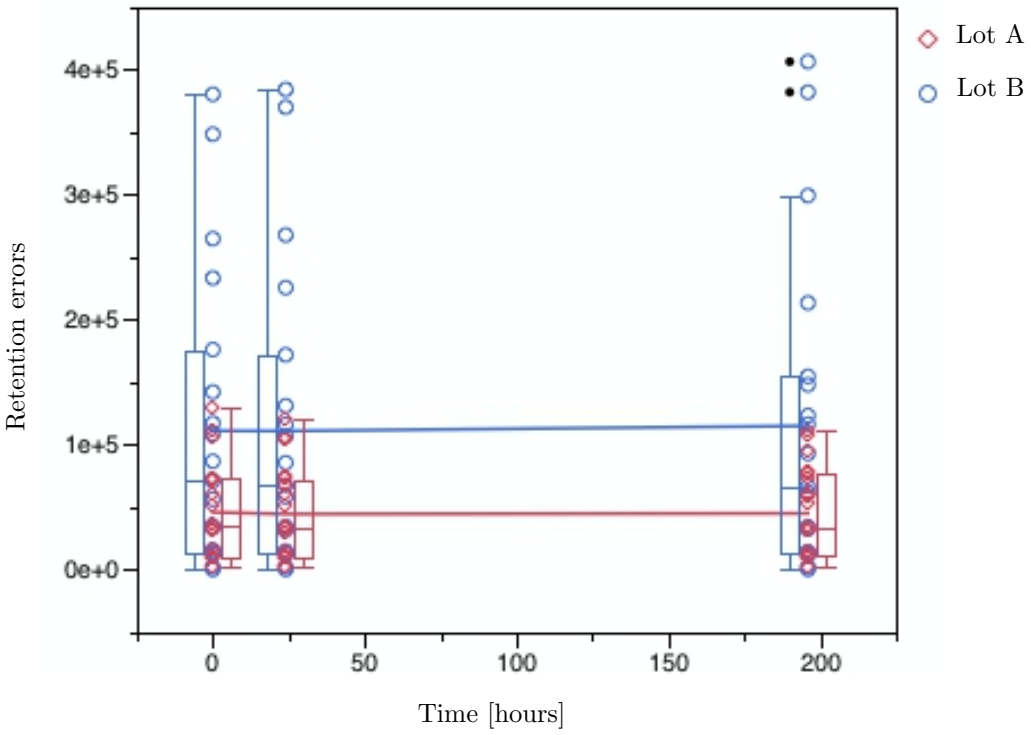


Figure 5.22: Retention errors during annealing, after TID exposure.

In addition, from this annealing data analysis, one sample belonging to Lot B has been eliminated (B1), since this device failed during the measurement 24 hours after the end of the total dose exposure. During the following readings it

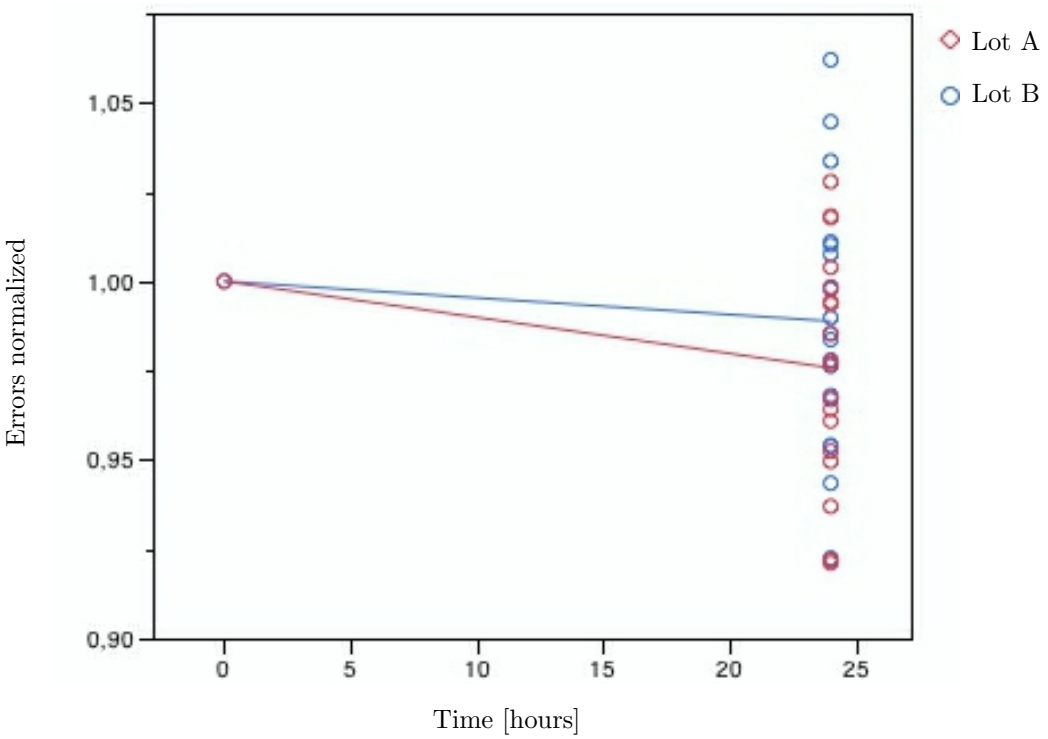


Figure 5.23: Retention errors, up to 24 hours after the gamma run stop, normalized over the number of errors detected just after TID exposure.

