# Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
"SCIENZE DELL'INGEGNERIA"

CICLO XXIV

COORDINATORE Prof. Stefano Trillo

# ELECTRICAL CHARACTERIZATION, PHYSICS, MODELING AND RELIABILITY OF INNOVATIVE NON-VOLATILE MEMORIES

Settore Scientifico Disciplinare ING/INF-01


**Dottorando**
Dott. Zambelli Cristian

*(firma)*

**Tutore**
Prof. Olivo Piero

*(firma)*


Anni 2009/2011

# Abstract

Enclosed in this thesis work it can be found the results of a three years long research activity performed during the XXIV-th cycle of the Ph.D. school in Engineering Science of the Universitá degli Studi di Ferrara. The topic of this work is concerned about the electrical characterization, physics, modeling and reliability of innovative non-volatile memories, addressing most of the proposed alternative to the floating-gate based memories which currently are facing a technology dead end. Throughout the chapters of this thesis it will be provided a detailed characterization of the envisioned replacements for the common NOR and NAND Flash technologies into the near future embedded and MPSoCs (Multi Processing System on Chip) systems. In Chapter 1 it will be introduced the non-volatile memory technology with direct reference on nowadays Flash mainstream, providing indications and comments on why the system designers should be forced to change the approach to new memory concepts. In Chapter 2 it will be presented one of the most studied post-floating gate memory technology for MPSoCs: the Phase Change Memory. The results of an extensive electrical characterization performed on these devices led to important discoveries such as the kinematics of the erase operation and potential reliability threats in memory operations. A modeling framework has been developed to support the experimental results and to

validate them on projected scaled technology. In Chapter 3 an embedded memory for automotive environment will be shown: the SimpleEE p-channel memory. The characterization of this memory proven the technology robustness providing at the same time new insights on the erratic bits phenomenon largely studied on NOR and NAND counterparts. Chapter 4 will show the research studies performed on a memory device based on the Nano-MEMS concept. This particular memory generation proves to be integrated in very harsh environment such as military applications, geothermal and space avionics. A detailed study on the physical principles underlying this memory will be presented. In Chapter 5 a successor of the standard NAND Flash will be analyzed: the Charge Trapping NAND. This kind of memory shares the same principles of the traditional floating gate technology except for the storage medium which now has been substituted by a discrete nature storage (i.e. silicon nitride traps). The conclusions and the results summary for each memory technology will be provided in Chapter 6. Finally, on Appendix A it will be shown the results of a recently started research activity on the high level reliability memory management exploiting the results of the studies for Phase Change Memories.

# Abstract (Italiano)

Racchiusi in questo lavoro di tesi si possono trovare i risultati della triennale attivitá di ricerca eseguita durante il XXIV-esimo ciclo del Dottorato di Ricerca in Scienze dell'Ingegneria svolto presso l'Universitá degli Studi di Ferrara. L'argomento di questa trattazione riguarda la caratterizzazione elettrica, la fisica, la modellistica e l'affidabilitá di memorie non-volatili innovative, indirizzando l'argomentazione verso le tecnologie proposte a rimpiazzare nel prossimo futuro le tradizionali memorie a gate flottante, ormai proiettate verso un naturale declino tecnologico. Attraverso i capitoli di questa tesi verrá fornita una caratterizzazione dettagliata dei possibili rimpiazzi delle NOR e NAND Flash nei prossimi sistemi embedded o a multi-processore. Nel Capitolo 1 verrá introdotta la tecnologia di memorie non-volatili con un chiaro riferimento alla tecnologia Flash, fornendo indicazioni e suggerimenti ad un ipotetico system designer verso il passaggio alla nuova generazione di memorie. Nel Capitolo 2 verrá presentata una delle memorie della post-floating gate generation piú studiate nell'ultimo decennio: le memorie a cambiamento di fase (PCM). I risultati di una esaustiva caratterizzazione elettrica su questa tecnologia hanno permesso di scoprire fenomeni molto importanti quali la cinetica dell'operazione di cancellazione e altre potenziali minacce all'affidabilitá delle operazioni di scrittura sulle stesse. Un'intesa at-

tivitá di modeling ha poi permesso di validare i risultati sperimentali in proiezione verso una tecnologia scalata. Nel Capitolo 3 verrá mostrata una memoria per ambiente automotive: la SimpleEE p-channel. La caratterizzazione elettrica di questa memoria ha permesso di provarne la robustezza e allo stesso tempo ha aiutato a capire meglio il fenomeno dei bit erratici giá presente nelle passate NOR e NAND. Il Capitolo 4 mostrerá gli studi eseguiti su una memoria basata su micro sistemi meccanici (Nano-MEMS) provandone la capacitá di integrazione in ambienti estremamente ostili per l'affidabilitá come sistemi militari, geotermici e avionica spaziale. Uno studio dettagliato sui principi fisici alla base di questa memoria verrá proposto. Il Capitolo 5 mostrerá il successore naturale della tecnologia Flash: le NAND Flash a intrappolamento di carica. Questo tipo di memorie usa gli stessi principi fisici delle memorie a floating gate per la memorizzazione dell'informazione, cambiando peró la natura del mezzo in cui i dati vengono immagazzinati (i.e. trappole di nitruro di silicio). Le conclusioni e il riassunto dei risultati sulle ricerche eseguite sulle varie tipologie di memoria verranno presentati nel Capitolo 6. In conclusione, nell'Appendice A verranno mostrati i risultati di un'attivitá di ricerca recentemente iniziata sulla gestione dell'affidabilitá delle memorie ad alto livello, sfruttando i risultati degli studi delle memorie a cambiamento di fase.

*To Margherita, Marzia and Marco*

# Contents


**1 Introduction** **1**

1.1 Intrinsic Flash reliability limits . . . . . . . . . . . . . . . . . . . . . 2

1.2 Emerging issues of Flash technology . . . . . . . . . . . . . . . . . 10

**2 Phase Change Memories** **17**

2.1 Physics of chalcogenide-based memories . . . . . . . . . . . . . . 18

2.2 Electrical characterization of PCM arrays . . . . . . . . . . . . . . 19

2.3 Optimization of the writing operations . . . . . . . . . . . . . . . . 22

2.4 Modeling the SET kinetics . . . . . . . . . . . . . . . . . . . . . . . 33

2.5 The SET seasoning and the secondary shunt phenomena . . . . . 45

**3 P-channel SimpleEE Memories** **67**

3.1 Read performance . . . . . . . . . . . . . . . . . . . . . . . . . . . 69

3.2 Program/Erase speed . . . . . . . . . . . . . . . . . . . . . . . . . . 71

3.3 Endurance and data retention . . . . . . . . . . . . . . . . . . . . . 73

3.4 Disturbs robustness . . . . . . . . . . . . . . . . . . . . . . . . . . . 76

3.5 Evidence of the erratic bits . . . . . . . . . . . . . . . . . . . . . . . 79

**4 NanoMEMS Memories** **90**

4.1 Device physics . . . . . . . . . . . . . . . . . . . . . . . . . . 91

4.2 Memory architecture . . . . . . . . . . . . . . . . . . . . . . . . 95

4.3 Endurance characterization . . . . . . . . . . . . . . . . . . . . . 97

4.4 Data retention characterization . . . . . . . . . . . . . . . . . . . 101

4.5 Environmental and mechanical characterization . . . . . . . . . . . 103

**5 Charge Trapping NAND Flash Memories 106**

5.1 Electrical characterization . . . . . . . . . . . . . . . . . . . . . 107

5.2 Program operation characterization . . . . . . . . . . . . . . . . . 111

5.3 Room temperature retention . . . . . . . . . . . . . . . . . . . . 114

5.4 Disturbs characterization . . . . . . . . . . . . . . . . . . . . . . 116

5.5 Edge Wordline Disturb (EWD) phenomenon . . . . . . . . . . . . 121

**6 Conclusions 138**

**A System-level reliability of non-volatile memories 140**

A.1 General idea . . . . . . . . . . . . . . . . . . . . . . . . . . . . 140

A.2 The physical view on PCM . . . . . . . . . . . . . . . . . . . . . 142

A.3 A case study . . . . . . . . . . . . . . . . . . . . . . . . . . . . 143

A.4 Experimental results . . . . . . . . . . . . . . . . . . . . . . . . 146

# List of Figures


1.1 FLOTOX device and its equivalent capacitance model. . . . . . . 3

1.2 Band diagram sketch of tunneling effect. . . . . . . . . . . . . . . 4

1.3 Band diagram during a program operation: without traps (solid lines in the oxide region) and with traps (dashed lines in the oxide regions). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

1.4 Threshold voltage degradation during cycling of NAND Flash with different geometrical features. . . . . . . . . . . . . . . . . . . 7

1.5 Band diagram of the cell when programmed and not biased. The main mechanism for data loss is tunneling through the tunnel oxide. 9

1.6 Cumulative distribution of a NAND array on Program state. Both detrapping and SILC effects are appreciable on the time evolution. 10

1.7 Two traps assisted tunneling (2TAT) band diagram (left) and comparison between the SILC-2TAT model result versus the classical Fowler-Nordheim theory (right). . . . . . . . . . . . . . . . . . . . 10

1.8 Condition for the depletion region near the gate-drain overlap region of a nMOS transistor when the surface is accumulated with a low negative gate bias (a), and n+ region is depleted or inverted with high negative gate bias (b). . . . . . . . . . . . . . . . . . . . 13
1.9 Leakage current measured in a nMOS transistor for different Drain-Source voltages VDS in the subthreshold regime. When VGS approaches 0 V, for high VDS values the leakage current is due to GIDL. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
1.10 Bias conditions possibly activating GIDL effects on SSL transistors belonging to columns $BL_{i-1}$ and $BL_{i+1}$. . . . . . . . . . . . . 14
1.11 Measured program disturbs characteristics for 3 cells belonging to WL0, WL15 and WL31, respectively. NOP indicates the number of partial programming when multiple writing of a world line is allowed for specific applications. It can be observed that, for higher values of Vpass, the disturb on cell WL0 becomes significant. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
1.12 Electrostatic effect on the injection process: the Floating Gate (FG) potential changes when an electron is injected from the substrate (a) to the FG (b), reducing the tunnel oxide field and the tunneling current. . . . . . . . . . . . . . . . . . . . . . . . . . 16
1.13 Probability distributions of the injection statistics with and without the contribution of the RTN (Random Telegraph Noise). . . . 16

2.1 Cross section of a PCM cell (left) and its relative graphic representation (right). . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2.2 Typical PCM array topology with $pnp$ bipolar transistors as a selecting element. . . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.3 Waveforms applied to each cell in the array to perform the dynamic $I-V$ characterization and the extraction of $R_{set}$ and $V_{th}$. . 22
2.4 Result of the dynamic $I-V$ characterization. The two curves on display are the average over the whole cell population. . . . . . . 23
2.5 $R-I$ characterization. The curve on display represents the average over the whole cell population. . . . . . . . . . . . . . . . . . 24
2.6 Distribution of $V_{T1}$, $V_{T2}$, $V_s$, parameters in a Gaussian probability plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
2.7 Average value of $V_s$ versus the wordline index. Each point is the average over the columns belonging to the considered wordline. . 25
2.8 Average values of $V_{th}$, $V_{T1}$, $V_{T2}$ and $V_s$ during cycling. . . . . . . 25
2.9 Average values of $I_{max}$, and $R_{set}$ during cycling. . . . . . . . . . 26
2.10 a): MCW and BCW waveforms; b): SCOW and COW waveforms. 31
2.11 Logical flow used for waveforms optimization process. . . . . . . 31
2.12 $I_{cell}$ distribution dependency on $V_p$ applied within a MCW. . . . . 32
2.13 a): Variation of $\mu$ criterion on MCW, in relation to $V_p$ b): Variation of $\sigma$ criterion on MCW, in relation to $V_p$. . . . . . . . . . . . . . 32
2.14 $I_{cell}$ distribution dependency by applied $V_p$ within SCOW. Similar distribution are obtained using COW. . . . . . . . . . . . . . . . 33
2.15 a): Variation of $\mu$ criterion on SCOW, in relation to $S$ (log-scaled) b): Variation of $\sigma$ criterion on SCOW, in relation to $S$ (log-scaled). 33
2.16 a): Variation of $\mu$ criterion on COW, in relation to $t_H$ b): Variation of $\sigma$ criterion on COW, in relation to $t_H$. . . . . . . . . . . . . . 34

2.17 Sequence of pulses used for the experiments. The SET pulse conditions have been: $\Delta t$=(10ns), 25ns, 50ns and 100ns; $V_{SET}$=3.5V, 3.75V, 3.9V. A total number of SET pulses $N$ has been applied. . . 35
2.18 Average read current for each operating condition. . . . . . . . . . 36
2.19 Picture of the stages of the crystalline shunt formation and development. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
2.20 Read current distributions after the first 5 SET pulses. . . . . . . . 37
2.21 SET curves for the same cell measured 10 consecutive times ($\Delta t$=10ns). . . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.22 Average saturation current as a function of the operating conditions. 39
2.23 Read current distributions for each pulse duration. . . . . . . . . . 39
2.24 Read current distributions for each SET voltage. . . . . . . . . . . 39
2.25 Number of Slow cells as a function of their wordline (row) position within the array. . . . . . . . . . . . . . . . . . . . . . . . 40
2.26 Conductive GST model (left) and equivalent read path circuit (right). 40
2.27 Conductive GST model. A cylindric crystalline shunt approximation is used for model compactness. . . . . . . . . . . . . . . . . 42
2.28 Model output on different SET operative conditions. . . . . . . . 44
2.29 Seasoning effect as a function of cycles number, evidenced as an increase of the average (on 512 Kbits cells) $R_{RESET}$ and a decrease of $R_{SET}$. . . . . . . . . . . . . . . . . . . . . . . . . 45
2.30 Equivalent R-I characteristic of the PCM array. Characterization data were retrieved from cycle 1 to $10^6$. . . . . . . . . . . . . . . 46

2.31 PCM cell model for measurements and connection to the array. The voltage drop on the bitline MOS selector and on BJT cell selector is accounted. . . . . . . . . . . . . . . . . . . . . . . . . 47

2.32 Waveform sequences for cycling with different erasing schemes in seasoning investigation experiments. Used erasing schemes are depicted: MCW (solid), COW (dashed) and SCOW (dotted). Waveforms parameters such as $V_p$, $t_H$ and $S$ are also evidenced. . 48

2.33 Average variation of the array $R_{SET}$ calculated in modulus as the difference between two consecutive cycles measurements. . . . . 48

2.34 Seasoning effect evaluated on consecutive single cell measurements. 49

2.35 Average array $R_{SET}$ measured with a MCW erasing scheme. A comparison with the reference read only cycling is reported. . . . 50

2.36 Average array $R_{SET}$ measured with a COW erasing scheme. A dependance of the reduction of the $R_{SET}$ by the waveform parameter $t_H$ is evidenced. . . . . . . . . . . . . . . . . . . . . . . 50

2.37 Average array $R_{SET}$ measured with a SCOW erasing scheme. A dependance of the reduction of the $R_{SET}$ by the waveform parameter $S$ is evidenced. . . . . . . . . . . . . . . . . . . . . . . . . 51

2.38 Average array $R_{SET}$ dependency from $r_{sat}$ within a COW erasing scheme with two different $t_H$ used. Cycle dependence is also evidenced. . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

2.39 PCM cell structure evidencing the seasoning phenomenon both in RESET and SET state. . . . . . . . . . . . . . . . . . . . . . . 53

2.40 Example of RTN-like behaviors in four PCM cells observable by plotting $I_{read}$ after the Erase operation versus the number of cycles. Two levels are evidenced (i.e. $H$ and $L$ in fig. $d$) representing the presence/absence of the secondary path. . . . . . . . . . . . . 54
2.41 a) Equivalent read/write path circuit of a PCM cell. b) Program (RESET) and Erase (SET) waveforms exploited in this work. . . . 55
2.42 Average $I_{read}$ versus $V_{SET}$ characteristic. As shown, $V_{SET}$ = 3.75V holds the active material in a Partial-SET state. . . . . . . . 56
2.43 Schematic description of the secondary path creation kinetics. During the growth stage of the main percolation path, new GST crystallite grains may nucleate and subsequently join together, contributing to the secondary path formation. . . . . . . . . . . . 57
2.44 Two-states Markov chain used for modeling the secondary path presence/absence condition in PCM cells, where $\eta$ and $\theta$ represent the probability of remaining in the L and H status, respectively. . . 58
2.45 Distribution of the probabilities $\eta$ of remaining in the $L$ state for three pulse durations $t_H$ of the Erase operation. The cells population prone to create the secondary shunt is identified by the first Gaussian peak with $\eta < 0.5$. . . . . . . . . . . . . . . . . . . . . 59
2.46 Distribution of the probabilities $\theta$ of remaining in the $H$ state for three pulse durations $t_H$ of the Erase operation. . . . . . . . . . . 60
2.47 Scatter plot of $\eta$ versus $\theta$ state probabilities for $t_H$ = $2\mu$s (similar results are achieved for other pulse durations). The points accumulates in the center of the plot clearly evidencing a linear correlation between the two probability coefficients. . . . . . . . . 61

2.48 Log-normal probability plot of the $\Delta I$ distribution. . . . . . . . . 62

2.49 Numbers of state transitions for any set of cells characterized by the same average $L$ level within 200 program/erase cycles for the three erasing times $t_H$. Symbols denote the average number of state transitions, whereas the error bars represent the standard deviation from the mean value. . . . . . . . . . . . . . . . . . . . . . 63

2.50 $\Delta I$ shifts for any set of cells characterized by the same average $L$ level within 200 program/erase cycles for the three erasing times $t_H$. Symbols denote the average $\Delta I$ shift, where the error bars represent the standard deviation from the mean value. . . . . . . . 64

2.51 Measured (symbols) and calculated (lines) $I_{read}$ tail distributions after 200 Program/Erase cycles for the three pulse durations. $t_H$ = $1\mu$s: circles and dotted line; $t_H = 2\mu$s: squares and dashed line; $t_H$ = $4\mu$s: diamond and dashed/dotted line. The full line represents the Gaussian Model behavior for $t_H = 2\mu$s. Similar Gaussian behavior are found for other $t_H$ values. The inset shows experimental data and fitting of the main part of the distribution for $t_H = 2\mu$s (the two other $t_H$ cases are almost superimposed). . . . . . . . . . 66

3.1 Read characterization of SimpleEE modules lots. . . . . . . . . . 70

3.2 Read benchmark between Q3 and Q4 device families. . . . . . . . 71

3.3 Extracted SimpleEE array output gain at different temperatures. . 71

3.4 Program characteristics of the memory with respect to working temperature (left) and voltage exploited (right). . . . . . . . . . . 72

3.5 Erase characteristics of the memory with respect to working temperature (left) and voltage exploited (right). . . . . . . . . . . . . 72
3.6 Endurance characteristics of the memory modules at temperatures: -40°C, 25°C (room), 150°C and 170°C. . . . . . . . . . . . . 73
3.7 Data retention characteristics of the memory modules at bake temperatures: 100°C, 150°C (room), 200°C and 250°C. . . . . . . . . 75
3.8 Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a virgin device. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77
3.9 Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a cycled device (500K cycles at room temperature). . . . . . . . . . . 77
3.10 Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a cycled device (500K cycles at room temperature). . . . . . . . . . . 77
3.11 Distributions before and after disturb application for a sample cycled at hot temperature (150°C) for 500K cycles. . . . . . . . . . 78
3.12 p-EEPROM cell architecture (left) and table summarizing normal operating conditions (right). . . . . . . . . . . . . . . . . . . . . 80
3.13 Erased (left) and programmed (right) distribution of the Flash sample. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
3.14 Erased (left) and programmed (right) distribution of the p-EEPROM sample. . . . . . . . . . . . . . . . . . . . . . . . . . . 81
3.15 Example of erratic cell in the Flash sample. . . . . . . . . . . . . 81
3.16 Lognormal distribution of threshold voltage shift of erratic cells. . 82

3.17 Examples of erratic bits in p-EEPROM device after weak-program operation during cycling. . . . . . . . . . . . . . . . . . . 83
3.18 Examples of erratic bits in p-EEPROM device after weak-erase operation during cycling. . . . . . . . . . . . . . . . . . . . . . . 84
3.19 $V_T$ of an arbitrary p-EEPROM cell monitored during cycling using standard program waveforms. The same behavior can be observed on different cells and for more cycles. . . . . . . . . . . . 85
3.20 Shift distribution of the erratic bits in p-EEPROM device. . . . . . 85
3.21 Different values of energies are involved in the AHHI (Anode Hot Hole Injection) behavior of p-EEPROM and FLASH. In particular, AHHI on FLASH is fed by higher energies with respect to p-EEPROM, thus inducing more erratic behaviors. . . . . . . . . 86
3.22 Impact of the charge cluster on electron tunneling barrier. . . . . . 86

4.1 Simple cantilever element for physical principles evaluation. . . . 91
4.2 The teeter-totter concept used as a MEMS switch in the arrays characterized in this work (top) and a TEM picture of its integration (bottom). The geometrical dimension of the structure are evidenced. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.3 Read window of a Nanomech eNVM array. In this sample a read window of more than four orders of magnitude is shown. . . . . . 97
4.4 Benchmark of the $< R_{CNT} >$ between the room temperature endurance stress and the HTOL endurance stress. . . . . . . . . . 98

4.5 Insight of the $< R_{CNT} >$ characteristic under room temperature endurance stress (triangles) and under the HTOL endurance stress (squares). . . . . 99
4.6 $< R_{CNT} >$ monitoring during cumulative endurance experiment. 100
4.7 $< R_{CNT} >$ measured for different architectures of Nanomech eNVM. The values were calculated before and after the liquid-liquid experiment. . . . . 102
4.8 Reliability tests executed on NanoMEMS memories. No failing bits. . . . . 104
4.9 Relative resistance drift, average across full array, for all reliability tests performed on 6 alternative MEMS switch architecture variations. . . . . 105
4.10 Contact resistance behavior on Liquid-Liquid test with -55C - 150C temperature range. The distribution shows an improvement of the $R_{CNT}$ after the test. . . . . 105

5.1 Signals applied to the array for the program operation. . . . . 109
5.2 Signals applied to the array for the erase operation. . . . . 110
5.3 Program and erase distributions measured using $I_{ref}$=200nA. . . . . 110
5.4 Program and erase distributions measured using $I_{ref}$=400nA. . . . . 111
5.5 I-V characteristics on the programmed state of an arbitrary array string. . . . . 112
5.6 I-V characteristics on the erased state for same string. . . . . 113
5.7 Wordline average threshold voltage dependency. . . . . 114

5.8 Bitline even/odd behavior evidenced on wordline average threshold voltage. . . . 115
5.9 Staircase programming waveform. . . . 116
5.10 Single block average program characteristic. Even/Odd bitlines separate contributions have been evidenced. . . . 117
5.11 Evolution of the threshold voltage distributions of an entire array block during a programming ramp. . . . 118
5.12 Evolution of the threshold voltage distributions of cells belonging to even bitlines only during a programming ramp. . . . 119
5.13 Evolution of the threshold voltage distributions of cells belonging to odd bitlines only during a programming ramp. . . . 119
5.14 Cells number belonging to a threshold voltage bin after application of the post-analysis data compactation. . . . 120
5.15 Threshold voltage distribution using the ramped waveform with 1 V steps on even bitlines. . . . 120
5.16 Threshold voltage distribution using the ramped waveform with 1 V steps on odd bitlines. . . . 121
5.17 Retention characteristics of the array. The average retention of the whole array, the average retention of even bitlines only and the average retention of odd bitlines only are evidenced by black, green and red curves respectively. . . . 122
5.18 Threshold voltage distribution of the full device in RTB experiments. 123
5.19 Threshold voltage distribution of the array even bitlines in RTB experiments. . . . 124

5.20 Threshold voltage distribution of the array odd bitlines in RTB experiments. . . . 125
5.21 Impact of read disturb on both programmed and erased cells. . . . 125
5.22 Threshold voltage measured after 1, 2, 5, ..., 200 program pulses applied simultaneously on WL1, WL16, and WL30. . . . 126
5.23 Net effect (disturb) of a program disturb measured after 1, 2, 5, ..., 200 disturb pulses applied on WL1. . . . 126
5.24 Evolution of disturb on WL0 and WL2 with the number of program disturb pulses applied to WL1. . . . 127
5.25 Block to block interference evaluated by monitoring both programmed and erased cells while applying program disturb pulses on WL33 of an adjacent block. . . . 128
5.26 Standard NAND array architecture. The figure represents a single block of the CT array considered in this analysis. . . . 129
5.27 ISPP Pulse characteristics exploited in this work. The duration per pulse is 4$\mu$s. . . . 130
5.28 Bias conditions possibly activating the GIDL effect from GSL transistors belonging to columns BL$_{i-1}$ and BL$_{i+1}$. . . . 131
5.29 Bias conditions possibly activating the GIDL effect from GSL transistors belonging to columns BL$_{i-1}$ and BL$_{i+1}$ and the $V_{pass}$ disturb in unselected cells belonging to columns BL$_i$. . . . 131

5.30 Band structure of the CT cells (excluding the TaN/Ti/TaN metal gate) considered in this work. For high biases electron B is free to tunnel into $SiO_2$ and then gets trapped into $Si_3N_4$ layer. For low biases electron A is still able to get trapped, although the energy levels of the bands are shifted upwards. . . . . . . . . . . . . . . . 132

5.31 EWD dependence on the exploited ISPP voltages during a program operation. $V_{pass}$ has been fixed at 8 V for the unselected wordlines. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134

5.32 EWD dependence on the number of performed NOP on WL1. $V_{pass}$ has been fixed at 8 V for the unselected wordlines. . . . . . . 135

5.33 EWD as a function of the NOP number. $V_{pass}$ has been varied in order to evaluate the dependence of the phenomenon on the pass voltage. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136

5.34 EWD dependence on the number of write cycles performed on the CT NAND array . . . . . . . . . . . . . . . . . . . . . . . . . . 137

5.35 EWD as a function of the NOP number. A dependence on the device aging is evidenced. $V_{pass}$ has been fixed to 8 V for the unselected wordlines. . . . . . . . . . . . . . . . . . . . . . . . . 137

A.1 Calculated write throughput comparison between the different erasing schemes applied within a PCM. . . . . . . . . . . . . . . . 147

A.2 Mean write throughput calculated with different writing scenarios. The uniform traffic condition is compared with 50 writings on the data reliable area and 100, 150, 200, 300 writings on the high speed area. . . . . . . . . . . . . . . . . . . . . . . . . . . 147

A.3 Error probability comparison between single PCM data partitions and PCM comprising the whole array . . . . . . . . . . . . . . . . 148

# List of Tables

2.1 Comparison resume of the analyzed waveforms . . . . . . . . . . . 32
2.2 Seasoning impact resume with different waveform . . . . . . . . . 51
2.3 Parameters used for the statistical modeling of the gaussian part of the $I_{read}$ distributions extracted from the experimental data at cycle 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
2.4 Parameters used for the fitting of the bimodal gaussian distributions of $\eta$ and $\theta$ state probability. . . . . . . . . . . . . . . . . . . . 65
2.5 Parameters used for the calculation of the $\Delta I_i$ shift. . . . . . . . . 65

A.1 Erasing schemes comparison resume. . . . . . . . . . . . . . . . . 144

# Chapter 1

# Introduction

The last decade of research in non-volatile memories, both from academic and industrial perspective, arose potential threats to the so-called ”happy scaling” era, where the only way to achieve memory density, performance and enhanced reliability exploited a furious geometrical shrinking of the memory cells dimensions.

The mainstream for non-volatile memory technology in these years has been represented by the floating gate-based memories such as NOR and NAND Flash. These memories relies on a manufacturing process and on a physical storage solution which reached so far the maturity level. Indeed, both kind of Flash architectures are reliably used either as a code memory (NOR Flash) for storing BIOS, applications and operating systems for mobile environments or as a data memory (NAND Flash) relying on the high storage capacity offered by these memories. The NAND Flash market segment gained particular momentum during these decades thanks to the blast of the portable applications such as digital cameras, smartphones, mp3 players, an so on. Unfortunately, these largely used components in embedded systems and MPSoCs (Multi Processing System on Chip) are

facing difficulties in keeping up the performances and the reliability requested by a non-volatile memory components.

That explain the gained interest in these years on the so-called universal memories which represent the intersection between the performances of the traditional semiconductor volatile memories such as DRAM and SRAM and the reliability offered by a non-volatile technology. The 2010 global market for emerging non-volatile random access memory products was projected to have reached 115 million USD. This market will increase to 1,590 million USD by 2015 showing an average annual growth rate of 69% per year from 2010 to 2015. In this thesis several innovative memory concepts will be analyzed as potential alternative to Flash technology. To name few of them it has been studied the Phase Change memories, SimpleEE memories, NanoMEMS memories and Charge Trapping memories.

Now it will provided a quick overview on what are the intrinsic limit of the floating gate technology and what are the emerging issues that are forcing system designers to change their mind in favor of such innovative memory technologies.

## 1.1 Intrinsic Flash reliability limits

The concept element of traditional Flash memory cell is a metal oxide semiconductor device with a floating gate electrically isolated by means of a tunnel oxide and of an interpoly oxide as sketched in Fig.1.1 [1]. The former oxide plays a basic role for the control of the device threshold voltage whose value represents, from a physical point of view, the stored information. Electrons transferred into the floating gate give a threshold voltage variation $\Delta VT = \frac{-Q}{Cpp}$. In quiescent conditions, thanks to the two oxides, the charge stored should not leak away, thus granting

the nonvolatile paradigm fulfillment. Oxides are available in different material depending on the Front End of Line (FEOL) process. The common materials are: pure silicon dioxide ($SiO_2$) for tunnel oxides and a stack of Oxide-Nitride-Oxide ($SiO_2$-$Si_3N_4$-$SiO_2$) for interpoly oxides.

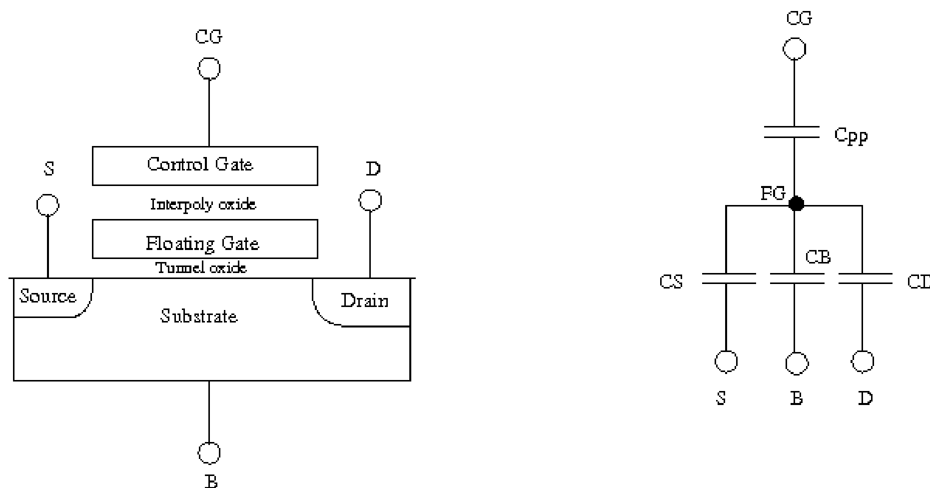



Figure 1.1: FLOTOX device and its equivalent capacitance model.

The cells are rearranged into an array organization [2] in which we can define three basic subdimensions: string, page and block. The physical mechanism used for both injecting and extracting electrons to/from the floating gate is the Fowler-Nordheim (FN) tunneling [3]. High electrical field applied to the tunnel oxide (FOX is almost 10 MV/cm) allows for electron transfer across the thin insulator to the floating gate. In NAND architectures the electronic tunneling involves the MOS channel/substrate and requires appropriate biasing of control gate and bulk terminals (see Fig.1.2), while drain and source are left floating. With respect to the Channel Hot Electron mechanism exploited for cell programming in NOR architectures, FN tunneling requires higher voltages, therefore much more complex charge pumps and higher programming times. In addition, the use of the same

mechanism for both programming and erasing the cells exposes the tunnel oxide to a larger degradation, with possible reliability effects. These drawbacks, however, are compensated by the much lower current (orders of magnitudes) required by the writing operations significantly improving power consumption and/or programming parallelism.

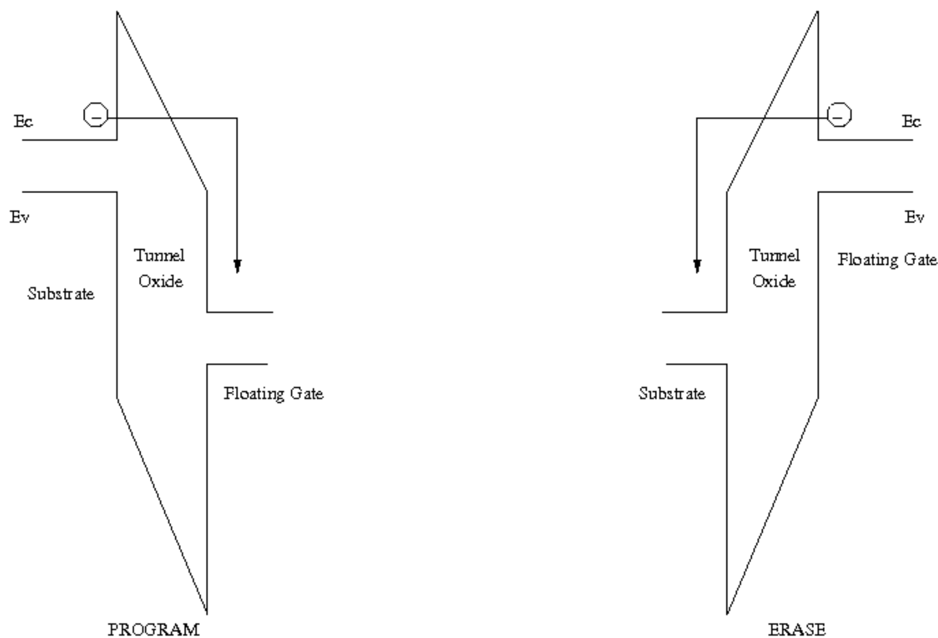



Figure 1.2: Band diagram sketch of tunneling effect.

During its lifetime a Flash module undergoes a large number of Program/Erase cycles. Every cycle involves very high electrical fields applied to the tunnel oxide. The reliability of the entire memory requires that the tunnel oxide is able to operate correctly under stress conditions. It is obvious that huge efforts are to be spent to determine the right process for the tunnel oxide creation (in terms of thickness, material, growth, defectivity, interface, ) in order to achieve a successful and reliable Flash technology. However, the scaling of the geometries is not helping these efforts, most of the time forcing to consider new radical design

approaches.

The main reliability issues for Flash technology are represented by the endurance and the data retention features.

The endurance of a memory module is defined as the minimum number of Program/Erase cycles that the module can withstand before leading to a failure. The erased and programmed distributions must be suitably separated, in order to correctly read the logical state of a cell. The difference between EV (Erase reference Value) and PV (Program reference Value) is defined as the "read margin window". However, keeping a correct read margin is not sufficient to guarantee a correct read operation: if during its lifetime the threshold voltage of an erased cell exceeds the EV limit and approaches 0 V, the current flowing through the cell may be not high enough to be identified as "erased" by the reading circuitry, thus producing a read error. Similarly, a programmed cell could be read as "erased" if its threshold voltage becomes lower than PV and approaches 0 V. As for the programmed distribution, it is also important that the upper threshold limit does not increase significantly with time, since a too high threshold can block the current flowing through the strings during reading operations. FN tunneling leads intrinsically to oxide degradation [4]. As a result of consecutive electron tunneling, traps are generated into the oxide [5]. When filled by electrons, charged traps can increase the potential barrier thus reducing the tunneling current, as shown in Fig.1.3. Since the programming and erasing pulses feature constant amplitude and duration, less charge is transferred to and from the floating gate causing an efficiency reduction of both the program and erase operations. A narrowing of the read margin window is then expected. The charge trapped inside the oxide also produces a threshold shift $\Delta VT$ directly proportional to its amount. The VT

shift is symmetrical and increases the threshold voltage of both erased and programmed cells (see Fig. 1.4). Writing waveform optimization can help in limiting the trapped charge. For example, it has been shown that the window closure can be reduced by using low voltage erasing pulses able to remove the charge accumulated in the oxide. As shown in Fig.1.4, the threshold voltage shifts increase with the number of program/erase operations until an endurance failure occurs. Threshold voltage shifts could be recovered by applying specific procedures, that however result not convenient when comparing their effects with the required architectural overheads and time consumption. As evidenced in Fig.1.4, the most critical effect is the increase towards 0 V of the erased threshold. To check whether all cells of a sector have been correctly erased (all threshold voltage must be below 0 V), an erase verify procedure is applied after any erase operation. It consists in a particular read operation performed by driving simultaneously all the word lines of the sector at 0V: if the read current in a bitline is 0, it means that at least a cell blocked the current flow because its threshold voltage was higher than 0 V. The entire block is marked as bad block by the memory controller and no longer addressed [2].

The retention concept instead, is the ability of a memory to keep a stored information over time with no biases applied. Electron after electron, charge loss could slowly leading up to a read failure: a programmed cell can be read as erased if its VT shifts below 0 V. The intrinsic retention is mainly limited by tunneling (see Fig.1.5) through the oxide even if thermionic emission mechanisms over the barrier can be considered. Recent studies [6] demonstrated that an oxide thickness of 4.5 nm is enough for granting theoretical intrinsic retention of 10 years, which is the minimum limit imposed by present standards. The cell retention worsen with

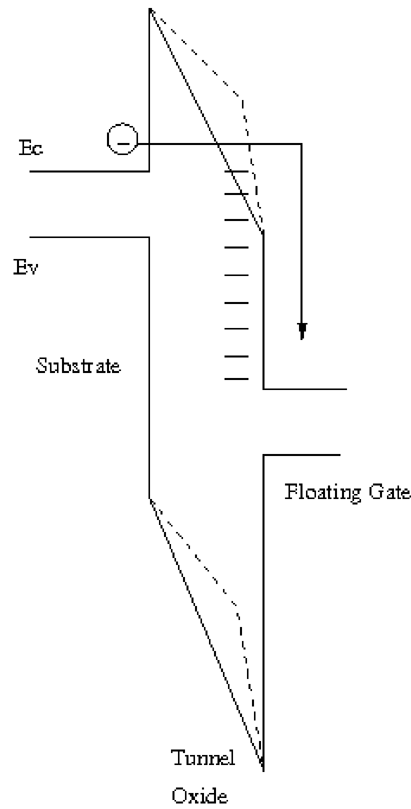



Figure 1.3: Band diagram during a program operation: without traps (solid lines in the oxide region) and with traps (dashed lines in the oxide regions).

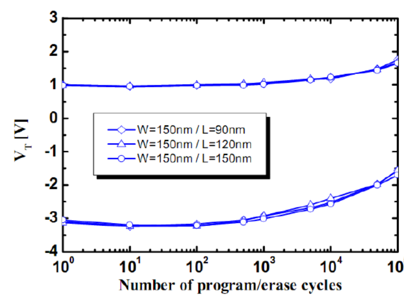



Figure 1.4: Threshold voltage degradation during cycling of NAND Flash with different geometrical features.

memory cycling and this effect is appreciable as a reduction of the VT levels as sketched in Fig.1.6 showing the time evolution of the cumulative VT distribution of programmed cells. Charge loss from the floating gate moves the VT distribu-

tion towards lower values. In additions, a tail in the lower part of the distribution indicates that a small percentage of cells is losing charge faster than average. The rigid shift of the cumulative VT distribution can be related to the oxide degradation within the oxide and at the Si - $SiO_2$ interface. As described early, successive electron tunneling leads intrinsically to oxide degradation, characterized by traps generation. These traps may be responsible for charge loss from the floating gate towards the silicon substrate. In fact, an empty trap suitably positioned within the oxide can activate trap assisted tunneling (TAT) mechanisms characterized by a significantly higher tunnel probability with respect to a triangular barrier unmodified by the trap presence. In addition, an electron trapped within the oxide during writing operations and responsible for the VT increase leading up to endurance failures may be detrapped when the program pulse is switched off, when the cell is read or even when the cell is not addressed. As a result, the empty trap may enhance the TAT phenomenon (assuming a positive charged trap) and, in addition, it can increase the electron field at the Floating gate-tunneling oxide interface thus raising the electron tunnel probability. The required activation energy for detrapping has been calculated in several experiments about 1.1 eV [7]. It is clear that these mechanisms are strongly related to the oxide degradation and therefore data retention decreases with the number of applied writing pulses. Results of retention stresses showed that the cells contributing to the tails of the VT distribution are characterized by a leakage current larger than the average at the same stress level [8]. The Stress Induced Leakage Current (SILC) of these cells is attributed to TAT process of carriers through the tunnel oxide traps. By modeling the behavior of the tail cells it has been shown that a single TAT model is not consistent with the observed leakage current. The charge loss of these cells is attributed to a

tunneling process assisted by two traps (2TAT) - see Fig. 1.7. The characteristics of this phenomenon are the opposite with respect to detrapping: low activation energy (0.1 - 0.3 eV) and strong field acceleration. SILC affects a larger number of cells after each writing cycle, thus confirming that also this abnormal charge loss is driven by oxide degradation. The position of leaky cells within the array, however, does not show any clusterization which could be related to a technological process. In addition, the abnormal leakage is not permanent: a cell can exhibit a SILC effect after a cycle and behave as normal when erased and reprogrammed. This result is consistent with a model requiring 2 traps with suitable locations within the oxide to activate a SILC effect. Some of these cells can suffer an erratic behavior [9] due to a trap annihilation/reactivation process. It has been proved, however, that traps are annealed over 85C so high temperatures partially mitigate this effect [10].

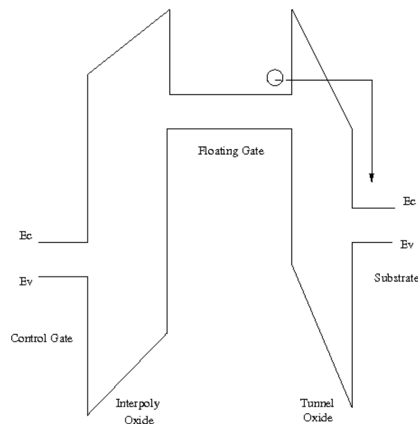



Figure 1.5: Band diagram of the cell when programmed and not biased. The main mechanism for data loss is tunneling through the tunnel oxide.

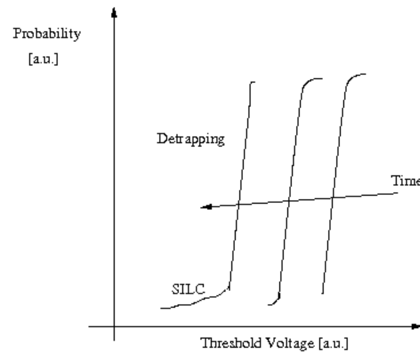



Figure 1.6: Cumulative distribution of a NAND array on Program state. Both detrapping and SILC effects are appreciable on the time evolution.

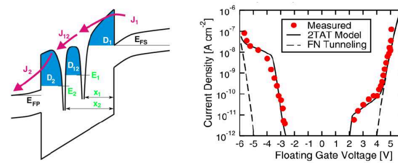



Figure 1.7: Two traps assisted tunneling (2TAT) band diagram (left) and comparison between the SILC-2TAT model result versus the classical Fowler-Nordheim theory (right).

## 1.2 Emerging issues of Flash technology

As the technology scaling of Flash proceeds, emerging reliability threats have to be considered in addition to the traditional issues presented in the previous sections. The influence of phenomena like the Gate-Induced Drain Leakage and the injection statistics cannot be neglected during the validation phase of a technology since they introduce new disturb types during the common operations of a module.

The Gate Induced Drain Leakage (GIDL) is a major leakage mechanism occurring in OFF MOS transistors when high voltages are applied to the Drain and it is attributed to tunneling taking place in the deep-depleted or even inverted region underneath the gate oxide [11]. When the gate is biased to form an accumulation

layer at the Silicon surface, the surface behave like a p region more heavily doped than the substrate. This effect causes the depletion layer of the junction at the surface to be more narrow than elsewhere. The narrowing of the depletion layer near the surface increases the local electric field thus enhancing high field effects near that region. When the gate voltage in a nMOS transistor is 0 V (or below) and the Drain is biased at a high voltage, the n+ drain region under the gate can be depleted or even inverted (see Fig.1.8). This effect causes a peak field increase leading up to high field phenomena such as avalanche multiplication and band-to-band tunneling. The band-to-band tunneling probability can also be increased by the presence of surface traps, resulting in a band-trap-band tunneling. As a result of these effects, minority carrier are emitted in the drain region underneath the gate and swept laterally to the substrate which is at a lower potential, thus completing a path for the leakage current. Fig.1.9 shows the sub-threshold characteristics in an nMOS transistor for different Drain voltages [12]. On NAND arrays the GIDL has been found to produce erroneous programming in specific cells, that is those belonging to WL0, adjacent to the SSL transistor. The SSL transistor is OFF during programming and GIDL effects may be present if its drain is driven at high voltages. This situation occurs because of the self-boosting techniques adopted to prevent programming. As shown in Fig.1.10, if WL0 is driven at Vpgm to program the cell in the central column, the channel voltage of the cells sharing the WL0 line are to be raised to prevent their programming. Because of the self-boosting technique, the source terminal of those cells (therefore also the drain node of their adjacent SSL transistors) are raised to values higher than VCC, thus leading up to bias configurations activating GIDL effects. The electron-holes pair generation follows and the generated electrons are accelerated at the SSL - WL0

space region and can be injected as hot electrons in the floating gate of WL0 cells. Also the DSL transistors may trigger GIDL mechanims when boosting effects take place on cells sharing WL31, but since the voltage applied to the DSL gate line is VCC, the consequences suffered by cells belonging to WL31 are much lower with respect to those sharing WL0. Fig.1.11 shows experimental data for 3 different cells: WL0 and WL31, potentially affected by GIDL disturbs, and a reference one located in the middle of the string. This unwanted programming in specific cells is supposed to burden in scaled architectures, since a reduced separation between WL0 and SSL lines will increase the accelerating fields for hot electrons. Recently, to mitigate GIDL effects, it has been proposed to introduce two dummy world lines to separate the two select transistors and the effective string of cells. To reduce the impact of these two additional word lines, longer strings of 64 cells have been proposed to improve area efficiency. A higher number of cells in series (64 cells vs 32 of standard architectures), however, increase the string resistance so that word-line voltage modulation are required during read and programming thus ensuring that proper voltage levels are applied depending on the cell location within the string. For instance, a higher word-line voltage is used when accessing a cell near the top of the string (close to the bit line) to compensate for the string resistance.

The aggressive scaling of Flash memories, necessary to improve their performances, brings attention also to the discrete nature of the charge stored in the floating gate. The number of electrons determining the stored information, in fact, continuously decreases with the tunneling area dimension. When only few electrons control the cell state (i.e. setting the transistor threshold voltage), their statistical fluctuations determine a non-negligible spread. These fluctuations may

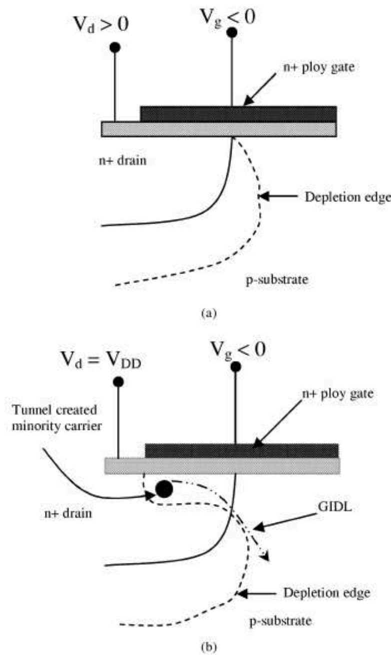



Figure 1.8: Condition for the depletion region near the gate-drain overlap region of a nMOS transistor when the surface is accumulated with a low negative gate bias (a), and n+ region is depleted or inverted with high negative gate bias (b).

