Politecnico di Milano
Dipartimento di Elettronica e Informazione
Doctoral Program In Information Technology

Investigation of the emerging physical mechanisms limiting the reliability of nanoscale Flash memories

Doctoral Dissertation of:
Carmine Miccoli

Supervisor:
Prof. Alessandro S. Spinelli

Tutor:
Prof. Angelo Geraci

The Chair of the Doctoral Program:
Prof. Carlo Fiorini

2012 – XXV

Preface

This Doctoral Dissertation concludes a period of more than four years that began when I joined the *Micro and Nanoelectronics Lab* as undergraduate and continued when I enrolled in the Ph.D. program in Information Technology at the *Dipartimento di Elettronica e Informazione* (DEI) of the *Politecnico di Milano*. I would like to thank the people I worked with: my advisor Prof. Alessandro Spinelli and Prof. Andrea L. Lacaita, for giving me the opportunity to join the *NanoLab* and for many fruitful discussions, Dr. Christian Monzio Compagnoni, who guided me trough my research activity and chose me as Teaching Assistant for the Nanoelectroncs course, Prof. Daniele Ielmini, Dr. Andrea Bonfanti, Dr. Guido Zambra, my Ph.D. *colleagues* Alessandro, Davide, Federico, Giovanni, Mattia, Michele, Niccolò, Nicola, Salvatore, Simone B., Simone L. and Ugo and all the undergraduate students who worked with us in the last years, with a special acknowledgement to Riccardo Mottadelli, Luca Crespi, Andrea Bertelè, Giovanni Paolucci and Luca Pasini for their valuable contribution to my research activities.

My research activities on Flash reliability would not have been possible without the close collaboration with *Micron* (former *Numonyx*) Process R&D, Agrate Brianza (MI): I wish to thank Angelo Visconti, Silvia Beltrami, Luca Chiavarone and Marcello Calabrese for their assistance and contribution, Paolo Fantini and Alessio Spessot for the helpful discussions and suggestions.

During the last year of my Ph.D. program I spent seven month with Micron Technology R&D Flash development in Boise (USA): this was a very intense, challenging and fruitful experience, thanks, in primis, to my supervisor Akira Goda. I have to thank all the Micron team members I worked with for their patience and assistance, in particular Jeff Kessenich, John Barber and Terri Beaubien for their contribution to the experimental work, Karhtik Sarpatwari and Randy Koval for the helpful discussions and Giusy and Andrea for their friendly assistance. I have also to acknowledge Mandy Kayelr and Kelly Knudesn for guiding me trough the bureaucracy and the paperworks I had to carry out. Moreover, my stay in Boise was warmed by the friendship of the local *Italian community*; I'm particular

grateful to Ugo, Ale and Matilde, who welcomed me in Boise and assisted me.

I also want to thank the reviewer of this thesis Prof. Piero Olivo for the careful reading and the useful suggestions.

Finally I want to acknowledge my parents and Ivana, who supported me in the most intense moments and shared with me the happiness for my achievements.

January MMXIII,

Carmine Miccoli

Abstract

Flash memory is today the leading solid-state non-volatile memory technology, allowing high integration density, low costs and good reliability. The continuous scaling has been the main driver of the success of this technology, pushing it, however, to its physical limits: the reduction of the array pitch is today limited by the increasing capacitive coupling among adjacent cells, the low number of electrons controlling cell state is raising issues related to the discrete nature of the charge flux from/to the floating gate, single electrons stored in the tunnel oxide result into more and more severe threshold voltage instabilities during read and data retention. Aim of this thesis is to study the emerging physical mechanisms limiting the reliability of ultra-scaled Flash memories, highlighting from a theoretical standpoint the fundamental limitations to the functionality of nanoscale memory arrays. All the work has been carried out with a scaling perspective, trying to assess the ultimate scaling limitations and to propose feasible solutions able to extend the success of the Flash technology to the future technology nodes. A particular attention, moreover, has been devoted to then analysis and assessment of qualifications schemes for ultra-scaled Flash arrays.

In Chapter 1, the fundamentals of Flash memory are presented and the major reliability issues of Flash technology are discussed in a scaling perspective. Finally, the future trends of Flash technology are considered, discussing possible evolutionary approaches which may be able to bring Flash into the Terabit regime and concluding with a glance at non-charge based technologies for innovative memory architectures.

A detailed investigation of threshold voltage instabilities after cycling on nanoscale NAND Flash is presented in Chapter 2, highlighting their dependence not only on the bake but also on the cycling conditions, in terms of temperature and time. Experimental characterizations and data analyses assess the validity of a universal model for damage recovery after distributed cycling, allowing the development of accelerated test schemes reproducing the real on-field usage of the devices.

Chapter 3 presents a detailed compact-modeling investigation of the string

current in decananometer NAND Flash arrays, enabling the assessment of the role of short channel effects on the string current. Thanks to the model, moreover, the basic properties of damage creation and recovery during cycling and bake, respectively, are studied in nanoscale Flash arrays, providing insight into the time dynamics and activation energies of charge trapping/detrapping and interface states creation/annealing occurring in the tunnel oxide.

Then, Chapter 4 is devoted to the investigation of the validity of the distributed-cycling schemes on scaled NOR Flash memory technologies. In particular, long gate-stresses required to gather the array threshold voltage (V_T) map are shown to give rise to parasitic V_T-drifts, which add to the V_T-loss coming from damage recovery during post-cycling bake. When the superposition of the two phenomena is taken into account, the effectiveness of the conventional qualification schemes is fully confirmed.