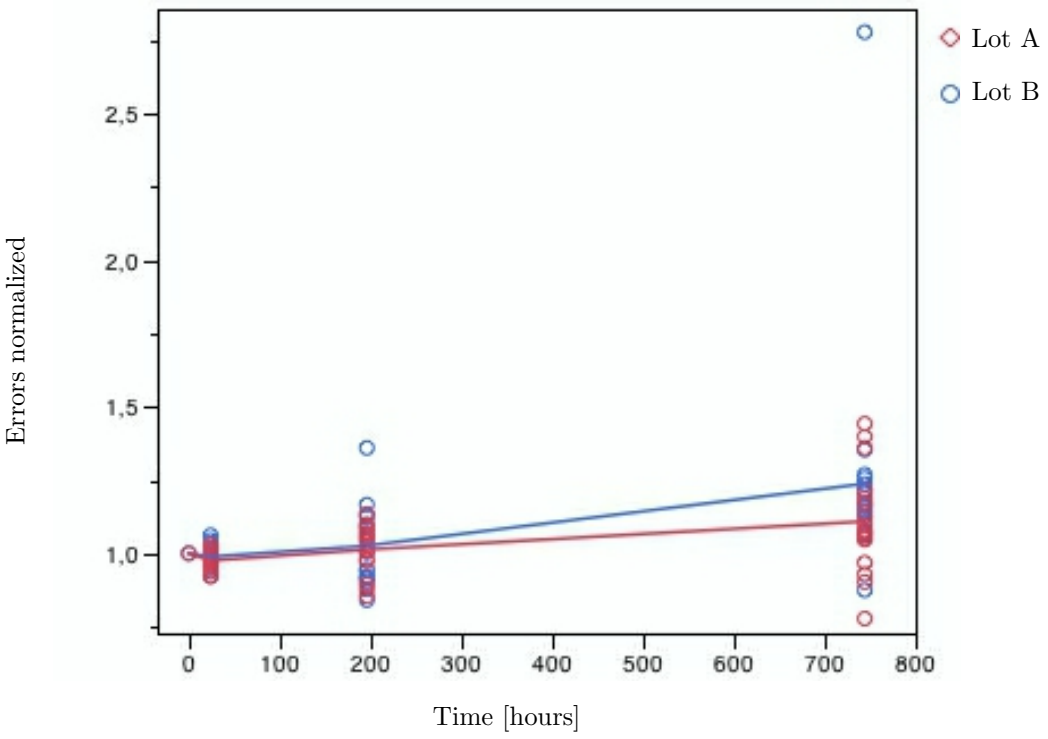


Figure 5.24: Retention errors normalized over the number of errors detected just after TID exposure.

seemed to work correctly, however, the annealing trend would be distort for the lack of one measurement.

The same considerations of the previous section hold true during annealing: Lot B has a mean errors value and errors variability across devices higher than Lot A, as shown in Fig. 5.22. The general trend is a slight reduction of errors in the hours just after the TID test, whereas they slowly increase over the time.

This FG errors evolution over time can by explained by the contribution of two physical mechanisms that go in the oppose direction. The annealing of the FG errors is predominated in the hours just before the beginning of the retention test. The charge trapped in the oxides surrounding the floating gate is removed or neutralized and some errors disappear. Forward in time, discharge of the FG cause the errors increase. This is an unwanted effect, since the device reliability is compromised. Thus it is interesting to investigate the physical mechanisms which causes the data loss.

Mainly two phenomena explain the charge loss from the floating gate. Although we are not able to exactly determine if one is predominant or if they occur in the same time, however, it is possible to give a qualitative description for the charge loss after TID exposure.

i) Electrons from the floating gate tunnel in the oxide to recombine with trapped holes.

ii) Formation of permanent neutral radiation-induced defects in the tunnel oxide, which allow electrons tunneling via oxide traps. Radiation Induced Leakage Current (RILC) mediates the charge loss from the floating gate. RILC is modeled on the Trap Assisted Tunneling (TAT).

In both cases, the decrease of the number of electrons stored in the FG makes the threshold voltage to shift towards lower value.

The efficiency of electron TAT depends on several factor, among which:

- amount of holes generated by ionizing radiation;
- the traps distribution in the oxide;
- the oxide field, whose increase, increases the RILC;

These considerations are useful to evaluate the graph in Fig. 5.24, where errors are normalized over the number of errors at the beginning of the retention test (t=0). Retention errors over time increase faster in Lot B. As considered previously, because of the small amount of cells detected with errors respect to more than 10^{12} irradiated cells, lot-to-lot errors variation arising during annealing

is attributed only to lot-to-lot parameter variability, and not to radiation-induced variability. As follows, parameter variability between the lots could be identified in different oxide qualities. Many oxide defects, which act as traps supporting the electron TAT mechanism, increase the electron tunnel probability from the floating gate into the substrate, thus negative charge is lost. It could be supposed that different process conditions during oxide growing may have produced a lower oxide quality for the devices belonging to Lot B.

Finally, Fig. 5.25 shows the evolution over time of the error distribution. It is interest to notice how the annealing over time acts on the errors distribution of Lot B, spreading it. On the contrary, for Lot A, the annealing time makes the errors distribution spreading compact. These histograms clearly show that devices belonging to Lot B are more prone to charge loss: the V_T distribution of devices characterized by higher bit error values shifts towards lower values faster than other devices of the same lot. On the other hand, in Lot A, devices which experienced more bit errors in the hours after the TID exposure tend to homogenize with the mean value of the lot.

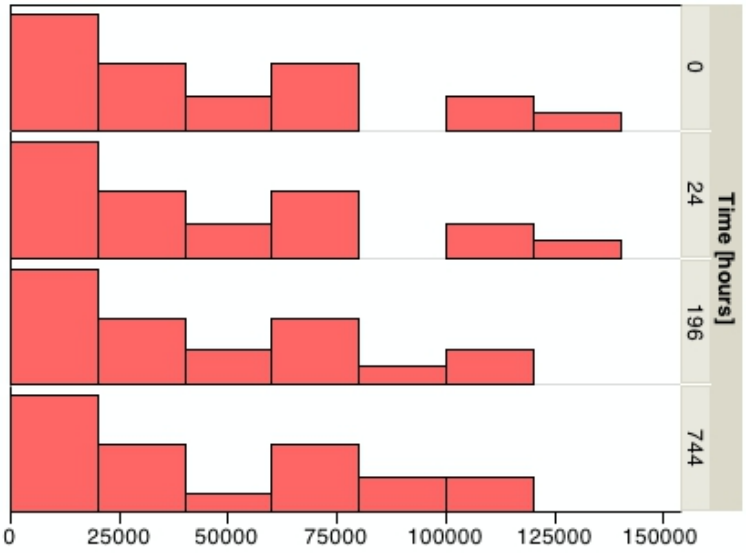
5.4 Current

During the read operation, performed after the two dose steps of 20 krad(Si) and 33 krad(Si), each sample was placed on the board connected with the multimeter, to check the current during an erase/program/read cycle of 30 blocks of the memory array. Fig. 5.26 shows the typical evolution in the time domain of the current measured during the E/P/R cycle. The shape is almost the same for all the devices, even if the current value slightly changes from device to device. As explained in the first chapter, in NAND architecture the erase operation performed at block level, while the program operation at page level. This is clearly evident in the short erasing (few milliseconds), and the longer time to complete the programming, about 10 seconds.