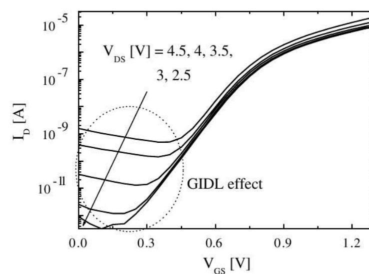



Figure 1.9: Leakage current measured in a nMOS transistor for different Drain-Source voltages VDS in the subthreshold regime. When VGS approaches 0 V, for high VDS values the leakage current is due to GIDL.

be attributed to the statistics ruling the electron injection into the floating gate during program or to the electron emission from the floating gate during erase or retention, both related to the granular nature of the current flow [13]. A slight variation of the number of electrons injected during programming may produce VT

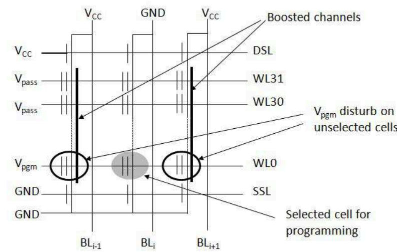



Figure 1.10: Bias conditions possibly activating GIDL effects on SSL transistors belonging to columns $\mathrm{BL}_{i-1}$ and $\mathrm{BL}_{i+1}$.

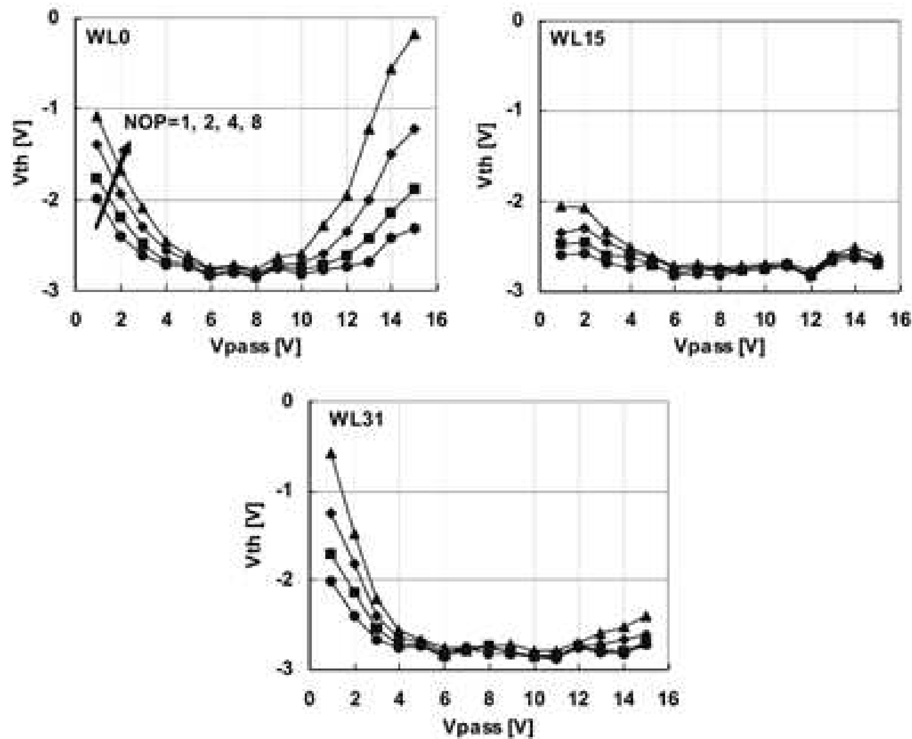



Figure 1.11: Measured program disturbs characteristics for 3 cells belonging to WL0, WL15 and WL31, respectively. NOP indicates the number of partial programming when multiple writing of a world line is allowed for specific applications. It can be observed that, for higher values of Vpass, the disturb on cell WL0 becomes significant.

variations possibly leading to errors in MLC architectures. Cells in NAND MLC architectures are programmed by using a staircase voltage on the control gate.

This kind of waveform allows achieving a constant current Fowler-Nordheim tunneling: by increasing the wordline voltage after each short pulse it is possible to compensate the field reduction that follows the electrons injection into the floating gate (see Fig.1.12). For sufficiently large step numbers, a linear VT increase is obtained, with a $\Delta VT$ per step almost equal to the applied voltage step but leading to a significant VT distribution enlargement (see Fig.1.13). This is due to the programming current convergence toward an equilibrium stationary value, corresponding to an average number of electrons transferred to and from the floating gate for each step [14]. The discrete nature of the charge flow introduces, for each cell, a statistical spread contribution to the resulting VT after each step. It has been evidenced that the threshold voltage variation spread, indicated as $\sigma_{\Delta VT}$, depends only on the parameter Vstep of the programming waveform (i.e. on the injected charge per step $qn$) and not on the pulse duration and on the number of pulses to achieve $\Delta VT$.

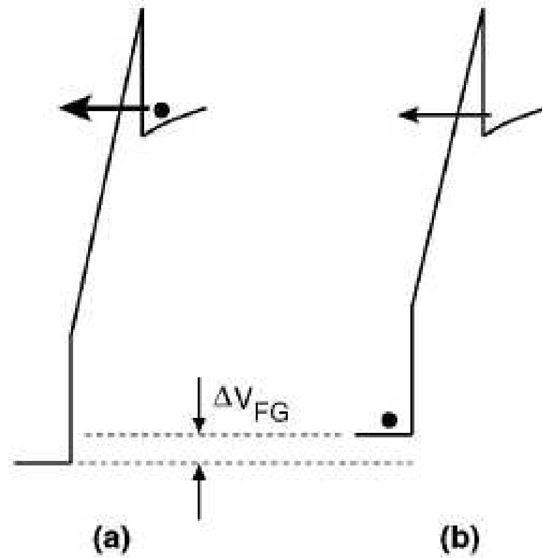



Figure 1.12: Electrostatic effect on the injection process: the Floating Gate (FG) potential changes when an electron is injected from the substrate (a) to the FG (b), reducing the tunnel oxide field and the tunneling current.

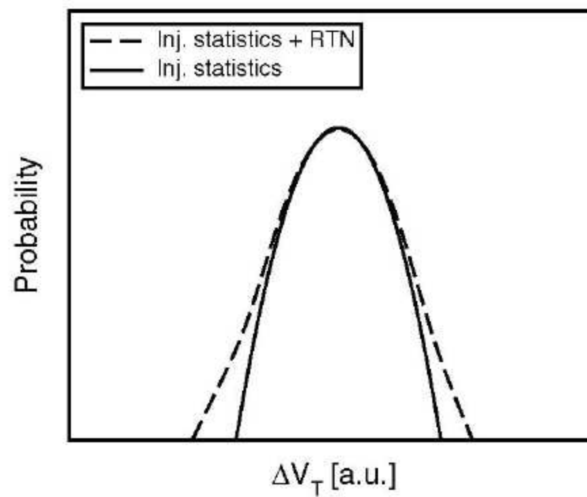



Figure 1.13: Probability distributions of the injection statistics with and without the contribution of the RTN (Random Telegraph Noise).

# Chapter 2

# Phase Change Memories

Phase Change Memory (PCM) is one of the most promising emerging Non-Volatile Memory (NVM) technology thanks to fast writing operations, long endurance and very good radiation hardness. PCMs offer direct write (any bit can be independently reprogrammed with no need for block erasing), improved write throughput versus NOR-based memories and random access time versus NAND-based memories, as well as the potential to be scalable beyond Flash technology [15]. Moreover, PCMs ensure high endurance and good compatibility with standard CMOS fabrication processes. Although a relevant effort is still being put on design optimization and material analysis, the PCM technology has today reached a significant level of maturity so that test chips containing multimegabit arrays of cells are commonly available for electrical characterization purposes. As a result, a significant amount of statistical information can be provided to the reliability/cell design engineer.

## 2.1 Physics of chalcogenide-based memories

In PCMs, the storage device is made of a thin film of chalcogenide alloy (in our case, Ge2Sb2Te5, GST) [16]. This material can reversibly change between an amorphous (high impedance, RESET state) and a polycrystalline (low impedance, SET state) phase when thermally stimulated via Joule heating, thus allowing information storage. The phase conversion of a storage element is obtained by appropriately heating (by means of electrical pulses applied to a suitable heater element) and then cooling a small, thermally isolated portion of the chalcogenide material. Fig. 2.1 shows a cross section of a PCM cell and its relative graphic representation. Once the chalcogenide material melts, it completely loses its crystalline structure. When rapidly cooled, the chalcogenide material is locked into its amorphous state (to this end, the cooling operation rate has to be faster than the crystal growth rate). To switch the memory element back to its crystalline state, the chalcogenide material is heated to a temperature between its glass transition temperature and its melting point temperature [17]. In this way, nucleation and micro-crystal growth occur in tens of nanoseconds, thus leading to a (poly)crystalline state. From above, it is apparent that the storage element can be modeled as a programmable resistor (high resistance = logic 0; low resistance = logic 1). Reading a cell basically consists in measuring the resistance of the addressed storage device. To this purpose, a predetermined voltage is forced across the storage element of the selected cell, and the resulting current flow is sensed.

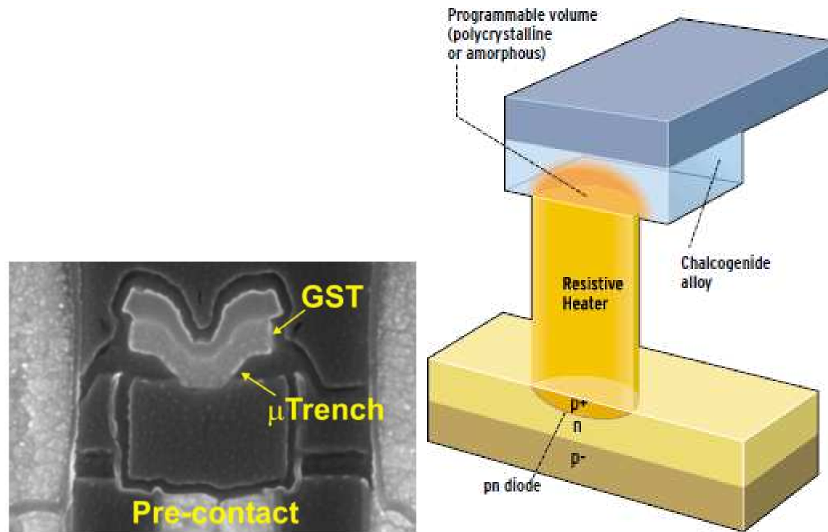



Figure 2.1: Cross section of a PCM cell (left) and its relative graphic representation (right).

## 2.2 Electrical characterization of PCM arrays

All characterization measurements have been performed on a 512 Kbit population of PCM cells with the aid of a dedicated testing equipment. Fig. 2.2 depicts a detail of the memory cell integrated in an array environment. A $pnp$ Bipolar Junction Transistor (BJT) is employed as the cell selector in order to minimize silicon area occupation, and, hence, improve data storage density. Read voltages and write voltages are transferred to the cells through a select nMOS transistor supplied by $V_{pp} = 4.8V$. A heater resistance $R_h$ is in series with $R_{GST}$, representing the resistance of the GST active element. Details on the 180 nm PCM technology characterized in this thesis can be found in [18]. The read current of the cell can be measured in a Direct Memory Access mode by applying a read voltage at the control gate of the select nMOS transistor. The SET and RESET operations are performed similarly by applying a SET/RESET pulse at the control

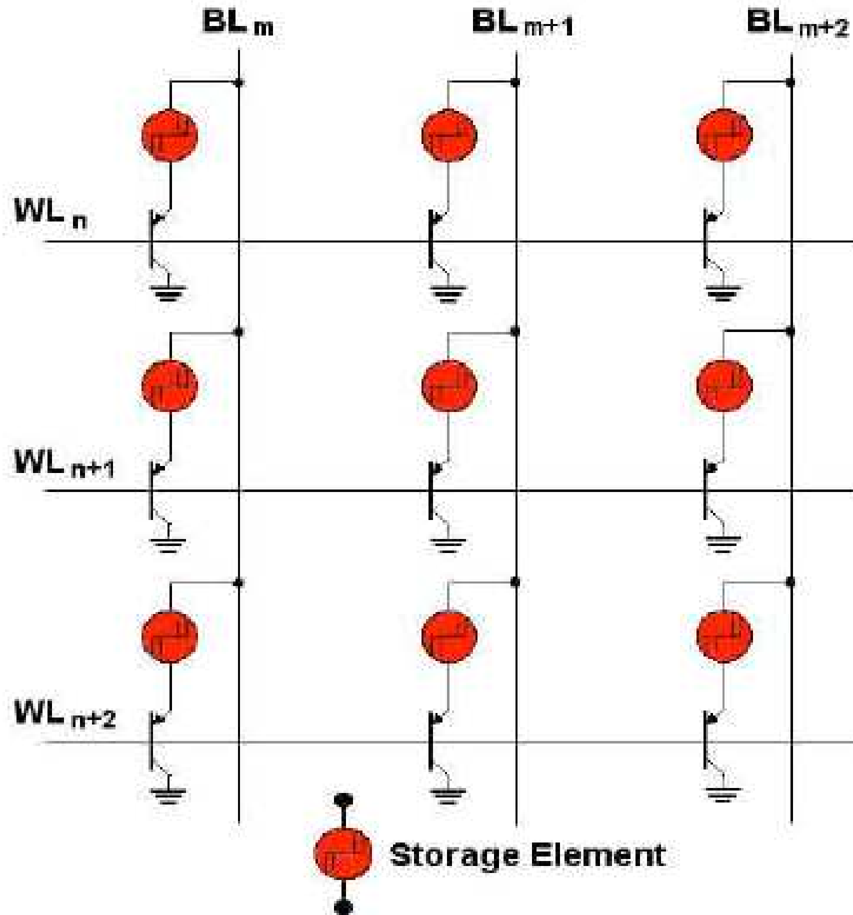



Figure 2.2: Typical PCM array topology with $pnp$ bipolar transistors as a selecting element.

gate of the select nMOS transistor.

A set of Relevant Electrical Parameters (REP) has been defined and characterized on test chip arrays [19]: $V_{th}$ , $V_{T1}$, $V_{T2}$, $V_s$, $I_{max}$, and $R_{set}$. REP are intended to provide a synthesis of results obtained from key characterization procedures which are commonly adopted on single cell study. The synthesis becomes important when the same procedures are applied to a population of cells. The dynamic $I - V$ characteristics describe both the SET curve and the RESET curve which

are measured by applying the sequence of pulses shown in Fig. 2.3a-b. Fig. 2.4 shows the $I - V$ curves obtained by averaging the $I - V$ curves of all the cells of the array. These curves can be described in terms of at least two parameters which can be easily extracted from the data: $R_{set}$ and $V_{th}$. When a complete IV characterization is performed for each cell of a 5I2K population, 80MB of data are collected, whereas the extraction of REP allows a more easy analysis on 4MB of synthetic data only. The $R-I$ curve describes the read current (or equivalently the resistance) as a function of the SET pulse amplitude. The $R-I$ curve is measured by applying the sequence of pulses shown in Fig. 2.3c. Fig. 2.5 shows the $R - I$ curve obtained by averaging the $R - I$ curves of all the cells of the array. These curves can be described in terms of four parameters which can be easily extracted from the data: $V_{T1}$, $V_{T2}$, $V_s$, $I_{max}$.

When a complete $R - I$ characterization is performed for each cell of a 512 Kb population, 60MB of data are collected, whereas the extraction of REP allows a more easy analysis on 8MB of synthetic data only. The usage of synthetic parameters allows the analysis of statistical behavior of a cell population as shown in Fig. 2.6. From these graphs it is possible to evidence some anomalous behavior like, for example, the one shown in the $V_s$ distribution which exhibits a tail at its lower left corner. A further analysis of $V_s$ over the whole cell population allows to show that this feature is related to the array architecture and in particular to the cell distance from the bitline strap contact (see Fig. 2.7). Standard reliability cycling tests can be performed and results effectively displayed by using REP. Fig. 2.8 shows the behavior of the cell population displaying the average values of $V_{th}$, $V_{T1}$, $V_{T2}$ and $V_s$ during a 1M SET/RESET cycle experiment. This figure summarizes significant results. The switching threshold value decreases during

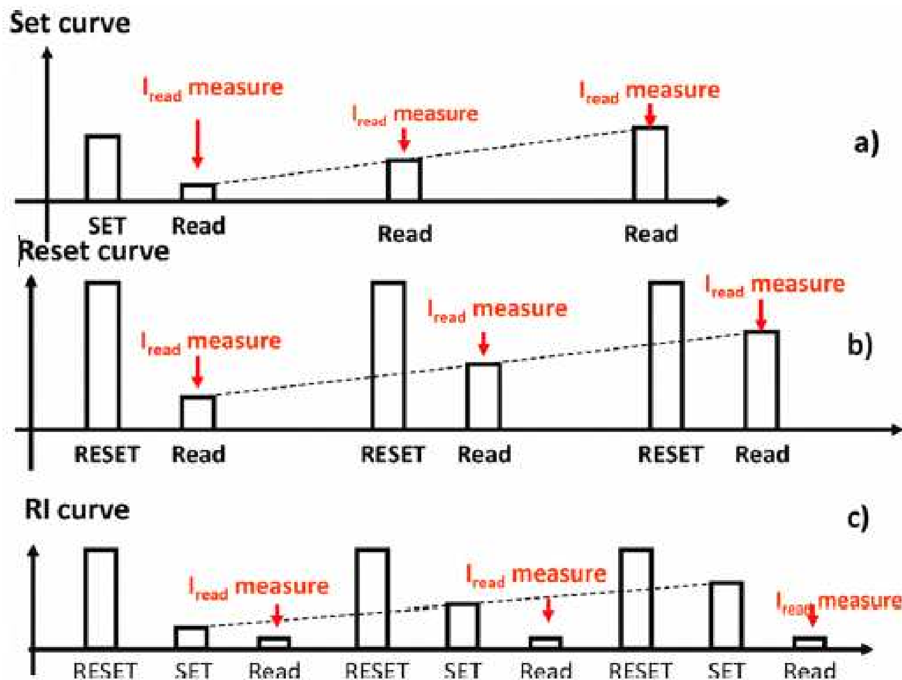



Figure 2.3: Waveforms applied to each cell in the array to perform the dynamic $I-V$ characterization and the extraction of $R_{set}$ and $V_{th}$.

cycling although only after some hundred thousands cycles. This phenomenon has to be taken into account in the design of the reading voltage. There is also an increase of $V_{T2}$ which can interfere with the RESET operation. Fig. 2.9 shows the increase of the maximum current of the $R-I$ curve and the decrease of the SET resistance with cycling. This result clearly indicates that cells undergo a better crystallization with cycles, a phenomenon sometimes called cell seasoning.

## 2.3 Optimization of the writing operations

Different voltage waveforms are used for both programming (or RESET) and erasing (or SET) operations in PCM, referring to the GST amorphization and crystallization processes, respectively. The main differences consist on the voltage, the

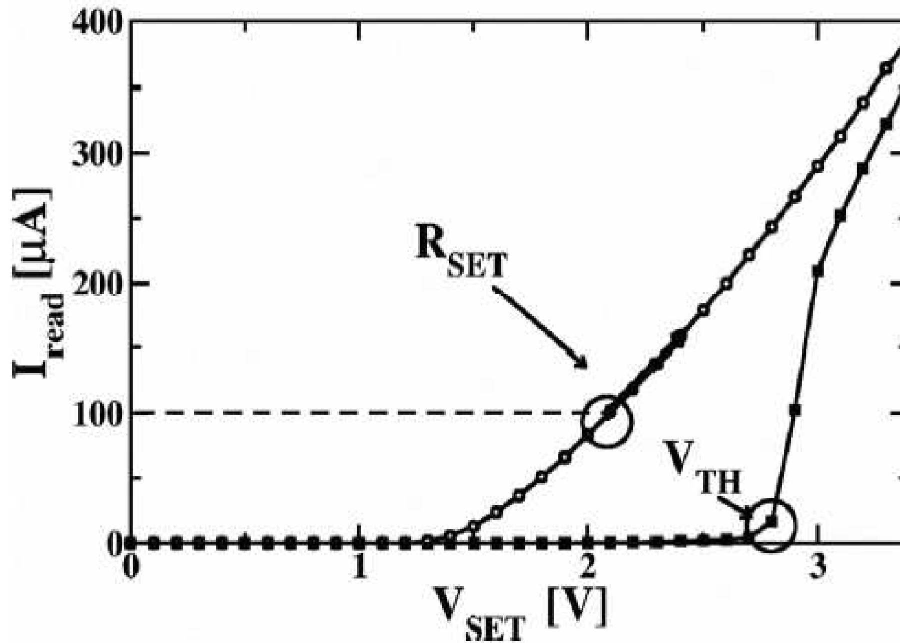



Figure 2.4: Result of the dynamic $I - V$ characterization. The two curves on display are the average over the whole cell population.

length, and the shape of the applied waveforms. Typically, the programming operation is achieved by using a voltage pulse, bringing the GST above its characteristic melting point, then cutting away the pulse rapidly, in order to melt and then amorphize the GST. The erasing operation, instead, can be achieved in many ways, due to multiple possibilities for achieving crystallization of active chalcogenide. Many reliability issues of multimegabit memory arrays have been tackled in the last few years, mainly concerning the used phase change materials and the cell geometrical parameters. However, a detailed analysis of the impact of waveform parameters on the overall reliability of large arrays is still missing, in particular for the erasing case where the randomness of the crystal nucleation physics in chalcogenide materials plays a basic role [17].

The waveform applied to erase a PCM cell has the objective to create linked

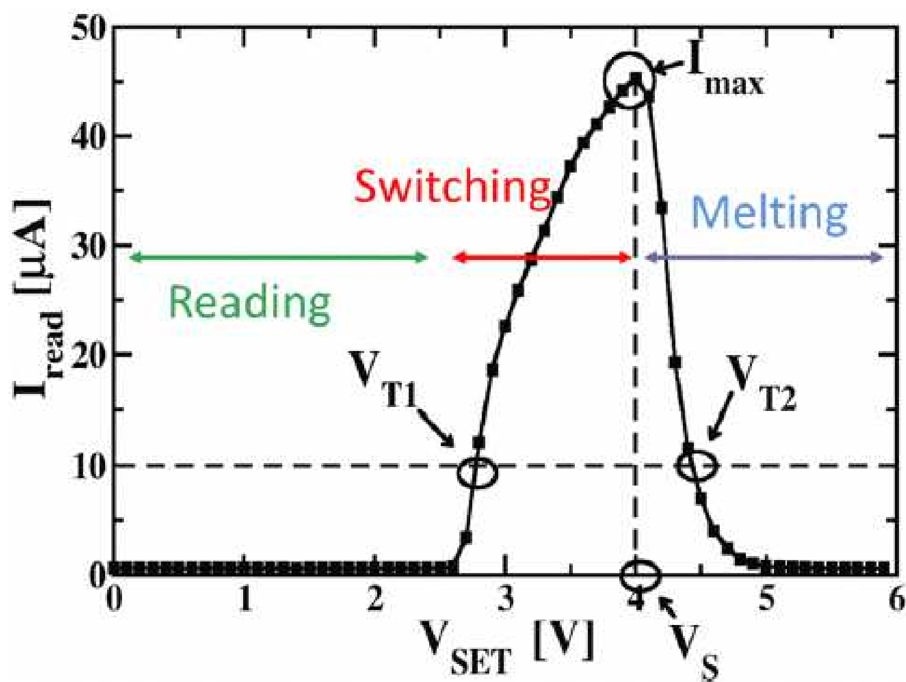



Figure 2.5: $R - I$ characterization. The curve on display represents the average over the whole cell population.

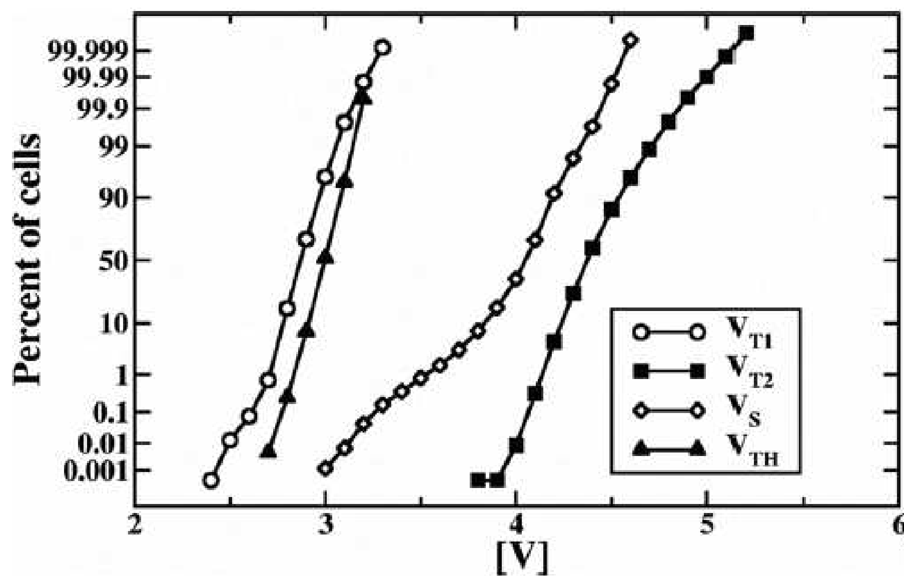



Figure 2.6: Distribution of $V_{T1}$, $V_{T2}$, $V_s$, parameters in a Gaussian probability plot.

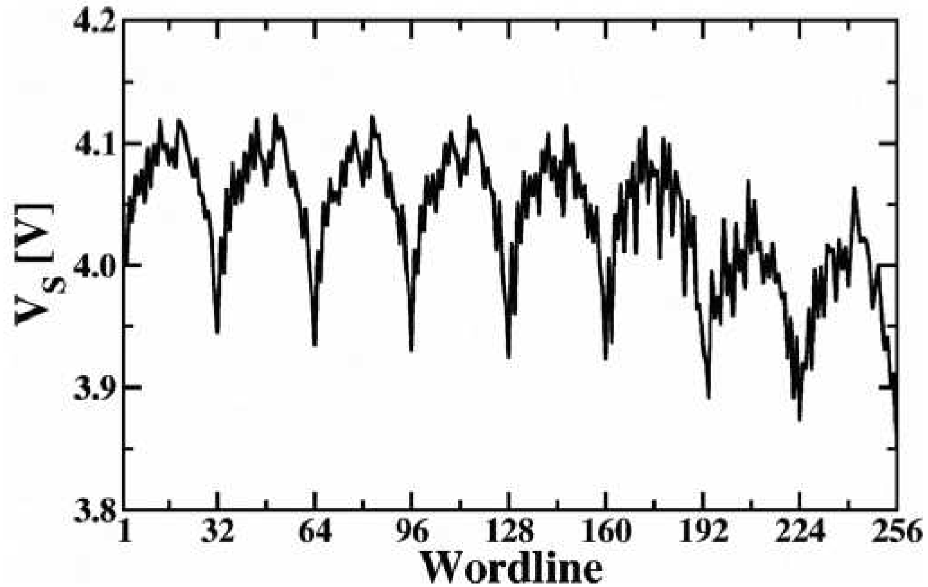



Figure 2.7: Average value of $V_s$ versus the wordline index. Each point is the average over the columns belonging to the considered wordline.

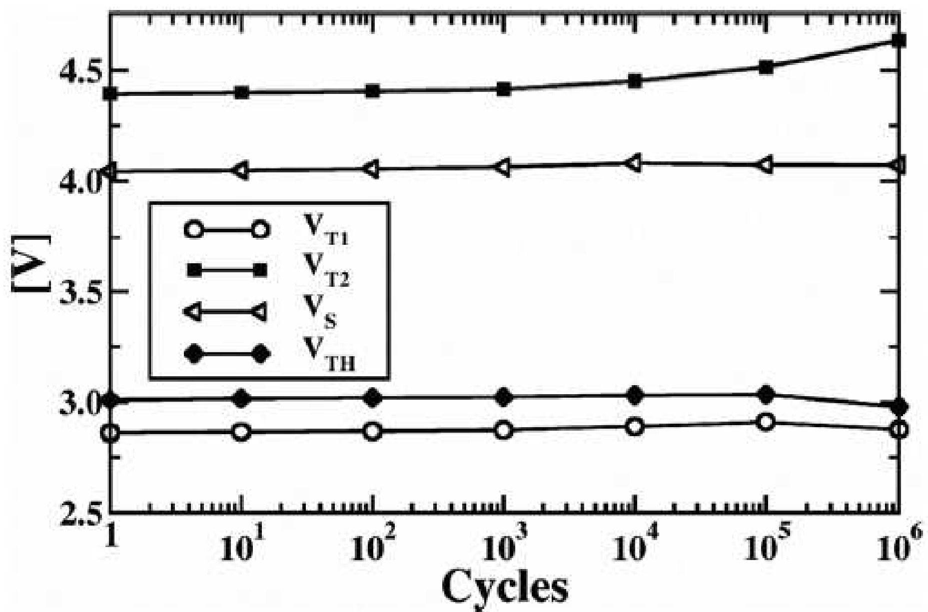



Figure 2.8: Average values of $V_{th}$, $V_{T1}$, $V_{T2}$ and $V_s$ during cycling.

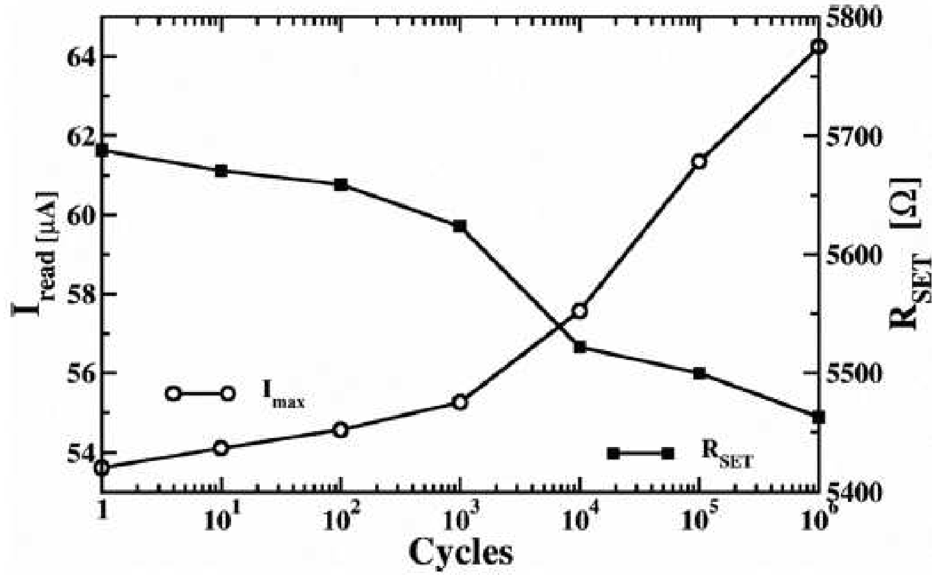



Figure 2.9: Average values of $I_{max}$, and $R_{set}$ during cycling.

crystal grains inside the programmable region (PR) and then ensure a low resistivity path between the cell polycrystalline GST and the heater. Crystal grain creation follows a nucleation dynamic [17], taking place pseudo-randomly inside the whole active area, following the electrical field distribution generated by the voltage waveform.

To compare and optimize erasing waveforms some comparison criteria are to be introduced:

- Number of waveform parameters. Each waveform has a typical set of parameters that can be varied in order to achieve the desired erasing target. Having more parameters, a greater complexity is introduced during the waveform analysis.

- Array current mean value $\mu$ [$\mu A$]. This parameter is a target for erase operation. It mainly represents the endpoint of the crystallization process.

- Array current standard deviation $\sigma$ [$\mu A$]. The distribution of the cells read currents after erase, $I_{cell}$, is found to follow a gaussian behavior. The standard deviation $\sigma$ is an indicator of how the erasing algorithm is able to keep the distribution compact.

There are three common approaches for achieving erasing operation:

- Melt and crystallize: the formation of the low resistive path is preceded by a melting operation in the PR.
- Crystallize once: the phase transition from amorphous to crystal is achieved by applying to the cell a voltage below the melting point $V_m$
- Burst crystallization: the grain nucleation is achieved by a sequence of short voltage pulses applied to PR.

Each method requires suitable waveforms. Optimization is achieved by following the flow represented in Fig.2.11.

Melt and Crystallize Waveform (MCW) belongs to the melt and crystallize approach. It is constituted by two parts: the former is a linear drop from a peak voltage $V_p$ higher than $V_m$ to 2 Volts; the latter is a rapid ($\approx$ 10 ns) voltage cut to 0 V (see Figs.2.10a). The initial part of the waveform must provide a sufficient heat inside PR, driving it to melt. Once the GST is completely molten the successive crystallization process starts from a fully amorphized material where pre-crystallized spots within PR have been certainly removed. Crystallization starts when the waveform is above $V_m$ and ends slightly above 2 Volts, so that the grain formation and the percolation path creation is a slow process. MCW allows getting low $\sigma$ values (thus compact distributions), but it is difficult to control the $\mu$

value. MCW is also called a *one parameter waveform*, since only $V_p$ can be optimized. As can be seen in Fig.2.12, by lowering this parameter the $I_{cell}$ distribution broadens significantly and a large tail for low $I_{cell}$ values is observed. This is related to the erasing approach followed by MCW: if $V_p$ is not sufficiently larger than $V_m$ the crystallization process does not necessarily starts from a fully molten material and the subsequent dynamics can vary from cell to cell. A too large $V_p$, however, may overstress the GST structure due to the melting process. Since MCW may produce the tightest distribution with respect to the other waveforms, the limitation of $\sigma$ can be the optimization goal. Therefore $V_p$ could be increased until the required $\sigma$ has been reached. Fig.2.13a and Fig.2.13b confirm the poor ability to control the crystallization dynamic, showing the rapid saturation of $\mu$ and $\sigma$ values.

Sloped Crystallize Once Waveform (SCOW) is used for a *crystallize once* approach. Its shape is that of a rectangular trapezium, with $V_p < V_m$ (see Figs.2.10b). The crystallization process starts from the previous amorphous state. After a *hold time* $t_H$, in which the voltage is kept at $V_p$, the voltage is linearly driven to 0, waiting for the complete crystal grains formation. The controlled parameters are: $V_p$, that influences the crystal grains formation speed, $t_H$, the time $t_s$ required to reach 0 V, which also define the slope ($S$) parameter as $S = V_p/t_s$. The availability of 3 adjustable parameters makes SCOW also suitable for multi-level applications, since the average current value can be controlled while keeping a tight distribution. Since $V_p$ is mandatory for knowledge of crystallization speed, is therefore requested a determination of its optimum value, thus granting an acceptable trade-off between compactness of $I_{cell}$ and higher read window margin. As shown in Fig.2.14, by applying $V_p$ closer to $V_m$, $I_{cell}$ exploits large tails into

distribution, due to unwanted melting, that can occur in cells characterized by a $V_m$ value lower than the expected for the array. Same tails appears when $V_p$ is not sufficiently larger to grant correct crystallization. The best solution evaluated, that will be furthermore used for BCW and COW, is $V_p$ = 3.75 V. $t_H$ has been found to affect only the $\mu$ parameter, so that it can be conveniently adjusted to shift the entire $I_{cell}$ distribution in multilevel application. The slope $S$, on the contrary, influences also the $\sigma$ parameter and therefore it must be optimized in order to tighten the $I_{cell}$ distribution. Fig.2.15a and Fig.2.15b show the $\mu$ and $\sigma$ dependency on $S$. An optimized value $S = 1$ is found.

Crystallize Once Waveform (COW) and Burst Crystallize Waveform (BCW) are used for a *crystallize once* approach and are characterized by a rectangular shape (see Figs.2.10b). With respect to SCOW the crystallization process ends abruptly. The key parameter is $t_H$, that allows controlling the average read current after erase, making COW also suitable for Multi-levels architectures. A too short $t_H$ may otherwise provoke parasitic programming phenomena, especially if

$$t_H \leq t_d \tag{2.1}$$

where $t_d$ is the characteristic delay requested by PR to create the percolation path. It has been decided to group the COW and the BCW optimization, due the evident similarity of the waveforms. The main difference consists in the presence of the $R$ parameter on the BCW, which allows to implement the burst crystallization process using the crystallize-once mechanism as base. The COW waveform is characterized by two parameters: $V_p$ and $t_H$. BCW can replace COW when a deeper control on parasitic programming phenomena (such as verify algorithms on Flash

memories) is required in multi levels architectures (see Figs.2.10a). This waveform realizes the *burst crystallization* approach and it is constituted by a sequence of short box pulses. The physical mechanism behind the crystal grains formation is a slight variation of that achieved by COW, since the pulse is removed before the complete grain formation. A burst of COWs is applied, thus creating a sort of temperature elastic effect, in which the temperature rise (well below the melting point) increases the crystallization process speed, while the temperature fall induces a stabilization of the grain size. The three BCW characteristic parameters are: $V_p$, the pulse duration $t_H$ and the repetition number $R$ of the standard pulse. The first two parameters are the same of COW, so that they influence the $I_{cell}$ in a similar way. The last parameter is only used to reach longer exposure times. Critical issues related to the use of BCW may arise if (2.1) is not respected or if the typical RC delays of the chip circuitry are not correctly taken into account, which may force crystal grains to be reabsorbed into amorphous structure. It has been first analyzed the dependance on $V_p$ with the same methodology used for SCOW, finding an optimal value of 3.75 V, resulting in the lowest $\sigma$ and in the smallest $I_{cell}$ distribution tail (see Fig.2.14). As for $t_H$, a strong dependance has been found for both $\mu$ and $\sigma$ criteria, only for very short times (see Fig.2.16a and Fig.2.16b). This result can be explained by the nucleation dynamics: the final $I_{cell}$ value saturates after the end of the crystal grains growth phase [19] (occuring, in this study case, after $\simeq$ 500 ns).

The performed study can be further extended to non-common erasing waveforms such as full exponential (ECOW) and ramped exponential (RECOW). Table 2.1, resume the details of the analyzed waveforms.

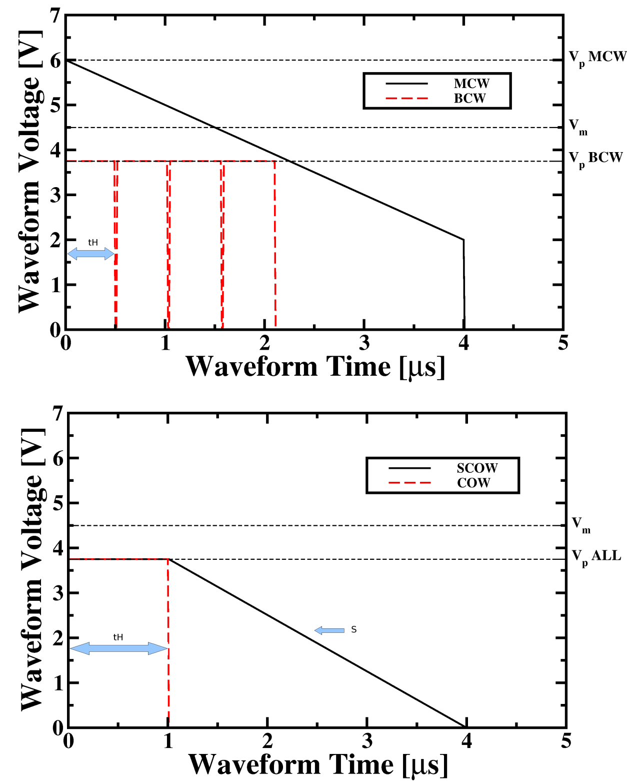



Figure 2.10: a): MCW and BCW waveforms; b): SCOW and COW waveforms.

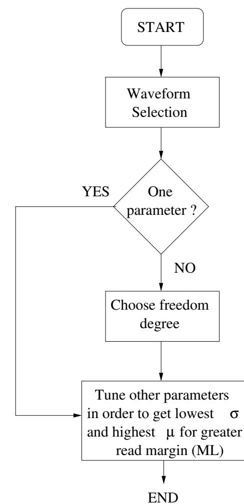



Figure 2.11: Logical flow used for waveforms optimization process.

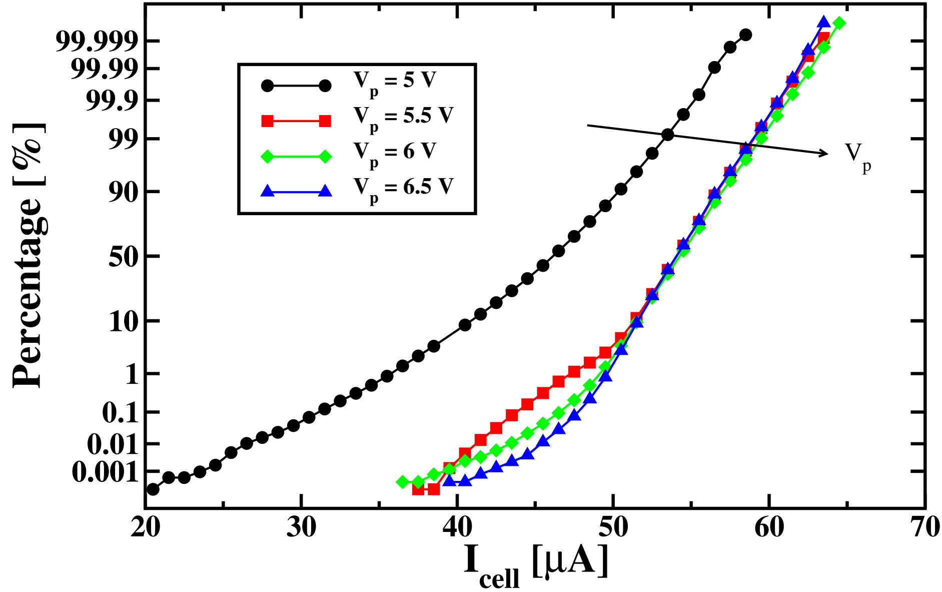



Figure 2.12: $I_{cell}$ distribution dependency on $V_p$ applied within a MCW.

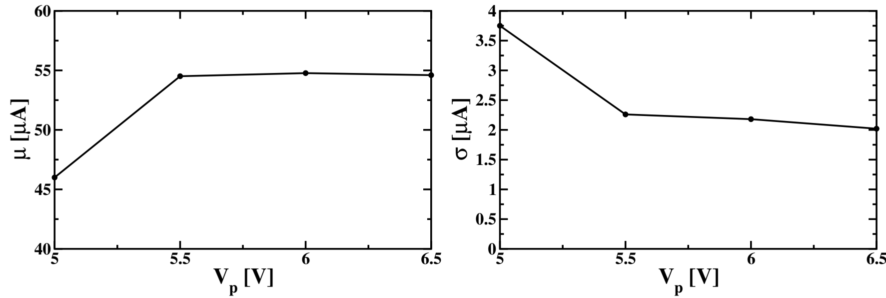



Figure 2.13: a): Variation of $\mu$ criterion on MCW, in relation to $V_p$ b): Variation of $\sigma$ criterion on MCW, in relation to $V_p$.

Table 2.1: Comparison resume of the analyzed waveforms

| | MCW | SCOW | COW | BCW |
|---|---|---|---|---|
| Parameters Number | 1 | 3 | 2 | 3 (2) |
| $\mu$ controllability | None | Coarse | Excellent | Excellent |
| $\sigma$ reduction | Excellent | Good | Bad | Bad |
| ML suitable | No | Difficult | Yes | Yes |

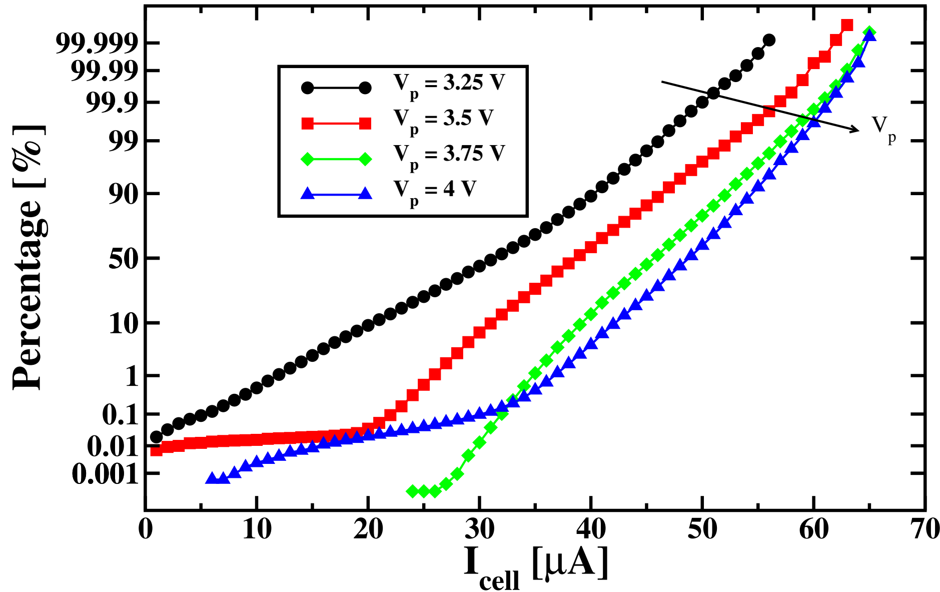



Figure 2.14: $I_{cell}$ distribution dependency by applied $V_p$ within SCOW. Similar distribution are obtained using COW.

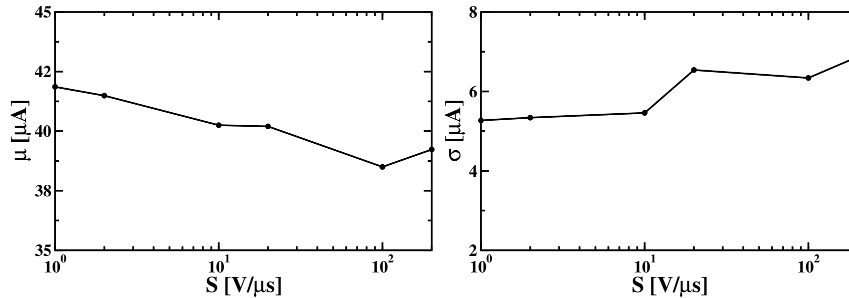



Figure 2.15: a): Variation of $\mu$ criterion on SCOW, in relation to $S$ (log-scaled) b): Variation of $\sigma$ criterion on SCOW, in relation to $S$ (log-scaled).

## 2.4 Modeling the SET kinetics

Electronic switching is the fundamental mechanism of Phase Change Memory governing the transition from the RESET state to the SET state [20, 21]. The physics behind this mechanism still remains a puzzle and, although various physical models [20, 21] have already been proposed in literature, their validation has been carried out mostly through single cell electrical characterization. On the

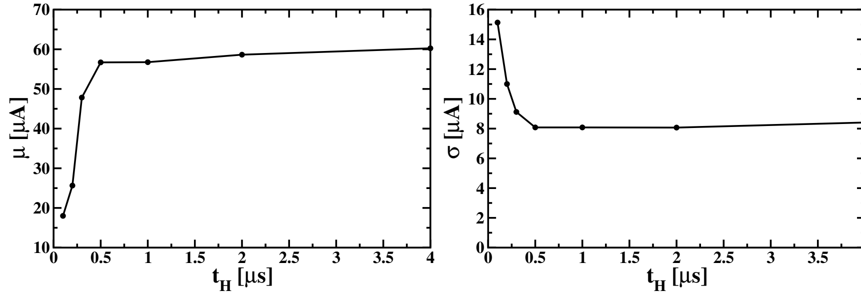



Figure 2.16: a): Variation of $\mu$ criterion on COW, in relation to $t_H$ b): Variation of $\sigma$ criterion on COW, in relation to $t_H$.

other side, the statistical nature of the electronic switching process would require a relatively larger cell population to be analyzed in order to give a more complete picture of the whole phenomenon. In particular, it has been found [17] that a certain delay time ($t_d$) is necessary in order to form a polycrystalline grain percolation path (shunt) between the bottom electrode and the top electrode across the Ge2Sb2Te5 (GST) material. When this condition occurs, cell resistance rapidly drops to small values. It is possible to model the delay time in terms of simple analytical formulae which take into account the average behavior and the parameter dependence on both technological parameters and operating conditions [17]. However any detailed statistical description of the shunt formation time has never been presented and, more in general, also any statistics related to the whole crystallization process involving nucleation, growth and saturation, has never been shown. Experimental results provide a statistical description of the delay time and are consistent with the delay time model proposed in [17], also revealing some interesting features of crystal growth dynamics and its saturation.