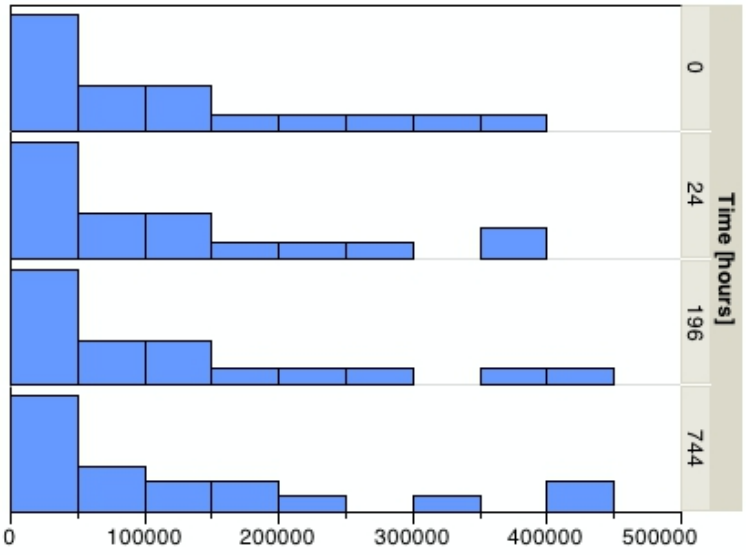
Averaging the program current for each device, it has been obtained the graph in Fig. 5.27. The program current slightly increases as a function of the total dose, probably because of the reduction of the program pump output voltage. In fact the thickness of the tunnel oxide ($\sim$8 nm) does not allow to justify the increase of the current, because the positive charge trapped in the oxide is small.

The average program time increases as total dose raises (Fig. 5.28). This effect may be ascribed to the program pump output voltage degradation too. During the program algorithm implementation, a greater number of program pulses is need to reach the program verify level.

The large number of errors and their variability observed in Lot B are followed



(a) Lot A



(b) Lot B

Figure 5.25: Histograms of the errors distribution of Lot A and B over time.

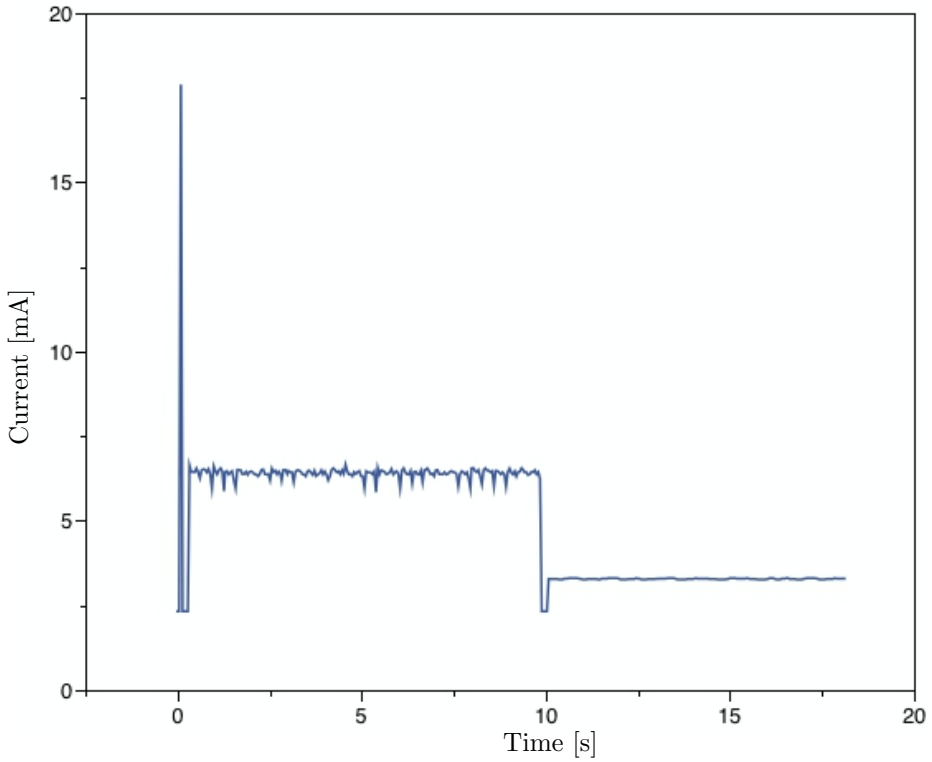


Figure 5.26: Current during E/P/R cycle measured after a total dose of 33 krad(Si)

by an higher program current. Also the time needed to perform the program operation is longer in this lot (Fig. 5.28). This behavior can be ascribed to wider threshold voltage distribution respect to Lot A, thus program operation needs more iterations of the program algorithm.

5.5 Conclusions

In this chapter the variability observed in detected errors has been studied. Data collected before, during and after the radiation exposure have been combined to identify the bit error trend among cells belonging to the same lot or not. Previous studies has been taken into account, in order to provide a physical explanation for the output value ("0" or "1"), related to the unknown threshold voltage distribution.

Interesting characteristics of the spatial distribution of errors in pages and blocks emerged, for which the device decapping has been useful to give a possible explanation.

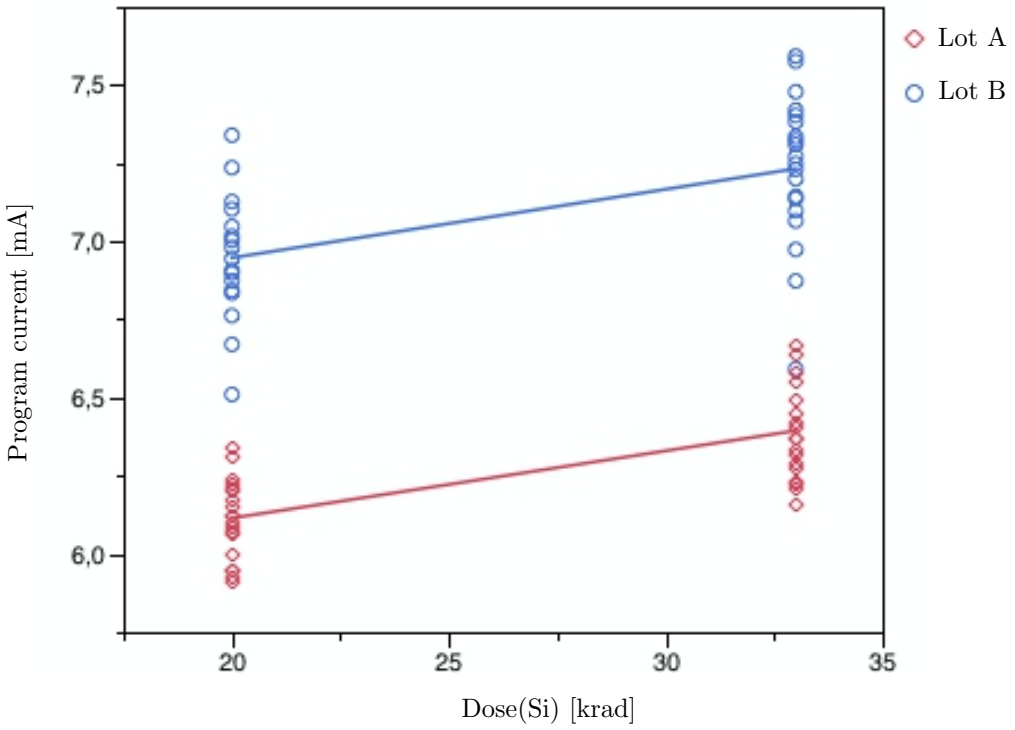


Figure 5.27: Program current measured after 20 krad(Si) and 33 krad(Si).

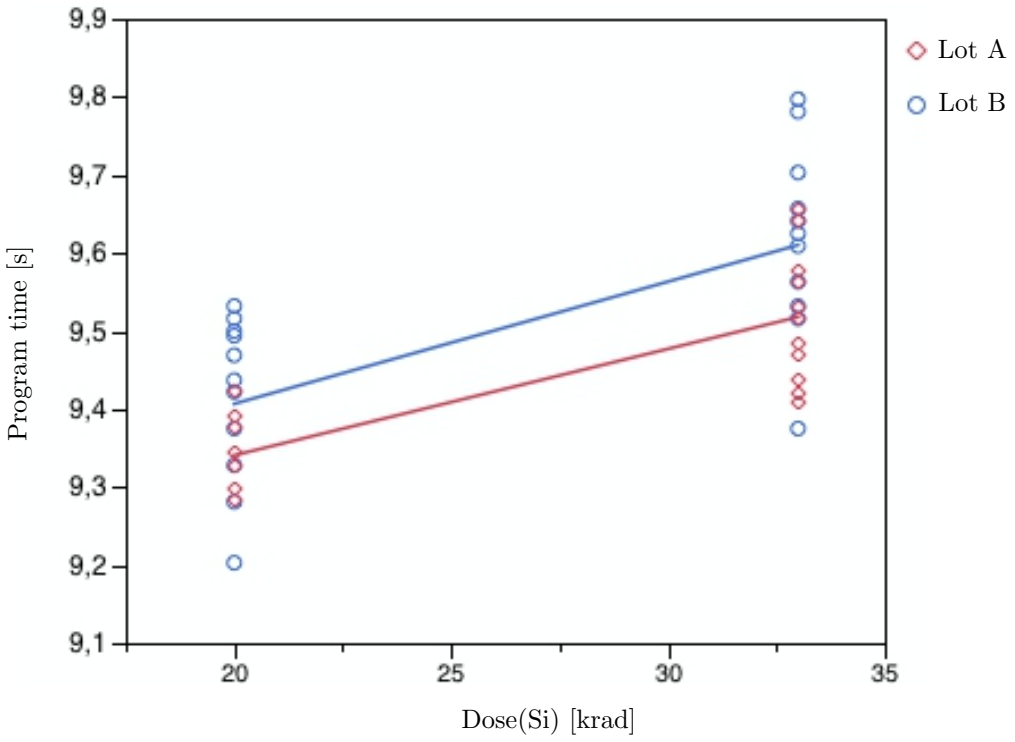


Figure 5.28: Program time measured after 20 krad(Si) and 33 krad(Si) .

Chapter 6
Final Considerations

The irradiation of a large set of samples allows to extrapolate useful information on 25-nm NAND Flash memory for space applications. With a total dose of 33 krad, split in two dose steps, in order to perform an intermediate measurement, we have been able to collect a considerable number of FG errors, avoiding control circuitry failure.

The first observation, from calibration results, is that FG errors occur only in programmed cells, i.e. storing electrons, and at a total dose level lower than previous devices. This can be due to a slight reduction in the tunnel oxide thickness, as well as new paths for cell discharge.

The 40 tested devices were taken from two different lots, thus it is possible to derive some conclusions on lot-to-lot variability. First of all, devices belonging to different lots evidently differ for the average number of detected errors, already before the TID irradiation, after the program operation. In addition, during and after irradiation, Lot B is more prone to retention errors, and TID increases this tendency. The lot-to lot variability is mainly ascribed to programming variability. In fact, the wider threshold voltage distribution, which appears as a higher number of bit errors, is the result of electron injection statistic (EIS), i.e the variability of the number of electrons injected at each program pulse, performed during program algorithm. This phenomenon, related to the miniaturization of the FG cell, increases the threshold voltage spread. A further variability source during programming can be identified in the slight variability of the tunnel oxide thickness (t_{ox}), in cells belonging to the same array, because of process control issues. The result is the programmed threshold voltage distribution spreading.

During annealing, several measurements have been performed after the total ionizing dose irradiation, to check the bit error rate over time. After a slight decrease of FG errors in the hours following the gamma run stop, retention errors progressively increase over time, for both lots. However, devices belonging to Lot B are characterized by a more rapid FG errors increase. This confirms that Lot B is more prone to retention errors, probably because of the presence of

oxide defects which act as traps, allowing the trapped assisted tunneling, and consequently the floating gate discharge, throughout Radiation Induced Leakage Current (RILC) mechanism. This observation points out how the oxides quality could strongly affects the post-irradiation behavior of Flash devices.

Chip-to-chip variability may be mainly ascribed to phenomena due to few electrons, since the injected/removed charge from the floating gate must be considered as the sum of stochastic events, because of the small feature size of devices under study. Between several sources of variably belonging to this category, the most affecting are Random Discrete Dopants (RDDs) and Interface Trapped Charge (ICT). The Poisson distribution of FG errors agrees with the Poisson process which governs the distribution of dopants and charge. The discrete nature of dopants is also involved in Random Telegraph Signal disturbs, which may produce an erroneous output value.