All experimental results have been measured and collected by using a dedicated testing equipment called Rifle developed by Activetechnologies. All waveforms have been externally applied by the instrument on a 8Mb PCM test chip

featuring 180nm technology. All data shown in this work refer to 512Kb populations. Fig.2.17 shows the waveform and the sequence of operations performed during the measurement. We used sequences of short SET pulses with durations $\Delta t$=(10ns), 25ns, 50ns and 100ns and amplitudes VSET=3.5V, 3.75V, 3.9V. The number of pulses $N$ has been adjusted in order to keep constant $Nx\Delta t$=1$\mu$s for each condition. After each pulse we measured the read current ($I_{read}$) of each cell of the array, thus obtaining a SET curve for each cell of the population ($I_{read}$ versus time).

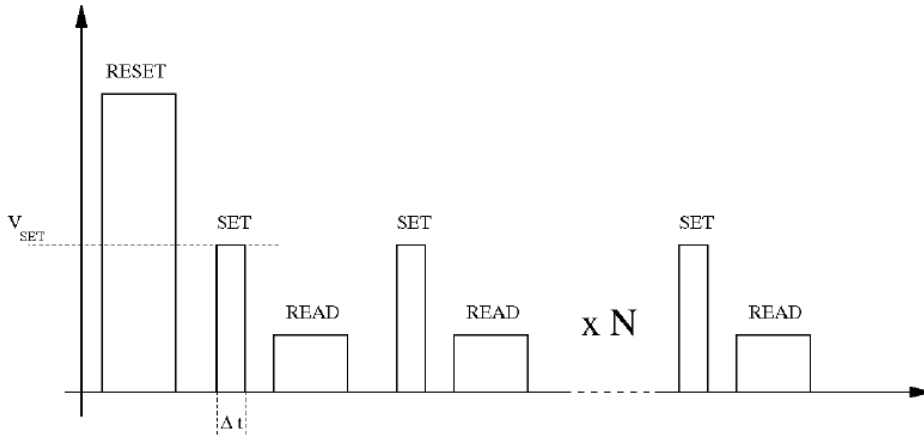



Figure 2.17: Sequence of pulses used for the experiments. The SET pulse conditions have been: $\Delta t$=(10ns), 25ns, 50ns and 100ns; $V_{SET}$=3.5V, 3.75V, 3.9V. A total number of SET pulses $N$ has been applied.

During nucleation stage 2 shows the average SET curves measured in each operating condition. A significant dependence on both $\Delta t$ and $V_{SET}$ is clearly visible in Fig.2.18. Each set curve can be divided into three parts each corresponding to specific stages of the crystallization process (see Figure Fig.2.19): nucleation, growth and saturation. During nucleation crystals of radius $R_c$ nucleate and align vertically until, after a delay time $t_d$, a stable percolation path is

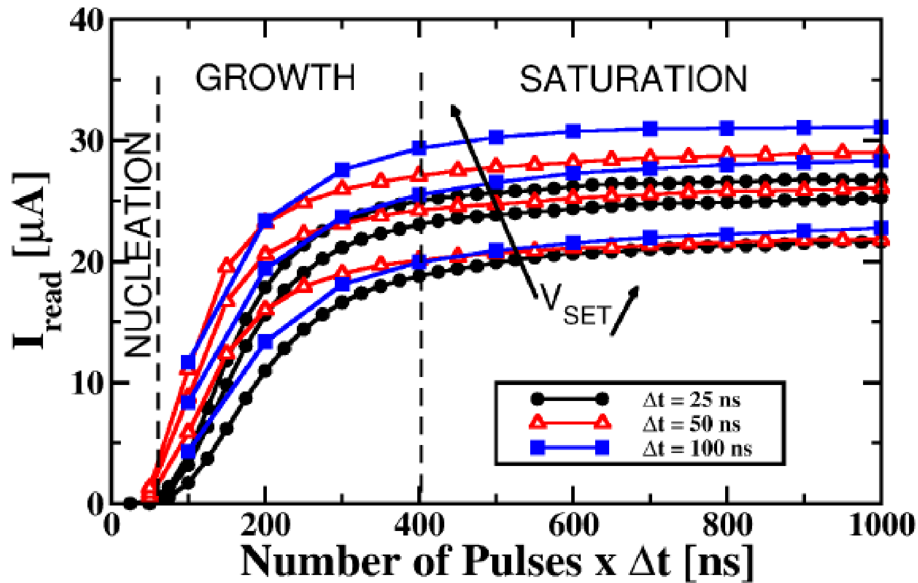



Figure 2.18: Average read current for each operating condition.

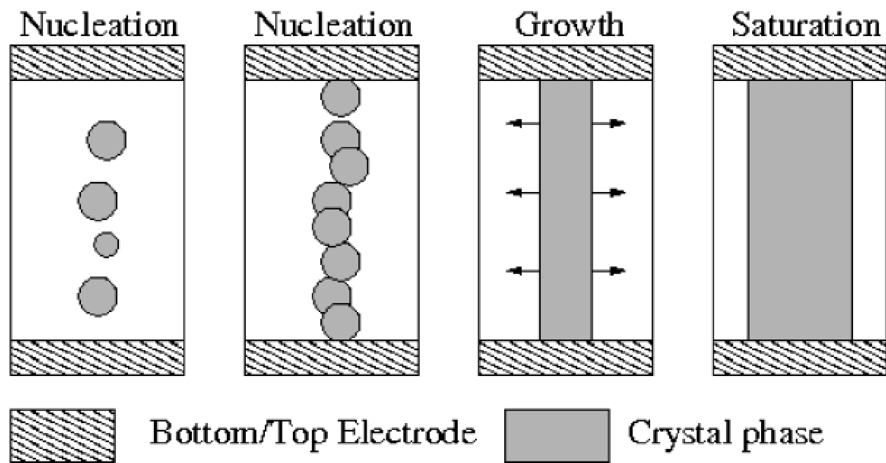



Figure 2.19: Picture of the stages of the crystalline shunt formation and development.

formed. From that on the percolation path radius starts growing in size until saturation occurs. Figure 4 shows the distribution of the read currents of the 512Kb cell population during the first 5 pulses in the $\Delta t$=25ns and $V_{SET}$=3.5V condition. Similar results can be obtained in the other conditions. An upper tail shows

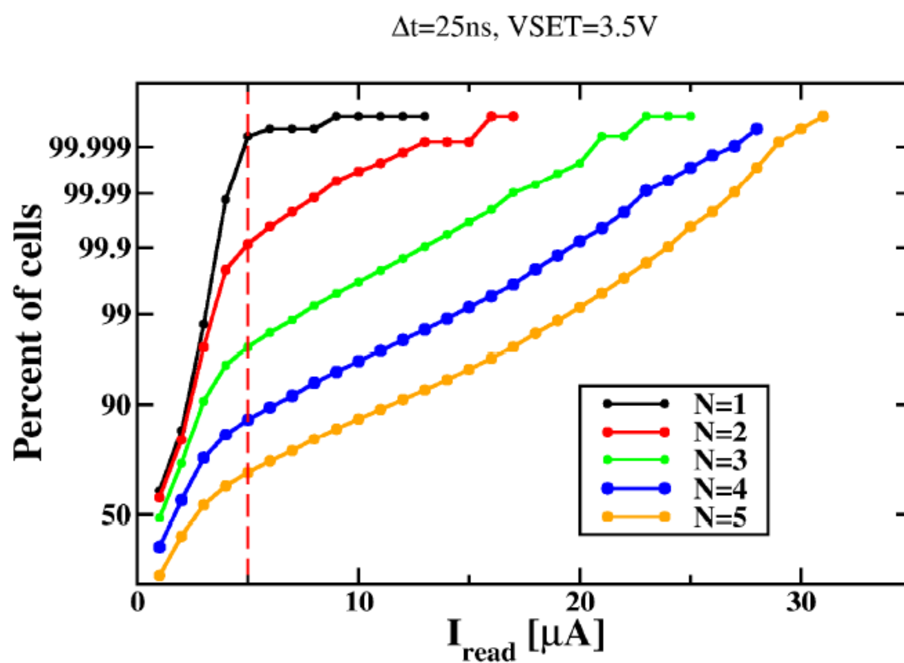



Figure 2.20: Read current distributions after the first 5 SET pulses.

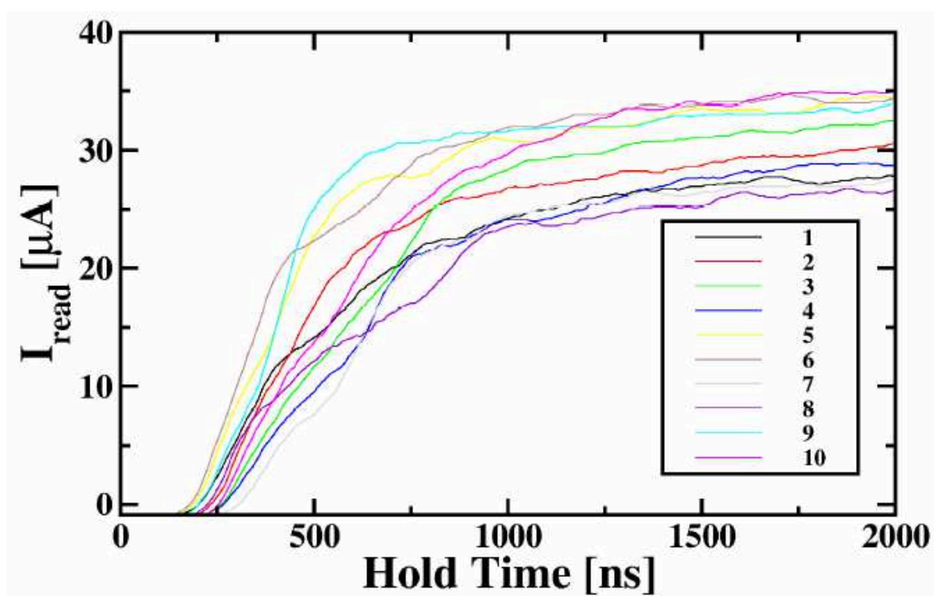



Figure 2.21: SET curves for the same cell measured 10 consecutive times ($\Delta t$=10ns).

up and grows in size after each pulse. Cells with $I_{read}$ <$5\mu A$ (dashed red line) are to be considered in the RESET state. Therefore, the number of cells exhibiting electronic switching after each pulse could be simply calculated by checking the condition $I_{read}$ >$5\mu A$ for each cell of the array. Note however that a tail of cells with $I_{read}$ >$5\mu A$ can already be present after RESET and therefore this method may not be sufficiently accurate in estimating the number of electronic switched cells as a function of the number of pulses. However, from the results shown in Fig.2.18 it is still possible to draw some important conclusions. In particular the increase of tail cell population with the number of applied pulses is an evidence that not all cells of the array behave at the same way: some cells may switch after the first pulse, whereas others may require more pulses. This behavior may depend on the statistical distribution of cell technological parameters and/or to the intrinsic statistical nature of the phenomenon. The presence of the latter can be shown by observing Fig.2.18 where the SET curves of the same cell have been measured in 10 consecutive experiments.

During saturation Fig.2.22 shows the average saturation current, i.e. the average read current measured after the very last SET pulse of the sequence. A small dependence on $\Delta t$ and a significant dependence on $V_{SET}$ can be observed. Fig.2.23 shows the read current distributions for each pulse duration. It can be verified that all cells can switch after a sufficient number of pulses. All distributions exhibit the same shape of the distribution which is not perfectly Gaussian. The SET voltage simply shifts the entire distribution to the right and similarly does the pulse duration although in a minor fashion. These results show that some topological effects related to signal propagation issues through the array can be present. Fig.2.24 confirms these results by evidencing that the influence of the

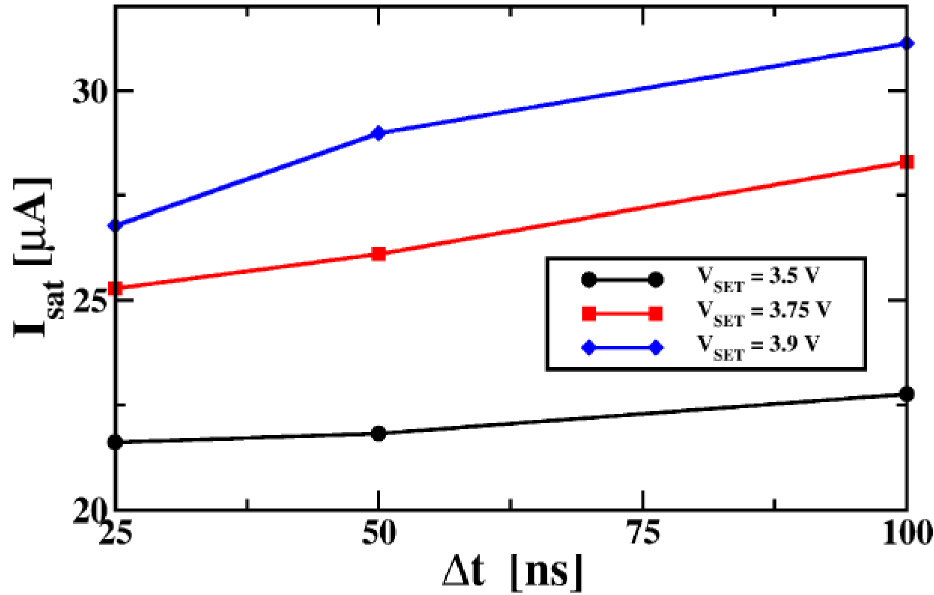



Figure 2.22: Average saturation current as a function of the operating conditions.

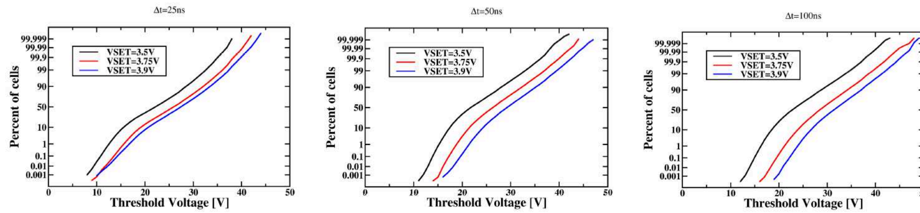



Figure 2.23: Read current distributions for each pulse duration.

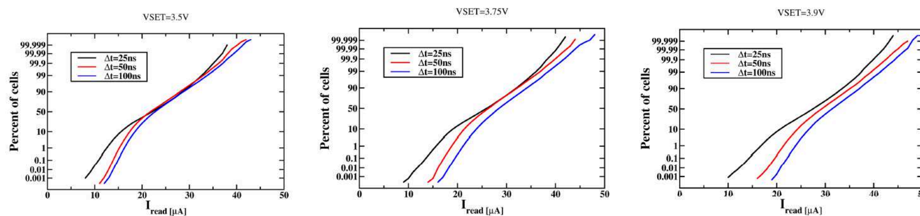



Figure 2.24: Read current distributions for each SET voltage.

pulse duration is more significant at larger SET voltages. Almost $10\%$ of cells (Slow cells) exhibits a larger dependence on $\Delta t$. As shown in Fig.2.25, a large number of these Slow cells is located in the first rows of the sector. As expected,

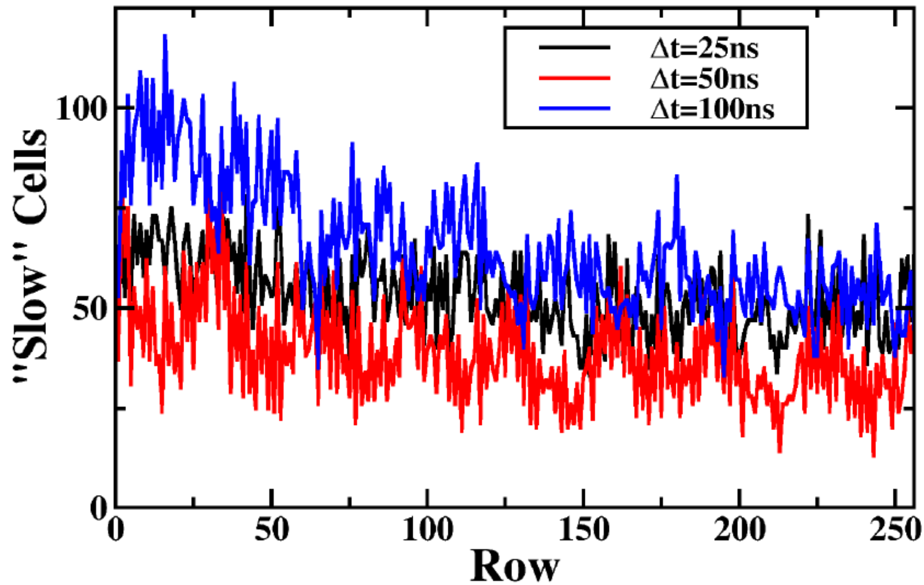



Figure 2.25: Number of Slow cells as a function of their wordline (row) position within the array.

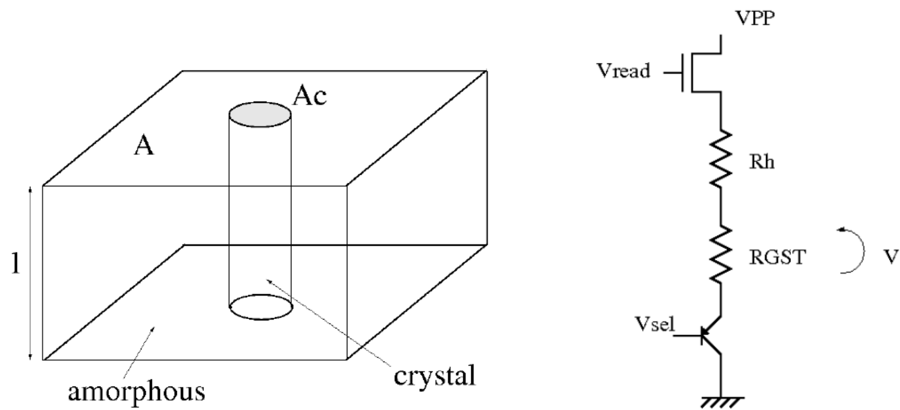



Figure 2.26: Conductive GST model (left) and equivalent read path circuit (right).

the bitline resistance variation along the wordline coordinate can play an important role in determining the signal propagation behavior through the array. It can be shown that for $\Delta t >$10ns these effect are not so significant with respect to the following discussion.

For a further analysis of the read current data we used a GST resistance model in which a crystalline cylinder with area $A_c$ is surrounded by the amorphous material. A parallelepiped volume of $\alpha - GST$ is used to identify the programmable active area with thickness $l$ and area $A$ on which a cylindrical crystalline shunt grows with area $A_c$, as shown in Fig. 2.27. The expression for $A_c$ is therefore:

$$A_c = \frac{\frac{l}{R_{GST}} - \frac{A}{\rho_a}}{\frac{1}{\rho_c} - \frac{1}{\rho_a}} \tag{2.2}$$

where $l$ assumes a value of 30 nm, $A$ is equal to 800 nm$^2$ whereas $\rho_a$ and $\rho_c$ assume the typical values for $\alpha - GST$ resistivity (3 $\Omega$cm) and $x - GST$ resistivity (20 m$\Omega$cm) [22]. The resistance $R_{GST}$ value is extracted directly from the read current value from the memory cell. From (2.2) it is possible to derive the expression for the crystalline shunt radius by considering the cylindrical approximation resulting in:

$$r_c = \sqrt{\frac{A_c}{\pi}} \quad . \tag{2.3}$$

It has been evidenced in [22] that $r_c$ has a strong dependance on time. In particular, consistently with [23], it has been found that the growth of $r_c$ can be subdivided in three domains: the nucleation of the GST crystallites originating the conductive percolation path, the growth of the path, and its final saturation, as shown in Fig 2.19. The superposition of the three mechanisms leads to the following equation:

$$r_c(t) = r_{sat}\left(1 - e^{-\frac{t-t_0}{\tau}}\right) \tag{2.4}$$

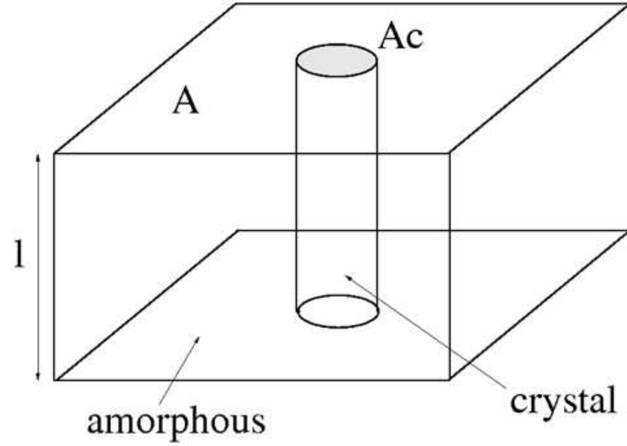



Figure 2.27: Conductive GST model. A cylindric crystalline shunt approximation is used for model compactness.

where the three characteristic parameters describing the erase kinetics can be evidenced: the saturation radius of the percolation path $r_{sat}$, the initial growth time (IGT) $t_0$ and the kinetic constant $\tau$.

From the electrical characterization data of 512kbit PCM cells we were able to model the statistical spread of these parameters in order to simulate the erase operation. We found that $r_{sat}$ and $t_0$ correctly fit a lognormal probability distribution, whereas $\tau$ fits better with a Weibull probability distribution [22]. In such way, by including the equations of the aforementioned distributions of the parameters inside a numerical analysis tool, it is possible to simulate the erase kinetic of an array of PCM cells. The equations for the parameters distributions are:

$$F(r_{sat}) = \frac{1}{2}erfc\left[-\frac{\ln(r_{sat}) - \mu_{r_{sat}}}{\sigma_{r_{sat}} * \sqrt{2}}\right] \tag{2.5}$$

$$F(t_0) = \frac{1}{2}erfc\left[-\frac{\ln(t_0) - \mu_{t_0}}{\sigma_{t_0} * \sqrt{2}}\right] \tag{2.6}$$

$$F(\tau) = 1 - e^{-\left(\frac{\tau}{\beta_\tau}\right)^{\alpha_\tau}} \tag{2.7}$$

where $\mu_{r_{sat}}$, $\sigma_{r_{sat}}$, $\mu_{t_0}$ and $\sigma_{t_0}$ represent the mean value and the standard deviation value for the lognormal distribution of $r_{sat}$ and $t_0$ respectively, whereas $\alpha_\tau$ and $\beta_\tau$ represents the scale factor and the shape factor for the Weibull distribution of $\tau$ respectively.

Exploiting (2.5), (2.6) and (2.7), the expression for the $I_{read}$ current has been derived [22]:

$$I_{read} = \frac{x * (V_{read} - V_{drop})}{1 + R_h * x} \tag{2.8}$$

where $V_{read}$ is the voltage applied on the circuit of Fig. 2.31 for read operation, $V_{drop}$ is the voltage accounting for the drop on both bitline and wordline selector elements, $R_h$ is the resistance of the heater element and $x$ is a coefficient which includes the statistical distributions of the erase kinetic parameters expressed as:

$$x = \frac{1}{l} * \left[ A_c * \left( \frac{1}{\rho_c} - \frac{1}{\rho_a} \right) \right] + \frac{A}{\rho_a} \quad . \tag{2.9}$$

The aforementioned model equations have proven to correctly fit electrical characterization data erase kinetics with several crystallization approaches featuring different erasing voltages ($V_{SET}$) and timings [22].

However the previous expression for $I_{read}$ merges the crystalline component and the amorphous component of $R_{GST}$, hence a separation of the two is needed for accounting only the SET seasoning effect which we are modeling. The $R_{GST}$ resistance can be expressed accordingly to the crystalline fraction model [24] as:

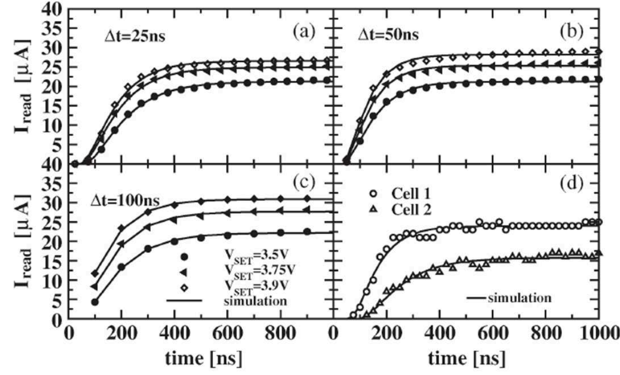



Figure 2.28: Model output on different SET operative conditions.

$$R_{GST} = \frac{l}{A} \frac{\rho_c \rho_a}{\rho_c + C(t)(\rho_a - \rho_c)} \tag{2.10}$$

where C(t) is the crystalline fraction coefficient expressed as the ratio of the crystalline shunt area and the amorphous area:

$$C(t) = \frac{A_c(t)}{A} \quad . \tag{2.11}$$

For crystalline fraction equal to 1 (a complete erase operation) $R_{GST}$ tends to $R_{SET}$. When simulating the array of PCM cells, we assumed to deal with a complete erase operation, hence meaning the account of the kinetics regime $I_{read}$ value from which we can efficiently extract $R_{SET}$. Model outcome is shown in Fig.2.28.

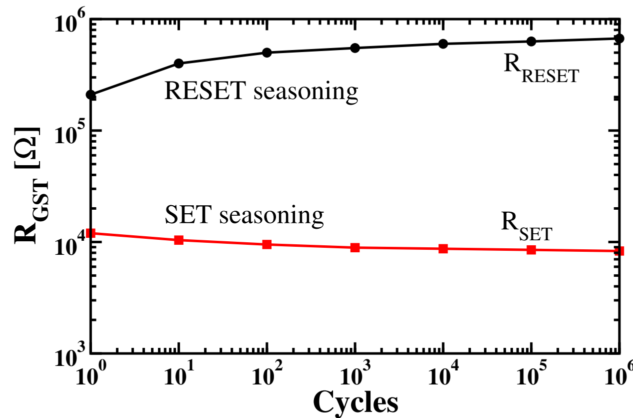



Figure 2.29: Seasoning effect as a function of cycles number, evidenced as an increase of the average (on 512 Kbits cells) $R_{RESET}$ and a decrease of $R_{SET}$.

## 2.5 The SET seasoning and the secondary shunt phenomena

The phase change in PCM can be controlled by a careful design of writing waveform shape and parameters [25]. Although the phase transition is fully reversible, an alteration of the properties of the active material may occur under SET/RESET cycling stress. This leads to a combined effect of both $R_{RESET}$ increase and $R_{SET}$ decrease, hereafter named seasoning, which produces a widening of the read window (see Fig. 2.29). Although this phenomenon positively affects the read operation performed within a SLC (Single Level Cell) architecture, on MLC (Multi Level Cell) it should be carefully taken into account. In addition, measurements also reveal that seasoning can be accompanied by an increase of the minimum programming voltage, as shown in Fig. 2.30. A complete picture of RESET seasoning was given in [26], and the SET characterization was only mentioned but not investigated.

Electrical characterization data have been obtained on 512 kb cell populations

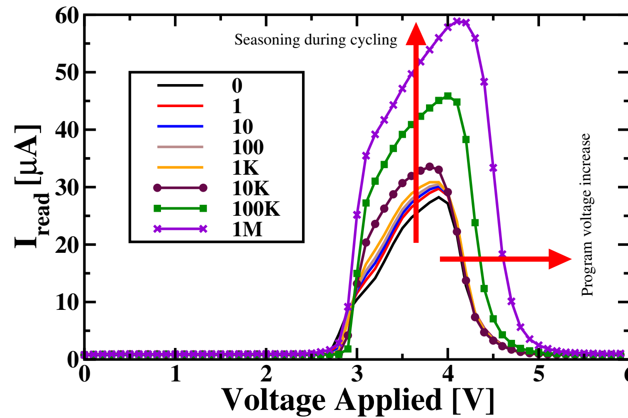



Figure 2.30: Equivalent R-I characteristic of the PCM array. Characterization data were retrieved from cycle 1 to $10^6$.

of 8 Mb PCM test chips in 180 nm technology with a $\mu$-trench structure connected with a BJT selector [27]. Writing operations on the memory array have been performed with a dedicated ATE capable of applying fully arbitrary waveforms ($V_{SET}$ and $V_{RESET}$ in Fig. 2.31). Reading operations on the memory array have been performed with a Direct Memory Access, which enables fast measurements of the read current $I_{read}$ drained the by the cell in the SET state when a read voltage $V_{READ}$ is applied. The measured current values are then converted into resistance $R_{GST}/R_{SET}$, by taking into account the structure of the PCM cell and its connection on the array, as shown in Fig. 2.31 [22].

Different erasing schemes, which are described in detail in [25], have been used for the seasoning characterization: MCW (Melt and Crystallize Waveform), COW (Crystallize Once Waveform) and SCOW (Sloped Crystallize Once Waveform).

A representation of the waveform sequence used in seasoning experimental characterization has been given in Fig. 2.32, with parameters shown in Tab. 2.2.

Seasoning effect is stronger during the very first operative cycles, as evidenced

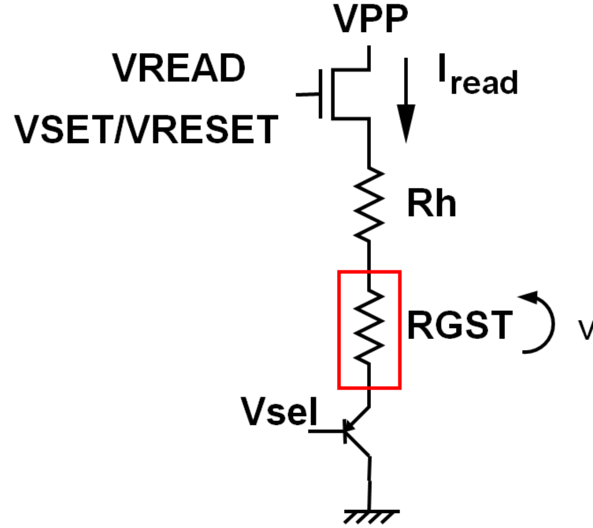



Figure 2.31: PCM cell model for measurements and connection to the array. The voltage drop on the bitline MOS selector and on BJT cell selector is accounted.

in Fig. 2.33 depicting the average cycle by cycle SET resistance variation $\Delta R_{SET}$. We therefore focused our analysis on the very first 200 cycles performed on fresh devices.

The study of the phenomenon on a large array enables statistical considerations that would not be possible on single cells measurements. As shown in Fig. 2.34, seasoning on single cell data would be masked, or strongly affected, by measurements noise.

The analysis of the data requires a metric for a quantitative comparison of the impact on seasoning of erasing waveforms. It has been defined the $R_{SET}$ percentile reduction $\delta$ on 200 operative cycles as:

$$\delta = (R_{SET}(1) - R_{SET}(200))/R_{SET}(200) \tag{2.12}$$

As a baseline, we first characterized $\delta$ on a sequence of 200 read-only cycles.

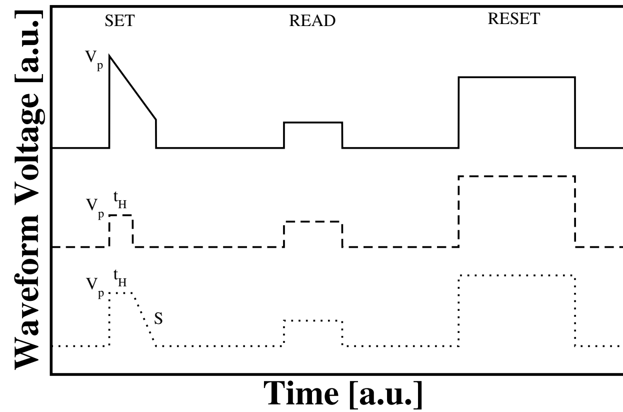



Figure 2.32: Waveform sequences for cycling with different erasing schemes in seasoning investigation experiments. Used erasing schemes are depicted: MCW (solid), COW (dashed) and SCOW (dotted). Waveforms parameters such as $V_p$, $t_H$ and $S$ are also evidenced.

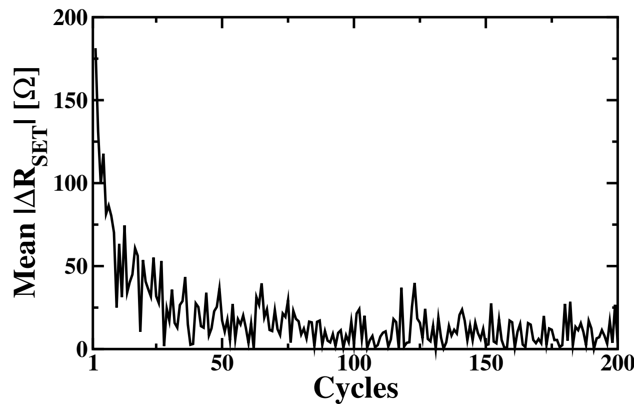



Figure 2.33: Average variation of the array $R_{SET}$ calculated in modulus as the difference between two consecutive cycles measurements.

Under this operating condition seasoning should not be present [26]. Yet the experimental data show a weak presence of seasoning even under these conditions.

By cycling the memory array with MCW it is possible to appreciate a reduction of the $R_{SET}$ in 200 cycles of about 300 ohms, as shown in Fig. 2.35. This leads to a $\delta$ value of 5% as shown in Tab. 2.2. By using MCW a more uniform and compact SET distribution can be obtained, but its particular shape does not

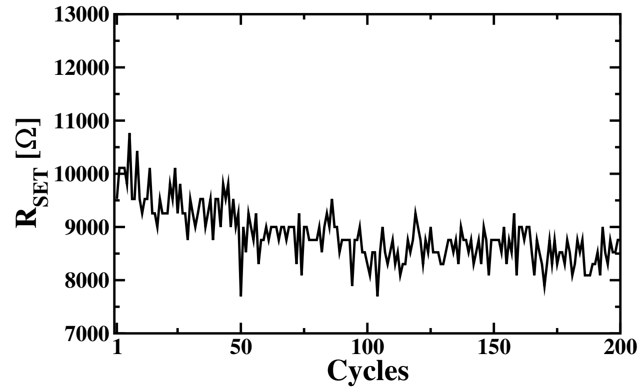



Figure 2.34: Seasoning effect evaluated on consecutive single cell measurements.

allow a further parametric analysis [25].

COW needs to be analyzed in detail since we have the possibility to vary the hold time parameter $t_H$ of the waveform. The starting level of R$_{SET}$ is inversely proportional to $t_H$ as experimentally shown by prior results [22]. Calculated $\delta$ values show that lower hold times induce high seasoning impact (see Tab. 2.2). Fig. 2.36 displays the mean behavior of the array SET resistance during cycling.

SCOW has two parameters: the slope $S$ and the hold time $t_H$. We decided to fix a value of 1 $\mu$s in order to evaluate only the impact of the waveform slope. Fig. 2.37 shows that with slope values greater than 10 V/$\mu$s the seasoning impact starts to decay. SCOW tends to COW for large values of $S$ and this is confirmed by the curve with $S$ = 200 V/$\mu$s, which is similar to that shown in Fig. 2.36. Values of $\delta$ as a function of the $S$ parameter depict a negative trend, as shown in Tab. 2.2.

A physical interpretation of SET seasoning will be given by assuming a correlation between this phenomenon and its counterpart occurring on the RESET state (see Fig. 2.29). It has been shown that the amorphous volume in the GST gradually expands due to cycling-induced stress [26]. Under these conditions, during the SET operation, the growth of the crystalline filament radius becomes more en-

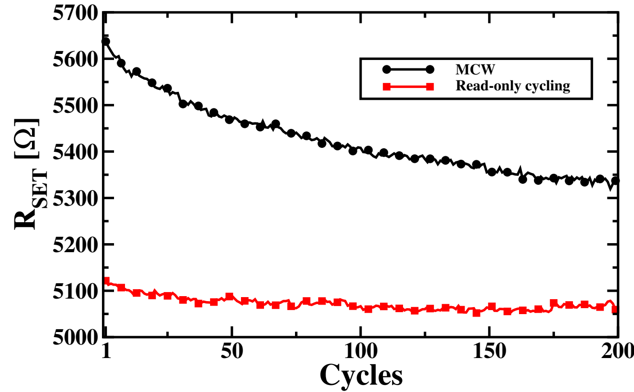



Figure 2.35: Average array $R_{SET}$ measured with a MCW erasing scheme. A comparison with the reference read only cycling is reported.

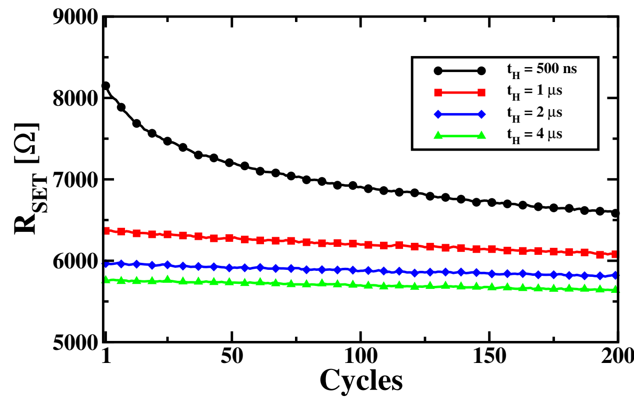



Figure 2.36: Average array $R_{SET}$ measured with a COW erasing scheme. A dependance of the reduction of the $R_{SET}$ by the waveform parameter $t_H$ is evidenced.

ergy efficient. In fact, the lateral heat flux dispersion gradually decreases during cycling, being the conductivity of the amorphous GST much smaller than that of the crystalline GST (0.19 W/mK versus about 1 W/mK respectively [28]). As a consequence, a larger energy is used for phase change and the saturation radius of the crystalline shunt will assume increasing values with cycles. Results shown in Fig. 2.36 can support this physical interpretation. The SET resistance can be written as:

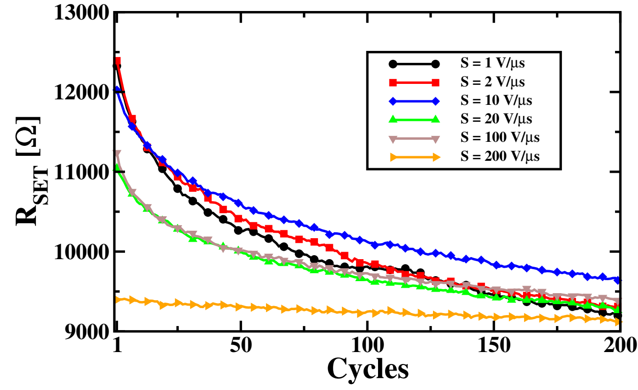



Figure 2.37: Average array $R_{SET}$ measured with a SCOW erasing scheme. A dependance of the reduction of the $R_{SET}$ by the waveform parameter $S$ is evidenced.

Table 2.2: Seasoning impact resume with different waveform

| | $V_p[V]$ | $t_H$ [$\mu$s] | $S$ [V/$\mu$s] | $\delta$ [%] |
|---|---|---|---|---|
| Read-only | 1.8 | | | 1 |
| MCW | 6.5 | | | 5 |
| COW | 3.75 | 0.5 | $\infty$ | 23 |
| | | 1 | $\infty$ | 3 |
| | | 2 | $\infty$ | 2 |
| | | 4 | $\infty$ | 2 |
| SCOW | 3.75 | 1 | 1 | 34 |
| | | | 2 | 33 |
| | | | 10 | 25 |
| | | | 20 | 19 |
| | | | 100 | 19 |
| | | | 200 | 4 |

$$R_{SET} = \frac{L}{A} \frac{\rho_a \rho_c}{\rho_c + \frac{\pi r_{sat}^2}{A}(\rho_a - \rho_c)} \tag{2.13}$$

where $L$ is the shunt length, $A$ is the active material area, $\rho_a$ and $\rho_c$ are the resistivity of the GST on the amorphous and crystalline state respectively and $r_{sat}$ is the saturation radius that has shown to depends on $t_H$ [22]. In particular, for large $t_H$ values ($\geq 1\ \mu$s) the saturation radius reaches a maximum value [22]. In this case:

$$R_{SET} \approx \frac{L\rho_c}{A} \tag{2.14}$$

Results in Fig. 2.36 show that with $t_H > 1\ \mu$s, SET seasoning is relatively small. This means that L, $\rho_c$ and A in (2.4) are almost constant. The more significant SET seasoning behavior shown when $t_H$ = 500 ns can therefore be explained in terms $\rho_a$ or $r_{sat}$ variation as indicated by (2.13). Indeed, $\rho_a$ has to be ruled out and considered quasi constant [26, 29, 30]. Therefore, SET seasoning for small $t_H$, can be almost entirely attributed to an increase of $r_{sat}$ during cycling. Fig. 2.38 shows the relation between $R_{SET}$ and $r_{sat}$ for different pulse durations used to perform the SET operation as obtained by using (2.13). As shown, a small increase of few nanometers in the saturation radius of the filament during cycling can properly take into account the seasoning effect.

It has been observed that, after the Erase operations, PCM cells may randomly exhibit $I_{read}$ shifts in a Random Telegraph Noise (RTN) fashion [22, 31] (see Fig. 2.40), with the effect of broadening the read current distribution after cycling operations. The physical and analytical models developed to describe the Erase operation [17, 22, 24, 32] evidenced that a cause of this phenomenon may reside on the possible creation of a secondary percolation path. In particular, the nucleation/growth framework described in [22] has been extensively used as a starting

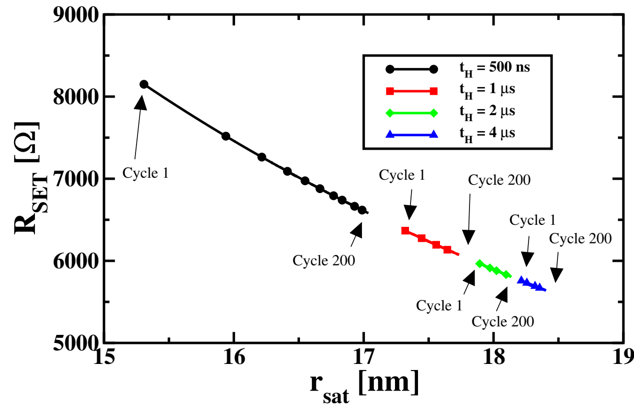



Figure 2.38: Average array $R_{SET}$ dependency from $r_{sat}$ within a COW erasing scheme with two different $t_H$ used. Cycle dependence is also evidenced.

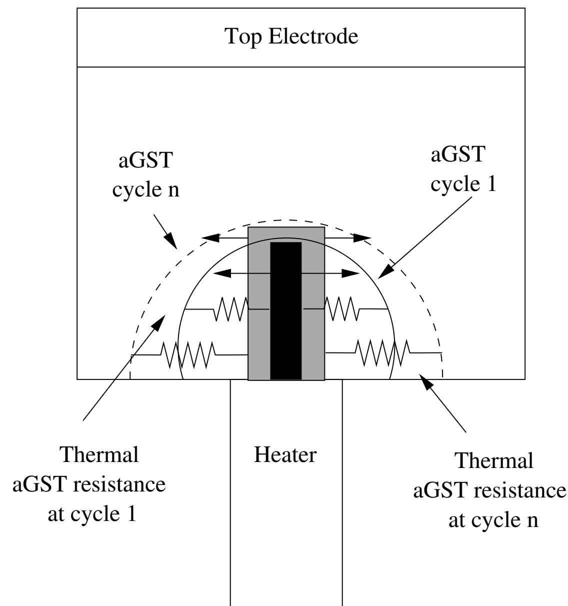



Figure 2.39: PCM cell structure evidencing the seasoning phenomenon both in RESET and SET state.

point in this discussion. The $I_{read}$ shift magnitude due to this phenomenon does not impact the reliability of Single-Level cell architectures, thanks to the wide read margin, but it may be a potential issue in Multi-Level Cell architectures where a tight control of the current distributions is required.

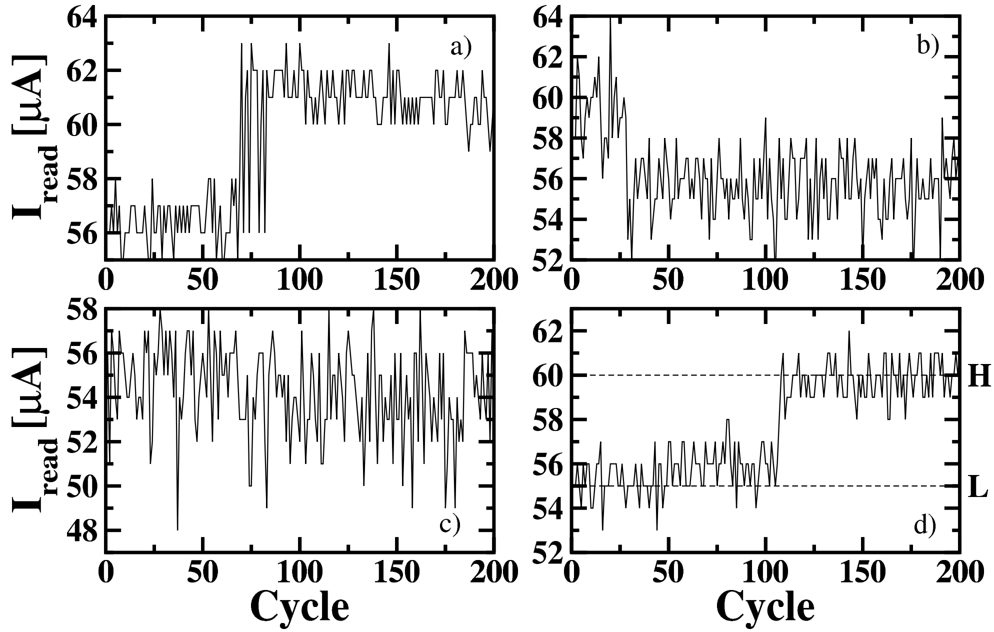



Figure 2.40: Example of RTN-like behaviors in four PCM cells observable by plotting $I_{read}$ after the Erase operation versus the number of cycles. Two levels are evidenced (i.e. $H$ and $L$ in fig. $d$) representing the presence/absence of the secondary path.

In order to characterize the secondary path dynamics, a sequence of 200 Program-Erase (RESET/SET) operations has been applied on virgin PCM arrays, and the current sinked by each cell after each Erase operation has been tracked. The voltage waveforms used for Program and Erase, applied to the selected memory cell through a low threshold voltage nMOS pass transistor, are depicted in Fig. 2.41b. As many Erase waveforms are commonly exploited in PCM arrays in order to obtain the GST crystallization [25], we decided to use the simple *Crystallize Once Waveform* (COW) scheme, which is characterized by two parameters: the peak voltage $V_p$ and the hold time $t_H$. The former one has been fixed at 3.75V, whereas the latter has been varied (1$\mu$s, 2$\mu$s and 4$\mu$s) in order to evaluate any dependance on the pulse duration. The COW scheme allows achieving

the GST crystallization without enduring a prior melting of the material, which would require higher voltage bias and longer erasing time. This scheme leaves the active material in a Partial-SET state, as required for Multi-Level Cell architectures. Fig. 2.42 shows the average $I_{read}$ vs $V_{SET}$ characteristic of the PCM array evidencing the distance between the Partial-SET state and a Full-SET state.

The Read operation is performed by applying $V_{read} = 1.8V$ to the nMOS pass transistor. This voltage has been chosen since it is sufficiently far from the material switching threshold voltage while guaranteeing good sensing performance at the same time.

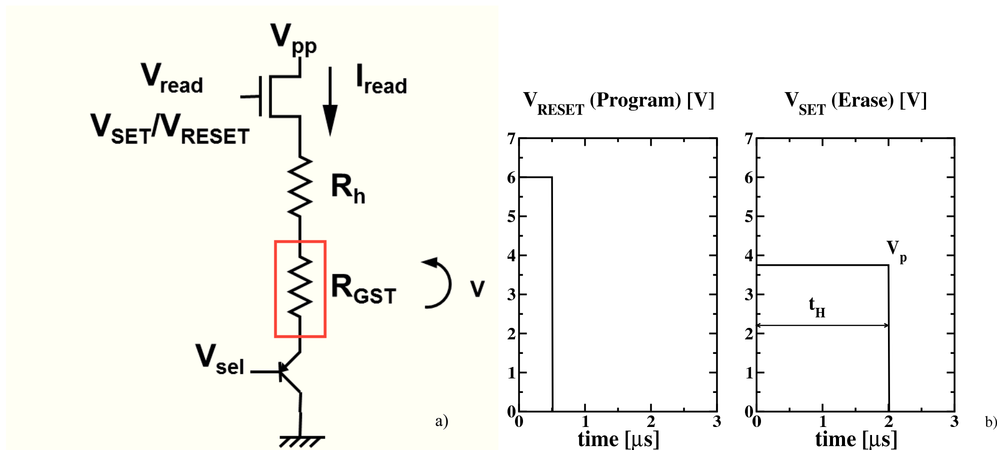



Figure 2.41: a) Equivalent read/write path circuit of a PCM cell. b) Program (RESET) and Erase (SET) waveforms exploited in this work.

The RTN-like behavior evidenced by PCM cells during cycling allows exploiting a methodology for extracting the parameters related to the secondary path creation [33]. By plotting $I_{read}$ for a selected cell after each Erase operation, it is possible to evidence, at first glance, two levels (see Fig. 2.40d). The lower, $L$, is related to a normal Erase operation in which the expected percolation path has been created in the amorphous region and eventually saturates [17]. The upper

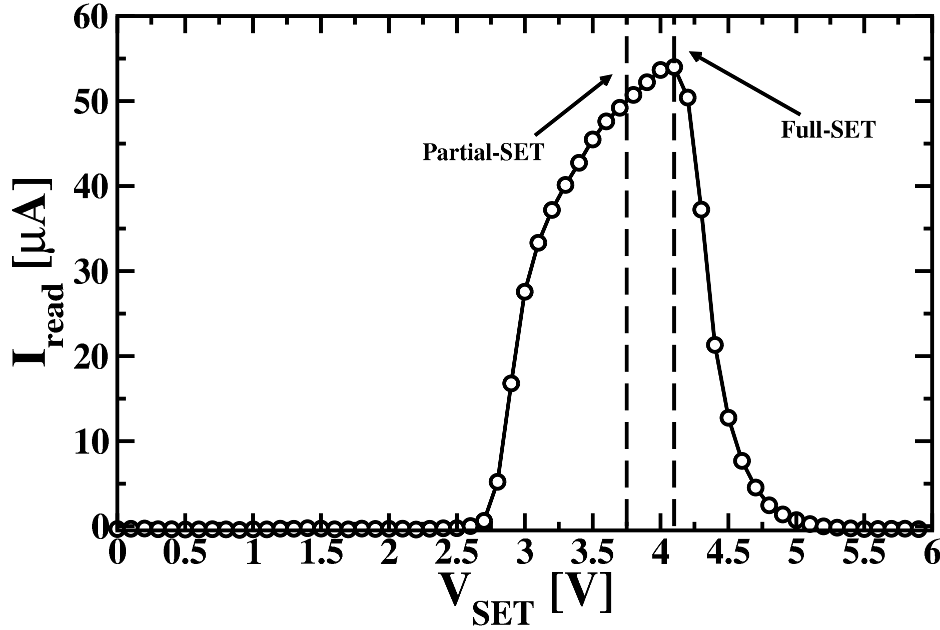



Figure 2.42: Average $I_{read}$ versus $V_{SET}$ characteristic. As shown, $V_{SET}$ = 3.75V holds the active material in a Partial-SET state.

level, $H$, on the contrary, is related to the creation of a secondary percolation path in parallel to the main one, resulting on a slightly increase of $I_{read}$. As described in [22], $L$ and $H$ represent, for each cell, the maximum likelihood estimation for the minimum and the maximum $I_{read}$ values, using a two-states model. The difference between the two levels is indicated as $\Delta I = |H - L|$, representing the current increase due to the secondary path.

The physical mechanism responsible for the creation of the secondary path has been attributed to the possible nucleation and subsequent growth of new crystal grains outside the main percolation path [19, 23, 32], as schematically described in Fig. 2.43. The proposed physical interpretations rely on the time allocated to the Erase operation, since longer times are supposed to favor the creation process of a secondary crystalline path within the amorphous region.

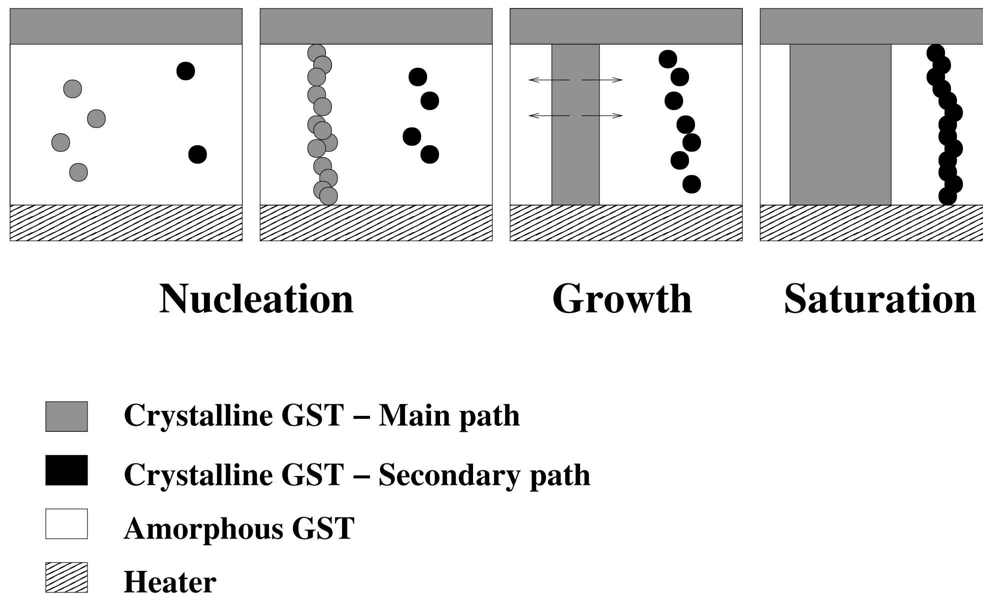



Figure 2.43: Schematic description of the secondary path creation kinetics. During the growth stage of the main percolation path, new GST crystallite grains may nucleate and subsequently join together, contributing to the secondary path formation.

The presence/absence of a secondary crystalline path after Erase in PCM cells has been statistically modeled by a two-states Markov chain (see Fig. 2.44) [22], [33]. In this model $\eta$ and $\theta$ represent the probability of remaining in the level $L$ and $H$, respectively. The transition probabilities from level $L$ to $H$ and from $H$ to $L$ are indicated as $1 - \eta$ and $1 - \theta$, respectively. The model also presented a robustness criterion to correctly identify the two levels $L$ and $H$ in the presence of an inherent noise superposed to the measured $I_{read}$ [22].

The analysis of the state probabilities $\eta$ and $\theta$ in a PCM array evidences a bimodal Gaussian distribution (see Fig. 2.45 and Fig. 2.46). A clear correlation is evidenced by scatter plotting the $\eta$ and $\theta$ probability coefficients for all the array cells (see Fig. 2.47). The points accumulates on the center of the plot as a signature of a linear correlation between the two state probabilities. In particular, by

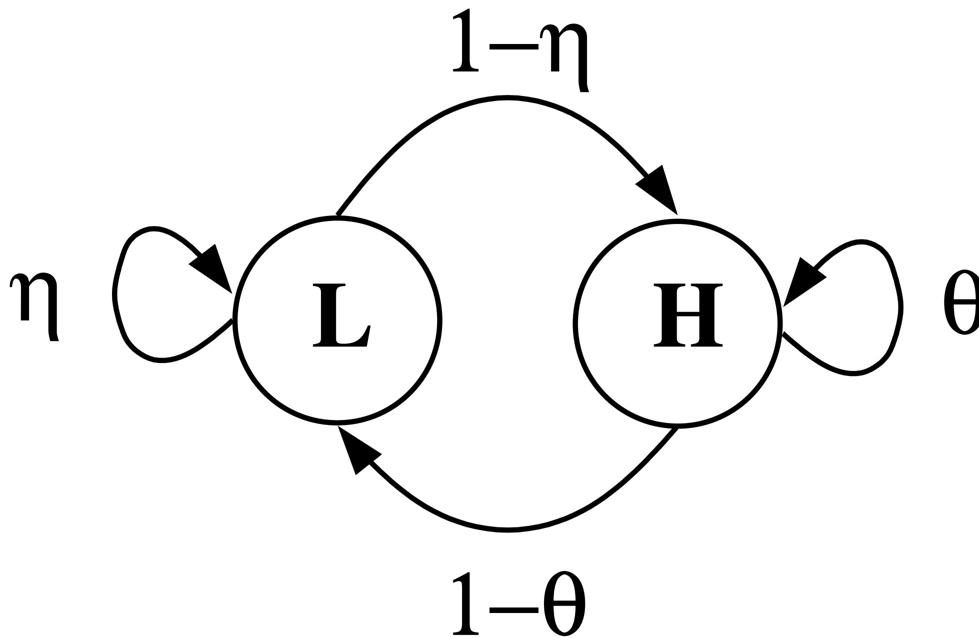



Figure 2.44: Two-states Markov chain used for modeling the secondary path presence/absence condition in PCM cells, where $\eta$ and $\theta$ represent the probability of remaining in the L and H status, respectively.

focusing on the $\eta$ distribution, a sub-population of cells which is more prone to create a secondary path can be observed. In the rest of the discussion i will concentrate our attention on these cells, characterized by $\eta < 0.5$. The dependency of the probability of creating a secondary conductive path on the erase pulse duration $t_H$ is clearly evidenced: as expected, longer erase times increase the creation probability of the secondary path. The effect of the Erase duration can be observed also in the $\theta$ distribution, showing that short erase times reduce the probability of creating a secondary conductive path. In this case too, we will concentrate our analysis on those cells characterized by $\theta < 0.5$.

The shift entities $\Delta I$, on the contrary, slightly depend on the Erase time $t_H$, as evidenced in Fig. 2.48, consistently with the theory that the geometrical dimen-

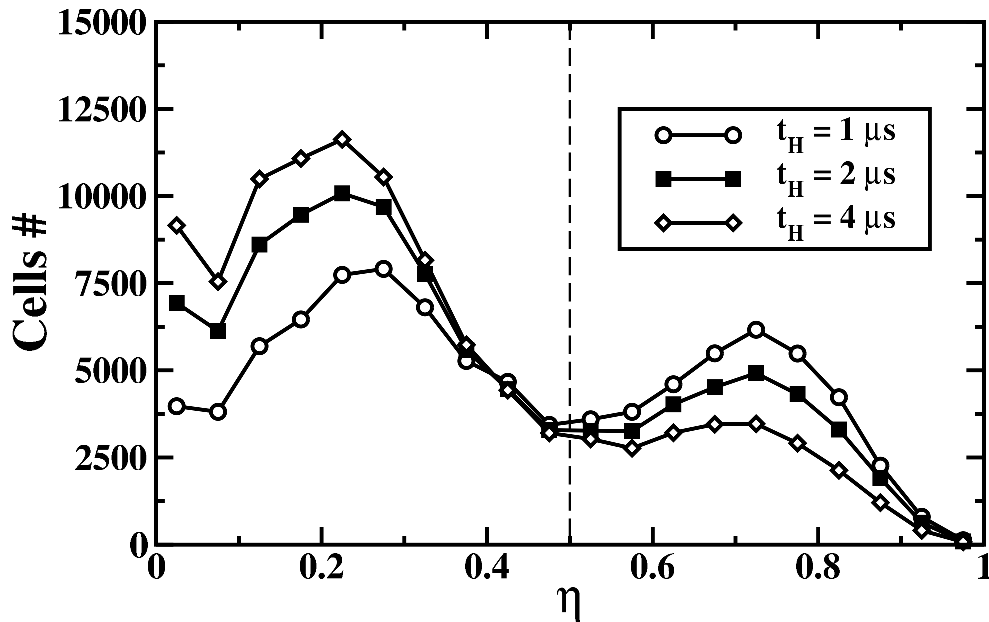



Figure 2.45: Distribution of the probabilities $\eta$ of remaining in the $L$ state for three pulse durations $t_H$ of the Erase operation. The cells population prone to create the secondary shunt is identified by the first Gaussian peak with $\eta < 0.5$.

sions of the secondary path are principally voltage dependant [23] and that longer erasing pulses just give sufficient time for the crystal grains to join together.

Figs 2.49 and 2.50 show the dependency of the number of state transitions and of the $\Delta I$ shifts on the average $L$ levels for the cells characterized by $\eta < 0.5$, respectively. In the two figures, for all cells with a specific average $L$ value within 200 Program/Erase cycles, symbols represent the average number of state transitions and of $\Delta I$ shifts for the three different Erase times, whereas the error bars represent the standard deviation from the mean values. As for the number of state transitions (Fig. 2.49), the only observable result is the expected increase of the average number of state transitions from $L$ to $H$ with $t_H$, whereas no other correlations between the number of state transition and the $L$ levels can be determined. On the contrary, the $\Delta I$ shifts of Fig. 2.50 are fully consistent with the basic

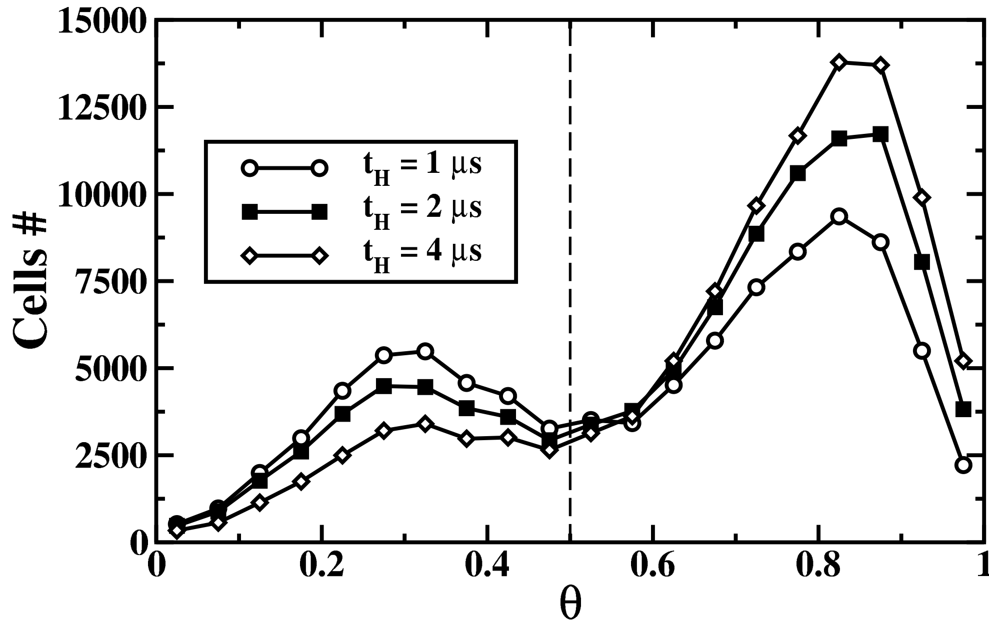



Figure 2.46: Distribution of the probabilities $\theta$ of remaining in the $H$ state for three pulse durations $t_H$ of the Erase operation.

theory of two conductive paths in parallel: the increment of the current flowing through two conductive paths in parallel with respect to the case of only one path depends on the ratio between the conductance of the added path with respect to the already present one. For these reasons, when a cell presents a large average $L$ value within cycles, the probability of creating a secondary path does not show pronounced differences with respect to other cells, while the measured average $\Delta I$ is limited since a large $L$ value corresponds to a highly conductive primary crystalline path.

A statistical model of $I_{read}$ after the Erase operation during cycling can be built as follows. A vector of 512K $I_{read}$ values has been randomly generated with a Gaussian probability distribution with $\mu_{Iread}$ and $\sigma_{Iread}$ experimentally evaluated after the first erase operation (see Table 2.3). Through a Monte Carlo simulation

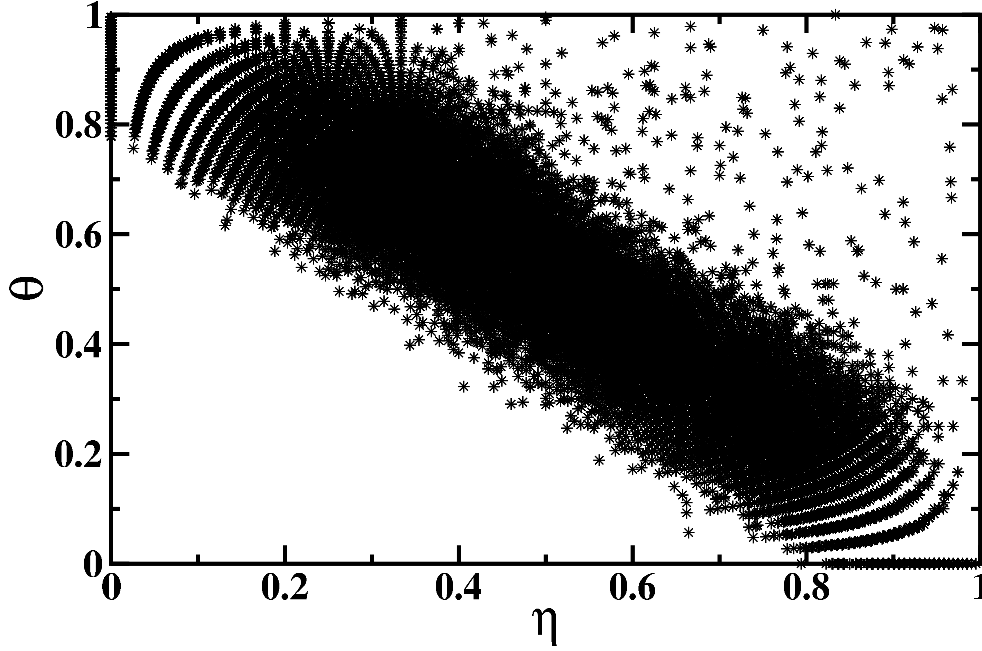



Figure 2.47: Scatter plot of $\eta$ versus $\theta$ state probabilities for $t_H$ = 2$\mu$s (similar results are achieved for other pulse durations). The points accumulates in the center of the plot clearly evidencing a linear correlation between the two probability coefficients.

it is then possible to estimate the behavior of the current $I_{read}$ cycle after cycle.

According to the two-states Markov model, each cell $i$ within the vector has a probability $\eta_i$ of remaining on state $L$ and a probability $\theta_i$ of remaining on state $H$. State probabilities are assigned to the PCM cells by fitting the experimental distributions of Fig. 2.45 and 2.46 with bimodal Gaussian distributions:

$$\begin{cases} \eta_i = a_1 * e^{-\left(\frac{x-b_1}{c_1}\right)^2} + a_2 * e^{-\left(\frac{x-b_2}{c_2}\right)^2} \\ \theta_i = a_3 * e^{-\left(\frac{x-b_3}{c_3}\right)^2} + a_4 * e^{-\left(\frac{x-b_4}{c_4}\right)^2} \end{cases} \tag{2.15}$$

where $x$ is a pure number in the [0,1[ interval and $a_j$, $b_j$, $c_j$ (j = 1,..,4) are the fitting parameters of the distributions (see Table 2.4).

The shift $\Delta I_i$ can be calculated as a function of $t_H$ according to the following

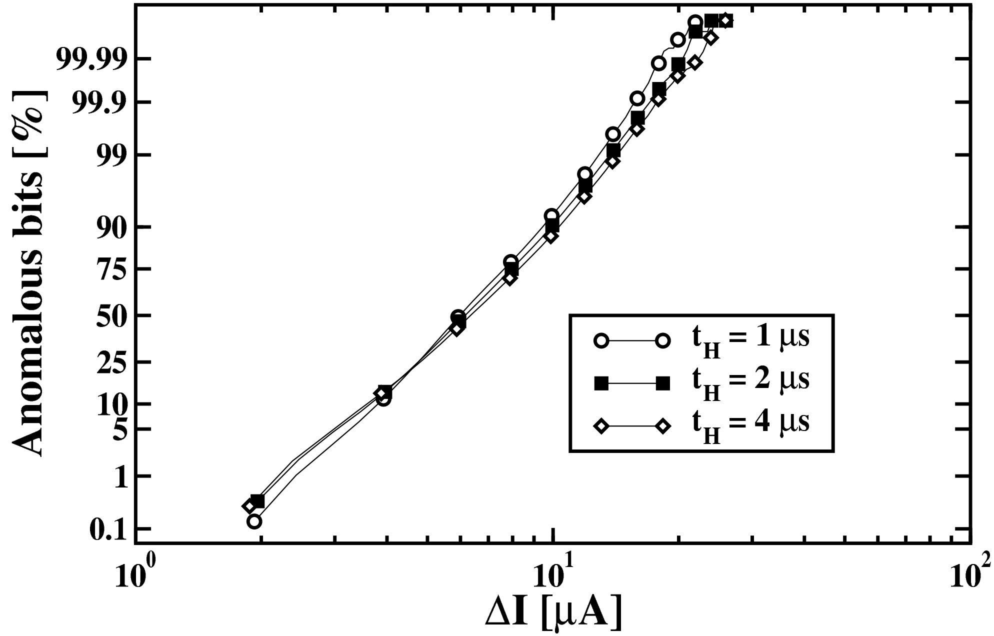



Figure 2.48: Log-normal probability plot of the $\Delta I$ distribution.

Table 2.3: Parameters used for the statistical modeling of the gaussian part of the $I_{read}$ distributions extracted from the experimental data at cycle 1.

| | $t_H$ [$\mu$s] | | |
|---|---|---|---|
| | 1 | 2 | 4 |
| $\mu_{Iread}$ [$\mu A$] | 44.5 | 44.8 | 45.3 |
| $\sigma_{Iread}$ [$\mu A$] | 1.24 | 1.18 | 1.27 |

equation:

$$F\left(\Delta I_i(t_H)\right) = \frac{1}{2} erfc\left[\frac{ln(\Delta I_i(t_H)) - T_{50}(t_H)}{\sigma(t_H)}\right] \tag{2.16}$$

where $T_{50}(t_H)$ and $\sigma(t_H)$ are the characteristic parameters of the log-normal distributions fitting the data of Fig. 2.48 obtained through Maximum Likelihood Estimation (see Table 2.5).

Finally $I_{read,i}$ can be calculated by adding or subtracting the $\Delta I_i$ shift according to the state transition probabilities of a considered cell and dependently on its

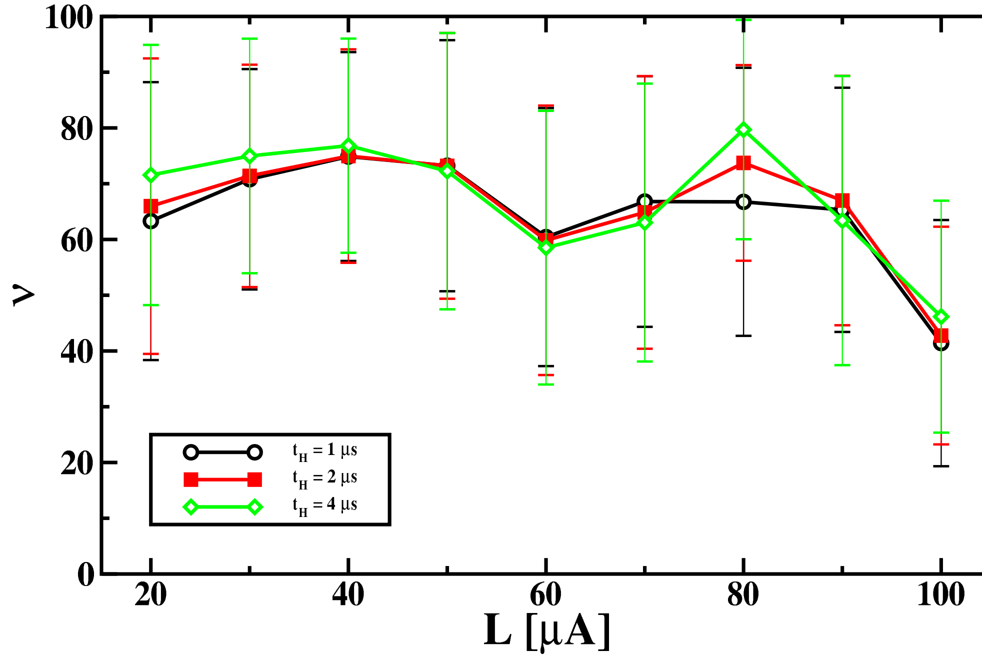



Figure 2.49: Numbers of state transitions for any set of cells characterized by the same average $L$ level within 200 program/erase cycles for the three erasing times $t_H$. Symbols denote the average number of state transitions, whereas the error bars represent the standard deviation from the mean value.

actual state:

$$\text{L-state cells: } I_{read,i} = \begin{cases} I_{L,i} + \Delta I_i & \text{with probability } 1\text{-}\eta_i \\ I_{L,i} & \text{with probability } \eta_i \end{cases} \tag{2.17}$$

$$\text{H-state cells: } I_{read,i} = \begin{cases} I_{H,i} - \Delta I_i & \text{with probability } 1\text{-}\theta_i \\ I_{H,i} & \text{with probability } \theta_i \end{cases} \tag{2.18}$$

Fig. 2.51 shows the agreement between the calculated and the measured $I_{read}$ distribution after 200 program/erase cycles for the three different $t_H$. The inset

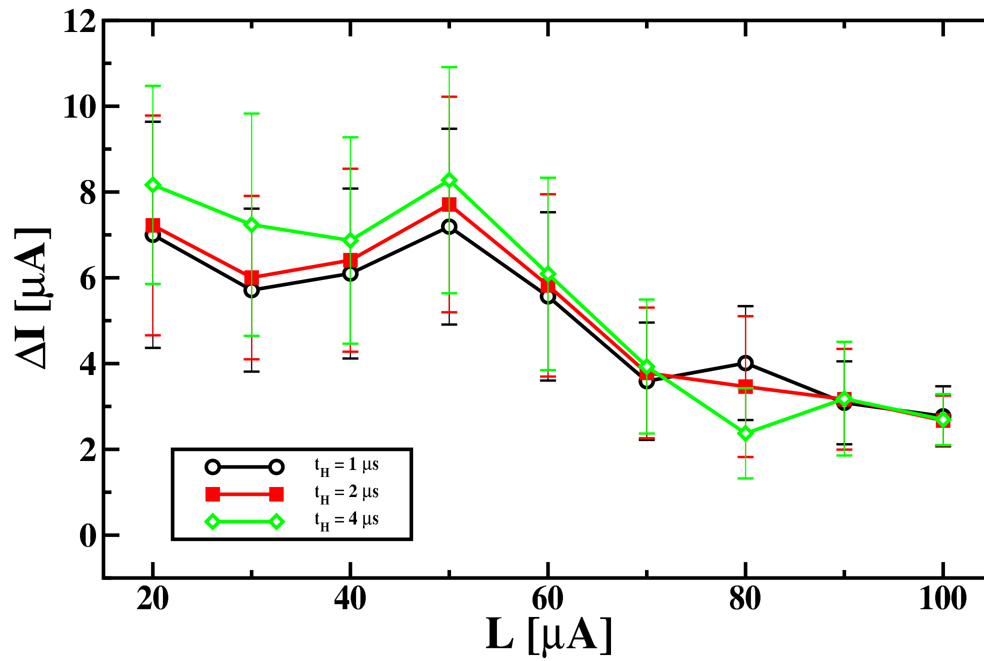



Figure 2.50: $\Delta I$ shifts for any set of cells characterized by the same average $L$ level within 200 program/erase cycles for the three erasing times $t_H$. Symbols denote the average $\Delta I$ shift, where the error bars represent the standard deviation from the mean value.

of Fig. 2.51, in particular, shows the ability of proposed model in tracking the distribution tails.

Table 2.4: Parameters used for the fitting of the bimodal gaussian distributions of $\eta$ and $\theta$ state probability.

| | $t_H$ [$\mu$s] | | |
|---|---|---|---|
| | 1 | 2 | 4 |
| $a_1$ | 7488 | 4834 | 5728 |
| $b_1$ | 0.25 | 0.23 | 0.23 |
| $c_1$ | 0.24 | 0.11 | 0.11 |
| $a_2$ | 5971 | 6291 | 18040 |
| $b_2$ | 0.72 | -0.12 | -1.07 |
| $c_2$ | 0.15 | 0.99 | 1.25 |
| $a_3$ | 9180 | -19620 | 13240 |
| $b_3$ | 0.80 | 1.04 | 0.83 |
| $c_3$ | 0.17 | 0.10 | 0.14 |
| $a_4$ | 5260 | 4561 | 3430 |
| $b_4$ | 0.33 | 0.25 | 0.46 |
| $c_4$ | 0.21 | 4.21 | 0.34 |

Table 2.5: Parameters used for the calculation of the $\Delta I_i$ shift.

| | $t_H$ [$\mu$s] | | |
|---|---|---|---|
| | 1 | 2 | 4 |
| $T_{50}(t_H)$ [$\mu A$] | 5.80 | 5.87 | 5.98 |
| $\sigma(t_H)$ [$\mu A$] | 0.33 | 0.35 | 0.36 |

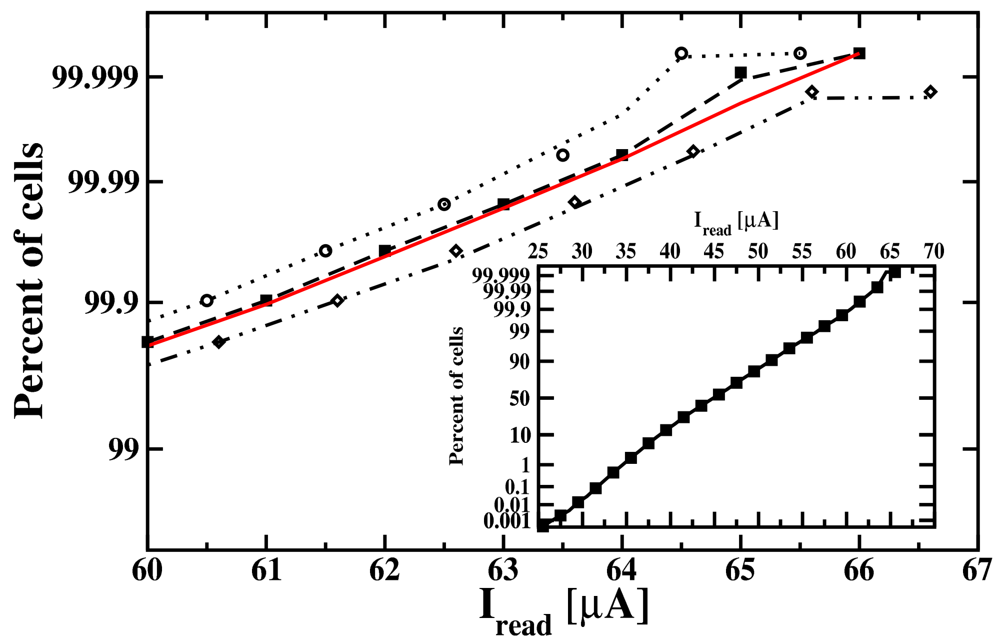



Figure 2.51: Measured (symbols) and calculated (lines) $I_{read}$ tail distributions after 200 Program/Erase cycles for the three pulse durations. $t_H$ = $1\mu$s: circles and dotted line; $t_H$ = $2\mu$s: squares and dashed line; $t_H$ = $4\mu$s: diamond and dashed/dotted line. The full line represents the Gaussian Model behavior for $t_H$ = $2\mu$s. Similar Gaussian behavior are found for other $t_H$ values. The inset shows experimental data and fitting of the main part of the distribution for $t_H$ = $2\mu$s (the two other $t_H$ cases are almost superimposed).

# Chapter 3

# P-channel SimpleEE Memories

The existing embedded nonvolatile memory technologies have failed to deliver a cost effective solution for SoC (System on Chip) applications. The major reason has been that most of these technologies were not designed specifically for the embedded applications. There have been two approaches for the embedded nonvolatile memories. One is to take the high density stand alone memory technology and use it for embedded applications and the other is to use the basic logic technology and make a nonvolatile element on it without adding additional masks. Both these approaches have their pros and cons. The first approach gives a very small cell size, but it has high cost, low yield, and compatibility with base CMOS process problems, while the other approach gives a very large cell which can not satisfy the density requirements. The C/H35 device of Austriamicrosystems [34] developed on the framework of the European project FP7-ATHENIS, better known as SimpleEE is a non-volatile EEPROM organized in a 4kx16 bit array designed for automotive and high temperature application based on 0.35m CMOS-technology. The technology differs from the standard EEPROM cells or-

ganization due to the presence of two Side Select Transistors (SSTs) per cell and by using p-channel technology both for floating gate devices and SSTs. The embedded nonvolatile memory technology and modules described in this chapter is designed from ground up by keeping the embedded applications and their requirements. The module size is 1.22 mm $x$ 1.12 mm and the cell size is 2.925 $\mu$m $x$ 1.95 $\mu$m. The SimpleEE cell, shown in Fig.3.12 is designed for very high reliability with more than 80$\%$ coupling ratio. The module functionality has been verified from -40°C up to 170°C. Consistent yields of more than 97$\%$ have been seen on several lots. The SimpleEE technology is based on the principle of storing charge on a floating-gate, which is the traditional methods used by EEPROM and flash memories. This method is normally sensitive to higher temperature and within the SimpleEE technology special measures have been taken to operate it at higher temperatures. Different lots were used for the characterization:

- Q3 Lot (Qualification Round 3 devices) both in Plastic and Ceramic package
- Q4 Lot (Qualification Round 4 devices) both in Plastic and Ceramic package

The main difference between the two lots consisted in a different oxide thickness used on the MOS transistors belonging to peripheral circuits of the memory (e.g. decoders and select transistors), in particular Q4 featured a thinner oxide with respect to the Q3 standard of 24 nm.

## 3.1 Read performance

The electrical characterization of the read operation provides information on the reading capabilities of the memory modules under different bias conditions and operating temperature. Several tests were carried in order to capture the substantial differences between Q3 and Q4 devices, thus providing an initial estimate of the devices performances to keep into account on further tests (e.g. endurance, disturbs, etc.). All the used devices were programmed using a Bulk programming mode (whole array gets programmed in one clock cycle) and then the current map of the arrays were retrieved at different temperature. Additional tests were performed to monitor the change of the cell output gain in relation to temperature and to evaluate the degradation of the cell conductivity path (floating gate transistor + SSTs series) in cycling.

Both device families proven great readout capabilities even at 200°C limit, as shown in Fig.3.1. Benchmarking the two lots shows that Q4 devices are characterized by a larger current level on PROGRAMMED state with respect to Q3 (see Fig.3.2). The output gain of the SimpleEE cells within the array has been extracted from previous data at different read voltage values, showing a strong dependence with temperature in accord to extracted values from read benchmark tests. The gain (i.e. $IV slope$ of the $I - V$ cell characteristic) has been calculated with the following equation:

$$IV slope = \frac{I_{read}}{(V_{read} - V_T)} \tag{3.1}$$

Indeed, temperature increase causes a decrease of the output gain as expected in traditional CMOS technology.

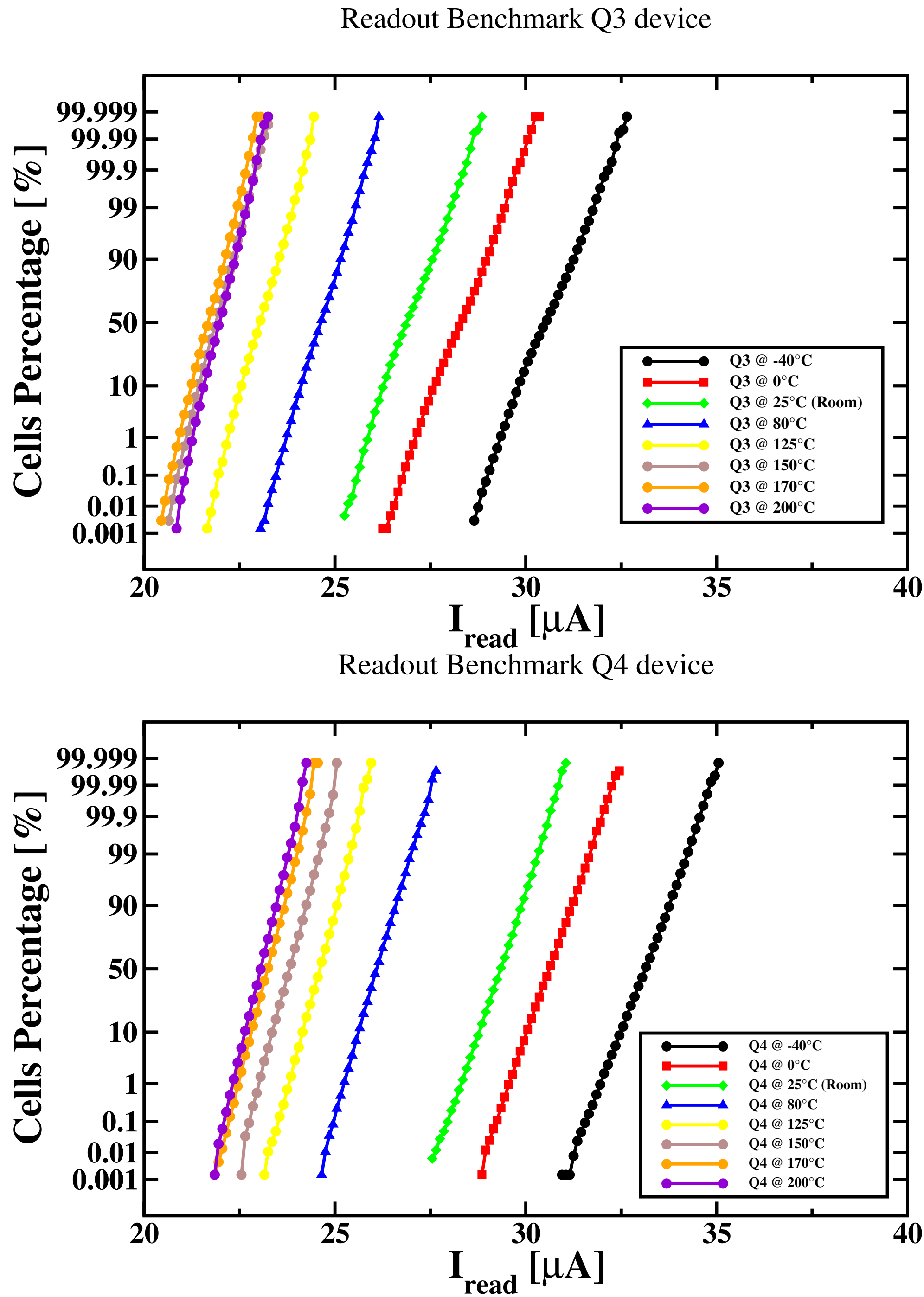



Figure 3.1: Read characterization of SimpleEE modules lots.

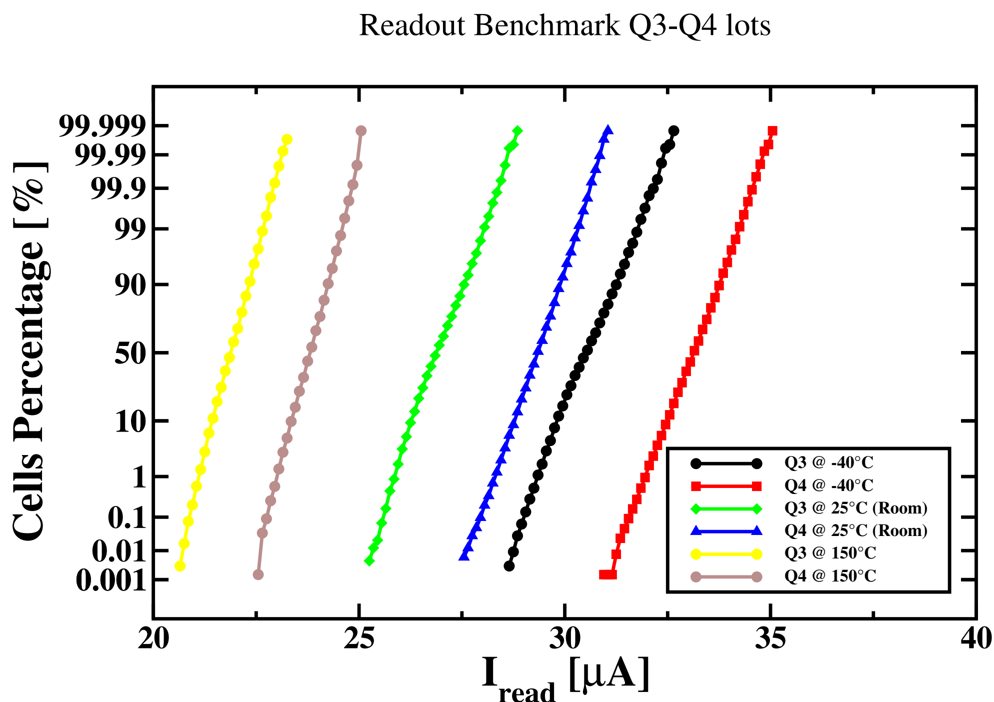



Figure 3.2: Read benchmark between Q3 and Q4 device families.

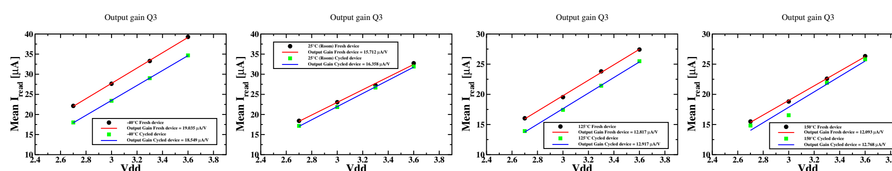



Figure 3.3: Extracted SimpleEE array output gain at different temperatures.

## 3.2 Program/Erase speed

The electrical characterization of the program and erase operations of the memory modules enable further considerations on the technology such as the performance validation at various operating temperatures. By analyzing the program and erase characteristics it is also possible to evaluate the speed of the memory in relation to the application of certain working conditions and to estimate how the degradation mechanisms (principally due to an endurance wear-out) can interfere with the

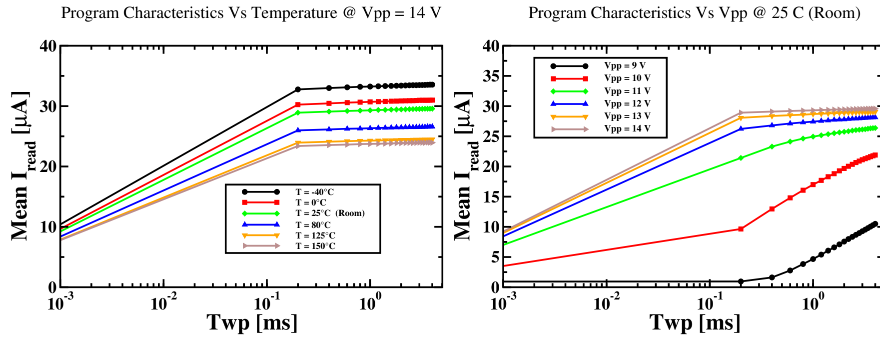



Figure 3.4: Program characteristics of the memory with respect to working temperature (left) and voltage exploited (right).

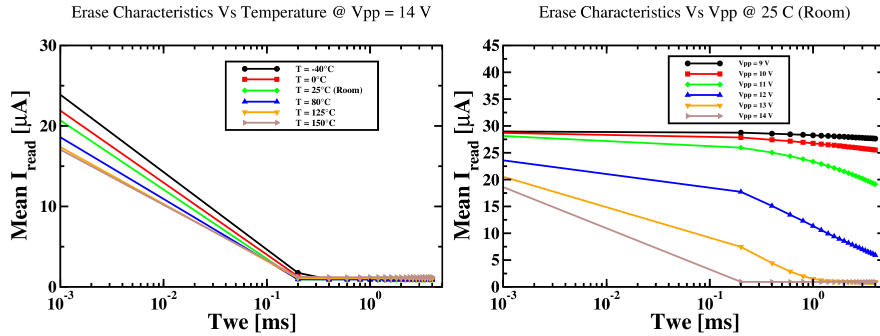



Figure 3.5: Erase characteristics of the memory with respect to working temperature (left) and voltage exploited (right).

basic functionalities of the memory. Both writing and erasing operations relies on the Fowler-Nordheim tunneling mechanism for moving electrons to/from the floating gate, thus providing low-power required during either program or erase.

The characterization of the program versus the voltage applied on the cells terminal shows the speed features of the memory module. With $V_{pp}$ voltages higher than 11 V only 200 $\mu$s of pulse width are required for correctly programming the SimpleEE array. The characterization versus the operating temperature shows the same results retrieved on measurements for evaluating temperature effects on reading. The characterization of the erase versus the voltage applied on the cells terminal shows the asymmetry on the speed features of the memory module. For

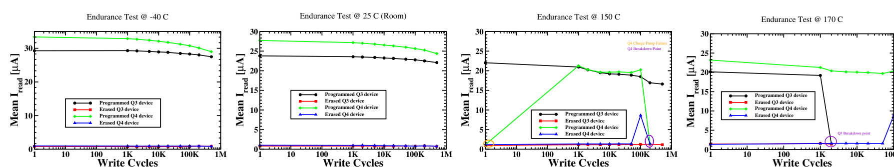



Figure 3.6: Endurance characteristics of the memory modules at temperatures: -40°C, 25°C (room), 150°C and 170°C.

a correct erasing of the SimpleEE array a higher Vpp voltage (at least 13 V) with respect to program is required due to physical effects to take into account (i.e. bulk is under strong inversion condition while on the program is on accumulation of majority carriers at interface). With $V_{pp}$ voltages higher than 13 V, 1 ms of pulse width are required for correctly erasing the SimpleEE array.

## 3.3 Endurance and data retention

For all non-volatile memory technologies is mandatory to investigate the capabilities of the memory modules to withstand a defined amount of program/erase cycles through endurance tests. The amount of cycles is determined typically by the target automotive requirements and is fixed as 500k write/erase cycles. In order to evaluate the endurance properties of the memory, the tests were performed at critical operating temperatures other than room. The endurance testing methodology is based on an application of 500k sequences of program and erase operations reading the memory content at regular intervals (quasi-logarithmic span).

No significant issues appears on both devices at -40°C. Normal wear-out of the memory cell is appreciable especially on the programmed state distribution (high currents). Read margin window closure is not significant up to 500k cycles. Q4 devices appear to be more sensible to degradation. At room tempera-

ture again no significant issues appears on both devices. Array current levels on programmed state are slightly lower than the measurements performed at -40°C. Normal wear-out of the memory cell is appreciable especially on the programmed state distribution (high currents). Read margin window closure is not significant up to 500k cycles. Q4 devices appear once again to be more sensible to degradation. At 150°C Q3 devices withstanded the 500k target cycling without any issues. Degradation of the read window margin for Q3 due to normal wear-out is still not significant. Q4 suffered a problem at cycle 1 due to a failure (temporary) of the internal charge pump for $V_{pp}$ voltage generation and then broke-down permanently at cycle 200k. Testing has been extended also on a non target automotive working temperature by reducing the cycle number to 100k. At 170°C Q3 has experienced a complete break-down after only 2k cycles. Until cycle 1k no significant issues appeared on the memory array. Normal wear-out were ongoing with no sensible impact on the read window margin. Q4 withstanded the 100k target cycling for this operating temperature even if at cycle 100k the erased distribution showed large tails due to a charge pump problem in generating the $V_{pp}$ internal signal for the erase operation. Programmed distribution showed large tails either. Fig.3.6 graphically summarizes the results.