It has been quite interesting to notice the distribution of FG errors among pages and blocks. A peculiar behavior emerged, a higher bit error rate in odd pages and blocks than in even ones. However this characteristic errors disposition appears only after the TID exposure, thus it is ascribed to radiation-induced effects on the control circuitry. It is an unavoidable behavior, which appears also at low dose rate during calibration. The asymmetrical position of the control circuitry respect to even and odd pages/blocks, added to a slight degradation of the output voltage of the read pump, has been identified has the most likely cause of this particular errors disposition.

The reliability analysis of 25-nm Flash memories for TID effects demonstrates that these devices can be used for spatial applications, working on orbits with a total dose level up to 20 krad(Si). Furthermore, since the radiation induced FG errors are less than 1% the amount of irradiated cells, the errors correction code (ECC) allows to correct them. However, 25-nm Flash memories are very sensible to small parameters fluctuations. The low number of electrons per bit (about 125 electrons) induces a quite large errors variability, which could considerably differ from lot to lot. To develop valid qualification methods for spatial components, lot-to lot test is necessary, analyzing at least more than one component per lot.

Appendix A

Program and Read Scripts

Programing script

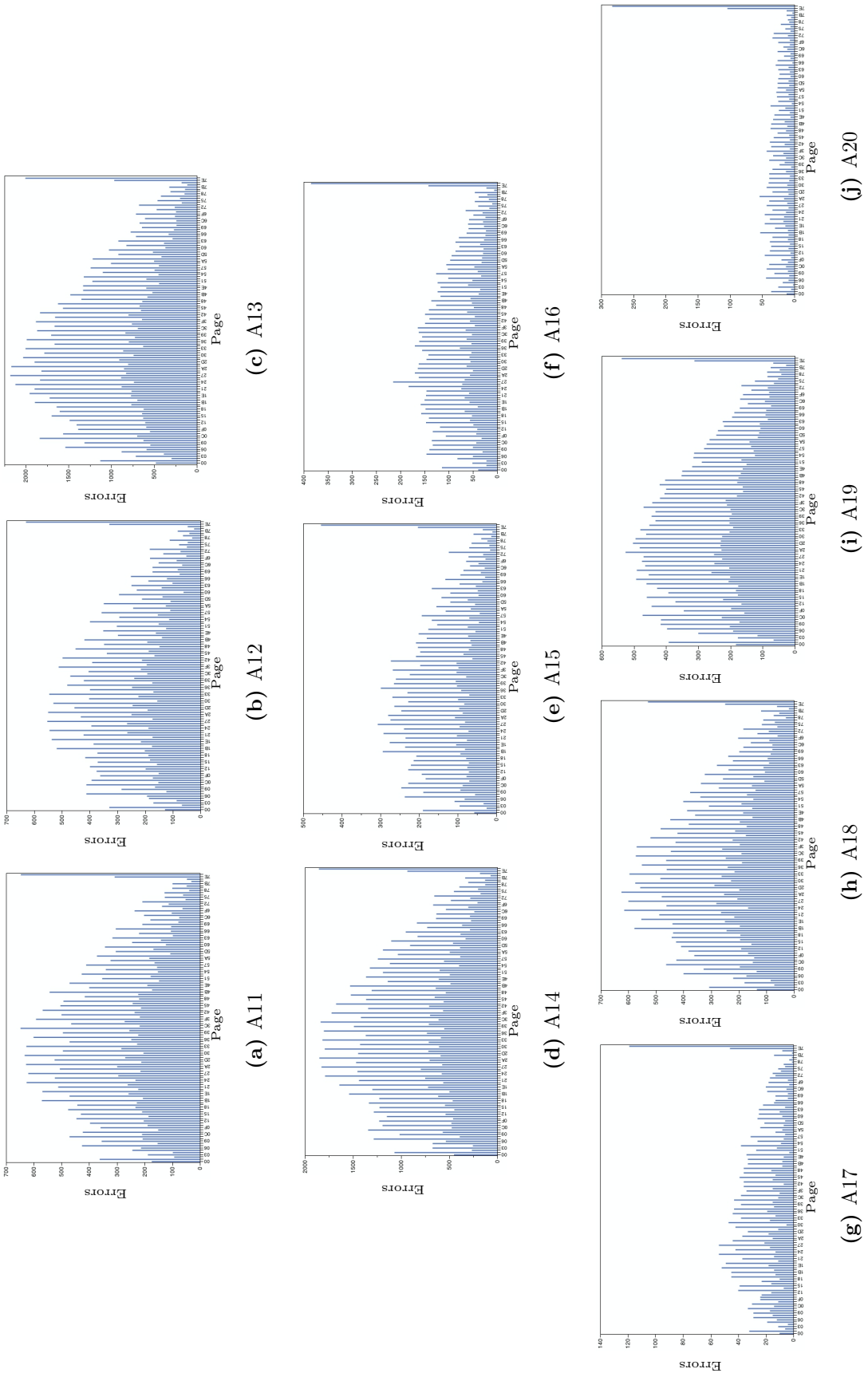
```
array set blocksToSkip { 1 0x5A 2 0x5B }
set dir [ChooseDirectory]
SetCurrentDirectory $dir
if { $dir == "" } { return }
# check if directory is empty
if { [CheckCurrentDirectory] == 0 } { return }
ClearLog
StreamLog "Log.txt"
SaveScript "Script.txt"
NAND_Configure 4
NAND_Select 1
NAND_Reset
NAND_ReadID
NAND_ReadSR
foreach { index block } [array get blocksToSkip] {
NAND_SkipBlockLoop $block
}
NAND_Erase 0x00 0xFFF
NAND_MLProgram 0x00 0x09 0x55 0x55 "r"
NAND_Program 0x0A 0xFFF 0x55 "s"
CloseStreamLog
```

Reading script

```
array set blocksToSkip { 1 0x5A 2 0x5B }
set dir [ChooseDirectory]
SetCurrentDirectory $dir
if { $dir == "" } { return }
# check if directory is empty
if { [CheckCurrentDirectory] == 0 } { return }
ClearLog
StreamLog "Log.txt"
SaveScript "Script.txt"
NAND_Configure 4
NAND_Select 1
NAND_Reset
NAND_ReadID
NAND_ReadSR
foreach { index block } [array get blocksToSkip] {
NAND_SkipBlockLoop $block
}
NAND_MLCheck 0x00 0x09 0x55 0x55 "r"
NAND_MLCheck 0x0A 0xFFF 0x55 0x55 "s"
CloseStreamLog
```