The data retention tests are aimed at measuring the state of the memory array at time intervals in order to evaluate the capabilities to retain the stored information at different operating temperatures. The goal, for this technology generation, has been established to investigate data retention degradation by baking devices at different temperatures up to 200°C. In order to have a comparison of the degradation due to temperature bake with respect to the degradation occurring at room temperature, two devices were measured at the same time: a reference device (left

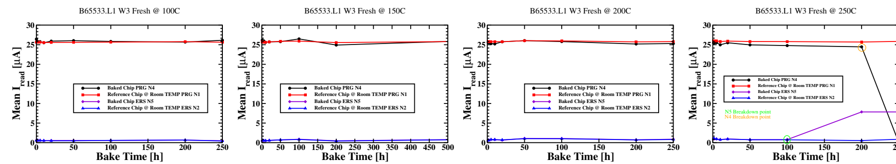



Figure 3.7: Data retention characteristics of the memory modules at bake temperatures: 100°C, 150°C (room), 200°C and 250°C.

outside the furnace for all test duration) and a baked one. For a complete picture of degradation, 6 devices (3 reference and 3 baked), were placed on a defined storage pattern before test start: 2 on programmed state, 2 on erased state and 2 with a checkerboard pattern. The stepped stress methodology was chosen for this type of measures: the bake starts at the first testing temperature (100°C) and every 250 hours the temperature increases of 50°C up to 250°C. The current distribution of the arrays were then retrieved periodically by measuring devices using a quasi-logarithmic time span.

The complete characterization required up to 1000 hours (approx. 6 weeks). Devices were marked using the following convention: N1, N2 are the reference devices respectively on programmed, erased state; N4, N5 are the bake devices respectively on programmed, erased state. From the plot of the current levels of the baked and reference arrays no significant degradation is appreciable, except for the test at 250C on which after 100h of baking at 250°C N5 device permanently broke-down. N4 device permanently broke-down after 200h of baking at 250°C. Before the failure the retention degradation was still not appreciable on both arrays state configuration. Checkerboard patterned devices did not show degradation too. Fig.3.7 graphically summarizes the results.

## 3.4 Disturbs robustness

Electrical disturbs in non-volatile memories are responsible of electrical failures occurring during the memory operating life and are defined as the undesired change of the state of a cell caused by writing and reading operations. One of the main goal of the disturbs characterization is to use a fast and complete methodology for disturbs identification and evaluation. Using the ”One-Shot” methodology proposed in [35] we can evaluate disturbs using this algorithm:

- The whole array is programmed with a random pattern and all read currents of the cells are measured and stored
- Disturb is applied to a random word within the array. The disturb consists in 10000 writing operations performed with internal waveforms on the same word address randomly selected. The same data is fed to the memory during each one of the 10000 writing operations. Before each disturb procedure a new random word is used as writing data.
- All read currents of the cells are measured and stored again
- Repeat from 1 for 512 times
- All possible disturb signal configurations are then checked

These tests required a characterization with different operating temperature in order to validate the robustness of the technology. Furthermore, for a complete picture of the disturbs characterization, tests were carried both on fresh and cycled devices. Disturbs performed at room temperature on the erased state of a virgin device are shown in Fig.3.8. Disturbs can be both positive and negative and are

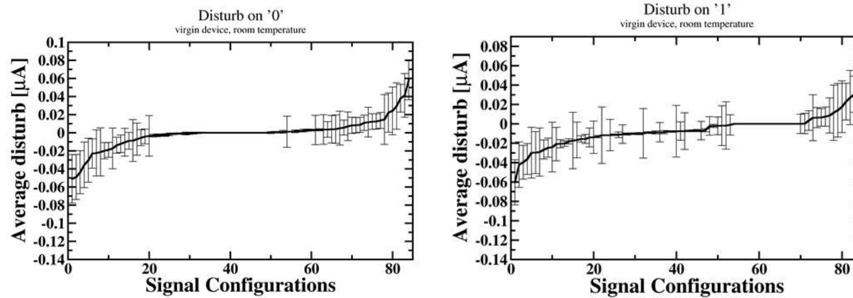



Figure 3.8: Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a virgin device.

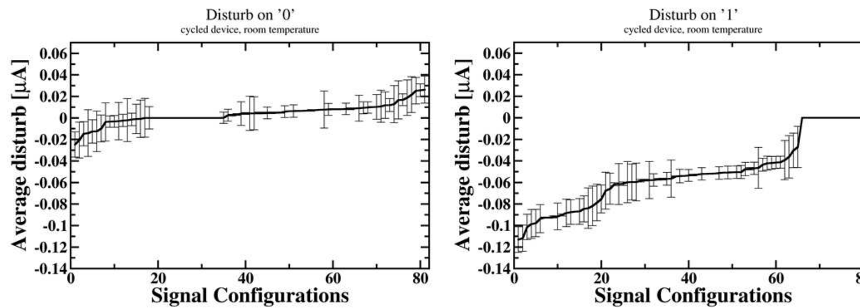



Figure 3.9: Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a cycled device (500K cycles at room temperature).

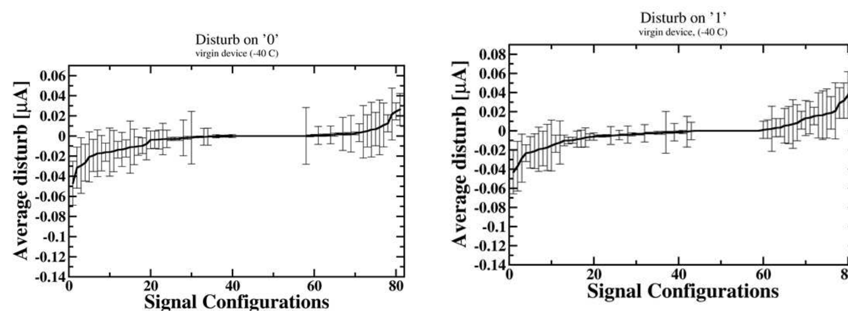



Figure 3.10: Average disturb on the erased (left) and programmed (right) state for each signal configuration sorted in ascending order for a cycled device (500K cycles at room temperature).

limited in the 50nA range. Disturbs performed at room temperature on the programmed state of a virgin device are shown in Fig.3.8 and are not very different

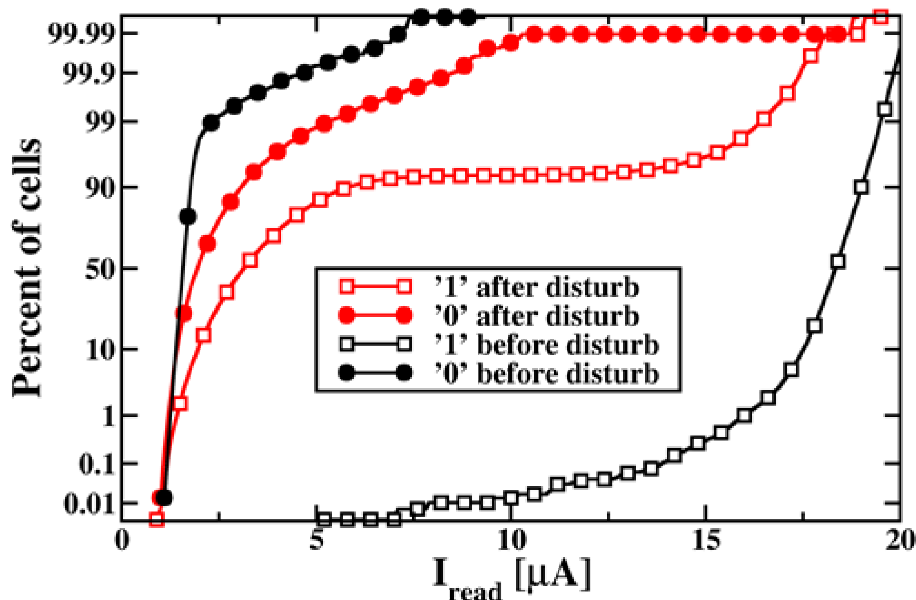


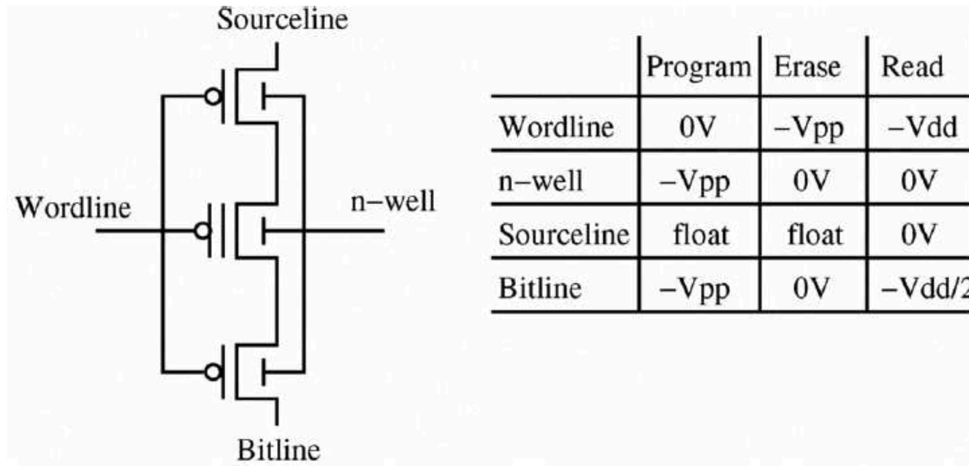

Figure 3.11: Distributions before and after disturb application for a sample cycled at hot temperature (150°C) for 500K cycles.

with respect to the previous case. Disturbs performed at room temperature on the erased state of a cycled device are shown in Fig.3.9. Disturbs can be both positive and negative and are limited in the 30nA range. Disturbs performed at room temperature on the programmed state of a cycled device are displayed in Fig.3.9. The measured disturb is negative only and a larger shift is measured with respect to the virgin device. The maximum average shift turns out to be 120nA which is still very low and is not an issue for reliability. Same results appears on devices tested at -40°C. The maximum wear-out and disturb sensitivity can be reached by cycling at hot temperature. We performed 500K cycles at 150°C in order to bring the memory degradation to the limit. The distributions of both program and erased states are shown in Fig.3.11. The degradation is so high that window closure is visible already before the disturb test. In these conditions one single

disturb (10000 pulses) can have a dramatic impact on the state of unselected cells as shown by the distributions after the disturb.

## 3.5 Evidence of the erratic bits

Erratic behavior consists in random threshold voltage shifts occurring in a cell during Fowler Nordheim (FN) tunneling operations [36]. As a result, neglecting measurement noise and threshold voltage changes due to cycling induced degradation, the final threshold voltage (VT ) after electron tunneling operation plotted versus the number of cycles can exhibit random shifts in a way which is similar to a Random Telegraph Signal (RTS) [37]. Up to now erratic phenomena have been observed and studied on n-channel floating gate devices only. In NOR architectures, the major risk related to the presence of erratic phenomena is the probability a memory cell becomes overerased. In this condition, due to the very low VT, even an unselected cell can drain large drain leakage currents which can interfere with the correct outcome of both a program or a reading operation [38]. Moreover, the presence of erratic bits is intrinsically related to a tail which is typically observed in the threshold voltage distributions after FN tunneling operation and it may therefore influence the threshold voltage distribution width and read window [39]. It has been shown, for the first time, that erratic phenomena are also present in p-channel technologies and that the considered p-channel SimpleEE (i.e. p-EEPROM) technology is intrinsically robust against them. From a scientific point of view, the observation of erratic behaviors in p-channel devices represents an important discovery in the puzzling physics of this mechanism. A comparison between the p-channel device and a conventional Flash shows also




| | Program | Erase | Read |
|---|---|---|---|
| Wordline | 0V | −Vpp | −Vdd |
| n−well | −Vpp | 0V | 0V |
| Sourceline | float | float | 0V |
| Bitline | −Vpp | 0V | −Vdd/2 |

Figure 3.12: p-EEPROM cell architecture (left) and table summarizing normal operating conditions (right).

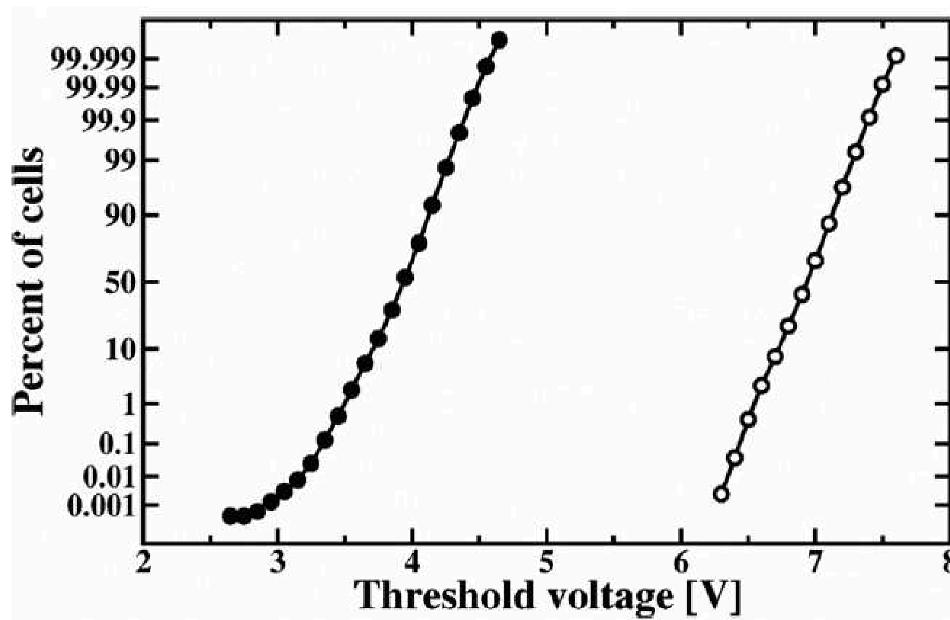



Figure 3.13: Erased (left) and programmed (right) distribution of the Flash sample.

that the former exhibits a minor degree of erratic behavior. A physical interpretation based on the Anode Hot Hole Injection [40] is suggested.

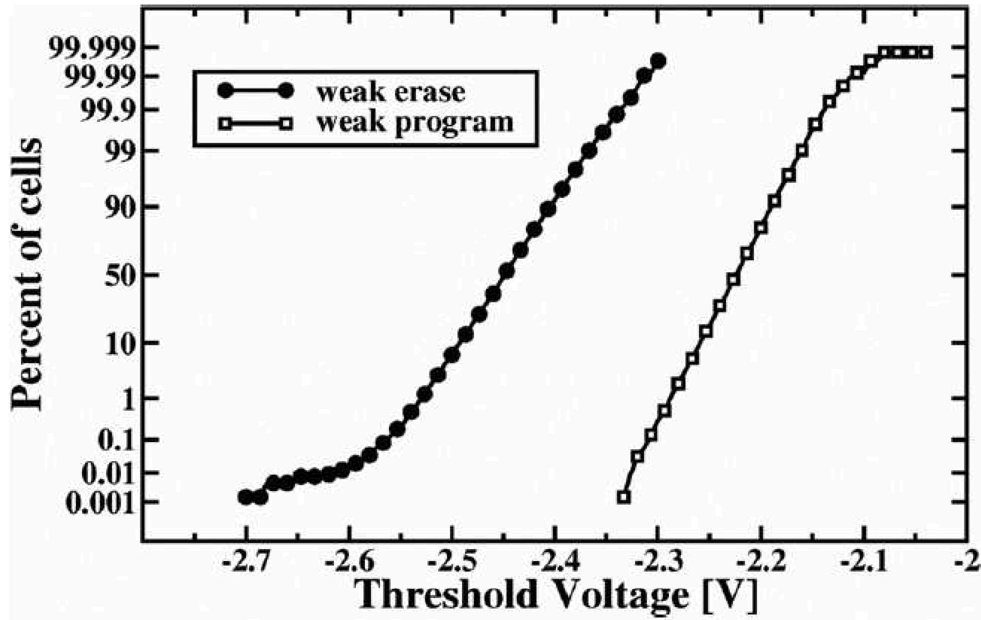



Figure 3.14: Erased (left) and programmed (right) distribution of the p-EEPROM sample.

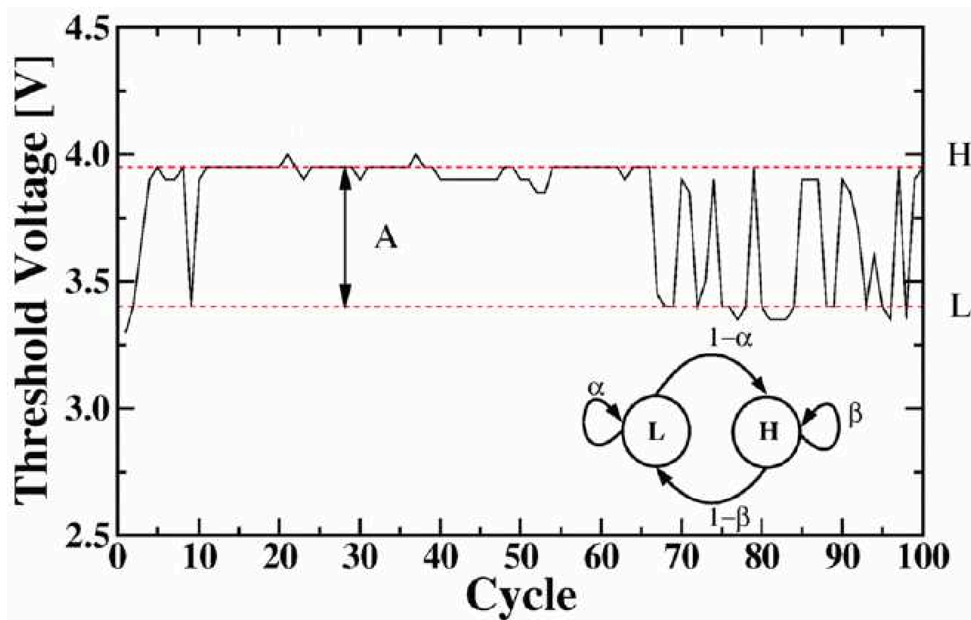



Figure 3.15: Example of erratic cell in the Flash sample.

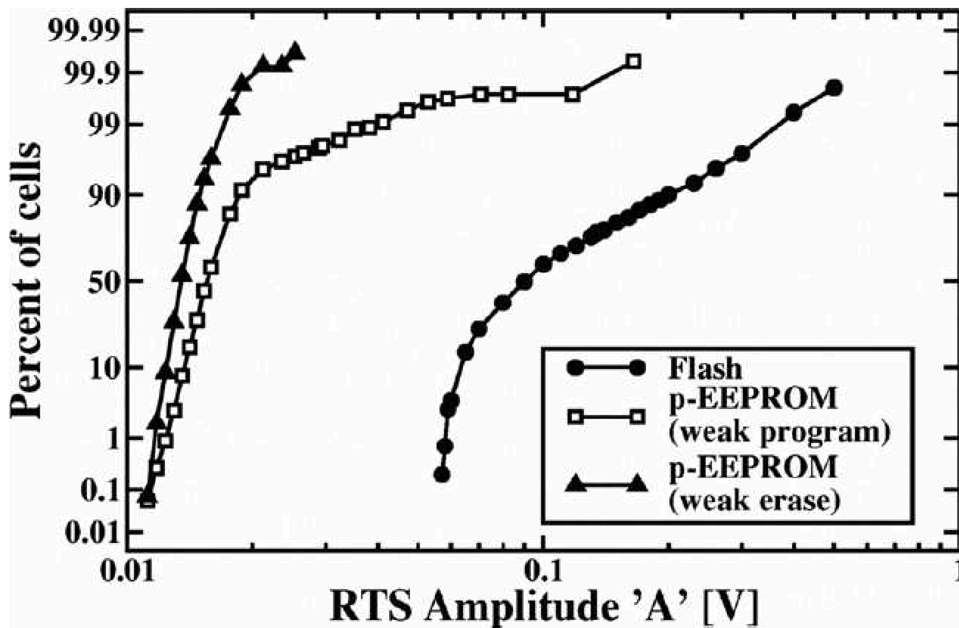



Figure 3.16: Lognormal distribution of threshold voltage shift of erratic cells.

Experiments have been carried out on 64Kbit p-EEPROM test chips featuring 8.5 nm tunnel oxide organized in one sector with 256 rows (wordlines) and 256 columns (bitlines) provided by austriamicrosystems AG. Cell architecture is shown in Fig.3.12. Reading operation is performed by measuring the cell read current ($I_{read} < 0$) in a Direct Memory Access mode with voltage configurations shown in the table in Fig.3.12 where $V_{dd}$ =3.3 V.

Standard program/erase operations are performed by applying constant voltage waveforms: $V_{pp}$=14 V to the control gate for 2ms during programming and $V_{pp}$=14V for 4ms to the n-well during erasing while keeping grounded other voltages. The experimental observation of the erratic behaviors in this p-channel floating gate technology has required the use of non-standard writing waveforms in order to overcome current saturation effects due to the select transistors: $V_{pp}$=1O V to the control gate for 2ms during weak-programming and $V_{pp}$=12V for 1ms

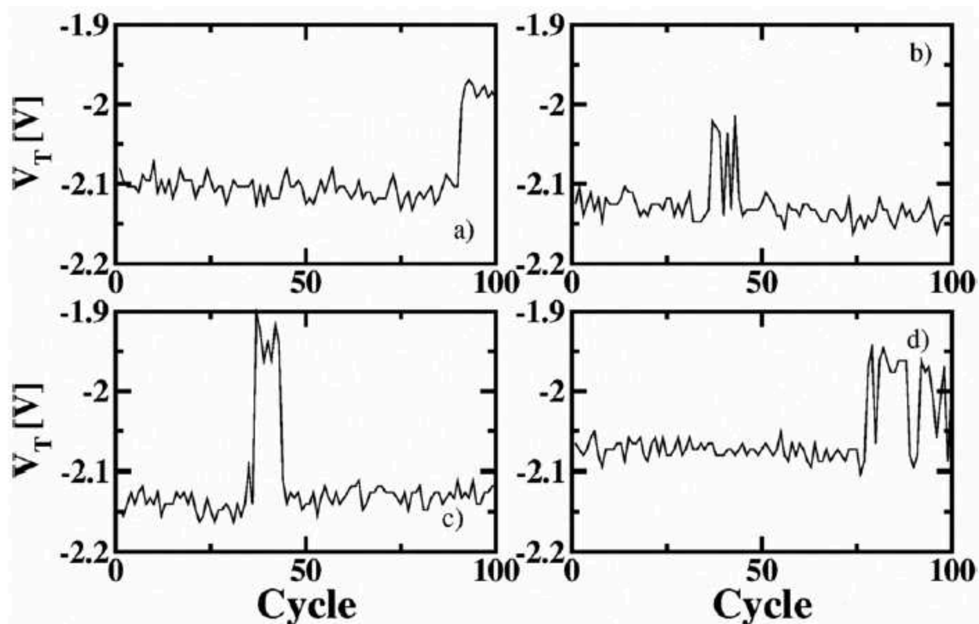



Figure 3.17: Examples of erratic bits in p-EEPROM device after weak-program operation during cycling.

to the n-well during weak-erasing. The measured read currents have been transformed into threshold voltages by taking into account the cell current voltage slope ($IVslope > 0$) inverting the (3.1).

Experimental results have been compared with a cell population of 64K cells of a conventional NOR Flash technology featuring 7 nm tunnel oxide thickness hereafter simply denoted as Flash. Erasing has been carried out by applying 5 V to the substrate and -5 V to the control gate of the selected cell for 50 ms. Programming has been carried out via Channel Hot Electron with 3 V to the drain and 8V to the control gate. It has been performed performed 100 program/erase cycles for both the Flash and the p-EEPROM devices. In the case of Flash we measured the threshold voltage after erase cycle after cycle whereas in the case of p-EEPROM we measured the read current after each program (and erase) opera-

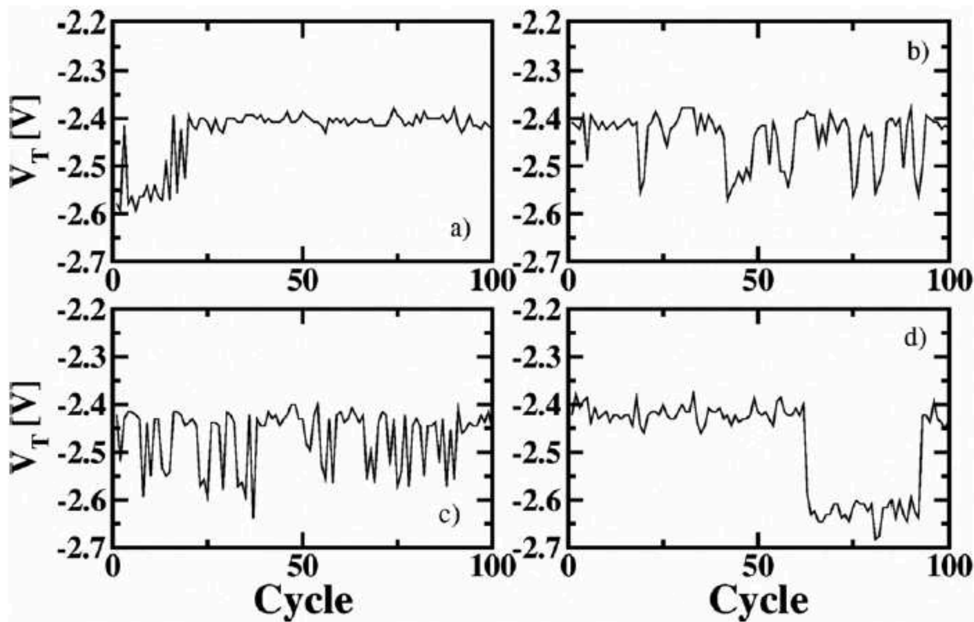



Figure 3.18: Examples of erratic bits in p-EEPROM device after weak-erase operation during cycling.

tion performed with nonstandard waveforms, i.e. weak program and weak erase. Erratic behaviors have been analyzed by using a new RTS analysis method [41] which is able to extract a complete set of RTS statistical parameters (see Fig.3.15): the two levels of the RTS signal (low level L and high level H) and the average distance between them, or erratic shift (A parameter). The statistical nature of the shift from one level to the other has been modeled with a two state Markov model featuring transition probabilities $\alpha$: and $\beta$ [41].

Fig.3.13 and Fig.3.14 show the programmed and erased threshold voltage distribution of Flash and p-EEPROM respectively. Fig.3.15 shows an example of erratic bit belonging to the Flash sample exhibiting a significant shift between the two levels (about 500mV). The distribution of the average threshold voltage shifts (A parameter) is shown in the lognormal probability paper in Fig.3.16 and it fea-

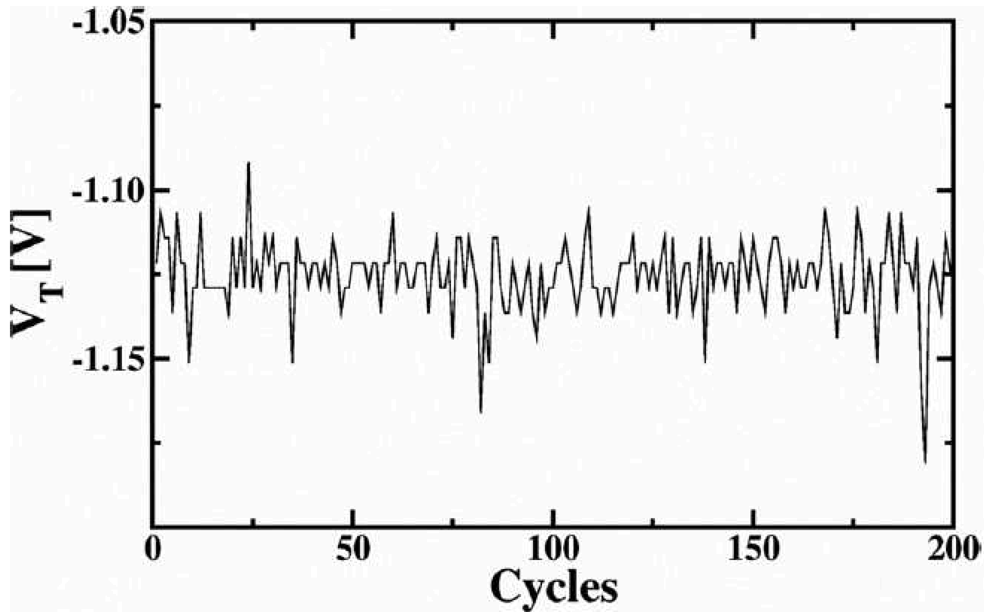



Figure 3.19: $V_T$ of an arbitrary p-EEPROM cell monitored during cycling using standard program waveforms. The same behavior can be observed on different cells and for more cycles.

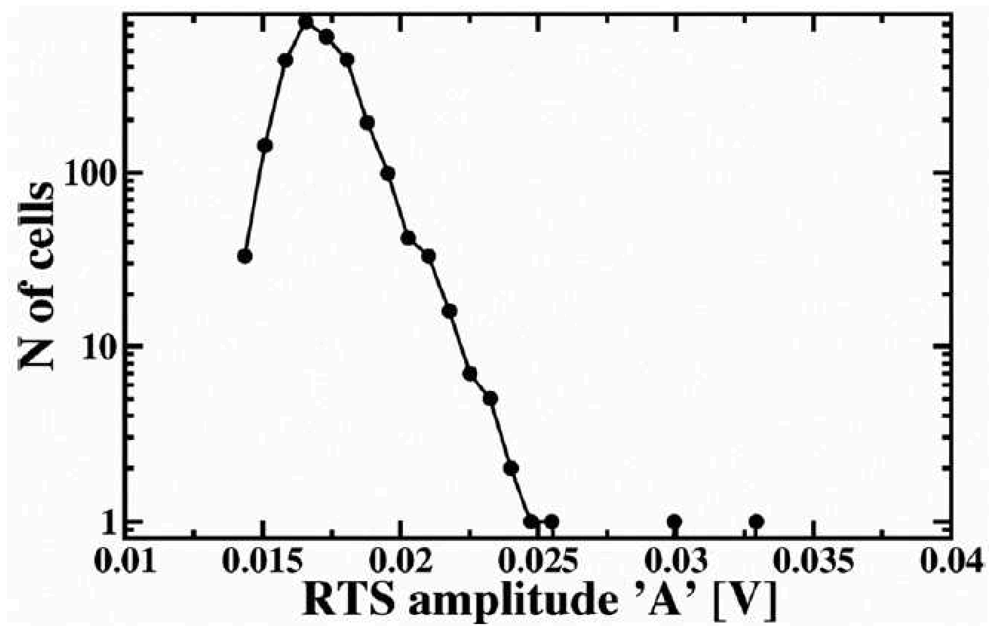



Figure 3.20: Shift distribution of the erratic bits in p-EEPROM device.

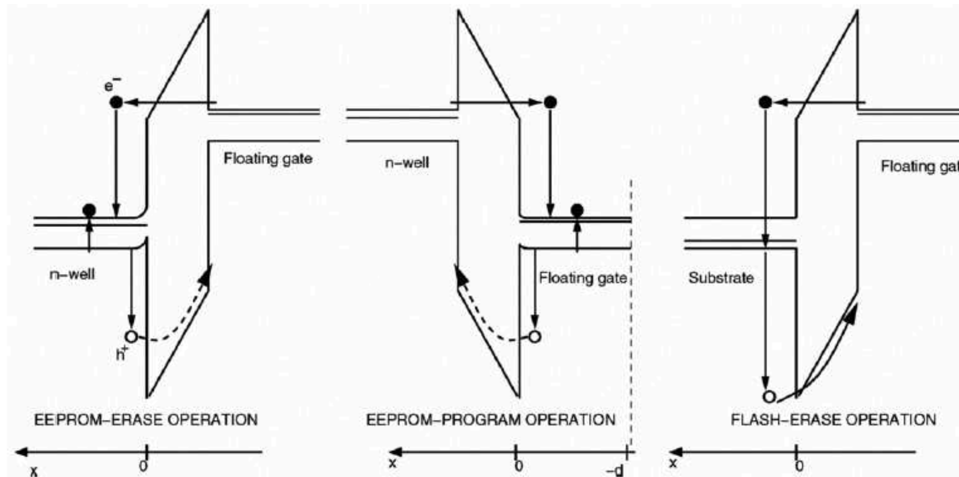



Figure 3.21: Different values of energies are involved in the AHHI (Anode Hot Hole Injection) behavior of p-EEPROM and FLASH. In particular, AHHI on FLASH is fed by higher energies with respect to p-EEPROM, thus inducing more erratic behaviors.

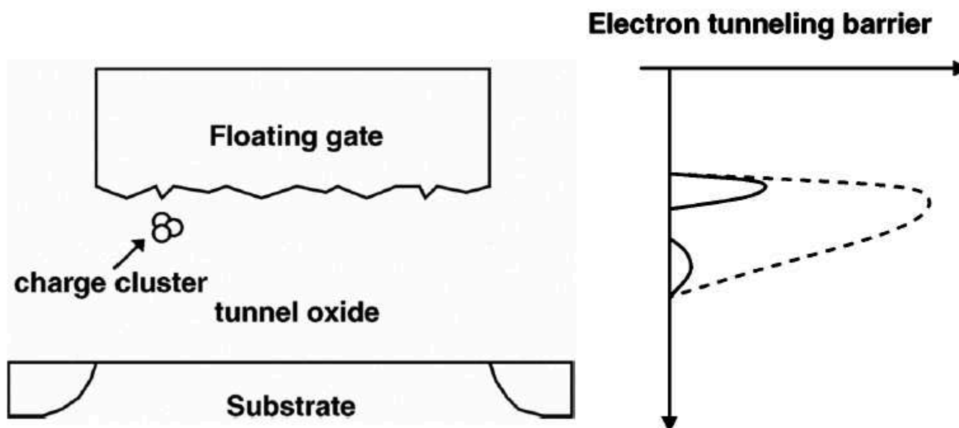



Figure 3.22: Impact of the charge cluster on electron tunneling barrier.

tures $T_{50}$=0.134 V and $\sigma$=0.54 V. Markov′s statistical parameters were $\alpha$=0.33, $\beta$=0.94 and a total number of 235 erratic bits has been detected with a high confidence level by the automated test. Fig.3.17 shows four examples of erratic bits observed in p-EEPROM during weak program operations. The distribution of threshold voltage shifts is shown in the lognormal probability paper in Fig.3.16

and it features $T_{50}$=0.7V and $\sigma$=0.6 V. Markov′s statistical parameters were $\alpha$=0.8 and $\beta$=0.5 and a total number of 44 erratic bits has been detected with a high confidence level by the automated test. Fig.3.18 shows four examples of erratic bits observed in p-EEPROM samples during weak erase operations. The distribution of threshold voltage shifts is shown in the lognormal probability paper in Fig.3.16 and it features $T_{50}$=0.9 V and $\sigma$=0.5 V. Markov′s statistical parameters were $\alpha$=0.4 and $\beta$=0.9 and a total number of 102 erratic bits has been detected with a high confidence level by the automated test. Previous results clearly show the presence of erratic behaviors in p-channel EEPROM when weak waveforms are used. On the contrary, when standard waveforms are used, the considered p-EEPROM technology proofs to be robust against erratic behaviors. As an experimental evidence we cycled the p-EEPROM samples with 5000 program/erase operations with standard waveforms and found no erratic bits, the behavior of the cells being like the one shown in Fig.3.19 for an arbitrary cell plotted over 200 cycles for the sake of clarity. The distribution of the most significant threshold voltage shifts is shown in Fig.3.20: an average shift of only 16mV has been measured. Similar results can be measured for the erasing operation using standard waveforms.

Results clearly show the presence of erratic behaviors under Fowler-Nordheim tunneling operation performed on p-channel floating gate devices. For the considered p-EEPROM technology, when standard waveforms are used, the erratic behavior of the cell including the select transistors becomes difficult to observe and a minor issue for this technology. In fact, when the p-EEPROM cell is programmed with standard waveforms, a larger charge is transferred to the floating gate. As a result, during a following reading operations, the larger overdrive volt-

age biases the floating gate device in a highly conductive state whose contribution to the series resistance with the two select transistors is negligible. In this way, the read current saturates with the number of program pulses. On the other side, in the case of the erasing operation, no significant current can be read using standard waveforms due to the high resistance contribution of the erased floating gate device in series with the select transistors. From a scientific point of view, the discovery of erratic behaviors in p-channel devices is very important for a deeper understanding of the physics behind the phenomenon. One of the most accredited models describes the erratic erase as the combined result of Anode Hot Hole Injection [40] and charge trapping within the tunnel oxide. Hot holes are generated during FN electron tunneling operation and some of them can have enough energy to be injected into the oxide (see Fig.3.21). Holes may then trap close to the cathode interface, eventually forming a positive charge cluster. The charge cluster can locally reduce the electron tunneling barrier which, in turn, causes an abrupt difference in the cell threshold voltage with respect to a normal condition (see Fig.3.22). Comparing these results with those obtained for a Flash array, it has been found, on the average, less erratic bits in p-channel devices: 44 erratic bits during programming and about 102 during erasing of p-EEPROM in contrast to 235 erratic bits during erasing of Flash. Also p-channel devices exhibit smaller erratic threshold voltage shifts with respect to Flash (see Fig.3.16). This different behavior can be the result of cell differences in terms of both technological and physical aspects. From the technological point of view the tunnel oxide quality plays an important role because it may influence hole trap density. However, it is reasonable to assume that oxide quality does not differ very much for the two technologies. Another important difference is the tunneling area. But since

the p-EEPROM features a larger tunneling area with respect to Flash (about 5 times larger), one would expect a larger number of hole traps in p-EEPROM and therefore more erratic behaviors, in contrast with the experimental data. Many physical aspects can be the cause of the different erratic behavior. In particular it is known that smaller tunnel oxide thickness are related to less erratic behaviors [42]. On the contrary, when comparing the p-EEPROM with respect to Flash, the former having a thicker oxide, the result is the opposite. Another important physical aspect is related to the electric field across the tunnel oxide during the tunneling operation. In fact, it is known that larger electric fields can induce more erratic behaviors [43]. A deeper analysis is required which properly takes into account the dynamic relation between the electric field and the hot hole generation during the considered writing operation, but first guess values show tunneling electric fields of the same order of magnitude for both technologies. Another root cause of the different behavior can be the different type of doping used for both well and floating gate in the two technologies. As shown in Fig.3.21, the Anode Hot Hole (AHHI) mechanism can be more efficient in a n-channel Flash with respect to a p-channel device, since electrons injected into the p-type substrate can release a larger energy with respect to the hole/electron pairs generation occurring in a p-channel device. During the Fowler-Nordheim tunneling operation in Flash, a larger number of hot holes can by injected in the oxide with respect to a p-EEPROM, and therefore more erratic bits are expected. This hypothesis, however, still requires deeper investigation and accurate modeling.

# Chapter 4

# NanoMEMS Memories

The embedded non-volatile memory (eNVM) market requires solutions able to withstand harsh environmental conditions and unconventional operative effectively over a wide range of working conditions. Take as example the under the hood automotive application market segment, that has been probably been the main driver contributor for the development of new eNVM concepts, followed by avionics, geothermal, and military. These examples are all the applications where the reliability is easily threatened by the applications critical environmental requirements (i.e. extreme temperatures, humidity, radiation, etc.). One of the technologies commonly used for embedded applications is the floating gate-based EEPROM. Several studies [44–46] have been performed in order to find the best tradeoff between the complexity of the architecture and the performances offered in terms of data retention and endurance. However, such a technology sometimes failed to provide enough safe margin for extremely high temperatures applications. An innovative solution to overcome the issues typical found with an eNVM has been proposed by the integration of an array integration of MEMS

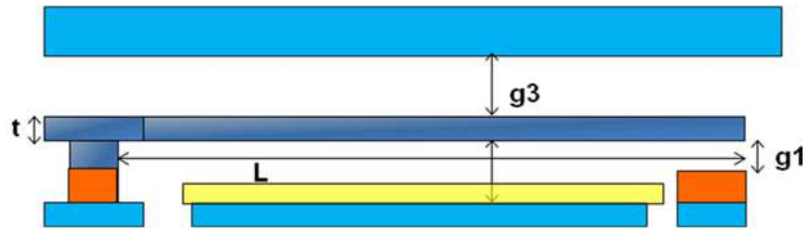



Figure 4.1: Simple cantilever element for physical principles evaluation.

switches into a CMOS circuit, the array being able to store the information using the position of the switch (i.e. being either on or off). The idea of using a MEMS cantilever switch as an individual non-volatile memory element was first patented by Cavendish Kinetics [47]. More recently larger cantilever switch devices have been developed for current switching and high frequency applications based on the seminal works [48, 49], but none of them was targeted embedded integration for non-volatile applications.

## 4.1 Device physics

The device physics underlying this MEMS memory developed on the framework of the European project FP7-ATHENIS will be first described by comparing the teeter-totter design with the equations governing the behavior of a more standard cantilever structure actuated by an electric field, whose structure is depicted in Fig. 4.1, with the following geometrical parameters: $t$ is cantilever thickness, $g_1$ is the gap between the cantilever tip and the contact plate, $g_2$ is the gap between the cantilever beam and the electrode plate and $g_3$ is the gap between the top of the cavity and the cantilever beam.

By applying a defined voltage on the electrode plate below the cantilever

beam, it is possible to drive the structure, which obeys to the classical formulation of the electrostatic force:

$$F_e(g) = \epsilon_0 \frac{WLV^2}{2g^2} \tag{4.1}$$

where $\epsilon_0$ is the vacuum permittivity constant, $W$ and $L$ are the width and the length of the contact plate, respectively, $V$ is the voltage applied on the electrode plate and $g$ is the gap between the cantilever beam and the electrode plate. When the gap consists of a near-vacuum and an oxide layer, this force can be calculated by using the effective gap formulation:

$$g = t_0 + \frac{t_{ox}}{\epsilon_{ox}} \tag{4.2}$$

where $t_0$ is the vacuum gap thickness, $t_{ox}$ is the oxide layer thickness and $\epsilon_{ox}$ is the dielectric constant of the oxide. To close the cantilever in order to Program the switch, the electrostatic pull-in force $F_e$ applied on the electrode plate needs to be larger than the spring like restoring force $F_r$ of the cantilever nearly approaching the contact plate. Hence, when the following inequality is reached:

$$F_r(g_1) < Fe(g_2) \tag{4.3}$$

the electrostatic force is larger than the restoring force and the cantilever will move to the contact electrode. The restoring force for a cantilever beam satisfies the following equation:

$$F_r(d) = \frac{EWt^3d}{4L^3} \tag{4.4}$$

where $E$ is the Young's modulus, $W$, $t$ and $L$ are respectively the width, the thickness and the length of the cantilever and $d$ is the displacement of the cantilever. Equation 4.4 is valid for $L$ values far greater than $d$. Therefore, when the cantilever is closed, by substituting $g_1$ into the eq. 4.4 it can be shown that:

$$\frac{EWt^3 g_1}{4L^3} < \epsilon_0 \frac{WLV^2}{2g_2^2} \tag{4.5}$$

In order to make the cantilever non-volatile, and thus keeping it on the contact electrode when no electrostatic charge is applied, the stiction force needs to overcome the spring force of the cantilever. Hence:

$$F_r(g_1) < F_s \tag{4.6}$$

where $F_s$ is the stiction force due to short range attractive forces at the contact including metal-to-metal bonding and Van der Waals forces [50].

Since it must be possible to open the cantilever (Erase operation or storage of a logical '0'), the electrostatic pull-off force from the top of the structure should be larger than the stiction force minus the restoring force.

$$F_e\left(g_3 + \frac{g_1}{2}\right) > F_s - F_r(g_1) \tag{4.7}$$

By substituting all the terms in eq. 4.3, 4.4 and 4.7, this yields to a set of constraints for dimensions $g_1$, $g_2$ and $g_3$, that must be fulfilled in order to obtain the features described previously:

$$\begin{cases} g1 < \frac{4F_sL^3}{Et^3W} \\ g2 < \sqrt{\frac{2\epsilon_0 L^4 V^2}{Eg_1 t^3}} \\ g3 < -\frac{g1}{2} + \sqrt{\frac{2\epsilon_0 L^4 V^2 W}{4F_s L^3 - Eg_1 t^3 W}} \end{cases} \tag{4.8}$$

This differs from the result for a teeter totter in that there is no spring like restoring force, so that the device is held in the program or erase state via the adhesion forces at the ends. For the above cantilever model there are theoretically an infinite set of $g_1$, $g_2$ and $g_3$ values satisfying the aforementioned equations, several Nanomech eNVM switch architectures with different feature sizes were investigated in this work. As $F_s$ will vary under different operating conditions, reliability requirements demand that the restoring force of the cantilever should be made as small as possible. A totally new architectural concept has been introduced by the teeter-totter design with a near-zero restoring force, which will obey eq. 4.6 under all operation conditions.

Another force that should be accounted for, in the physical analysis of the MEMS switches, is gravity, which would oppose the restoring force for a cantilever (depending on the chip orientation). However, the weight of the tiny cantilever has been found to be very small for these eNVM devices (approximately tens of femtograms), thus allowing to neglect the impact of the gravity. This has been proven by further mechanical acceleration tests.

In reality the picture is more complex than introduced in this paragraph as there are many non-linearities and fringe effects, whereas the forces will also modify during the movement of the cantilever. However, these equations are intended as a justification for the shift in design from a cantilever to a teeter-totter, based

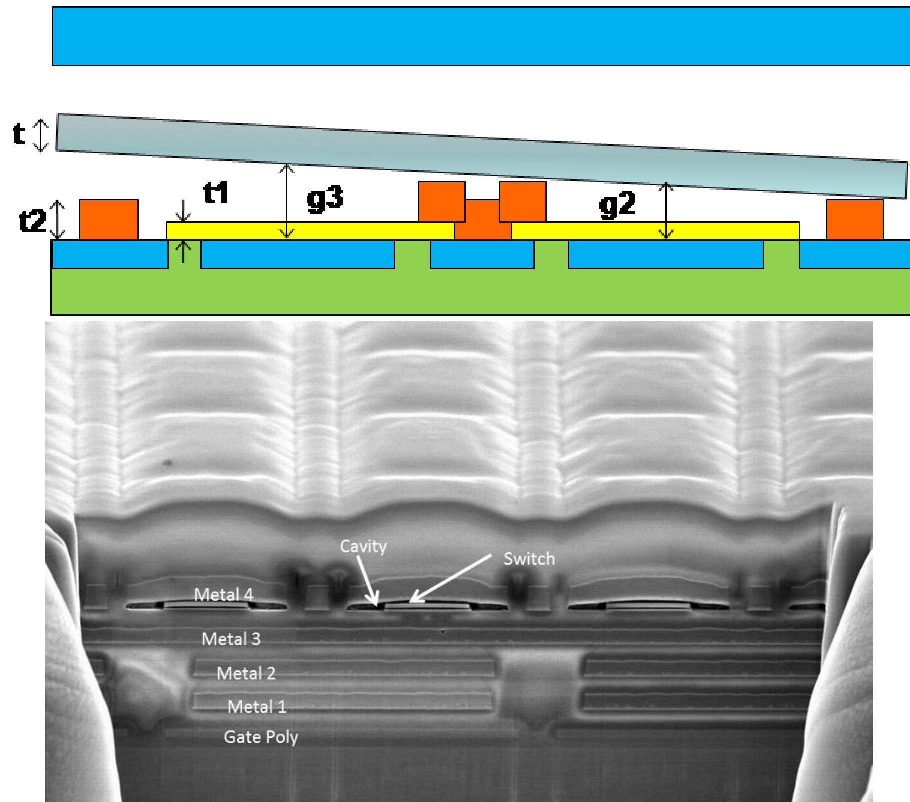



Figure 4.2: The teeter-totter concept used as a MEMS switch in the arrays characterized in this work (top) and a TEM picture of its integration (bottom). The geometrical dimension of the structure are evidenced.

on analysis of the physical principles underlying these MEMS switches.