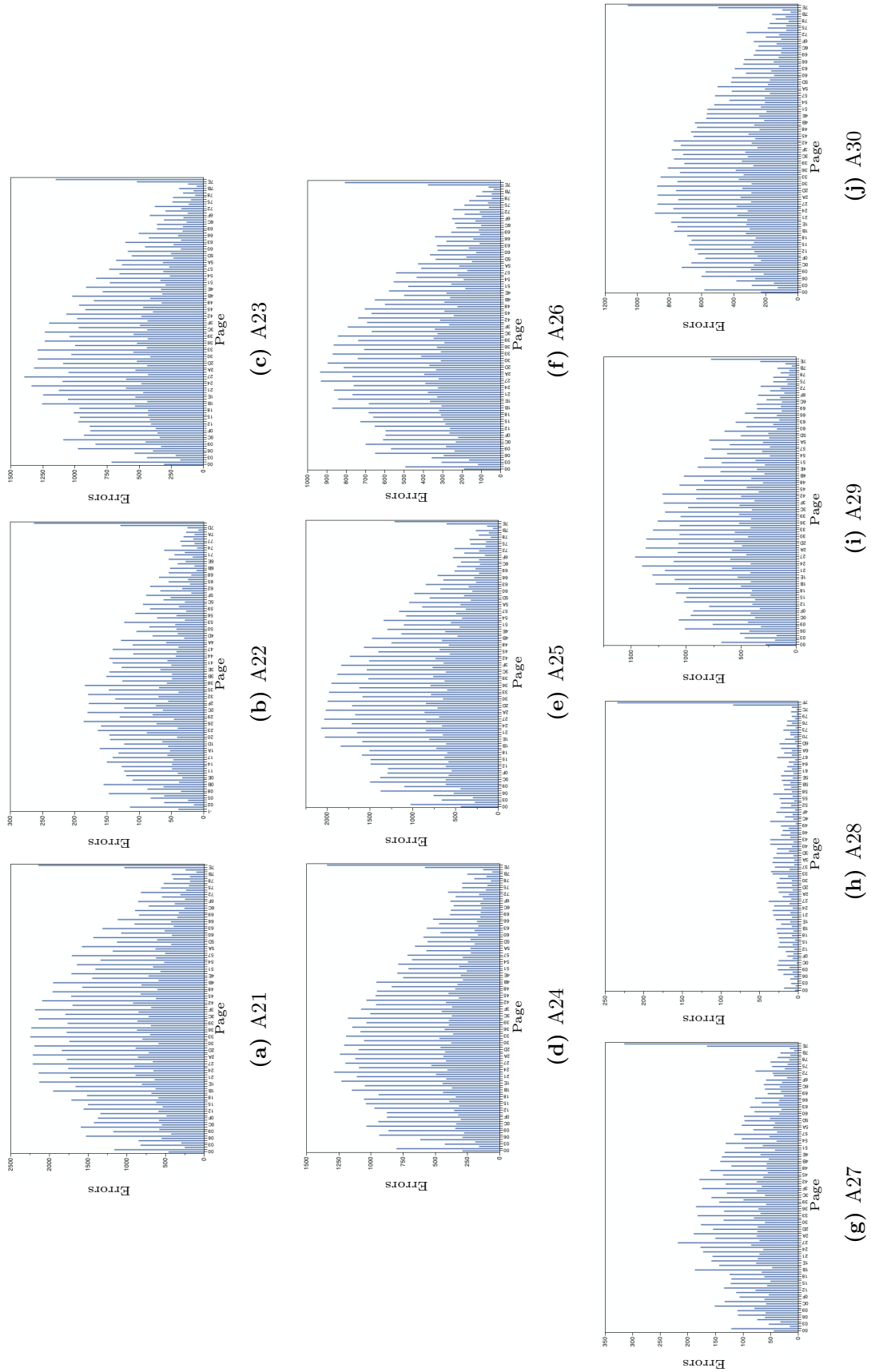
Appendix B
Errors distribution across pages

The following figures represent the errors distribution into page after a total dose of 33 krad.



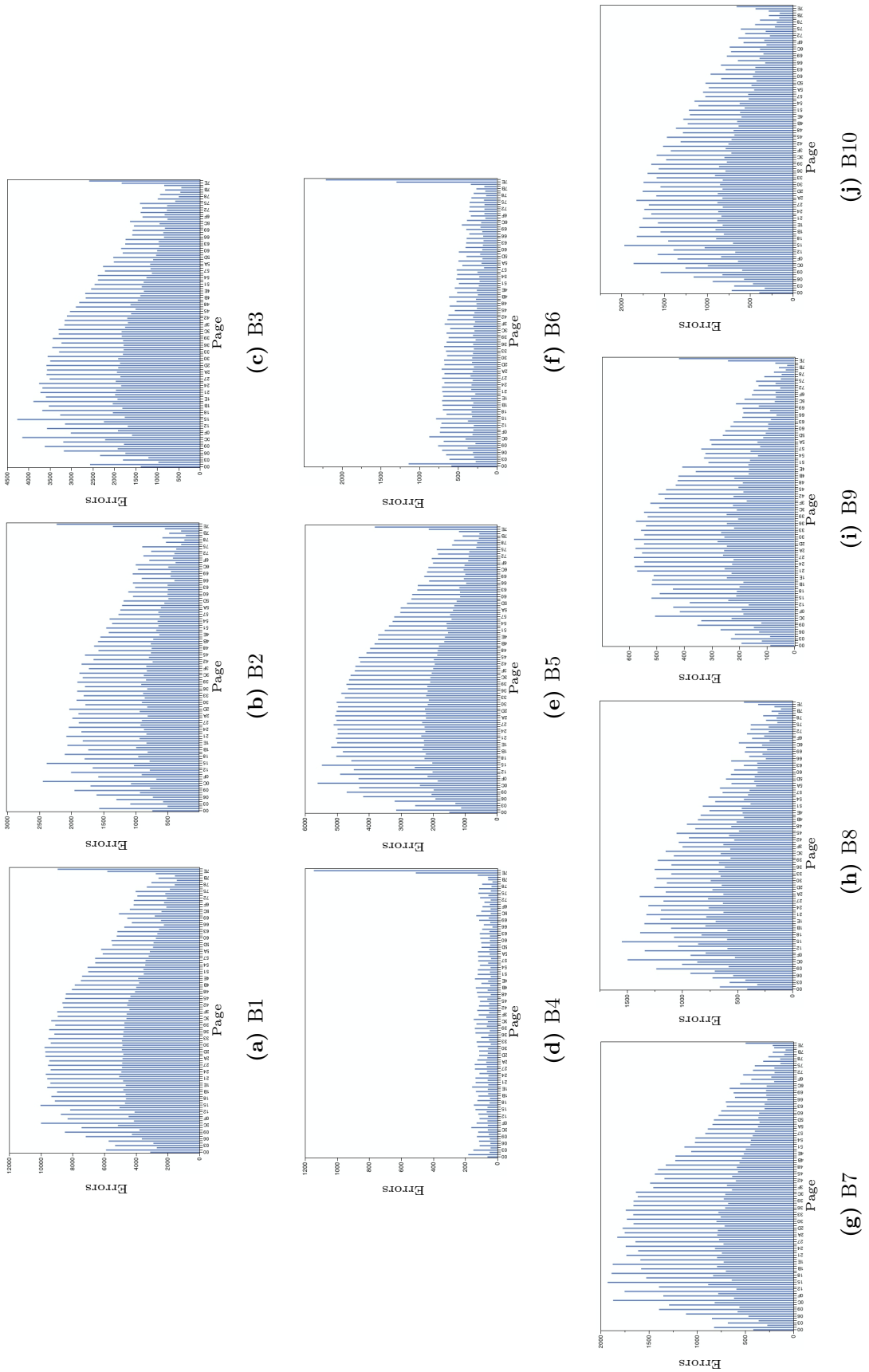
(a) A11 (b) A12 (c) A13 (d) A14 (e) A15 (f) A16 (g) A17 (h) A18 (i) A19 (j) A20