## 4.2 Memory architecture

The single storage element of the arrays characterized in this work is the Nanomech MEMS switch from Cavendish Kinetics, designed by using a AMS-0.35$\mu$m CMOS technology with 4 Metal Layers. The MEMS element exploits a teeter-totter shape, constituted by a twofold rocking cantilever actuated by electrostatic force, as shown in Fig. 4.2.

The memory cell consists of one teeter-totter MEMS-switch [51] and 2 nMOS transistors for the differential readout circuitry of the cell. Special prototype memories of 1 Kbits have been developed for executing reliability tests and memory characterization using the RIFLE-SE Automated Test Equipment (ATE) from ActiveTechnologies.

The data can be stored in a non-volatile fashion depending on the voltage waveform applied to the electrode plates. The stored logical '1' state (Program) is represented by the switch closure on one side of the teeter-totter, and therefore by a low contact resistance value retrieved during the read operation. The stored logical '0' state is represented by an open switch (Erase), measured by the memory control circuitry as a very high resistance. The eNVM features a byte Program operation and a double word Erase operation with very fast timings. Such a structure therefore supports both single ended and differential Read operations, if the monitoring of the contact resistance is performed either on one or on both contact electrodes, respectively.

The advantages of using the contact resistance sensing instead of the traditional voltage/current sense on which the charge-storage memory technology is based, is straightforwardly deducted from Fig. 4.3, where the measured logical '0' state is limited only by the noise in the test environment. The read window generally displays a gap of more than four orders of magnitude in resistance, allowing safe storage of the information.

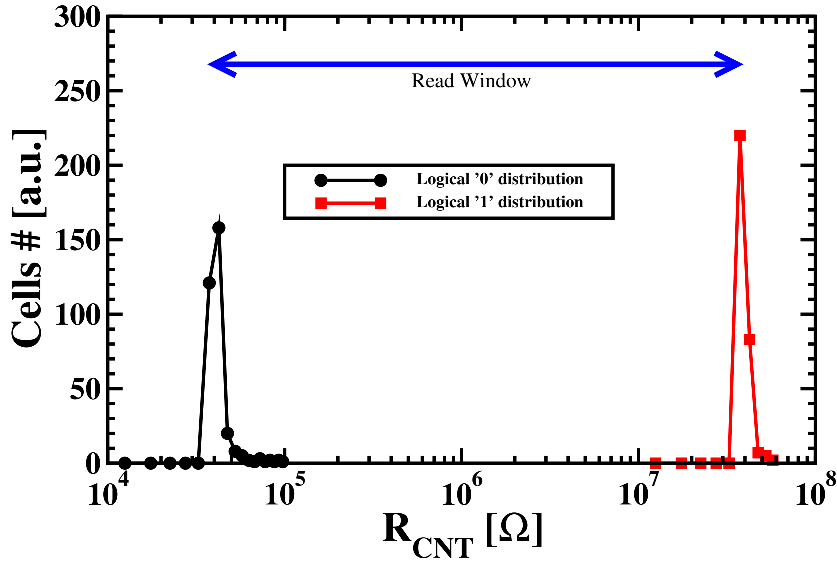



Figure 4.3: Read window of a Nanomech eNVM array. In this sample a read window of more than four orders of magnitude is shown.

## 4.3 Endurance characterization

The Nanomech eNVM arrays have been characterized from the standpoint of endurance, in order to analyze possible degradation mechanisms for the teeter-totter design with respect to Program/Erase cycle number. To this purpose, a set of HTOL (High Temperature Operating Life) cycling tests, room temperature cycling tests, and endurance tests, have been performed both on different cell architectures and designs. During the endurance tests the average contact resistance of the array $< R_{CNT} >$ has been monitored at quasi-regular intervals both for Programmed bits and Erased bits. The $< R_{CNT} >$ comprises the series resistance of the switch intrinsic contact resistance and control circuitry transistor resistance.

The global endurance test can be viewed as the sum of four stresses: 100 Program/Erase cycles for burn-in purpose, 10000 Program/Erase cycles at room temperature, 10000 Program/Erase cycles at 150°C (HTOL part) and the remain-

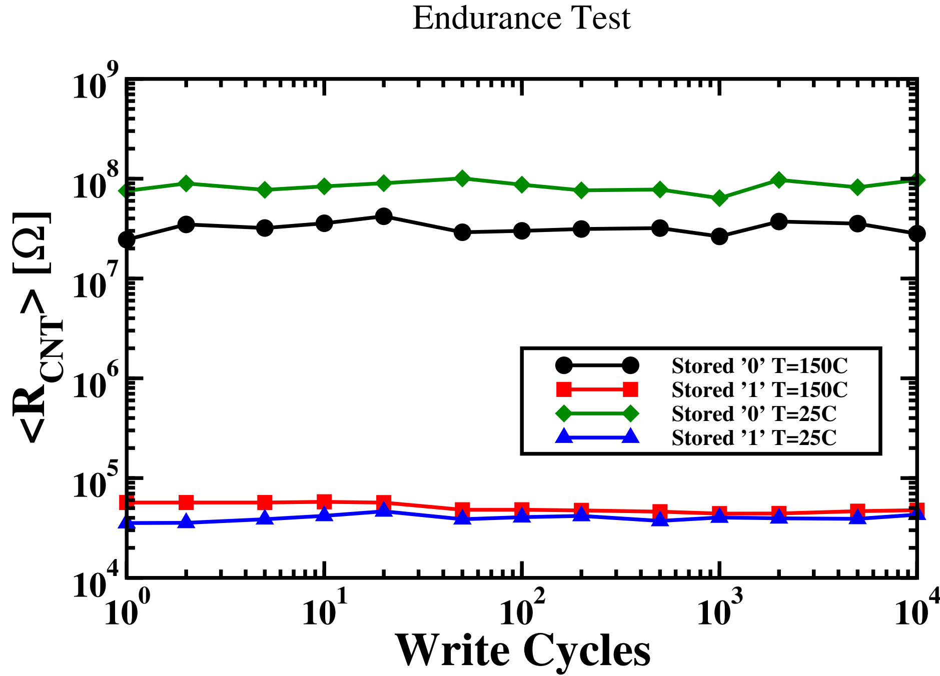



Figure 4.4: Benchmark of the $< R_{CNT} >$ between the room temperature endurance stress and the HTOL endurance stress.

ing cycles in order to demonstrate the survivability of $10^6$ Program/Erase cycles by the memory.

The HTOL part of the cumulative endurance stress serves a double purpose: the first is to investigate the effect of the temperature on the switches constituting the eNVM, the second is to compare the room temperature cycling performance with the high temperature cycling performance.

Fig. 4.4 demonstrates that by comparing the characterization data between the cycling at room temperature and the HTOL cycling there are no particular issues to be found, as the increase in resistance can be accounted for by the temperature dependence of the transistors in the measurement path. For the MEMS switch the stiction increases with temperature while the resistance decreases as has been described in [52]. The stored logical '0' bits average contact resistance level slightly

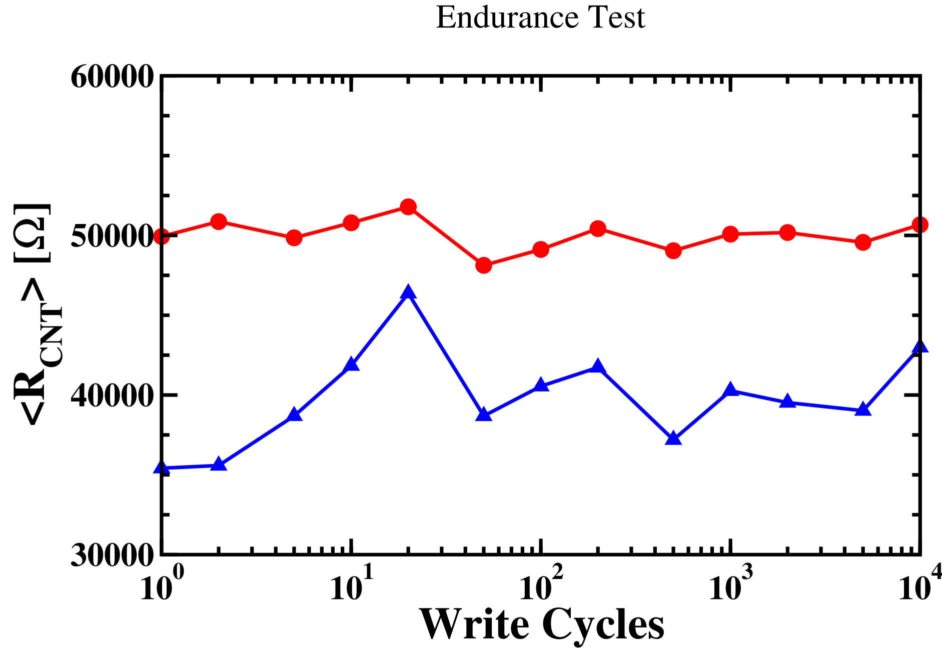



Figure 4.5: Insight of the $< \ R_{CNT} \ >$ characteristic under room temperature endurance stress (triangles) and under the HTOL endurance stress (squares).

lowers, whereas the stored logical '1' bits average contact resistance remains quite stable. The former effect is mainly ascribed to the noise increase in the measurements system due to the temperature, and is therefore not related to the MEMS switch. The latter effect can be better explained by considering the values of $< \ R_{CNT} \ >$ for Programmed bits, as shown in Fig. 4.5.

While cycling the memory, the switch may slam onto one of the contacts during the Program operation, which may have an effect on the surface roughness of both the contact electrodes as well as the underside of the teeter-totter beam. Also at release of the contact during the Erase operation, the separation will have some impact on the individual atoms which may even move from the contact plate to the beam. However, the memory design exploiting a near-zero restoring force, results in a tendency to stabilize the surfaces at a certain roughness during endurance

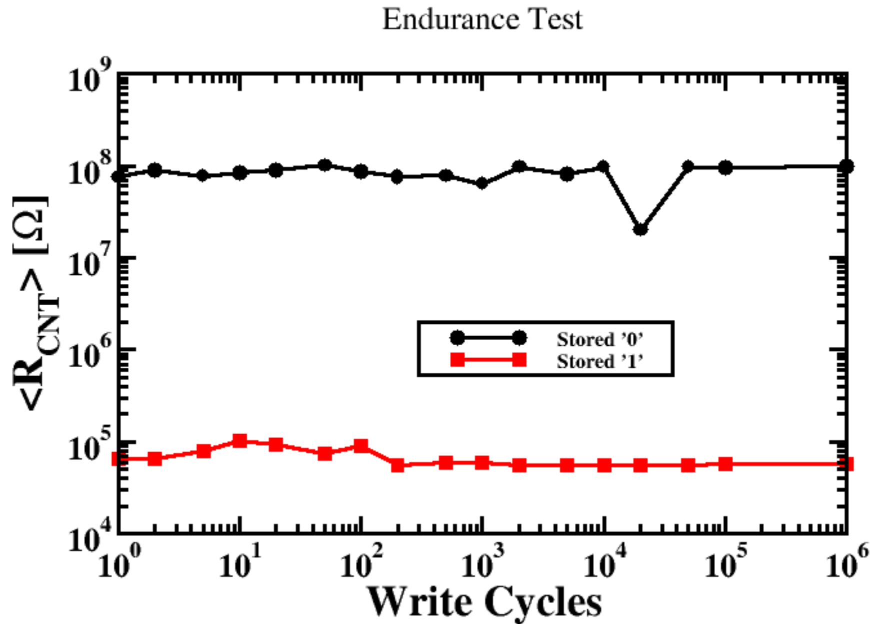



Figure 4.6: $< R_{CNT} >$ monitoring during cumulative endurance experiment.

cycling, providing also very little wear.

For high temperatures the picture is different, the contact becomes more intimate and the forces are strongly dependant on the temperature, so that the stiction will increase. Nevertheless, an appropriate tailoring of the voltage waveforms applied to the memory for electrostatic force control combined with near zero restoring force, allow us to neglect the effect of increased stiction due to temperature impact, thus leading to a better contact stability.

The global endurance experiment (see Fig. 4.6) shows that the contact resistance is stable over a range of $10^6$ cycles. The test has been performed on different Nanomech eNVM switch architectures, providing the same results. The NanoMech memory operating temperature is limited only by the CMOS transistors of the control circuitry. Individual MEMS teeter-totter switches have been

tested over a temperature range of -150°C to 300°C, as stated in [52].

## 4.4 Data retention characterization

The data retention is a critical aspect to investigate for eNVM, where the reliability constraints are tighter than for consumer NVM products. In Nanomech MEMS devices the study of the effect of temperature on the floating rocker should be considered: besides stiction which will keep the contact in vertical direction, there will also be a similar lateral friction force keeping the contact in place in the lateral position. The thermal expansion of the floating rocker will be different than the thermal expansion of the silicon substrate, as the relative thermal expansion coefficients are different due to the material differences. This leads to the generation of a further lateral force, which could overcome the friction force and temporarily break the contact (for Programmed bits) eventually moving to another position. During this movement the stiction force will be lower (but not zero) and a new sub-optimum contact will be formed. When this new contact point has a stronger stiction force, it will be less likely that the contact point will be broken. Thus, thermal cycling on average will result in a better contact.

To verify the above hypothesis, a High Temperature Storage Life (HTSL) experiments at 200°C were performed with a six weeks duration. In additions thermal cycling and liquid-liquid tests, have been executed. The test included different architectures of the teeter-totter device, in order to capture the particular behavioral for different MEMS designs. By monitoring the $< R_{CNT} >$ before and after the HTSL test for the Programmed bits on all the devices, it was found that not only is the memory immune from retention failures at high temperatures, but

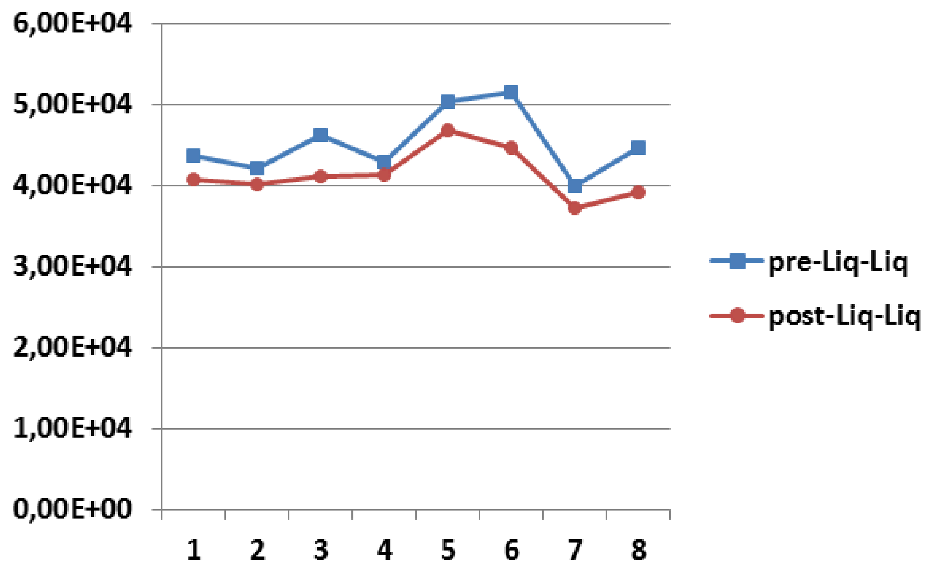



Figure 4.7: $< R_{CNT} >$ measured for different architectures of Nanomech eNVM. The values were calculated before and after the liquid-liquid experiment.

also temperature improves the quality of the stored information, hence retention will improve with temperature, unlike any other eNVM technology. This result is also supported by the decrease of the average contact resistance in Liquid-Liquid tests (see Fig. 4.7).

This phenomenon can be explained by a model where the stiction and friction forces increase due to the temperature gradient changes, leading to a more intimate contact between the contact plate and the floating rocker, and that this process occurs regardless of its geometrical dimensions and architecture.

## 4.5 Environmental and mechanical characterization

The table in Fig.4.8 provides a summary of the reliability tests. applied to the memory arrays. Fig.4.9 shows the average relative resistance-drift for different device MEMS switch architecture variations, for all the performed tests. It should be stressed once again that these percentage variations are still orders of magnitude smaller than the memory read window. Therefore, no loss of good bits were experienced on any of these tests. Remarkably, most tests result in a negative drift in resistance, which indicates an increase in the contact area and thus an increase in stiction and an improvement in retention. From the mechanical tests, consisting of a mechanical shock and vibration, a small resistance increase is observed, while a constant acceleration (positioned such that the acceleration counteracts the stiction force), results in a smaller resistance. The other tests clearly show that temperature has a positive effect. In any case, all resistance shifts have magnitudes way below the memory read window, indicating no loss of information during any of the stress tests. Fig.4.10 shows the difference in resistance of the measured average bit resistance ($R_{CNT}$) during Liquid-Liquid stress test. The test is performed on ceramic packaged dies with open lid, over a temperature window from -55C to 150C. The total sensed resistance is a measure of the MEMS switch contact resistance combined with 6 pass transistors. By analyzing the distributions of the bit resistances before and after testing one can observe a small improvement of the stored '1' state resistance values. This results in a larger read window, which also indicates stronger stiction and thus improved retention. This test was perceived to be the harshest test for these non-volatile MEMS switches.

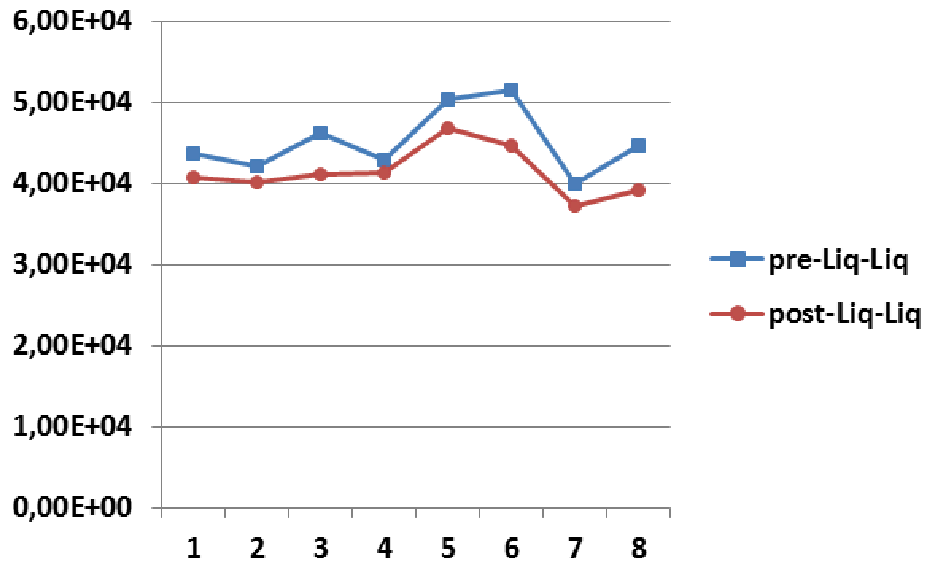



Figure 4.8: Reliability tests executed on NanoMEMS memories. No failing bits.

As the liquid-liquid test was supposed to have the highest chance to create 'cracks' in the passivation layers protecting the MEMS devices' cavities, it was decided to execute a post stress HAST test to see whether hermeticity was broken and humidity could enter the cavities. This test resulted still in no failing bits. Both liquid-liquid and HAST tests were performed with the lids of the ceramic packages removed to allow temperature and humidity to reach the bare die.

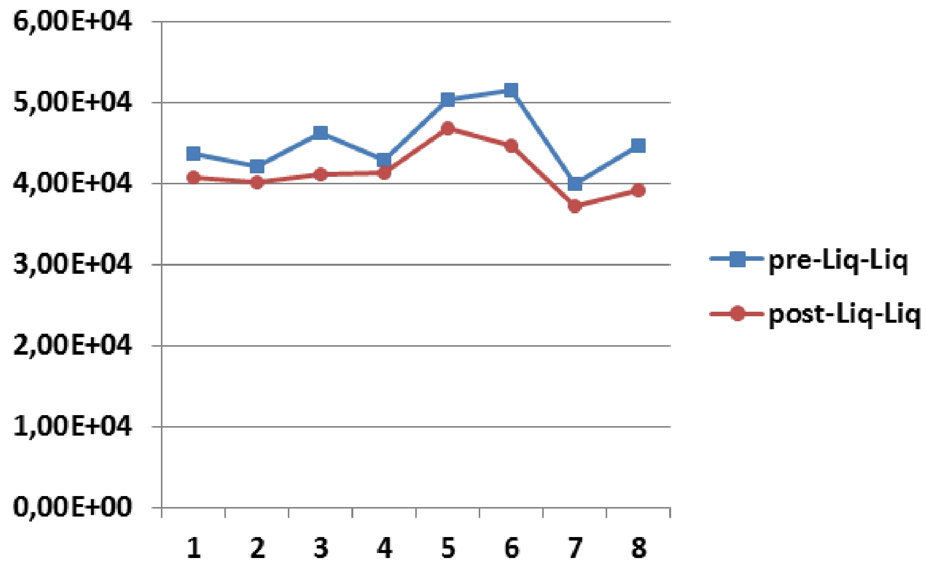



Figure 4.9: Relative resistance drift, average across full array, for all reliability tests performed on 6 alternative MEMS switch architecture variations.

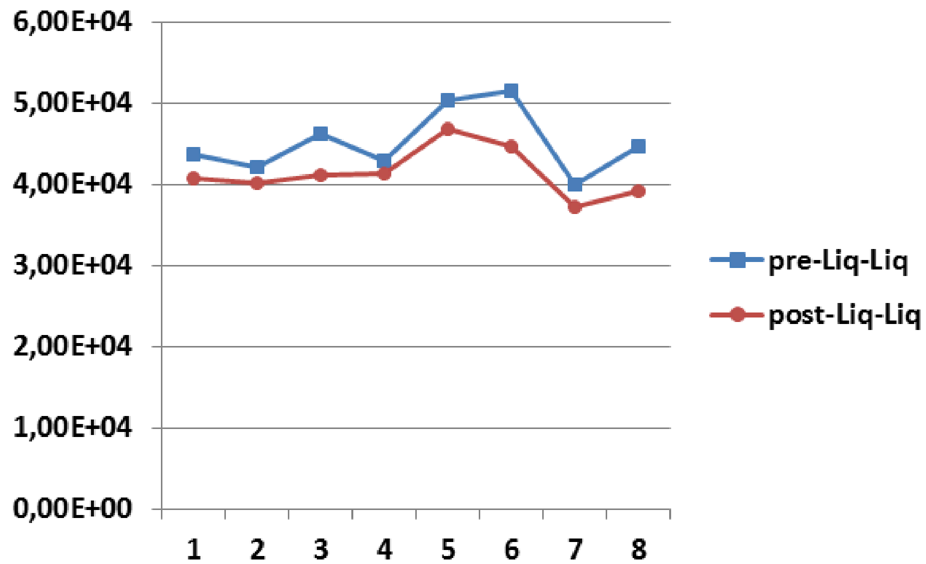



Figure 4.10: Contact resistance behavior on Liquid-Liquid test with -55C - 150C temperature range. The distribution shows an improvement of the $R_{CNT}$ after the test.

# Chapter 5

# Charge Trapping NAND Flash Memories

The possibility to use trap-rich dielectric layers as the storage medium in charge-trapping nonvolatile memory devices was recognized in the early semiconductor years. Today these devices are considered one of the most promising alternatives to the floating gate technology, especially for NAND applications. Charge Trapping devices have two main advantages with respect to the conventional floating gate transistor. First, they have an inherent immunity to retention loss mechanisms related to point defects in the tunnel oxide (e.g. Stress Induced Leakage Current). When a trap or percolation path is formed in the tunnel oxide, only the charge trapped in the portion of the nitride that is directly above it will discharge toward the substrate. As a consequence, the thickness of the tunnel oxide can be reduced well below the 6-7 nm limit of conventional floating gate devices. Moreover, this is the first technology to be seriously enabled toward a 3D integration thanks to the very low cell-to-cell interference. In this chapter it will be shown the results of

the study performed on a 4Mbits Charge Trapping NAND Flash array developed in the framework of the European project FP7-GOSSAMMER.

## 5.1 Electrical characterization

It has been started the array characterization with a set of preliminary measurements necessary for evidencing the capabilities of the memory in terms of program/erase operations. The program operation is achieved by applying a single pulse with 19 V amplitude and 80 $\mu$s duration on selected wordlines ($WLSEL$ signal on the array), and a 8 V pulse with same duration on the unselected wordlines ($WLUNSEL$ signal on the array), as evidenced in Fig.5.1. The erase operation is achieved by applying a pulse on the block common sourceline ($CSOURCE$ signal) with amplitude 19 V and duration 100 $\mu$s, and a pulse on the string side select transistors ($GSSL$ and $GDSL$ signals) with amplitude 12 V and same duration, as shown in Fig.5.2. It should be here noticed that these conditions actually give rise only to a ”weak” erase and not to a fully erased state. It has been operated this choice because the fully erased distribution is not completely observable by the instrument when measured in terms of threshold voltages. I therefore first focused on the ”weak” erase distribution in order to have a measurement of the features of the fully erased distribution as its width and uniformity, as it will be shown later, the two distributions are simply shifted as an effect of the ramped writing waveform. Fig.5.3 and Fig.5.4 show the Gaussian probability plots of the threshold voltage ($V_T$) cumulative distributions of the whole array measured after both program and erase. The read operation is performed on the whole array in a Direct Memory Access mode (DMA) with a selectable reference current in the

range of 200-400 nA, with $VGSSL$ and $VGDSL$ (voltage applied on the gates of string select transistors) equal to 3 V, $VDQSense$ (voltage forced on the string for retrieving a current) equal to 1 V, $WLUNSEL$ (over-programming voltage limit) equal to 5.5 V. As shown by the figures, a right value for the reference current has to be chosen as it may have a significant impact on the outcome of the measurement. On the average, the programmed and the erased levels have a good margin of separation (almost 2 V), although distributions are not very compact, especially on the very first block of the array. Preliminary measurements show that the array is capable of sustaining both a program and an erase operation, but a careful optimization of writing waveforms is needed in order to achieve more compact distributions and to suppress disturbs related to array topology. In particular Block 0, compared with other blocks, suffers from a significant number (almost $10\%$ of the array) of tail bits which shifts the entire erased distribution towards the programmed level. It has been investigated for possible issues related to array topology by analyzing both program and erase I-V characteristics of 32 cells belonging to a string of Block 0. The first wordline (WL0) shows both the better programmed characteristic and the worst erased characteristic, as evidenced in Fig.5.5 and Fig.5.6. Beside technological parameter dispersion also the presence of GIDL (Gate Induced Drain Leakage ) currents in relation to the proximity of the WL0 to the SSL (Source Side seLect) can play a significant role. The I-V measurements were performed with $VGSSL$=$VGDSL$=3V, $VDQsense$=1V and $WLUNSEL$=5.5V.

The dependence of the cell behavior on the cell topological position must be investigated for a correct interpretation of experimental results collected in endurance and retention experiments. We considered the behaviour of a memory

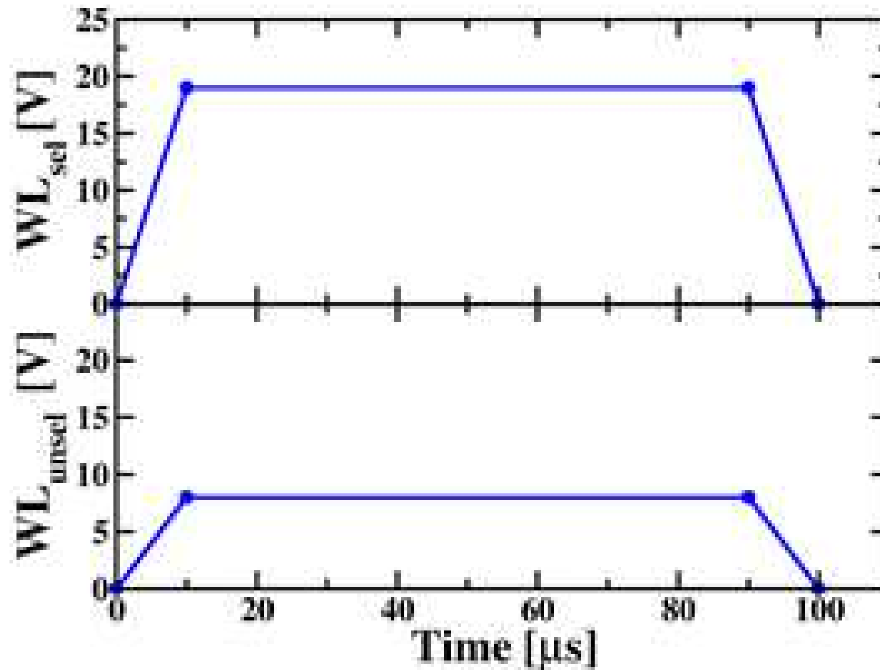



Figure 5.1: Signals applied to the array for the program operation.

block and found the presence of both a wordline and a bitline dependence. In particular wordline WL0, which has been investigated also in Figs.5.5 and 5.6, shows a fast programming feature, whereas WL1 and WL30 show an opposite behavior as evidenced in Fig.5.7, where the wordline average threshold voltage is calculated over the 32768 bitlines. The bitline dependence can be observed in Fig.5.8 where a periodic even/odd behavior is evident. As a consequence of the topological behavior both a non perfect Gaussian behavior of the $V_T$ distribution and a non optimized distribution width is expected.

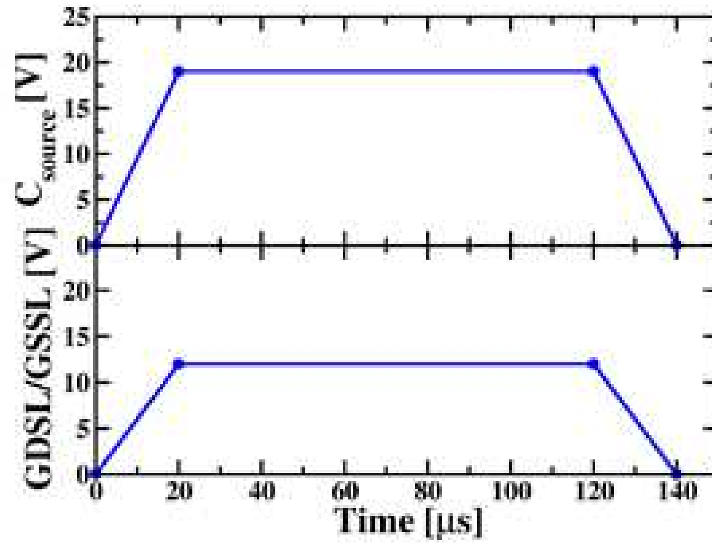



Figure 5.2: Signals applied to the array for the erase operation.

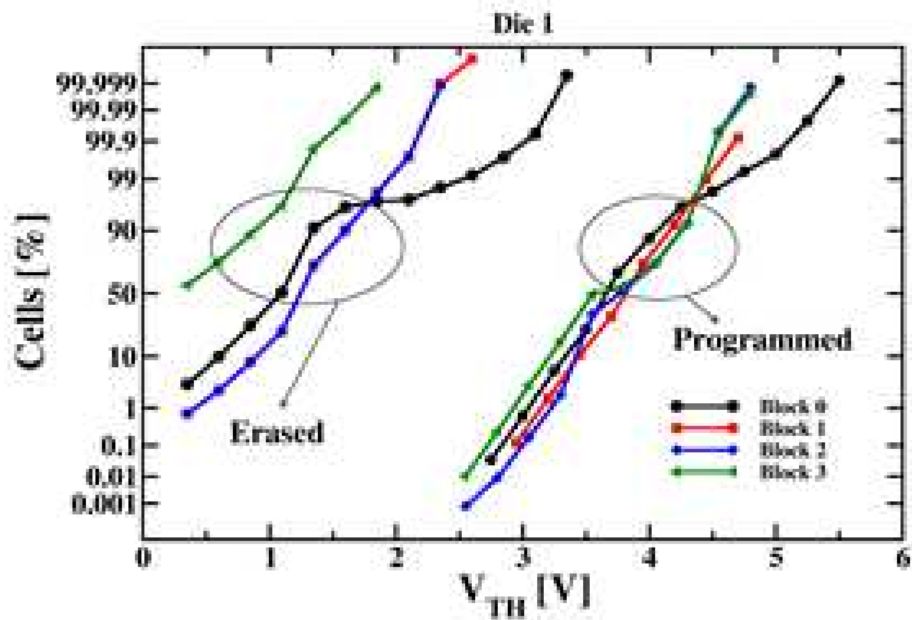



Figure 5.3: Program and erase distributions measured using $I_{ref}$=200nA.

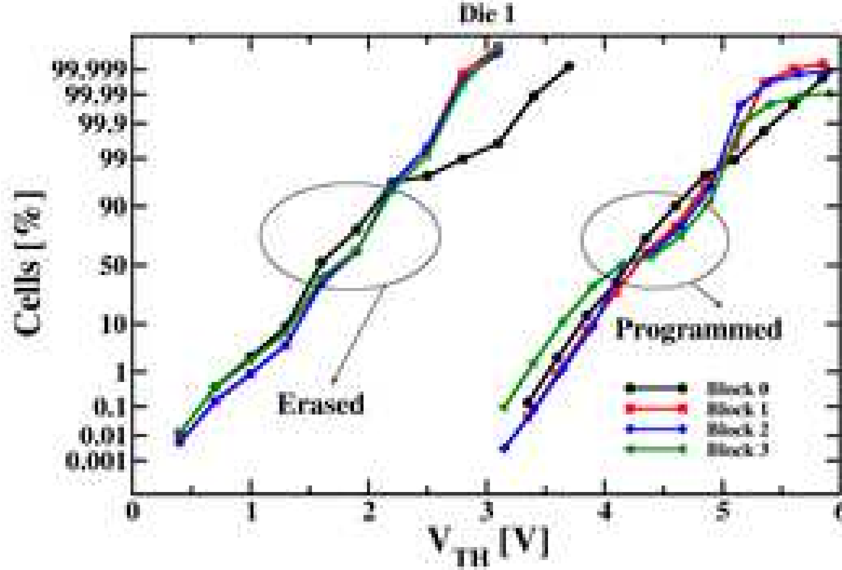



Figure 5.4: Program and erase distributions measured using $I_{ref}$=400nA.

## 5.2 Program operation characterization

Programming operation has been tested by using a ramped voltage from 10 to 22 V with single pulse duration of 10 $\mu$s and 2 V amplitude (see Fig.5.9). Reading conditions have made use of the following parameters: $VGSSL$ = $VGDSL$ = 3 V, $VDQSense$ = 1 V, $WLUNSEL$ = 5.5 V and a reference level $I_{ref}$ = 200 nA. The programming characteristic are shown in Fig.5.10. The slope is smaller than 2 V, i.e. the step amplitude, as expected in floating gate device structures. Moreover, it has been found that for programming voltages lower than 14 V, the silicon nitride trapping contribution in the array is negligible with no significant increase of cell $V_T$. Figs.5.11, 5.12 and 5.13 show the evolution of the $V_T$ distribution during a ramped programming operation where the $V_T$ distribution is measured after each programming pulse step. As expected, the entire $V_T$ distribution simply

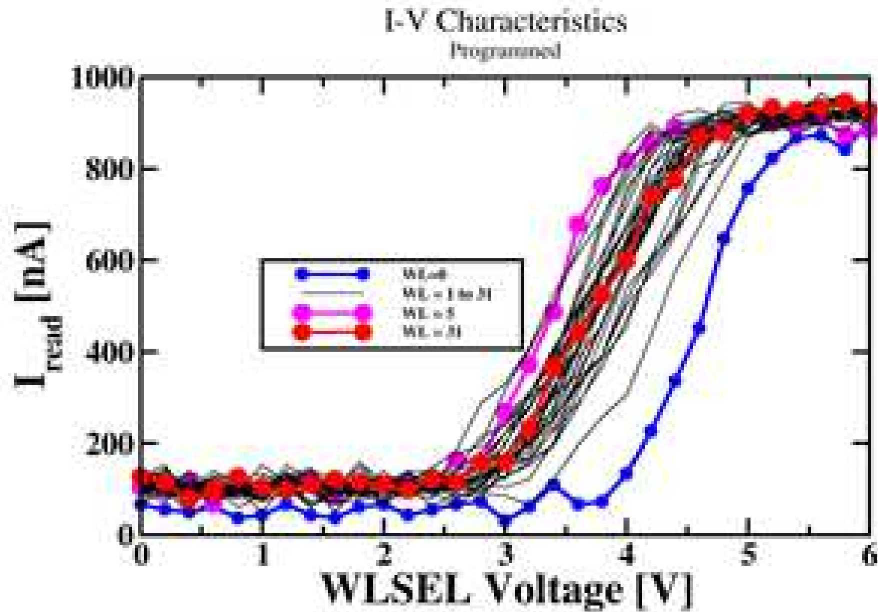



Figure 5.5: I-V characteristics on the programmed state of an arbitrary array string.

shifts to the right after each programming step thus evidencing that the ramped programming waveform can be used effectively in controlling the $V_T$ position. The effect of the topology, which is evident when comparing odd/even behavior with the entire array behavior, gives rise to an additional contribution of roughly 500 mV to the overall distribution width.

Ways for effectively compact the program $V_T$ distributions have been investigated. Beside process optimization which should aim at reducing the odd/even effect, or solve the wordline dependence issue, an effective way of improving the program distribution is to use Incremental Step Pulse Programming (ISPP) algorithms exploiting a verify phase performed after each programming step. Based on the ISPP algorithm, when a cell $V_T$ falls above a predefined verify limit, then programming pulses are no more delivered to the cell. The algorithm proceeds un-

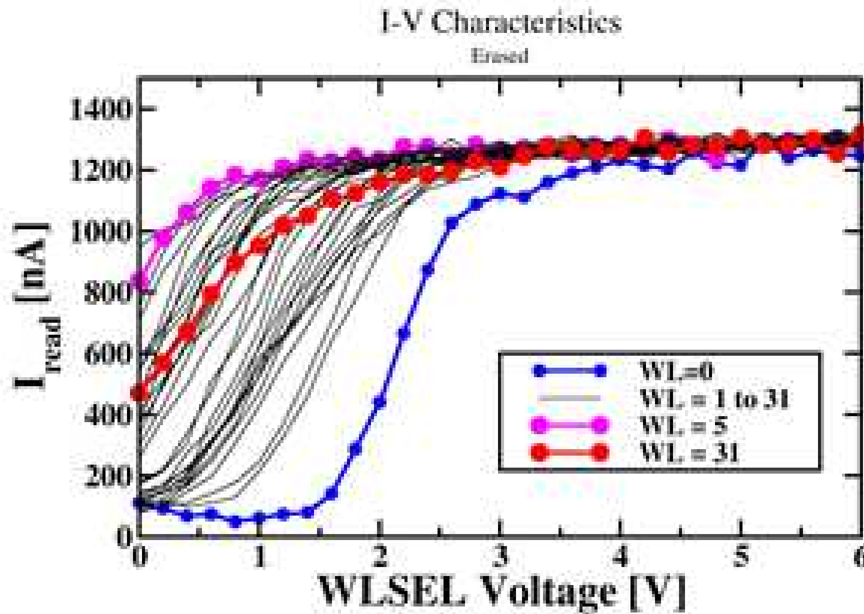



Figure 5.6: I-V characteristics on the erased state for same string.

til all cells have been programmed. Figs. 5.11, 5.12, and 5.13 show the cumulative distributions which can be obtained when an arbitrary limit of 2 V is used ( see 'virtual' distributions). Fig.5.14 shows the frequency distribution of the $V_T$ virtual distribution. Figs.5.15 and 5.16 show two more examples of $V_T$ distribution evolution during programming. The results confirm that the $V_T$ distribution shift is lower than the given pulse step which is here 1 V. Moreover, some statistical information such as the distribution $\sigma$, equal to 0.2 V, and the distribution width, equal to 1.5 V, can be retrieved. A small saturation effect can be observed for high programming voltages, consisting in a reduction of the net $V_T$ shift caused by each program pulse of the staircase waveform.

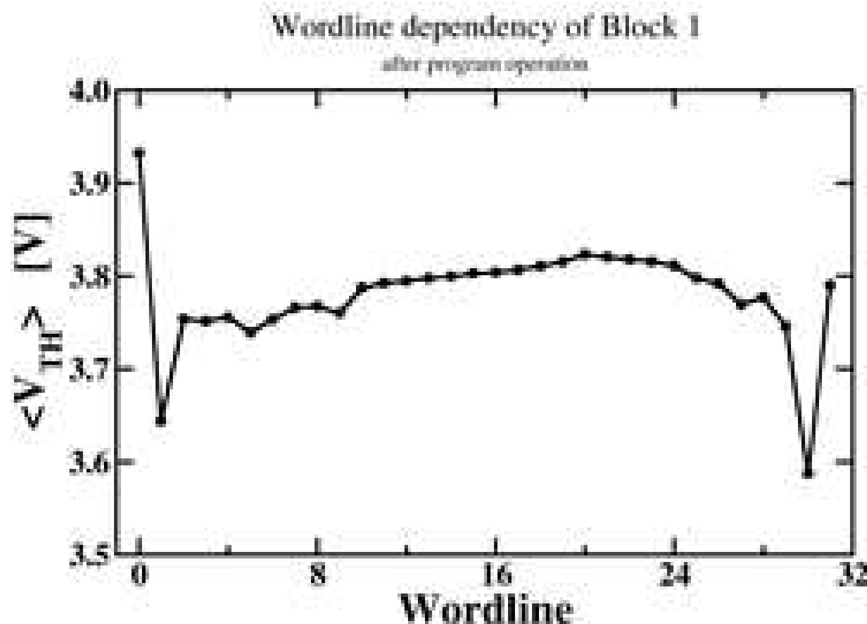



Figure 5.7: Wordline average threshold voltage dependency.

## 5.3 Room temperature retention

Room Temperature Bake (RTB) measurements have been performed in order to evaluate the retention capabilities of the memory array. The array has been erased by using the waveform of Fig.5.2 and then programmed by using a staircase waveform from 10 to 22 V with 2 V steps of 10 $\mu$s. In this condition we monitored the charge loss with respect to time by analyzing the variation of the threshold voltage at different time steps. All the readings have been performed in DMA with the following parameters: $VGSSL = VGDSL$ = 3 V, $VDQSense$ = 1 V, $WLUNSEL$ = 5.5 V and a reference level $I_{ref}$ = 200 nA. As depicted in Fig.5.17 the array shows relatively good retention capabilities, with saturation condition after about 5 hours of RTB. The major charge loss contribution is appreciable within the very

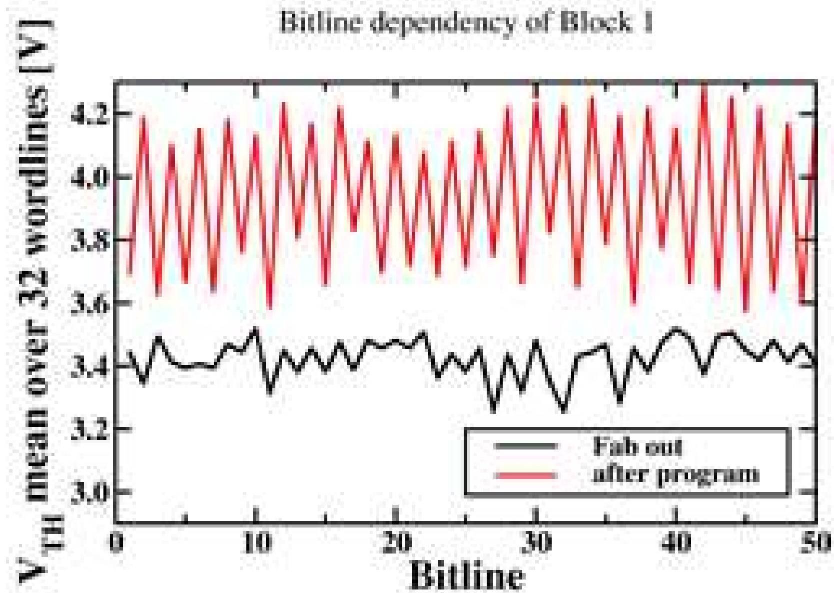



Figure 5.8: Bitline even/odd behavior evidenced on wordline average threshold voltage.

first hour of the experiment. Charge loss can be probably ascribed to shallow nitride traps emitting captured electrons to alumina layer via Poole-Frenkel emission. Figs.5.18, 5.19, and 5.20 show the statistical behavior of the charge loss. The threshold voltage distributions are measured after each time step. A lower tail is populated by a small number of cells belonging to stuck-to-erase bitlines (i.e. BL17467-BL17468) which failed to program, thus revealing some yield issues due to the novelty degree of the technology. An upper tail is also present. In this case, this tail is populated by cells belonging to the WL0 wordline.

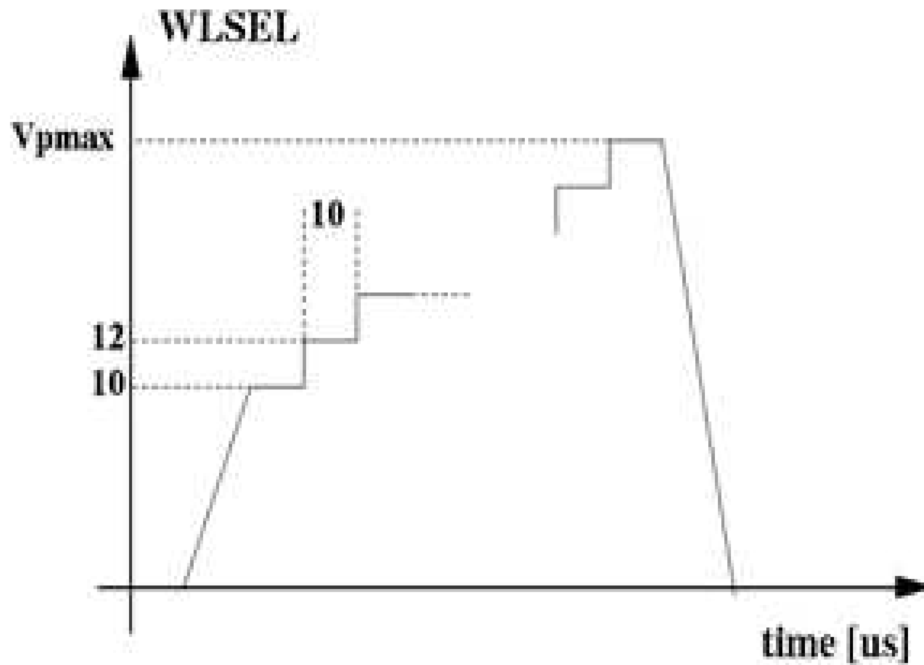



Figure 5.9: Staircase programming waveform.

## 5.4 Disturbs characterization

Different experiments have been carried out in order to verify the impact of read and program disturbs in the considered samples. Before each experiment, cells are erased by using a single 100 $\mu$s, 19 V pulse. A single program disturb pulse has been defined as a programming staircase from 10 to 20V with 2V voltage steps, each step lasting 10 $\mu$s. The disturb is measured as $V_T$ difference before/after the program disturb. Fig.5.21 shows the effect of 100 consecutive reading operations on both programmed and erased cells. The net effect of the read disturb is negligible, or at least hidden by the charge loss mechanisms which give rise to the exponential decrease of the threshold voltage with time. Fig.5.22 shows the impact of program disturbs applied simultaneously on three arbitrary wordlines, WL1, WL16, and WL32. The threshold voltage is measured after 1, 2, 5,

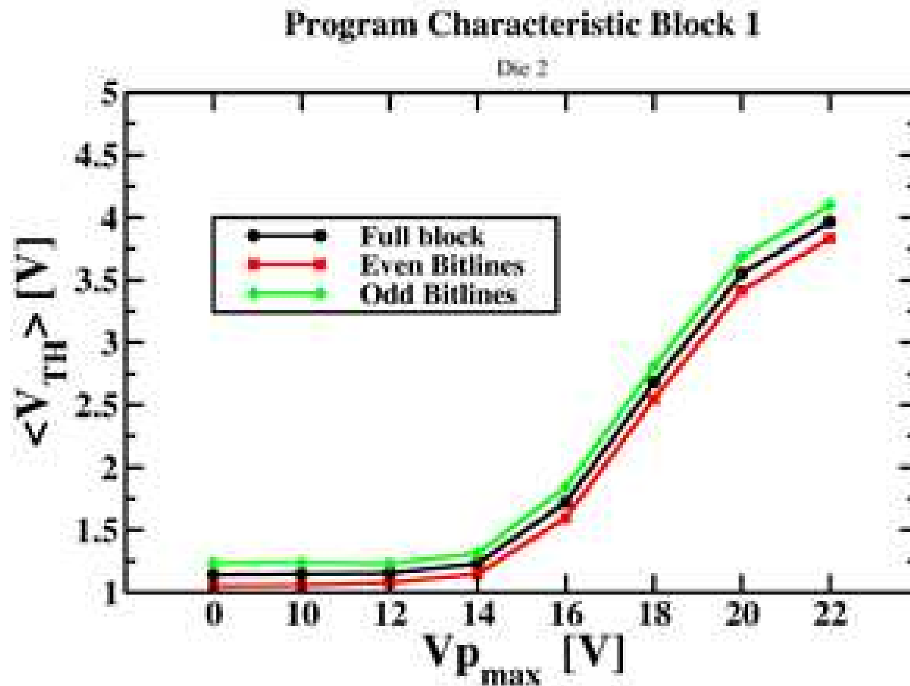



Figure 5.10: Single block average program characteristic. Even/Odd bitlines separate contributions have been evidenced.

10, 20, 50, 100, 200 program disturb pulses. Program disturb induces a small increase of the VT of unselected cells which worsens with the number of program disturb pulses. Fig.5.23 shows how a sequence of program disturb pulses applied on WL1 can disturb the neighboring wordlines. Interesting, the disturb is correlated with the distance from the wordline which subjected to programming so that WL0 and WL2 are the most affected one. Fig.5.24 shows the evolution of the disturb magnitude with the number of pulses for the two worst cases wordlines WL0 and WL2. A non-negligible disturb can be observed as 150 mV of $V_T$ difference are measured after 200 program pulses, with an almost linear trend with the number of program pulses. This measurement reveals that program disturbs can rise a potential issue in product operating conditions requiring a significant num-

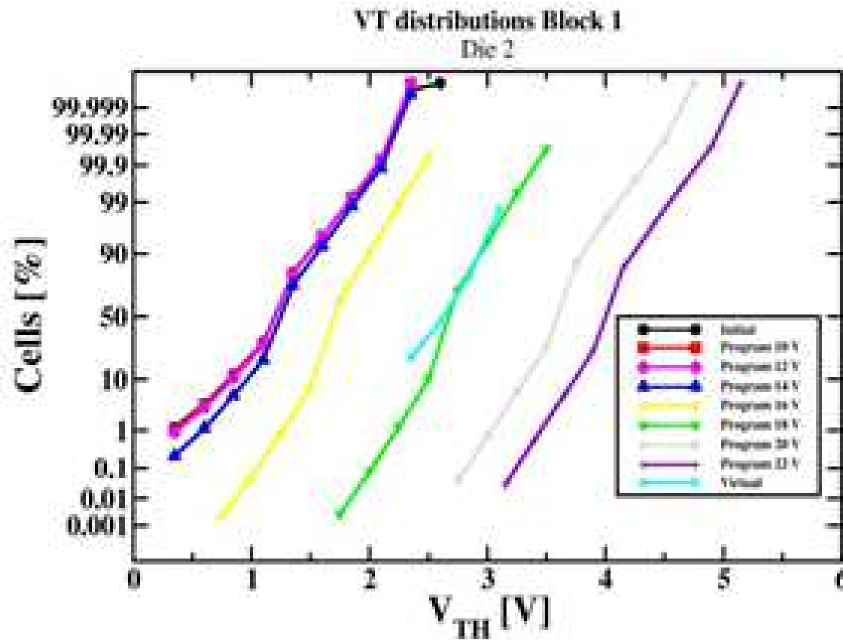



Figure 5.11: Evolution of the threshold voltage distributions of an entire array block during a programming ramp.

ber of writing operations, and/or when larger arrays with a significant number of wordlines are involved. The block to block interference has also been evaluated, as shown in Fig.5.25, by monitoring the disturb on both both programmed and erased cells while a certain number of program disturb pulses are applied to a wordline belonging to an adjacent block. The two curves refer to an average disturb calculated over the two populations of both programmed and erased cells. No significant correlation with the number of disturb pulses is measured for both programmed and erased cells. The quasi esponential decrease with the number of disturb pulses is very similar to the charge loss behavior measured in retention experiments showing that the block-to-block interference can indeed be neglected. In conclusion, results show that the sample devices seem to be affected by a significant program disturb occurring on cells belonging to the same block

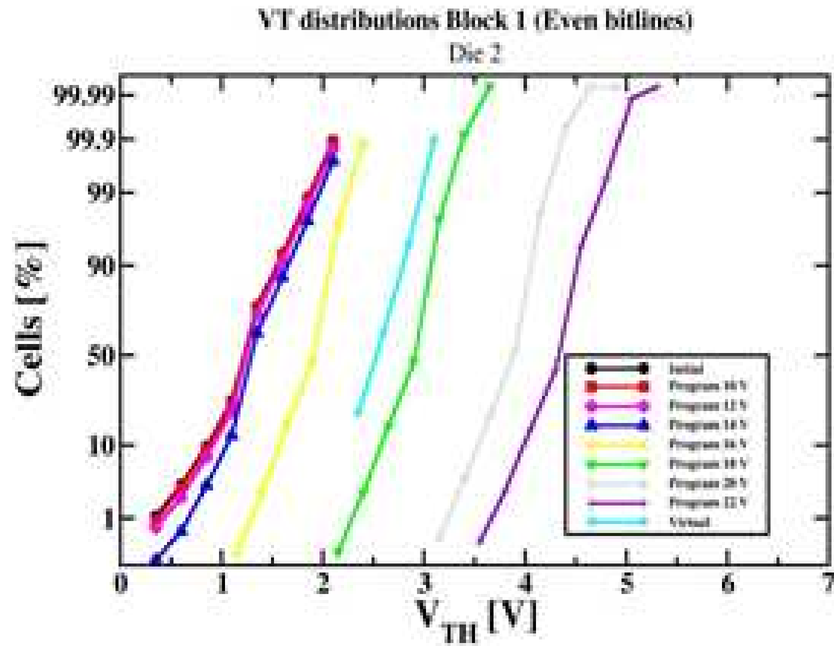



Figure 5.12: Evolution of the threshold voltage distributions of cells belonging to even bitlines only during a programming ramp.

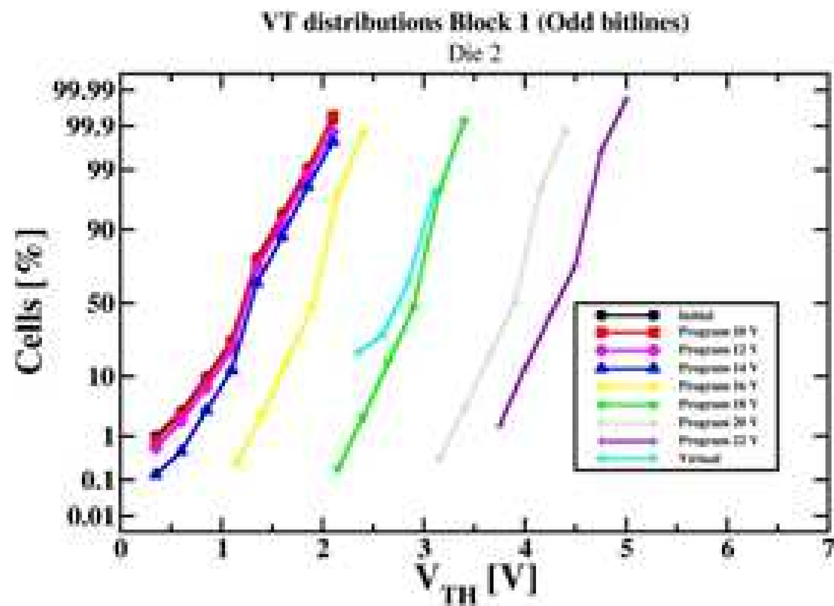



Figure 5.13: Evolution of the threshold voltage distributions of cells belonging to odd bitlines only during a programming ramp.

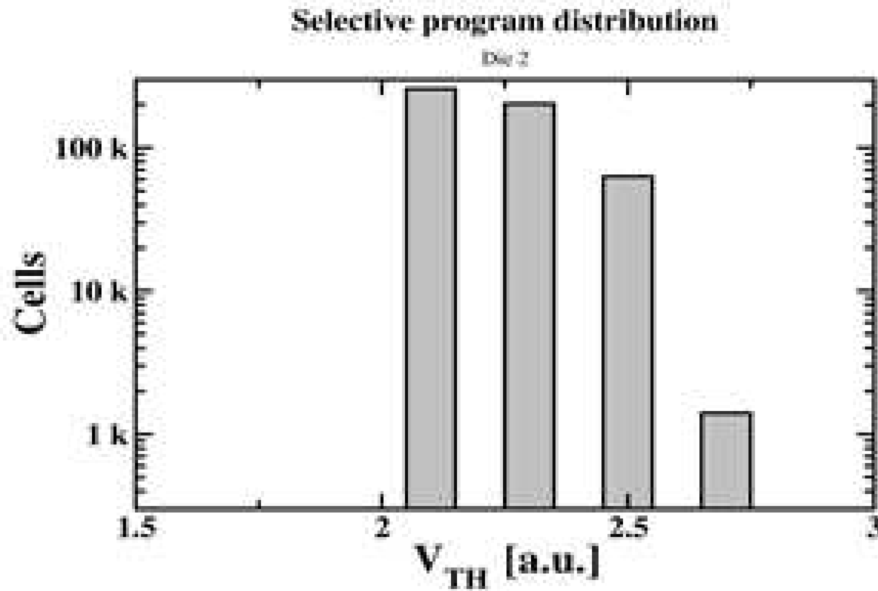



Figure 5.14: Cells number belonging to a threshold voltage bin after application of the post-analysis data compactation.

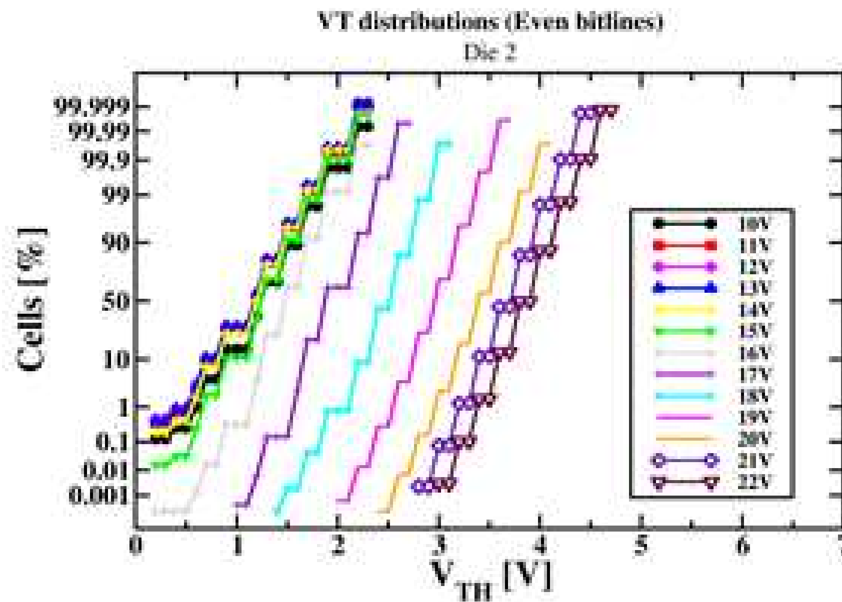



Figure 5.15: Threshold voltage distribution using the ramped waveform with 1 V steps on even bitlines.

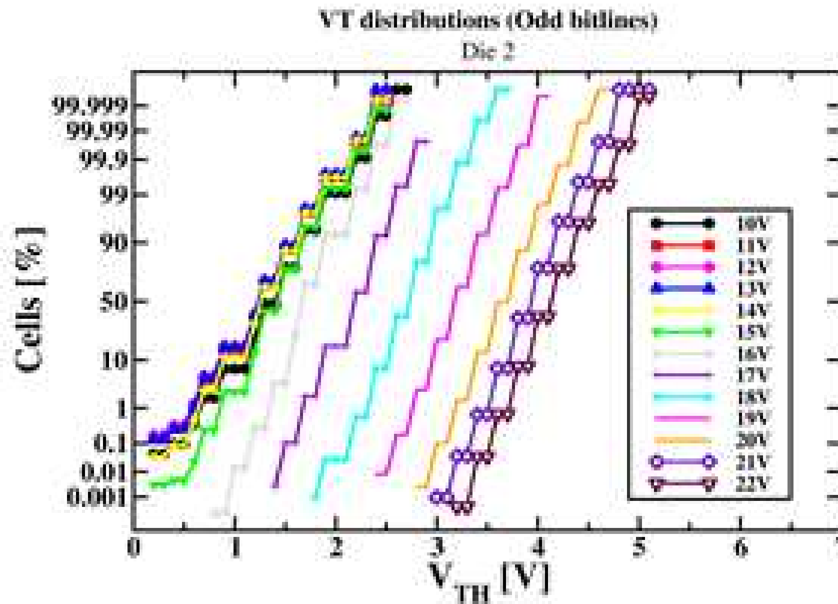



Figure 5.16: Threshold voltage distribution using the ramped waveform with 1 V steps on odd bitlines.

where wordlines are programmed. On a fully operating product such effects need to be taken into account, as array size and specific program algorithms may induce more severe disturb effects.

## 5.5 Edge Wordline Disturb (EWD) phenomenon

The constant scaling of traditional Floating Gate (FG) Flash memories is facing several limitations, ascribed both to architectural and physical issues [53]. One of the promising candidates for replacement of high density FG Flash NAND memories is represented by the Charge-Trapping (CT) memories [54]. However, such a technology has to overcome some reliability issues that have been a sore for the traditional FG memories, such as the so-called Edge Wordline Disturb (EWD) [30] occurring during programming.

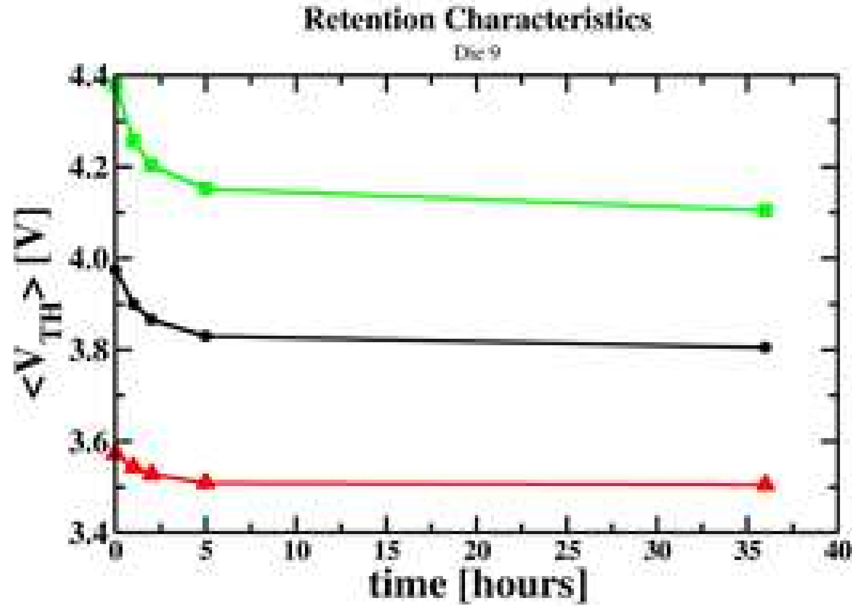



Figure 5.17: Retention characteristics of the array. The average retention of the whole array, the average retention of even bitlines only and the average retention of odd bitlines only are evidenced by black, green and red curves respectively.

In fact, in a standard NAND architecture (see Fig. 5.26), a disturb affects the cells belonging to the first wordline (WL0) connecting the cell strings to the string selector GSL. This disturb is evidenced as a difference between the average threshold voltage $< V_{T_WL0} >$ of the cells belonging to WL0 and the average threshold voltage $< V_T >$ of all other cells.

The difference between cells belonging to WL0 and the other cells can be ascribed to three effects: $i)$ different potentials at their terminals with respect to the other cells depending on the specific WL selected for programming; $ii)$ a different cell geometry due to the fact that these cells are located between a cell and a transistor (differently from cells belonging to WL1 $\div$ WL30), therefore with a different field underneath their channels and a modified programming dynamics;

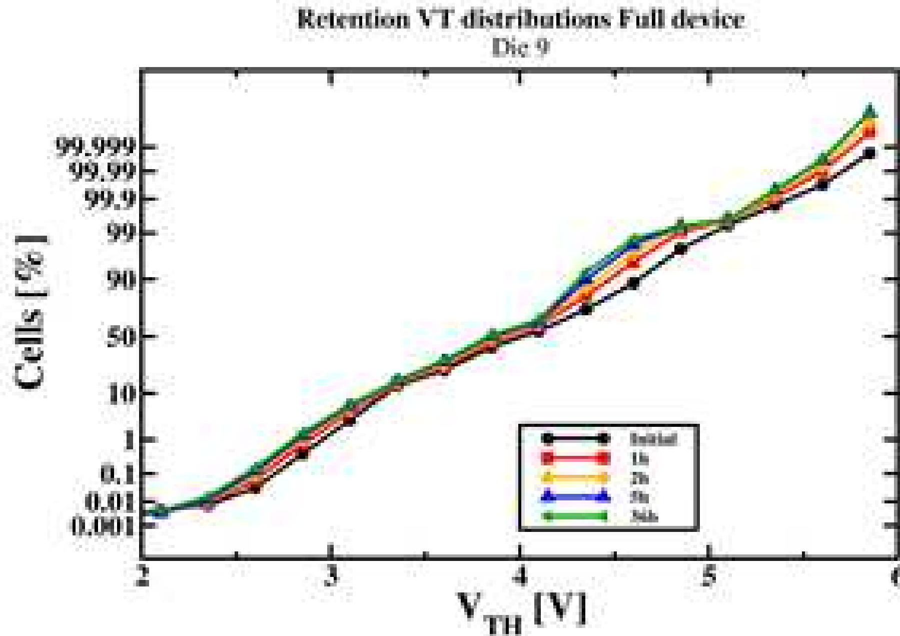



Figure 5.18: Threshold voltage distribution of the full device in RTB experiments.

$iii)$ the possible presence of a large GIDL (Gate Induced Drain Leakage) current generated at the drain edge of GSL transistors due to their drain potential raised by channel boosting [55]: such a field can efficiently trigger electron-hole pair generation followed by an acceleration of the electrons toward the channel of WL0 cells. These electrons can be injected into the floating gate of these cells, thus provoking an undesired increase of their threshold voltages.

In this section it has been experimentally analyzed the EWD in CT NAND Flash arrays during programming. In particular EWD has been measured by comparing $< V_{T_WL0} >$ with respect to $< V_T >$ as a function of the programming voltages applied to the selected WL, the device aging represented as the number of write/erase cycles, the number of partial programming operations (NOP) and the voltage applied to unselected cells to inhibit their programming.

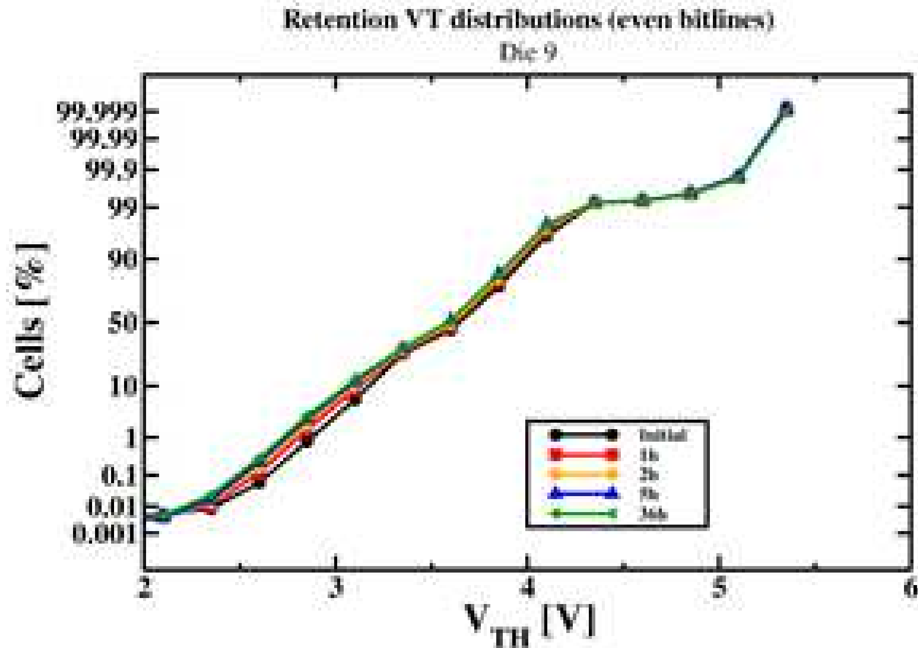



Figure 5.19: Threshold voltage distribution of the array even bitlines in RTB experiments.

The experimental data have been obtained by testing 4Mbits CT NAND Flash arrays with a CT version of the Active Technologies RIFLE-SE Automated Test Equipment (ATE) suitable for applying arbitrary waveforms, thus fully controlling the voltages applied during write and read operations. The array is arranged into four blocks, each one sized 1Mbits and organized with 32 wordlines and 32768 bitlines. The memory cells feature a p-Si/$SiO_2$/$Si_3N_4$/$Al_2O_3$ stack overwhelmed by a high work function TaN/Ti/TaN metal gate. The program operation is performed page-wide (mapped as an entire 32Kbits wordline) by using an Incremental Step Pulse Program (ISPP, see Fig. 5.27) algorithm with a starting voltage of 10V and 1V steps with 4$\mu$s duration, whereas keeping the unselected wordlines at a constant intermediate $V_{pass}$ voltage.

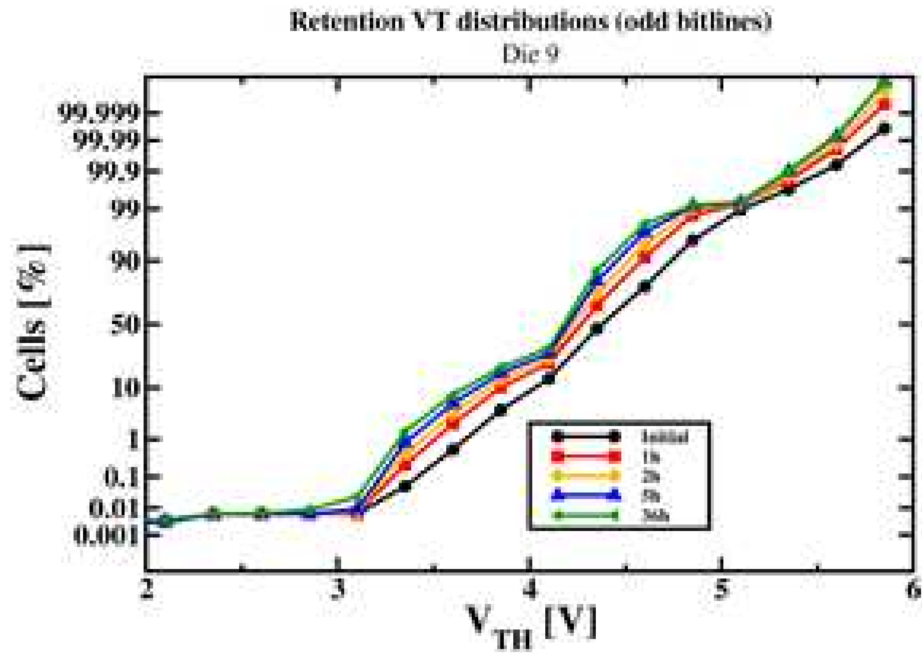



Figure 5.20: Threshold voltage distribution of the array odd bitlines in RTB experiments.

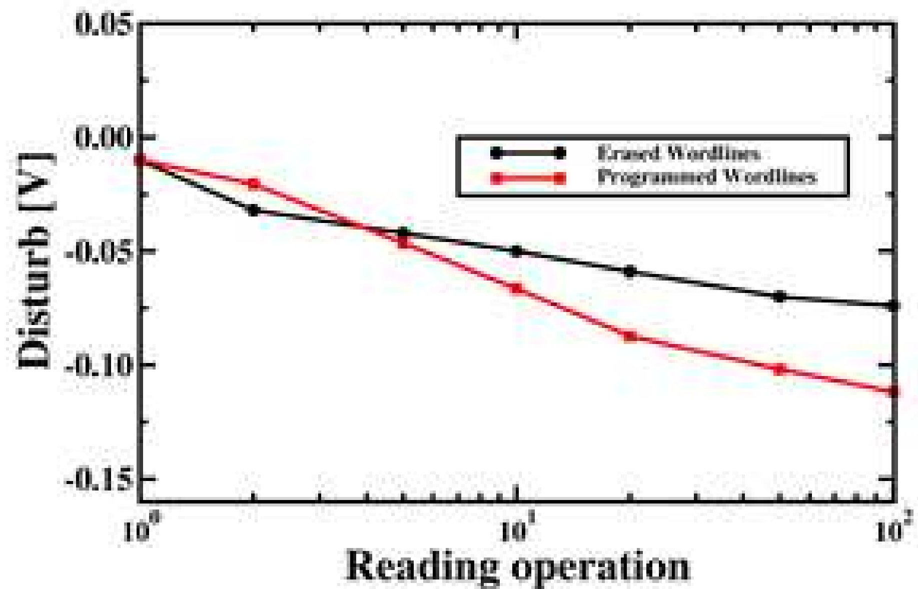



Figure 5.21: Impact of read disturb on both programmed and erased cells.

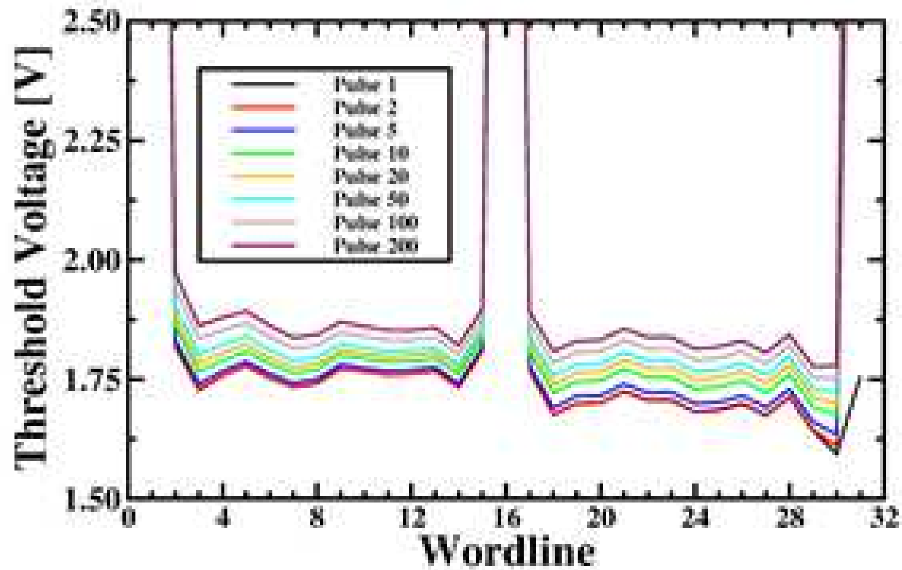



Figure 5.22: Threshold voltage measured after 1, 2, 5, ..., 200 program pulses applied simultaneously on WL1, WL16, and WL30.

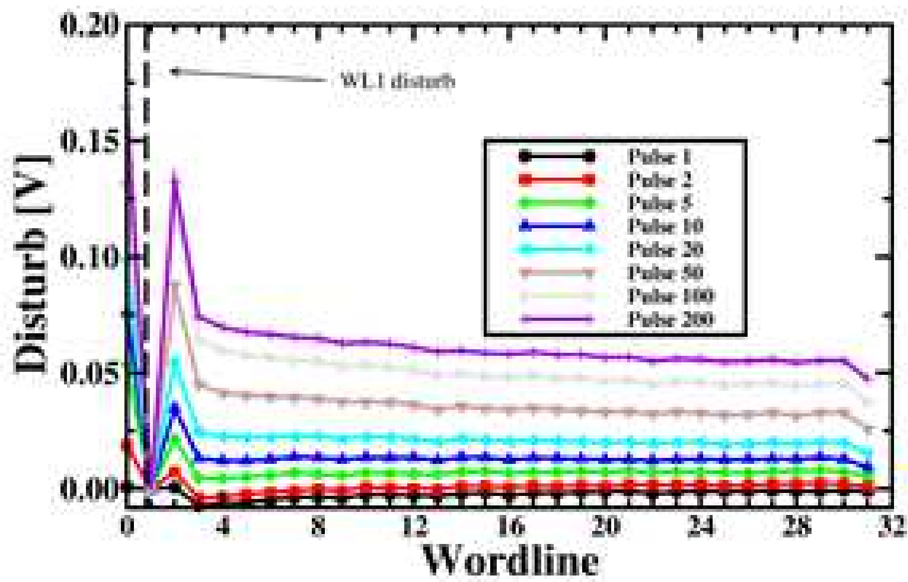



Figure 5.23: Net effect (disturb) of a program disturb measured after 1, 2, 5, ..., 200 disturb pulses applied on WL1.

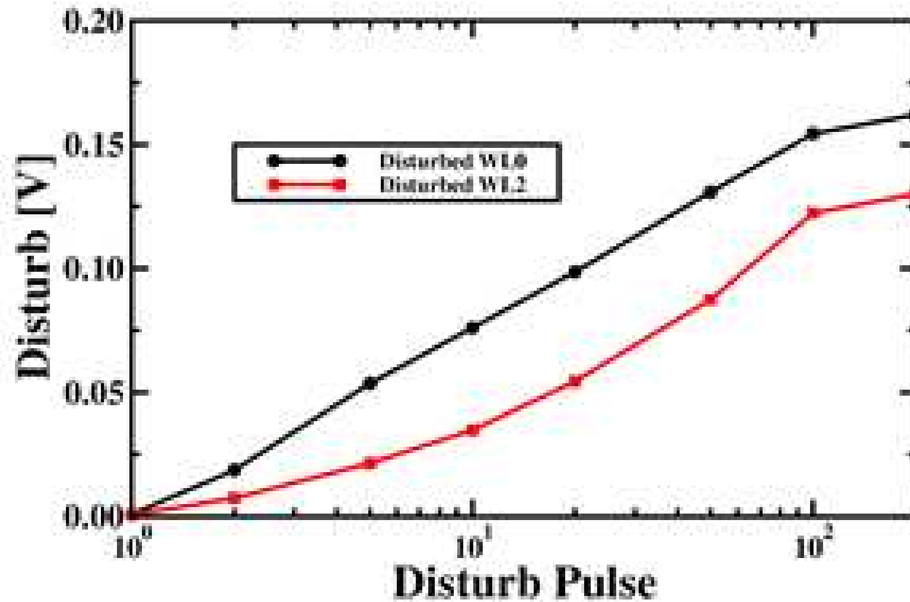



Figure 5.24: Evolution of disturb on WL0 and WL2 with the number of program disturb pulses applied to WL1.

In order to minimize interbitline capacitive coupling noise during a page Program, a dual bitline scheme has been used, in which only the even or the odd bitlines are activated simultaneously in the array. When WL0 is selected for programming, the bias situation is that depicted in Fig. 5.28 and the unselected cells are affected by the GIDL phenomenon. When other bitlines are selected for programming (see for example the case depicted in Fig. 5.29 where WL1 is selected for programming), the cells belonging to WL0 are affected either by the GIDL effect or by a $V_{pass}$ disturb. In the latter case, however, the boosted channel voltage may be different for cells above or below the selected one, depending on the actual voltage applied to the selected WL. These effects, controlling the actual program/inhibit dynamics, will be detailed in the next section.

The erase operation is performed block-wide (1 Mbits) with a single voltage

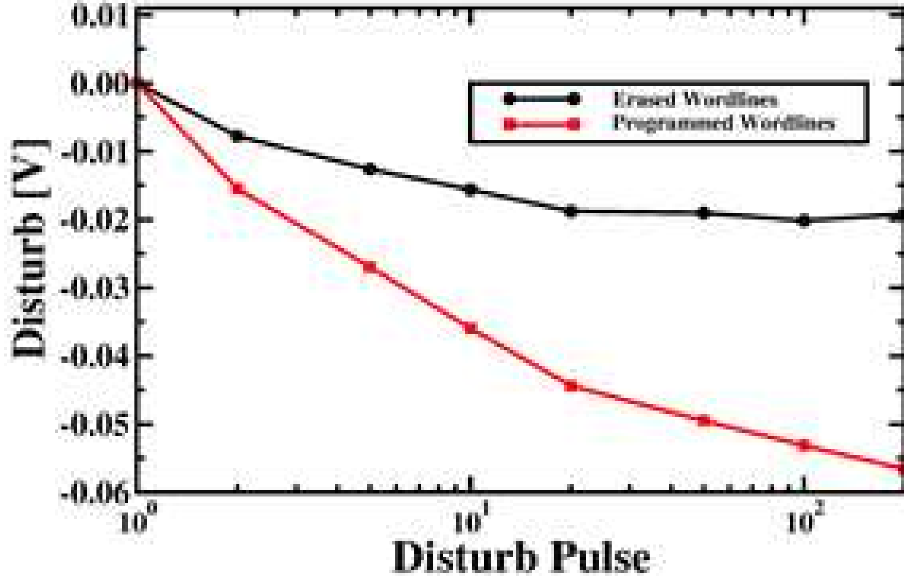



Figure 5.25: Block to block interference evaluated by monitoring both programmed and erased cells while applying program disturb pulses on WL33 of an adjacent block.

pulse featuring 19V amplitude and 100 $\mu$s duration. All the cells within the array are in the erased state before every experiment.

As introduced in the previous sentences, the Edge Wordline Disturb (EWD) magnitude, indicated as $\Delta V_{T_WL0}$, has been measured as:

$$\Delta V_{T_WL0} =< V_{T_WL0} > - < V_T > \qquad (5.1)$$

where $< V_{T_WL0} >$ is the average threshold voltage of all the cells belonging to row WL0, and $< V_T >$ is the average threshold voltage of all the other cells of the array.

As stated in the section beginning, a first contribution to EWD is attributed to the different programming dynamics affecting cells belonging to WL0 with respect to all other cells and related to their different geometries, therefore to a

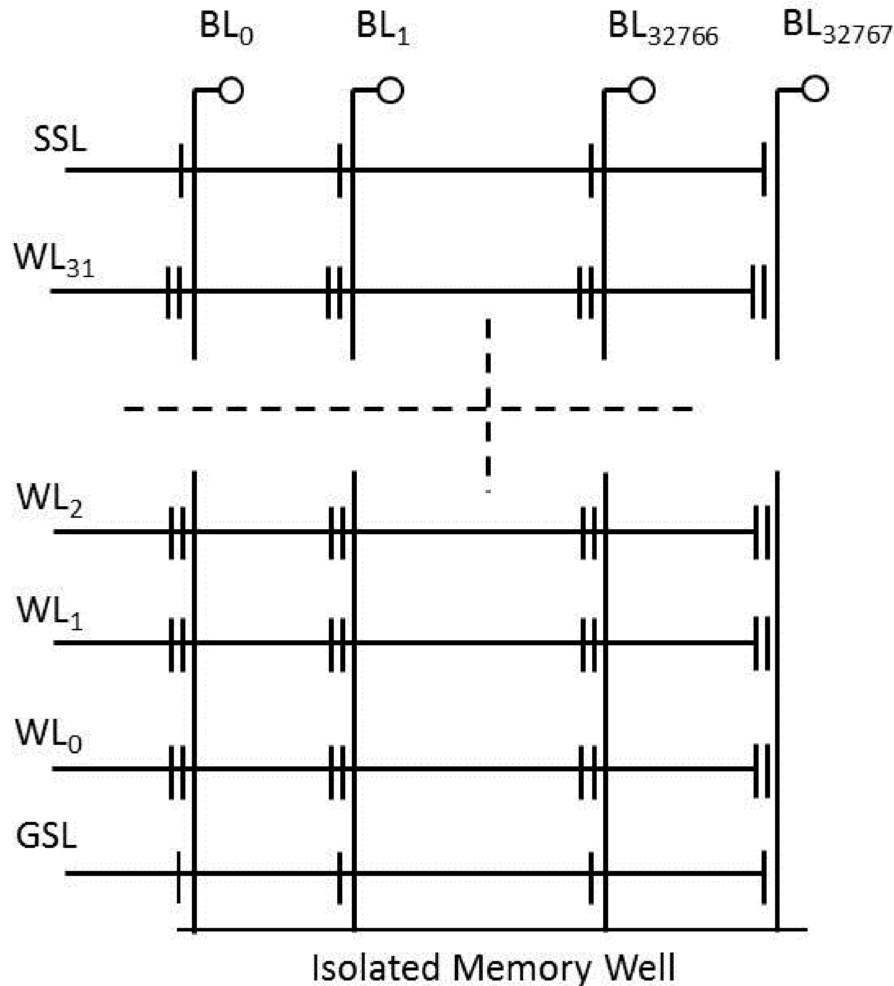



Figure 5.26: Standard NAND array architecture. The figure represents a single block of the CT array considered in this analysis.

different field distribution that allows cell programming with lower ISPP voltages.

Cells belonging to WL0 are also characterized by a peculiar field distribution in the space region between their source diffusion and the drain diffusion of the GSL selector making possible the presence of a GIDL current. The GIDL is a major leakage mechanism occurring in "OFF" MOS transistors when high voltages are applied to the Drain and it is attributed to tunneling taking place in the deep-

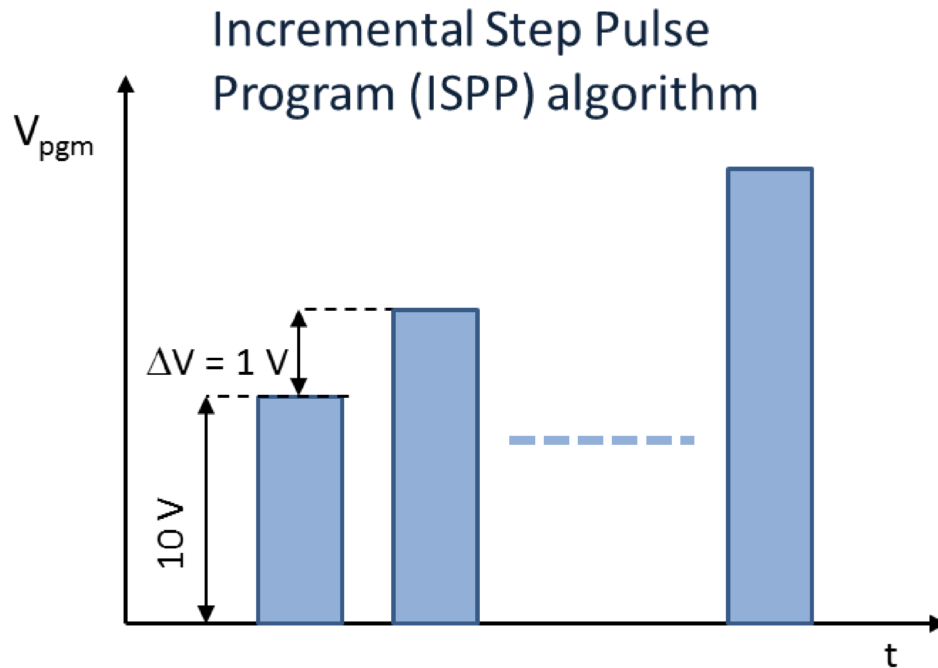



Figure 5.27: ISPP Pulse characteristics exploited in this work. The duration per pulse is $4\mu$s.

depleted or even inverted region underneath the gate oxide. Such a phenomenon has been revealed in FG NAND Flash during the Program operation [56] and it is theoretically present also in CT NAND Flash, as the basis of the Program operation are shared.

The GIDL effect may interest the cells belonging to WL0 in two different bias conditions: $i$) when WL0 is selected for programming and the channel of the inhibited bitlines is locally self-boosted (see Fig. 5.28); $ii$) when WL0 cells act as pass transistors and the bitlines are inhibited through channel self-boosting (see Fig. 5.29). Under these bias conditions large GIDL current may be generated at the drain edge of GSL transistor because the potential at the drain node of GSL transistor is raised by channel boosting [56]. Then, electron-hole pair generation

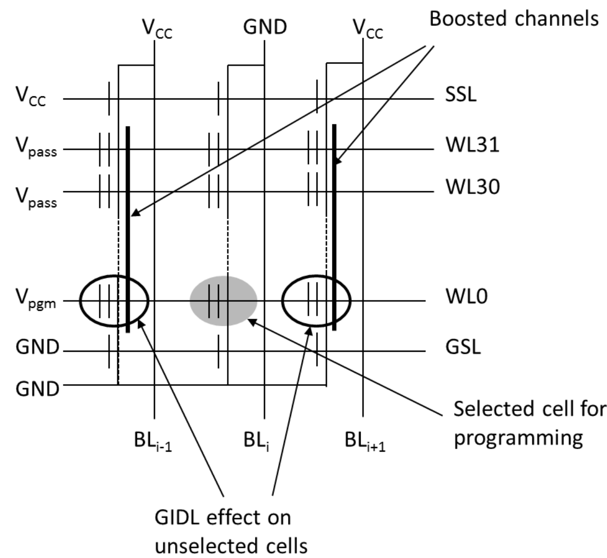



Figure 5.28: Bias conditions possibly activating the GIDL effect from GSL transistors belonging to columns $\mathrm{BL}_{i-1}$ and $\mathrm{BL}_{i+1}$.

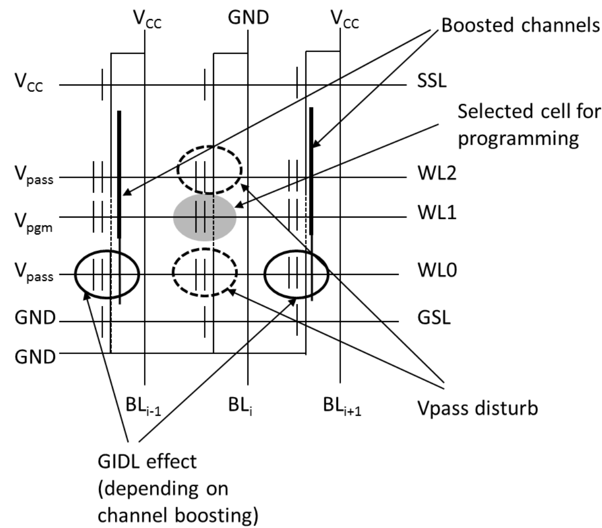



Figure 5.29: Bias conditions possibly activating the GIDL effect from GSL transistors belonging to columns $\mathrm{BL}_{i-1}$ and $\mathrm{BL}_{i+1}$ and the $V_{pass}$ disturb in unselected cells belonging to columns $\mathrm{BL}_{i}$.

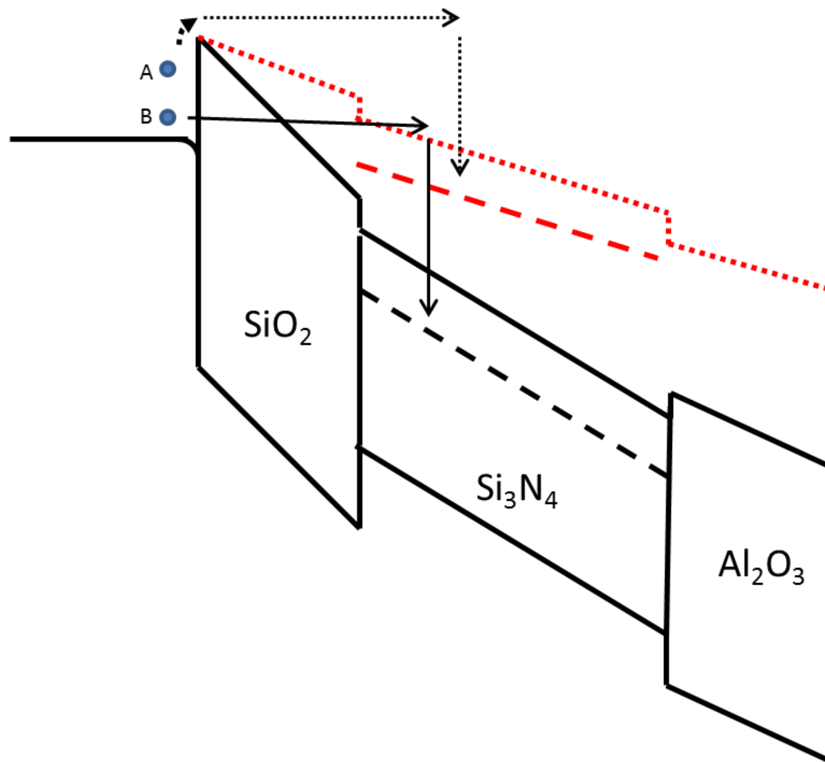



Figure 5.30: Band structure of the CT cells (excluding the TaN/Ti/TaN metal gate) considered in this work. For high biases electron B is free to tunnel into $SiO_2$ and then gets trapped into $Si_3N_4$ layer. For low biases electron A is still able to get trapped, although the energy levels of the bands are shifted upwards.

follows and the generated electrons are accelerated at the GSL-WL0 space region which can be hot enough to be injected into the floating gate in FG NANDs or trapped into the $Si_3N_4$ layer of the CT NANDs WL0 cells. This causes a positive shift of the average threshold voltage of the cells belonging to WL0, whereas keeping unmodified the average threshold voltage of all the other cells of the array. The electron injection and trapping depend on the relative values of the WL0 voltage and that of the self-boosted channel. In Fig. 5.30 it is shown the band diagram for the CT NAND devices considered in this work. For high gate biases (i.e. when WL0 is programmed) the electrons tunnel through the $SiO_2$ and get trapped

into the $Si_3N_4$ layer, whereas for low gate biases (i.e. when WL0 is inhibited) it is still possible for high energy electrons to pass the $SiO_2$ barrier and subsequently get trapped.

The difference on the impact of the EWD on the memory reliability between standard FG and CT Flash NANDs principally relies on the fact that FG NANDs depend only on tunneling efficiency, whereas CT NANDs depend also on trapping efficiency.

The last marginal contribution to EWD is caused by the $V_{pass}$ disturb that affects cells that are inhibited for programming (Gate voltage at $V_{pass}$ and bitlines at GND), as in Fig. 5.29. The impact of $V_{pass}$ disturb, that affects all cells acting as pass transistor because of an unwanted soft programming, is different for cells belonging to WL0 because of their different geometries and, therefore, coupling ratios.

Different experiments have been carried out to analyze the EWD in CT NAND arrays, and to expose its dependence on bias conditions.