Figure B.1: Errors spatial distribution into pages for the first ten tested devices of the A lot.



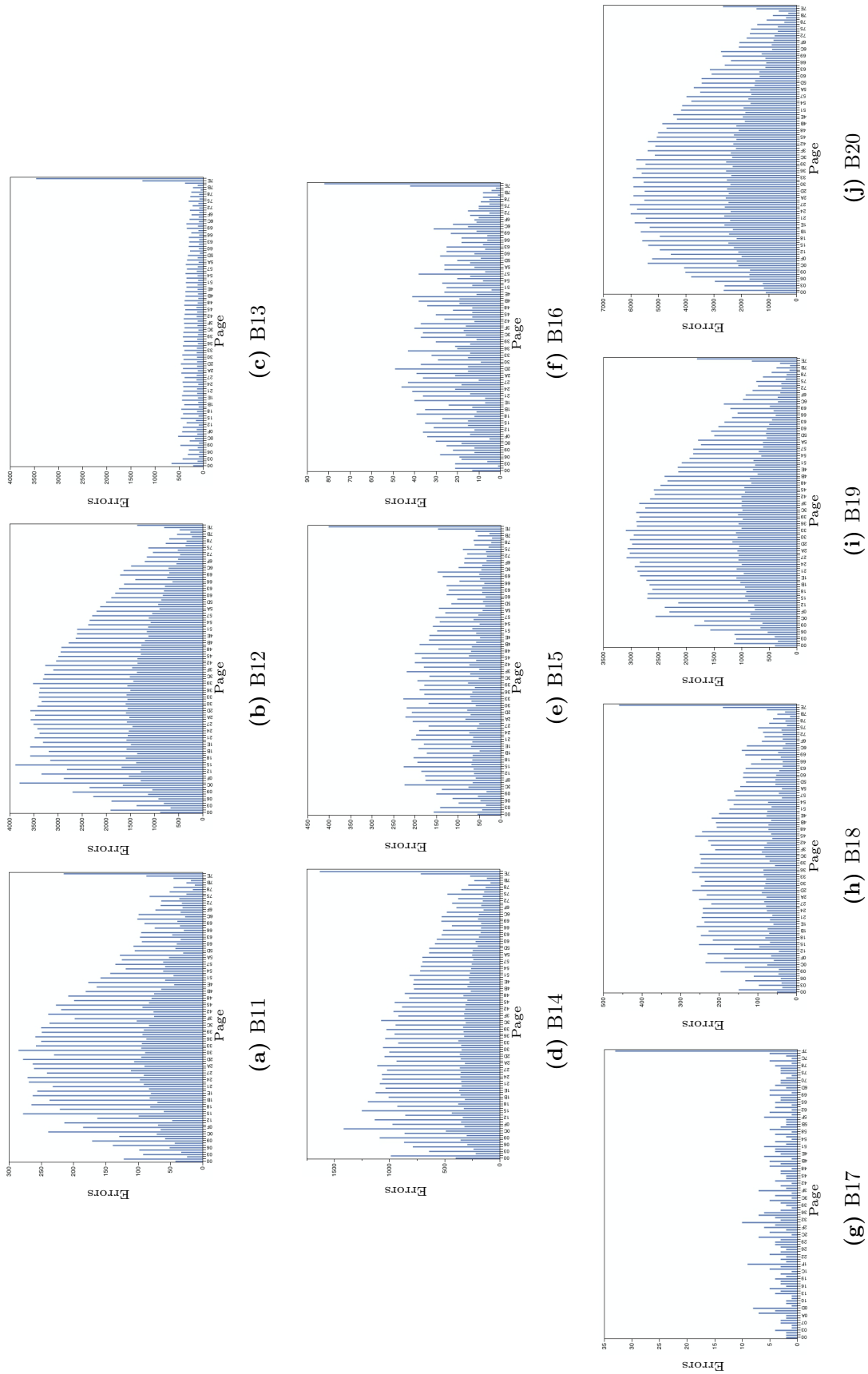
(a) A21 (b) A22 (c) A23
(d) A24 (e) A25 (f) A26
(g) A27 (h) A28 (i) A29 (j) A30

Figure B.2: Errors spatial distribution into pages for the second ten tested devices of the A lot.



(a) B1 (b) B2 (c) B3 (d) B4 (e) B5 (f) B6 (g) B7 (h) B8 (i) B9 (j) B10

Figure B.3: Errors spatial distribution into pages for the first ten tested devices of the B lot



(a) B11 (b) B12 (c) B13

(d) B14 (e) B15 (f) B16

(g) B17 (h) B18 (i) B19 (j) B20

Figure B.4: Errors spatial distribution into pages for the second ten tested devices of the B lot.

Bibliography

[1] S. Lai, "Non-volatile memory technologies: The quest for ever lower cost," in *IEDM Tech. Dig.*, 2008.