In order to discriminate from the presence of other indeterminate sources of measurements perturbation, a bulk Program operation has been executed on a whole memory block, by varying the final voltage of the ISPP algorithm.

Fig. 5.31 shows the dependency of EWD on the voltage steps exploited in ISPP retrieved by measuring the threshold voltage of all cells within a block. The plot evidences that for the first ISPP steps the disturb magnitude is in the range of 300mV, and it decreases for higher programming voltages. At low voltages indeed, the large EWD is justified by the geometrical and the coupling ratios mismatches that accelerate the program dynamics in WL0 cells. At high voltages, on the contrary, when the programming transient approaches a saturating trend,

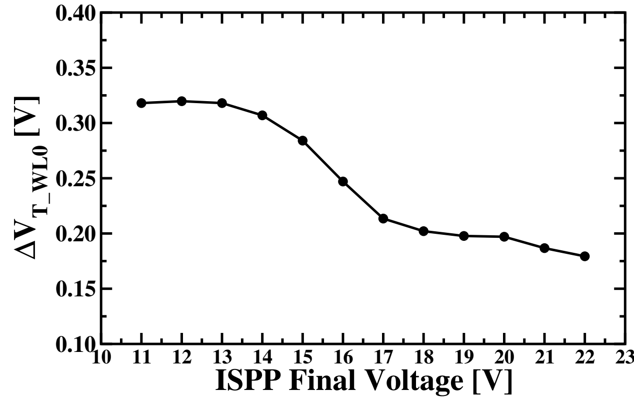



Figure 5.31: EWD dependence on the exploited ISPP voltages during a program operation. $V_{pass}$ has been fixed at 8 V for the unselected wordlines.

EWD is justified by the GIDL effect.

The presence of EWD has been analyzed when WL0 is driven into an inhibit condition. However, rather than analyzing all possible cases, the analysis has been focused on a repeated programming of WL1. In fact, in some applications such as multimedia and embedded systems, the multiple writing of a single wordline is allowed in NAND memories in order to optimize the usage of the memory addressing space. This policy is called *partial programming* and the number of multiple writing operations sustained by the memory is called NOP [56]. The impact of EWD on CT NAND Flash arrays has been analyzed with respect of NOP number, by applying the bias configuration indicated in Fig. 5.29.

By programming only cells belonging to WL1, while keeping all other wordlines at $V_{pass}$, it is possible to evaluate two different EWD causes: GIDL and $V_{pass}$-induced disturb. Fig. 5.32 shows EWD magnitude as a function of NOP number. In this case the EWD has been calculated as:

$$\Delta V_{T_WL0} = < V_{T_WL0} > - < V_{T_WL2} > \quad (5.2)$$

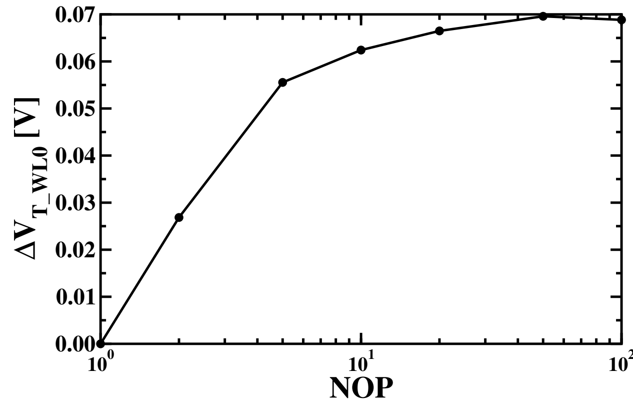



Figure 5.32: EWD dependence on the number of performed NOP on WL1. $V_{pass}$ has been fixed at 8 V for the unselected wordlines.

where $< V_{T_WL2} >$ is the average threshold voltage of the cells belonging to WL2. The choice of limiting the reference cells on which calculate EWD to those belonging to WL2 is justified by the need of maximizing the differences between the two bitlines adjacent to the selected one. Results in Fig. 5.32 show that by applying a number of 200 NOP on cells belonging to WL1, whereas keeping $V_{pass}$ voltage at 8V, an EWD of about 70mV is detected. The dependency on NOP number shows a logarithmic trend on the very first operations performed on WL1, whereas it reaches a steady state after a large number of operations.

Starting from the theoretical base of the previous experiments, the EWD can also be analyzed as a function of $V_{pass}$. As in the previous experiment, only the cells belonging to WL1 are programmed (up to 200 NOP) and a $V_{pass}$ bias is applied to the other wordlines. After each set of 200 NOP, the entire array is erased and the experiment is repeated by applying a different $V_{pass}$ voltage. The $V_{pass}$ ranges from 6.5 V to 10 V with 0.5 V step.

As shown in Fig. 5.33, a maximum EWD of 500 mV has been measured for the lowest $V_{pass}$ bias whereas it decreases for higher $V_{pass}$ voltages. A detailed

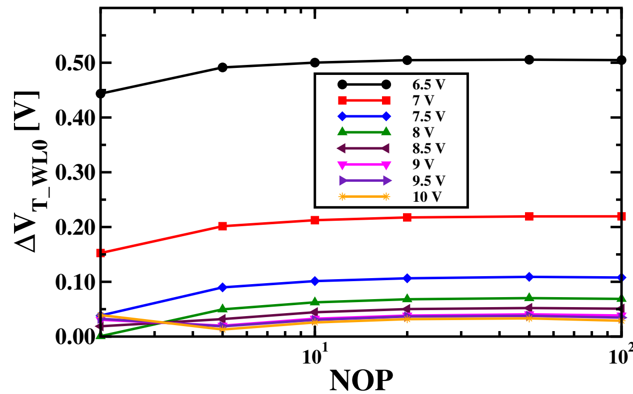



Figure 5.33: EWD as a function of the NOP number. $V_{pass}$ has been varied in order to evaluate the dependence of the phenomenon on the pass voltage.

EWD analysis must take into account the actual band structure within the region including the GSL selector and the 3 cells in series belonging to WL0, WL1 and WL2, together with the band structure modification induced by coupling effects and charge trapping.

The EWD dependency on cycling has been evidenced by stressing the CT array by a sequence of 1000 bulk Program/Erase cycles with intermediate Read operations for retrieving the threshold voltage values of all the cells. The EWD has been calculated by using equation (5.1), since bulk programming bias configuration has been considered. A measure of EWD is available both for the Programmed state and the Erased state of the cells belonging to WL0, as the performed Read measurements took place after each operation.

As shown in Fig. 5.34, the EWD initially increases the charge trapped in the $Si_3N_4$ layer: such charge cannot be efficiently removed by the subsequent Erase operation and a cumulative voltage shift in cells belonging to WL0 is evidenced. No saturating trend can be depicted after 1000 cycles.

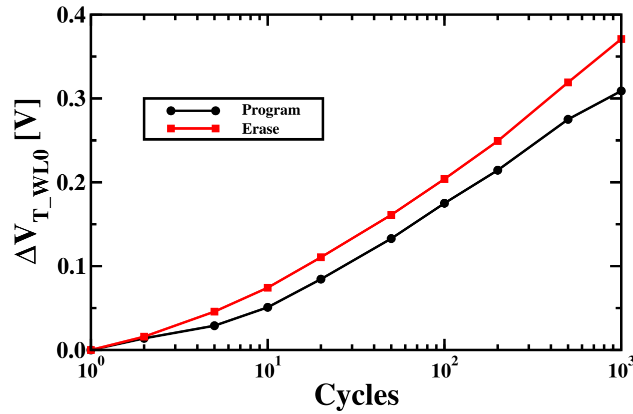



Figure 5.34: EWD dependence on the number of write cycles performed on the CT NAND array

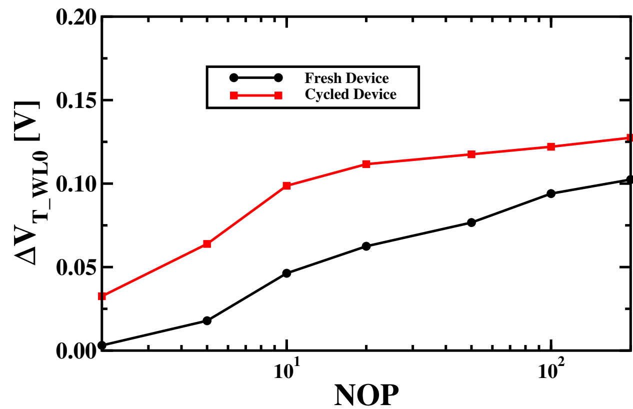



Figure 5.35: EWD as a function of the NOP number. A dependence on the device aging is evidenced. $V_{pass}$ has been fixed to 8 V for the unselected wordlines.

After cycling the array the experiment for evaluating the EWD dependency on the WL0 inhibition has been repeated, in order to compare the EWD impact on cycled devices with respect to fresh devices. As shown in Fig. 5.35, the EWD impact is greater on cycled devices. This can be attributed to the higher number of traps generated during cycling, which allows for a higher number of electron to be trapped thus causing a further upward threshold voltage shift.

# Chapter 6

# Conclusions

In this thesis it has been studied the most promising candidates of the upcoming future to become a serious replacement of the Flash memory technology. The topic of this work has been concerned about the electrical characterization, physics, modeling and reliability of innovative non-volatile memories, addressing most of the proposed alternative to the floating-gate based memories which currently are facing a technology dead end. Two possible replacement scenarios can be evidenced on the market for the innovative memories studied in this work: the consumer application and the embedded application market segments.

For the consumer application-oriented market segment it has been proposed an extended study for Phase Change Memory and Charge Trapping NAND Flash. The former technology proven its universal memory capabilities thanks to an extremely high read/write speed, providing in the same time a good reliability which promotes this technology to be largely integrated in MPSoCs. Thanks to this research study it has been possible to help the comprehension of many physical phenomenon such as the erase kinematics and the seasoning phenomenon, pro-

viding useful hints for future development of the PCM technology. The latter technology instead, proven the capability to open unprecedented integration scenarios thanks to the 3D integrability feature. By the reliability standpoint few concerns have been evidenced. In particular the natural similarities between the Charge Trapping technology and the standard floating gate Flash memories has been evidenced.

For the embedded application-oriented market segment two technologies have been studied: the SimpleEE and the NanoMEMS memories. Both have proven an extended reliability suitable for the integration in harsh environment such as automotive, military, space avionics, etc. Moreover, for the first time (this is the case for the NanoMEMS memories) it has been analyzed a memory which does not rely completely on the semiconductor theory principles expanding the knowledge towards other study disciplines.

Finally, a new research branch has been started on non-volatile memory technology by exploiting all the results and the experience gained on the physical characterization and on the reliability assessment of the described technologies.

# Appendix A

# System-level reliability of non-volatile memories

The need to improve non-volatile memories reliability in embedded systems is a key design concern. Here it is proposed a methodology, managed by the memory controller, that optimizes the data reliability at the physical level for critical data whereas exploiting the transaction performances for non-critical data. The reliability-performance trade-off is obtained by partitioning the memory addressable space in different functional blocks, each on written by means of a specific optimized writing algorithm. The method feasibility is demonstrated by a case study exploiting Phase Change Memories (PCM) features.

## A.1 General idea

Non-volatile memories (NVM) are generally recognized to be key components in embedded systems, since they allow storing permanently both code and data.

Depending on applications, different NVM features and/or technologies are exploited. Both for traditional NOR/NAND Flash memories and for emerging technologies, such as Phase Change memories (PCM) or resistive RAM (RRAM), many studies have been performed at the physical level to detect the optimal algorithm and voltage waveforms to be applied for memory writing [25, 35]. The optimal writing scheme guarantees the best trade-off between performances (in terms of data throughput), reliability (in terms of endurance and data retention), and noise immunity. In many applications, however, the write bandwidth become the designer's goal, since the targeted reliability can be reached by using self-correcting mechanisms such as Error Correction Codes (ECC). Such solutions, however, are very expensive in terms of area overhead and power/delay impact [57].

In this appendix it is shown that, rather than using a single writing methodology for the entire NVM, it is possible to partition the NVM addressable space in several functional blocks, each one characterized by a targeted data throughput and an expected inherent reliability. Each block is then associated to a different writing algorithm characterized by a specific voltage pulse amplitude, shape and duration. The memory partition is managed by the memory controller and it does not require any additional hardware at the controller level nor a modification of the NVM architecture. Being the method overhead limited to few code instructions controlling writing waveform generation, the partition can be reconfigured depending on the system application.

Since the data reliability for critical data is enhanced at the physical level by selecting the appropriate writing algorithm, the need for additional expensive ECC strategies can be relaxed.

In this appendix it will be demonstrated how the method can be applied to the case of a Multi-Processor Systems-on-Chip devoted to multimedia applications, such as video streaming, which integrates a PCM array as a primary memory solution. The choice of a PCM relies on the availability of an analytical model characterizing different writing schemes in terms of data throughput and estimated reliability [22].

## A.2 The physical view on PCM

Let us now delve into some details of PCM technology in an attempt to identify the degrees of freedom that the technology itself makes available when memory transactions are performed.

The PCM concept relies on the phase transition from crystalline into amorphous and vice versa of a $Ge_2Sb_2Te_5$ (GST) material as shown in Chapter 2 of this thesis. Different voltage waveforms are applied to the GST in order to achieve program/erase operations. The programming operation requires a voltage pulse, bringing the temperature of the GST above its melting point, then cutting away the pulse rapidly, in order to melt and then set the GST on the amorphous state. The erasing operation, instead, can be achieved in many ways [25] aimed at creating a percolation low resistive path inside the programmable area thus ensuring a read current one order of magnitude higher with respect to the programmed state.

In this work we consider only the following schemes (seen in Fig. 2.10):

- *Melt and Crystallize Waveform (MCW)*: the programmable area is completely molten before the creation of the low resistive path.

- *Crystallize Once Waveform (COW)*: the phase transition is obtained by applying a voltage pulse which does not melt the programmable area but creates linked crystal grains causing electrical conduction.
- *Sloped Crystallize Once Waveform (SCOW)*: it is a variation of the $COW$ waveform based on a smooth removal of the voltage necessary for crystal grains creation.

In terms of reliability and performance the MCW scheme is known for its high reliability in terms of distribution compactness and noise immunity [25], whereas its performance is poor in comparison with other schemes since a longer waveform is used, therefore resulting into a reduced write throughput. The characteristics of each erasing scheme are summarized in Tab. A.1, where the compactness factor indicates the ability of a particular erasing scheme to reduce the standard deviation of the erased distribution values, whereas the noise immunity indicates the capability of withstand events such as RTN (Random Telegraph Noise) [39].

The manyfold ways to perform erasing represent a degree of freedom which can be exposed to the upper layers of the design hierarchy to implement technology-based selective data protection. For instance, next section shows how the different features of an application data set perfectly match the flexibility enabled by the concurrent use of many memory erasing algorithms.

## A.3 A case study

In video streaming applications the real-time system capability to decode data compressed with a particular format is requested. The most commonly used for-

Table A.1: Erasing schemes comparison resume.

| Erasing scheme | Compactness factor | Noise immunity | Speed |
|---|---|---|---|
| MCW | Highest | Highest | Lowest |
| COW | Low | Low | Highest |
| SCOW | High | Medium | Medium |

mat is the MPEG4 in which imaging data are sent in stream of packets [58]. Each packet represents a portion of the image at a defined resolution. The first packet of the stream, also called the DC Sub-band information, contains the image data at a very low resolution. This packet, which features very small size with respect to other packets on the stream, is fundamental for the image reconstruction because if it gets either corrupted or lost, it would be impossible to eventually rebuild the image. The following packets of the stream, the so-called AC or Spatial Levels, represent the additional details about the image and in case they are lost the user will still have an image but at a lower resolution.

Based on the previous considerations, a part of the data must be securely written into a memory sector that sacrifices the write speed performances in favor of enhanced data stability, whereas the rest of the data can be stored in a less reliable sector but with high write throughput features.

Hence it has been decided to split a single PCM array into four different storage regions with different characteristics, each associated to a specific erasing scheme:

- *Boot Area*: it is responsible for memory initialization and system checking at power-up. This sector needs the highest level of reliability. Since the boot area is expected to be written once with a limited amount of data, no

particular constraints on write throughput are required. The best erasing scheme which fits the need is the $MCW$, featuring high data stability at lower write speed.

- *Code Area*: it is the area which stores the application code. This area differs from the previous one only by the amount of data to be stored, while the same reliability level is required. The $MCW$ can be used for this memory partition.

- *Reliable Data Area*: it stores the application data whose reliability is critical for the considered application. With our case study, critical data would be represented by the DC Sub-band data. $SCOW$ is the ideal erasing scheme since it provides a tradeoff between bit error rate and write throughput.

- *High Speed Data Area*: this is the memory area in which the data exchange/storage occurs frequently and the write speed is the basic goal. Within our case study these data are represented by the AC Spatial Levels data. $COW$ is the ideal erasing scheme.

This suggested partitioning scheme can be applied to any of the functional blocks that are multiple of the minimum memory addressable space (i.e. pages, sectors, blocks, etc.) for a defined NVM technology. It is worth to point out that since this methodology applies to the whole addressable memory space of a system, it can be applied on more than one integrated physical memory.

## A.4 Experimental results

In order to assess the validity of our approach for reliable system design, we carried a set of experiments aimed at exploiting the capabilities of PCMs to withstand the proposed partitioning scheme. For this purpose we adopted a simulation framework implemented on a numerical analysis tool, which models the characteristics of a PCM array. The model, based on an electrical characterization performed on a 8 Mbits PCM test chip manufactured in 0.18 $\mu$m technology [27], is able to reproduce the behavior of a memory cell in relation to different erasing schemes [22].

The PCM has been subdivided on 32 kbits for boot area sector, 128 kbits for code area, 2048 kbits for the reliable data area and the remaining 5984 kbits for the high speed data area. All sectors are written with random patterns.

The theoretical write throughput for each erasing scheme has been calculated by considering the following factors: the time length of the selected erasing waveform, the circuitry delay introduced by the on-memory peripherals (i.e. decoders, charge pumps, etc.), which has been fixed to 50 ns, and the parallelism of the data I/O which is equal to 8 [27].

Under these assumptions we calculated a write throughput of 247 KB/s for the $MCW$ erasing scheme, 645 KB/s for the $SCOW$ scheme and 6.7 MB/s for the $COW$ scheme (Fig. A.1).

The second experiment reproduces a typical situation of the case study: the boot and the code areas are written once at the beginning of the experiment, while the two data areas are written with random patterns more than once with different number of writing operations, thus providing an estimation of the mean write

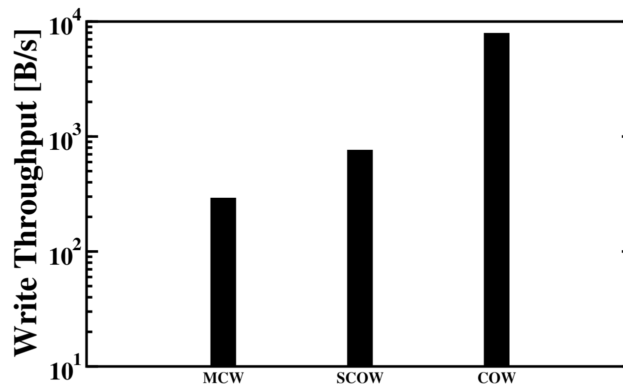



Figure A.1: Calculated write throughput comparison between the different erasing schemes applied within a PCM.

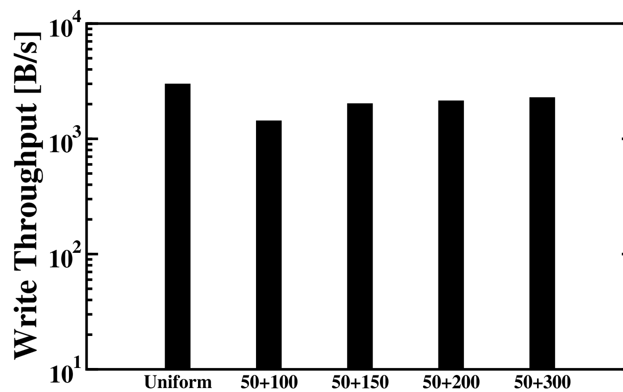



Figure A.2: Mean write throughput calculated with different writing scenarios. The uniform traffic condition is compared with 50 writings on the data reliable area and 100, 150, 200, 300 writings on the high speed area.

throughput under different load conditions (Fig. A.2). The write throughput is more than 1.5 MB/s for several experiment conditions, which represents a 40% reduction of the theoretical write throughput calculated under uniform writing of the memory (i.e. redistribution of the writings on all partitions) equal to 2.53 MB/s.

The last experiment analyzes the error probability of the partitioned memory by monitoring the number of bits in the array which fails to be erased (see

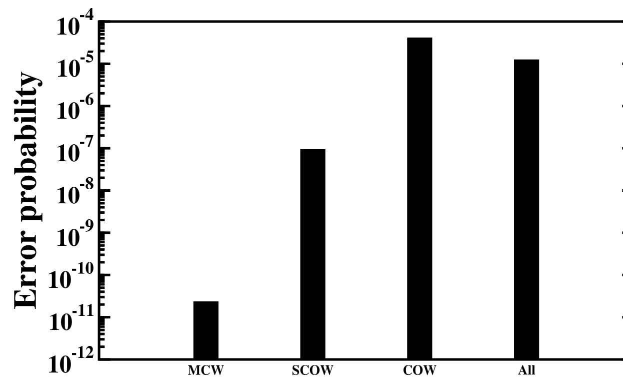



Figure A.3: Error probability comparison between single PCM data partitions and PCM comprising the whole array

Fig. A.3). These results show the physical reliability that can be achieved by using different writing schemes. Additional ECC polices applied at the system level can be used to reach the expected BER. Nevertheless, it is clear that the starting point is no longer the same for the whole memory array and that relaxed strategies can be applied to the most reliable blocks at the physical level.

The feasibility of the proposed method has been demonstrated on a simple case study integrating a PCM as a primary memory, whose writing speed and expected reliability as a function of the writing algorithm were analytically modeled. It has been shown that the use of specific algorithms for critical data, while impacting on the writing speed, may increase the inherent data reliability thus reducing the complexity of ECC at system level. The methodology can be directly applied to any NVM memory whose performances as a function of the erasing scheme are known.

# Bibliography


[1] F. Masuoka, M. Momodomi, Y. Iwata, and R.Shirota, "New ultra high density eprom and flash eprom cell with nand structure," *IEEE Tech. Dig.*, pp. 552–555, 1987.

[2] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, *Flash memories*. Kluwer, 1999.

[3] M. Lenzlinger and E. Snow, "Fowler-nordheim tunneling into thermally grown sio2," *IEEE Tech. Dig.*, pp. 273–283, 1969.

[4] Y. Park and D. Schroeder, "Degradation of thin tunnel gate oxide under constant fowler-nordheim current stress for a flash eeprom," *IEEE Trans. Electron Devices*, vol. 45, 1998.

[5] A. Modelli, A. Visconti, and R. Bez, "Advanced flash memory reliability," in *IEEE International Conference on Integrated Circuit Design and Technology*, pp. 211–218, 2004.

[6] D. Ielmini, A. Spinelli, and A. Lacaita, "Recent developments on flash memory reliability," *Microelectronic Engineering*, vol. 80, 2005.

[7] N. Mielke, H. Belgal, I. Kalastirsky, P. Kalavade, A. Kurtz, Q. Meng, N. Righos, and J. Wu, "Flash eeprom threshold instabilities due to charge trapping during program/erase cycling," *IEEE Transactions on Device and Materials Reliability*, vol. 4, 2004.

[8] P. Cappelletti, R. Bez, A. Modelli, and A. Visconti, "What we have learned on flash memory reliability in the last ten years," in *IEEE Tech. Dig. of IEDM*, 2004.

[9] A. Chimenton, P. Pellati, and P. Olivo, "Erratic bits in flash memories under fowler-nordheim programming," *Jpn. J. Appl. Phys.*, vol. 42, 2003.

[10] A. Spinelli, "Irps tutorial," 2009.

[11] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," in *Proceedings of the IEEE*, vol. 91, pp. 305–327, 2003.

[12] L. Lopez, P. Masson, D. Ne, and R. Bouchakour, "Temperature and drain voltage dependance of gate-induced drain leakage," *Elsevier Microelectronics Engineering*, vol. 72, 2004.

[13] C. Compagnoni, R. Gusmeroli, A. Spinelli, and A. Visconti, "Analytical model for the electron-injection statistics during programming of nanoscale nand flash memories," *IEEE Trans. on Electron Devices*, vol. 55, 2008.

[14] A. Chimenton, P. Pellati, and P. Olivo, "Constant charge erasing scheme for flash memories," *IEEE Trans. on Electron Devices*, vol. 49, 2002.

[15] A. Pirovano, A. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, and R. Bez, "Scaling analysis of phase-change technonlogy," in *Technical Digest IEEE Electron Devices Meeting*, pp. 29.6.1–29.6.4, 2003.

[16] S. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," *Physics Review Letters*, vol. 21, no. 20, pp. 1450–1453, 1968.

[17] I. Karpov, S. Savransky, and V. Karpov, "Mechanism of threshold swtiching in chalcogenide phase change memory devices," in *22nd IEEE Non-Volatile Semiconductor Memory Workshop*, pp. 56–57, 2007.

[18] F. Pellizzer, A. Pirovano, F. Ottogalli, M. Magistretti, M. Scaravaggi, P. Zuliani, M. Tosi, A. Benvenuti, P. Besana, S. Cadeo, T. Marangon, R. Morandi, R. Piva, A. Spandre, R. Zonca, A. Modelli, E. Varesi, T. Lowrey, A. Lacaita, G. Casagrande, P. Cappelletti, and R. Bez, "Novel $\mu$-trench phase-change-memory cell for embedded and stand-alone non-volatile memory applications," in *VLSI Symp. on Tech.*, pp. 18–19, 2004.

[19] A. Chimenton, C. Zambelli, A. Pirovano, and P. Olivo, "Set of electrical characteristic parameters suitable for reliability analysis of multimegabit phase change memory arrays," in *Non-Volatile Semiconductor Memory Workshop*, pp. 49–51, 2008.

[20] D. Adler *J. Vac. Sci. Technol.*, vol. 10, pp. 723–738, 1973.

[21] A. Pirovano, A. Redaelli, F. Pellizzer, F. Ottogalli, M. Tosi, D. Ielmini, A. Lacaita, and R. Bez, "Reliability study of phase-change nonvolatile memories," *IEEE Transactions on Device and Materials Reliability*, vol. 4, pp. 422–427, 2004.

[22] A. Chimenton, C. Zambelli, and P. Olivo, "A new analytical model of the erase operation in phase change memories," *IEEE Electron Device Letters*, vol. 31, pp. 198–200, 2010.

[23] V. G. Karpov, Y. A. Kryukov, I. V. Karpov, and M. Mitra, "Field-induced nucleation in phase change memory," *Phys. Rev. B*, vol. 78, p. 052201, 2008.

[24] D. Ventrice, P. Fantini, A. Redaelli, A. Pirovano, A. Benvenuti, and F. Pellizzer, "A phase change memory compact model for multilevel applications," *IEEE Electron Device Letters*, vol. 28, pp. 973–975, 2007.

[25] C. Zambelli, A. Chimenton, and P. Olivo, "Analysis and optimization of erasing waveform in phase change memory arrays," in *Solid State Device Research Conference, 2009. ESSDERC '09. Proceedings of the European*, pp. 213–216, 2009.

[26] J. Sarkar and B. Gleixner, "Evolution of phase change memory characteristics with operating cycles," *Appl. Phys. Lett.*, vol. 91, p. 233506, 2007.

[27] F. Bedeschi, C. Resta, O. Khouri, E. Buda, L. Costa, M. Ferraro, F. Pellizzer, F. Ottogalli, A. Pirovano, M. Tosi, R. Bez, R. Gastaldi, and G. Casagrande, "An 8mb demonstrator for high-density 1.8v phase-change memories," in *VLSI Circuits, 2004. Digest of Technical Papers. 2004 Symposium on*, pp. 442–445, 2004.

[28] H. K. Lyeo, D. G. Cahill, B. S. Lee, J. R. Abelson, M. H. Kwon, K. B. Kim, S. G. Bishop, and B. Cheong, "Thermal conductivity of phase-change material $ge_2sb_2te_5$," *Applied Physics Letters*, vol. 89, p. 151904, 2006.

[29] S. O. Ryu, S. M. Yoon, K. J. Choi, N. Y. Lee, Y. S. Park, S. Y. Lee, B. G. Yu, J. B. Park, and W. C. Shin, "Crystallization behavior and physical properties of sb-excess $ge_2sb_{2+x}te_5$ thin films for phase change memory (pcm) devices," *Journal of the Electrochemical Society*, vol. 153, pp. G234–G237, 2006.

[30] J. B. Park, G. S. Park, H. S. Baik, J. H. Lee, H. Jeong, and K. Kim, "Phase-change behavior of stoichiometric $ge_2sb_2te_5$ in phase-change random access memory," *Journal of The Electrochemical Society*, vol. 154, pp. H139–H141, 2007.

[31] U. Russo, D. Ielmini, A. Redaelli, and A. Lacaita, "Intrinsic data retention in nanoscaled phase-change memories - part i: Monte carlo model for crystallization and percolation," *IEEE Transactions on Electron Devices*, vol. 53, pp. 3032–3039, 2006.

[32] D. Fugazza, D. Ielmini, S. Lavizzari, and A. Lacaita, "Distributed-poole-frenkel modeling of anomalous resistance scaling and fluctuations in phase-change memory (pcm) devices," in *IEEE International Electron Devices Meeting (IEDM)*, pp. 1–4, 2009.

[33] A. Chimenton, C. Zambelli, and P. Olivo, "A new methodology for two-level random-telegraph-noise identification and statistical analysis," *IEEE Electron Device Letters*, vol. 31, pp. 612–614, 2010.

[34] J. Pathak, M. Thomas, J. Payne, G. Schatzberger, A. Wiesner, F. Leisenberger, E. Wachmann, and M. Schrems, "Embedded non-volatile memory

modules for low voltage and high temperature applications," in *9th Annual Non-Volatile Memory Technology Symposium*, pp. 27–30, 2006.

[35] A. Chimenton and P. Olivo, "Fast identification of critical electrical disturbs in nonvolatile memories," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2438–2444, 2007.

[36] T. Ong, A. Fazio, N. Mielke, S. Pan, N. Righos, G. Atwood, and S. Lai, "Erratic erase in etox flash memory array," in *VLSI Symp. on Tech.*, pp. 83–84, 1993.

[37] A. Chimenton and P. Olivo, "Erratic erase in flash memories (parti): Basic experimental and statistical characterization," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1009–1014, 2003.

[38] A. Chimenton, P. Pellati, and P. Olivo, "Overerase phenomena: An insight into flash memory reliability," *Proc. IEEE*, vol. 91, pp. 617–626, 2003.

[39] A. Chimenton, C. Zambelli, and P. Olivo, "A statistical model of erratic erase based on an automated random telegraph signal array characterization technique," in *Proc. IRPS*, pp. 896–901, 2009.

[40] K. Kobayashi, A. Teramoto, M. Hirayama, and Y. Fujita, "Metal-oxide-semiconductor-field-effect-transistor substrate current during fowler-nordheim tunneling stress and silicon dioxide reliability," *J. Appl. Phys.*, vol. 76, pp. 3695–3700, 1994.

[41] A. Chimenton, C. Zambelli, and P. Olivo, “A new automated methodology for random telegraph signal identification and characterization: a case study on phase change memory arrays,” in *Proc. IRPS*, pp. 128–133, 2009.

[42] A. Chimenton and P. Olivo, “Impact of tunnel oxide thickness on erratic erase in flash memories,” in *Proc. of the European Solid-State Device Conference*, pp. 363–366, 2002.

[43] A. Chimenton and P. Olivo, “Reliability of flash memory erasing operation under high tunneling electric field,” in *Proc. IRPS*, pp. 216–221, 2004.

[44] M. Seo, S. Sim, Y. Sim, M. On, S. Kim, I. Cho, H. Lee, G. Kim, and M. Kim, “A 0.9v 66mhz access, 0.13um 8m(256k x 32) local sonos embedded flash eeprom,” in *VLSI Circuits, 2004. Digest of Technical Papers. 2004 Symposium on*, pp. 68–71, 2004.

[45] J. V. Houdt, G. Groeseneken, and H. Maes, “Himos: an attractive flash eeprom cell for embedded memory applications,” *Microelectronics Journal*, vol. 24, pp. 190–194, 1993.

[46] A. Concannon, D. McCarthy, A. Mathewson, B. Guillaumot, C. Papadas, and C. Kelaidis, “A novel cmos compatible multi-level flash eeprom for embedded applications,” in *Device Research Conference Digest, 1998. 56th Annual*, pp. 78–79, 1998.

[47] C. Smith, “Bi-stable memory element.” Patent No. US5677823, 1994.

[48] J. Drake, H. Jerman, B. Lutze, and M. Stuber, "An electrostatically actuated micro-relay," in *Tech. Digest, 8th Int. Conf. Solid-State Sensors and Actuators (Transducers '95/Eurosensors IX)*, pp. 380–383, 1995.

[49] C. Goldsmith, T. Lin, B. Powers, W. Wu, and B. Norvell, "Micromechanical membrane switches for microwave applications," in *Proc. 1995 IEEE MTT-S Int. Microwave Symp.*, pp. 91–94, 1995.

[50] C. Smith, R. Kampen, J. Popp, D. Lacy, D. Pinchetti, M. Renault, V. Joshi, and M. Beunder, "Nanomechanical cantilever arrays for low-power and low-voltage embedded nonvolatile memory applications," in *Proc. SPIE 6464*, p. 646406, 2007.

[51] C. Smith, "Micro-mechanical elements." Patent No. US6441405, 2002.

[52] V. Joshi, R. Knipe, R. Kampen, D. Lacey, T. Nagata, D. Yost, and C. Smith, "A non volatile mems switch for harsh environment memory applications," in *International Conference and Exhibition on High Temperature Electronics Network (HiTEN)*, 2009.

[53] R. Micheloni, L. Crippa, and A. Marelli, *Inside NAND Flash memories*. Springer-Verlag, 2010.

[54] C. Lee, J. Choi, C. Kang, Y. Shin, J. Lee, J. Sel, J. Sim, S. Jeon, B. Choe, D. Bae, K. Park, and K. Kim, "Multi-level nand flash memory with 63 nm-node tanos (si-oxide-sin-al2o3-tan) cell structure," in *VLSI Symp. Tech. Dig.*, pp. 21–22, 2006.

[55] T. Melde, M. Beug, L. Bach, A. Tilke, R. Knoefler, U. Bewersdorff-Sarlette, V. Beyer, M. Czernohorsky, J. Paul, and T. Mikolajick, “Select device disturb phenomenon in tanos nand flash memories,” *Electron Device Letters, IEEE*, vol. 30, pp. 568–570, 2009.

[56] J. Lee, C. Lee, M. Lee, H. Kim, K. Park, and W. Lee, “A new programming disturbance phenomenon in nand flash memory by source/drain hot-electrons generated by gidl current,” in *Proc. NVSM Workshop*, pp. 31–33, 2006.

[57] R. Micheloni, A. Marelli, and R. Ravasio, *Error Correction Codes for Non-Volatile Memories*. Springer, 2008.

[58] I. Sodagar, H. Lee, P. Hatrack, and Y. Zhang, “Scalable wavelet coding for synthetic/natural hybrid images,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, pp. 244–254, 1999.

# Author's publications

## International journals

1. A. Chimenton, C. Zambelli, and P. Olivo,

   "**A New Analytical Model of the Erase Operation in Phase Change Memories**", *IEEE Electron Device Letters*, Vol. 31, pp. 198-200, Mar. 2010

2. A. Chimenton, C. Zambelli, and P. Olivo,

   "**A New Methodology for Two-Level Random Telegraph Noise Identification and Statistical Analysis**", *IEEE Electron Device Letters*, Vol. 31, pp. 612-614, Jun. 2010

3. C. Zambelli, A. Chimenton, and P. Olivo,

   "**Empirical Investigation of Set Seasoning effects in Phase Change Memories Arrays**", *Solid State Electronics*, Vol. 58, pp. 23-27, Jan. 2011

4. C. Zambelli, D. Bertozzi, A. Chimenton, and P. Olivo,

   "**Non volatile memory partitioning scheme for technology-based performance-reliability trade-off** ", *IEEE Embedded System Letters*, Vol. 3, pp. 13-15, Mar. 2011

5. A. Chimenton, C. Zambelli, P. Olivo,
   "**A Statistical Model of Erratic Behaviors in NAND Flash Memory Arrays**", *IEEE Trans. on Electron Devices*, Vol. 58, pp. 3707-3711, Nov. 2011

6. C. Zambelli, A. Chimenton, P. Olivo,
   "**Statistical Modeling of Secondary Path during Erase Operation in Phase Change Memories**", *IEEE Trans. on Electron Devices*, Vol. 59, pp. 813-818, Mar. 2012

7. C. Zambelli, A. Chimenton, P. Olivo,
   "**Modeling of SET Seasoning Effects in Phase Change Memory Arrays**", to appear on *Microelectronics Reliability*

# Book chapters

8. C. Zambelli, A. Chimenton, P. Olivo,
   "**Reliability of NAND Flash Memories"**,
   in Inside NAND Flash Memories, edited by R. Micheloni, L. Crippa, and A. Marelli, Springer, 2010

9. C. Zambelli, P. Olivo,
   "**SSD Reliability"**, *to be published by Springer*, 2012

# International conference proceedings

10. A. Chimenton, C. Zambelli, P. Olivo, and A. Pirovano,

    "**Set of electrical characteristic parameters suitable for reliability analysis of multimegabit Phase Change Memory arrays** ,

    in *Proc. IEEE Non Volatile Memory Workshop*, Opio (France), May. 2008

11. A. Chimenton, C. Zambelli, and P. Olivo,

    "**Impact of short SET pulse sequence on Electronic Switching in Phase Change Memory arrays** ,

    in *proc. Non-Volatile Memory Technology Symposium (NVMTS)*, Pacific Grove (Cal.), pp. 1 - 5 Nov. 2008

12. A.Chimenton, C. Zambelli, and P. Olivo,

    "**A new automated methodology for Random Telegraph Signal identification and characterization: a case study on Phase Change Memory arrays**,

    in *Proc. IEEE Int. Reliability Physics Symposium (IRPS)*, Montreal (Canada), pp. 128 - 133, April 2009

13. A.Chimenton, C. Zambelli, and P. Olivo,

    "**A statistical model of Erratic Erase based on an automated Random Telegraph Signal characterization technique**,

    in *Proc. IEEE Int. Reliability Physics Symposium (IRPS)*, Montreal (Canada), pp. 896 - 901, April 2009

14. C. Zambelli, A.Chimenton, and P. Olivo,

    "**Analysis and Optimization of Erasing Waveforms in Phase Change**

**Memory Arrays**,

in *Proc. IEEE European Solid-State Device Research Conf. (ESSDERC)*, Athens (Greece), pp. 213 - 216, Sept. 2009

15. A. Chimenton, C. Zambelli, P. Olivo, F.P. Leisenberg, A. Wiesner, G. Schatzberger, E. Wachmann, and M. Schrems

    "**Evidence of Erratic behaviors in p-channel floating gate memories and a cell architectural solution**,

    in *Non-Volatile Memory Technology Symposium (NVMTS)*, Portland (Oregon), Oct. 2009

16. C. Zambelli, A. Chimenton, and P. Olivo

    "**Modeling of Seasoning Effects in Phase Change Memory Arrays**,

    in *IEEE MOS-AK/GSA Workshop*, Rome (Italy), April 2010

17. C. Zambelli, A. Chimenton, P. Olivo,

    "**Experimental characterization of SET Seasoning on Phase Change Memory Arrays**,

    in *IEEE Int. Memory Workshop (IMW)*, Seoul (Korea), pp. 29 - 32, May 2010

18. A. Chimenton, C. Zambelli, and P.Olivo,

    "**Experimental Characterization of Phase Change Memory arrays**,

    in *International Symposium on Integrated Functionalities*, Portorico, June 2010

19. C. Zambelli, A. Chimenton, P. Olivo,

    "**Analysis of Edge Wordline Disturb in Multimegabit Charge Trapping**

**Flash NAND arrays**,

*Proc. IEEE Int. Reliability Physics Symposium (IRPS)*, Monterey (Cal.), pp. 2G.2.1 - 2G.2.4, April 2011

20. R. Gaddi, C. Schepens, C. Zambelli, A. Chimenton and P. Olivo,

    "**Reliability and Performance Characterization of a MEMS-based Non Volatile Switch**,

    *Proc. IEEE Int. Reliability Physics Symposium (IRPS)*, (**Invited Paper**) Monterey (Cal.), MY.4.1 - MY.4.4, April 2011

21. C. Zambelli, P. Olivo, R. Gaddi, C. Schepens, and C. Smith,

    "**Characterization of a MEMS-based Embedded Non Volatile Memory array for Extreme Environments**,

    *Proc. IEEE Int. Memory Workshop (IMW)*, Monterey (Cal.), pp. 1 - 4, May 2011

22. C. Zambelli, M. Indaco, M. Fabiano, S. Di Carlo, P. Prinetto, P. Olivo and D. Bertozzi,

    "**A Cross-Layer approach to the Reliability-Performance trade-off in MLC NAND Flash Memories**,

    to appear on *Proc. Design, Automation & Test in Europe (DATE) Conf.*, Dresden (Germany), March 2012

# Other publications

23. C.Zambelli, A.Chimenton, and P.Olivo,

    "**Statistical Modeling of Low-Probability Events in NAND/NOR Flash**

**and Phase Change Memory arrays**,

*2nd International Workshop on simulation and Modeling of Memory devices*, October 2011.

24. C.Zambelli, A.Chimenton, and P.Olivo,

    "**Reliability characterization of ATHENIS non volatile memory modules**,

    *ATHENIS dissemination workshop*, December 2010.

25. A.Chimenton, P.Olivo, and C.Zambelli,

    "**Reliability: statistical approach**,

    *Maratona delle Memorie conference*, September 2010.

26. C.Zambelli and D.Bertozzi,

    "**Performing Audio Processing by mean of XC161CJ/CS**,

    *Infineon Application Note Database*, December 2006.

# Acknowledgments

Being at the end of this journey it is time to acknowledge all the people close to me during this time. These three years have represented a wonderful experience both from professional and human perspective helping me to be more conscious of the potentiality that the academic research has to offer. Moreover i learned how to be open to the discussion of the research results and how to relate with other scientists in the electron devices field.

First of all i would like to thank my advisor Prof. Piero Olivo. He has been in charge of my academic education since the B.Sc. days. I am really honored to have worked under his guidance. His experience on the electron devices world has pushed me everyday to deepen my knowledge aspiring to reach his level. He has supported me on any situation, both inside and outside the university and i am really glad to say that i found more than a scientific advisor: above all i found a good friend.

The second person i would like to thank is my former advisor Dr. Andrea Chimenton. He friendly introduced me to the Ph.D. starting from the M.Sc. thesis guiding my researches on the first year of this path. He helped me to move my first steps into the academic research world and i am very grateful to him. Together we worked endless nights striving for the results facing a lot of technical difficulties,

but in the end we made it.

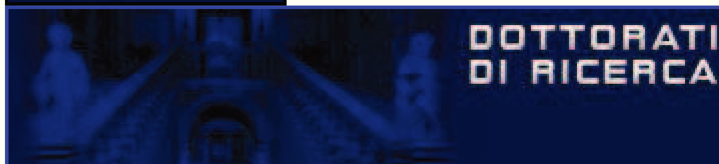
I would also like to thank all the good people found during my research visits: the SimpleEE team at Austriamicrosystems in Unterpremstatten (Austria), the reliability team of MASER Engineering in Enschede (The Netherlands), the NanoMEMS team of Cavendish Kinetics in 's-Hertogenbosch (The Netherlands) and the automotive Flash department at Infineon Technologies (Germany). I learned a lot during those periods spent away from home.

A special acknowledgement goes to all the guys at ActiveTechnologies in Ferrara for the continuous technical support in setting up the measurements leading to the results of this thesis. Thanks in particular to Ing. Michele Ramponi and Dr. Paolo Pellati for their friendship.

Thanks also to all the people at the Dipartimento di Ingegneria in Ferrara who supported me every day even when i was on a bad mood and in particular thanks to the administrative staff (Velia Margutti, Maria Rita Ferrari and Caterina Buosi) for the constant consults when i had to go abroad.

Most of all i would like to thank my parents Marzia, Marco and all my family. I started long time ago to study and now i finally concluded my formation. I hope that you would be proud of what i have done during these long years as i am of you. You have always been a constant in my life and i really could have done nothing without your unconditioned love.

Finally i would like to thank Margherita. No words can be used to explain the love for you and what you represent for me. You are my life...




Il tuo indirizzo e-mail
cristian.zambelli@unife.it
Oggetto:
Dichiarazione di conformità della tesi di Dottorato
Io sottoscritto Dott. (Cognome e Nome)
Zambelli Cristian
Nato a:
Copparo
Provincia:
Ferrara
Il giorno:
05/11/1983
Avendo frequentato il Dottorato di Ricerca in:
Scienze dell'Ingegneria
Ciclo di Dottorato
24
Titolo della tesi (in lingua italiana):
Electrical Characterization, Physics, Modeling and Reliability of Innovative Non-Volatile Memories
Titolo della tesi (in lingua inglese):
Electrical Characterization, Physics, Modeling and Reliability of Innovative Non-Volatile Memories
Tutore: Prof. (Cognome e Nome)
Olivo Piero
Settore Scientifico Disciplinare (S.S.D.)
ING-INF/01
Parole chiave della tesi (max 10):
Memory, Characterization, Physics, Modeling, Reliability
Consapevole, dichiara
CONSAPEVOLE: (1) del fatto che in caso di dichiarazioni mendaci, oltre alle sanzioni previste dal codice penale e dalle Leggi speciali per l'ipotesi di falsità in atti ed uso di atti falsi, decade fin dall'inizio e senza necessità di alcuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni; (2) dell'obbligo per l'Università di provvedere al deposito di legge delle tesi di dottorato al fine di assicurarne la conservazione e la consultabilità da parte di terzi; (3) della procedura adottata dall'Università di Ferrara ove si richiede che la tesi sia consegnata dal dottorando in 4 copie di cui una in formato cartaceo e tre in formato pdf, non modificabile su idonei supporti (CD-ROM, DVD) secondo le istruzioni pubblicate sul sito: http://www.unife.it/studenti/dottorato alla voce ESAME FINALE – disposizioni e modulistica; (4) del fatto che l'Università sulla base dei dati forniti, archivierà e renderà consultabile in rete il testo completo della tesi di dottorato di cui alla presente dichiarazione attraverso l'Archivio istituzionale ad accesso aperto "EPRINTS.unife.it" oltre che attraverso i Cataloghi delle Biblioteche Nazionali Centrali di Roma e Firenze; DICHIARO SOTTO LA MIA RESPONSABILITA': (1) che la copia della tesi depositata presso l'Università di Ferrara in formato cartaceo, è del tutto identica a quelle presentate in formato elettronico (CD-ROM, DVD), a quelle da inviare ai Commissari di esame finale

e alla copia che produrrò in seduta d'esame finale. Di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi; (2) di prendere atto che la tesi in formato cartaceo è l'unica alla quale farà riferimento l'Università per rilasciare, a mia richiesta, la dichiarazione di conformità di eventuali copie; (3) che il contenuto e l'organizzazione della tesi è opera originale da me realizzata e non compromette in alcun modo i diritti di terzi, ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura civile, amministrativa o penale e sarà da me tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi; (4) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati, non è oggetto di eventuali registrazioni di tipo brevettale o di tutela. PER ACCETTAZIONE DI QUANTO SOPRA RIPORTATO

Firma del dottorando

Ferrara, li __________16/01/2012_________________________ (data) Firma del Dottorando ____________________________________________

Firma del Tutore

Visto: Il Tutore Prof. Piero Olivo Si approva Firma del Tutore ____________________________________________________