[2] A. Visconti, "Flash memory reliability," Padova, 16 Gen. 2013.

[3] S. Gerardin and A. Paccagnella, "Present and future of non-volatile memories for space," *IEEE Transactions on Nuclear Science*, vol. 57, Dec. 2010.

[4] A. Spessot, C. M. Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, "Compact modeling of variability effects in nanoscale NAND Flash memories," *IEEE Transactions on Electron Devices*, vol. 58, August 2011.

[5] G. Roy, A. Ghetti, A. Benvenuti, A. Erlebach, and A. Asenov, "Comparative simulation study of the different sources of statistical variability in contemporary floating-gate nonvolatile memory," *IEEE Transactions on Electron Devices*, vol. 58, pp. 4155–4163, Dec. 2011.

[6] G. Molas, D. Deleruyelle, B. D. Salvo, G. Ghibaudo, M. Gély, L. Perniola, D. Lafond, and S. Deleonibus, "Degradation of floating-gate memory reliability by few electron phenomena," *IEEE Transactions on Electron Devices*, vol. 53, October 2006.

[7] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cell-an overview," *Proceedings of IEEE*, vol. 85, pp. 1248–1271, Aug. 1997.

[8] P. E. Cottrel, R. R. Troutman, and T. H. Ning, "Hot-electron emission in n-channel IGFET's," *IEEE Transactions on Electron Devices*, vol. 26, no. 4, pp. 520–532.

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

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

[11] C. Miccoli, C. M. Compagnoni, A. S. Spinelli, and A. L. Lacaita, "Investigation of the programming accuracy of a double-verify ISPP algorithm for nanoscale NAND Flash memories," in *IRPS, 2011 IEEE International*, 2011.

[12] C. M. Compagnoni, A. S. Spinelli, R. Gusmeroli, S. Beltrami, A. Ghetti, and A. Visconti, "Ultimate accuracy for the NAND flash program algorighm due to the electron injection statistics," *IEEE Transactions on Electron Devices*, vol. 55, pp. pp. 2695–2702, Oct. 2008.

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

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

[15] K. Sakakibara, N. Ajika, M. Hatanaka, and H. Miyoshi, "A quantitative analysis of stress inducedexcess current (SIEC) in SiO2 films," *Proceedings of IEEE International Reliability Physics Symposium (IRPS)*, pp. 100–107, 1996.

[16] K. Sakakibara, N. Ajika, M. Hatanaka, H. Miyoshi, and A. Yasuoka, "Identification of stress induced leakage current components and corresponding trap models in SiO2 films," *IEEE Transactions on Electron Devices*, vol. 44, pp. 986–992, 1997.

[17] K. Sakakibara, N. Ajika, and H. Miyoshi, "Influence of holes on neutral trap generation," *IEEE Trans. Electron Devices*, vol. 44, no. 12, pp. 2274–2280, 1997.

[18] Y. Kitahara, D. Hagishima, and K. Matsuzawa, "Reliability of NAND Flash memories induced by anode hole generation in floating-gate," *IEEE Transactions on Nuclear Science*, 2011.

[19] T. C. Chen, S. Li, S. Fung, C. D. Beling, and K. F. Lo, "Post-stress interface trap generation induced by oxide-field stress with FN injection," *IEEE Transactions on Electron Devices*, vol. 45, no. 1972-1977, 1998.

[20] A. Ghetti, M. Bonanomi, C. M. Compagnoni, A. S. Spinelli, A. L. Lacaita, and A. Visconti, "Physical modeling of single-trap RTS statistical distribution in Flash memories," in *IEEE 46th Annual International Reliability Physics Symposium*, 2008.

[21] D. Ielmini, "Overview of modeling approaches for scaled non volatile memories," in *International Conference on Simulation of Semiconductor Processes and Devices*, 2009.

[22] J. Barth, "Modeling space radiation environments," in *NSREC*, 1997.

[23] H. Barnaby, "Total-ionizing-dose effects in modern CMOS technologies," *IEEE Transactions on Nuclear Science*, vol. 53, pp. 3103–3121, December 2006.

[24] J. Schwank, "Total dose effects in MOS devices," in *NSREC*, 2002.

[25] E. S. Snyder, P. J. McWhorter, T. A. Dellin, and J. D. Sweetrnant, "Radiation response of floating gate EEPROM memory cells," *IEEE Transactions on Nuclear Science*, vol. 36, December 1989.

[26] G. Cellere, P. Pellati, A. Chimenton, A. Modelli, L. Larcher, J. Wyss, and A. Paccagnella, "Radiation effects on floating-gate memory cells," *IEEE Transactions on Nuclear Science*, vol. 48, pp. 2222–2228, 2001.

[27] G. Cellere, A. Paccagnella, A. Visconti, M. Bonanomi, P. Caprara, and S. Lora, "A model for TID effects on floating gate memory cells," *IEEE Transactions on Nuclear Science*, vol. 51, December 2004.

[28] M. Bagatin, S. Gerardin, G. Cellere, A. Paccagnella, A. Visconti, M. Bonanomi, and S. Beltrami, "Error instability in floating gate flash memories exposed to TID," *IEEE Transactions on Nuclear Science*, vol. 56, December 2009.

[29] M. Bagatin, G. Cellere, S. Gerardin, A. Paccagnella, A. Visconti, and S. Beltrami, "TID sensitivity of NAND Flash memory building blocks," *IEEE Transactions on Nuclear Science*, vol. 56, no. 4, 2009.

[30] F. Irom, D. N. Nguyen, G. Cellere, M. Bagatin, S. Gerardin, and A. Paccagnella, "Catastrophic failure in highly scaled commercial NAND Flash memories," *IEEE Transactions on Electron Devices*, vol. 57, pp. 266–271, Feb. 2010.

[31] A. Costantino, "Basic information about the Estec Co-60 facility." https://escies.org/webdocument/, December 2013.

[32] F. Irom, D. N. Nguyen, and G. R. Allen, "Single event effect and total ionizing dose results of highly scaled Flash memories," *IEEE Transactions on Nuclear Science*, 2013.

[33] S. Gerardin, M. Bagatin, A. Paccagnella, and V. Ferlet-Cavrois, "Degradation of sub 40-nm NAND Flash memories under total dose irradiation," *IEEE Transactions on Nuclear Science*, vol. 59, December 2012.

[34] A. Paccagnella, "Radiation response and reliability of oxides used in advances processes," in *IEEE NSREC*, 2003.