Chapter 5 is focused on the analysis of fundamental variability sources and emerging constraints to Flash reliability due to few electron phenomena. In particular, the discrete nature of the electron flow into the floating-gate during cell programming is discussed and design solutions are proposed to mitigate this issue, addressing the electron-injection spread scaling trend. Discrete electron emission from the floating gate or from the tunnel oxide in fresh and cycled arrays, respectively, are also considered during data retention, investigating the variability contributions for data retention due to charge granularity and cell parameters fluctuations.

Finally, in Chapter 6, the granular nature of the current flow to the floating gate, discussed in Chapter 5, is investigated in details as a fundamental source of programming noise during incremental step pulse programming (ISPP) of NAND arrays, studying the statistical dispersion of the programmed threshold voltage distribution. Using Monte Carlo numerical simulations, the possibility to increase programming accuracy by means of optimized double verify ISSP algorithms is considered, highlighting benefits and drawbacks of the discussed algorithms.

Riassunto

Le memorie Flash ad oggi costituiscono la principale tecnologia per dispositivi di memoria non volatili a stato solido, in quando permettono di ottenere una elevata densità di integrazione, bassi costi e una buona affidabilità. Il progressivo processo di scaling è stato il principale artefice del successo di tale tecnologia, spingendola, tuttavia, ai suoi limiti fisici: la riduzione delle dimensioni della matrice di memoria è ormai limitata dall'aumento degli accoppiamenti capacitivi tra celle adiacenti, il ridotto numero di elettroni che controllano lo stato della cella determina in maniera sempre più seria l'emergere di problematiche correlate alla natura discreta del flusso di carica da e verso la floating gate, singoli elettroni intrappolati nell'ossido di tunnel determinano instabilità di soglia in ritenzione sempre più gravi. Lo scopo di questa tesi è lo studio dei meccanismi fisici emergenti che limitano l'affidabilità di memorie Flash ultra scalate, evidenziano, da un punto di vista teorico, le limitazioni fondamentali alla funzionalità di matrici di memoria decananometriche. L'intero lavoro è stato condotto in una prospettiva di scaling, cercando di stabilire i limiti ultimi alla riduzione delle dimensioni dei dispositivi e di proporre delle soluzioni fattibili, in grado di estendere il successo della tecnologia Flash anche ai futuri nodi tecnologici. Particolare attenzione, inoltre, è stata rivolta allo studio e all'analisi delle metodologie di qualifica da adottare per dispositivi Flash ultra-scalati.

Nel Capitolo 1 sono presentati i principi di funzionamento delle memorie Flash e si discutono i principali problemi affidabilistici della tecnologia Flash, in una prospettiva di scaling. Alla fine, le prospettive future della tecnologia Flash sono prese in considerazione, discutendo i possibili approcci evolutivi che potrebbero essere in grado di portare le memorie Flash a raggiungere capacità dell'ordine del Terabit, concludendo con uno sguardo a tecnologie non basate sull'immagazzinamento di carica per architetture di memoria innovative.

Il Capitolo 2 presenta una investigazione dettagliata delle instabilità di soglia dopo ciclatura in dispositivi NAND Flash decananometrici, evidenziando la dipendenza non solo dalle condizioni di ritenzione ma anche da quelle di ciclatura, in termini di temperatura e durata. Le attività di caratterizzazione sperimentale

e analisi dei dati hanno dimostrato la validità di un modello universale per il recupero del danno dopo ciclatura distribuita, consentendo lo sviluppo di metodologie di qualifica accelerate, tali da riprodurre il reale utilizzo sul campo del dispositivo.

Il Capitolo 3 si occupa di una dettagliata attività di modellistica compatta della corrente di stringa in dispositivi NAND Flash scalati, permettendo di determinare il ruolo degli effetti di canale corto sulla corrente di stringa. Grazie a tale modello, inoltre, sono state studiate le proprietà fondamentali della creazione del danno durante ciclatura e del conseguente recupero durante ritenzione, fornendo una più approfondita comprensione delle dinamiche temporali e delle energie di attivazione dei fenomeni di intrappolamento/rilascio di carica e di creazione/annealing degli stati interfacciali che interessano l'ossido di tunnel.

Inoltre, il Capitolo 4 è dedicato all'investigazione della validità degli schemi di ciclatura distribuita su memorie NOR Flash scalate. In particolare, si mostra che i lunghi gate-stress richiesti per acquisire la mappa di tensione di soglia dell'intera matrice possono dar luogo a una deriva parassita della tensione di soglia, che va ad aggiungersi alla perdita di soglia derivante dal recupero del danno dopo ciclatura. Se la sovrapposizione dei due fenomeni è presa correttamente in considerazione, tuttavia, è possibile confermare l'efficacia dei convenzionali schemi di ciclatura distribuita.

Il Capitolo 5 è focalizzato sull'analisi delle fonti fondamentali di variabilità e sui vincoli affidabilistici emergenti che derivano da fenomeni di singolo elettrone. In particolare, la natura discreta del flusso di carica verso la floating gate durante l'operazione di programmazione è presa in analisi e sono proposte soluzioni per mitigare tale problematica, con riferimento alle proiezioni di scaling per la statistica di iniezione. Inoltre, anche il processo di emissione discreta di elettroni durante ritenzione dall'ossido di tunnelo dalla floating gate (nel caso di matrici di memoria ciclate o non ciclate) è analizzato dettagliatamente e sono investigati i contributi alla variabilità della ritenzione del dato derivanti dalla granulrità di carica e dalle fluttuazioni dei parametri di cella.

In conclusione, nel Capitolo 6, la natura granulare del flusso di carica verso la floating gate, discussa nel Capitolo 5, è presa in considerazione in quanto fonte di rumore di programmazione durante algoritmi di programmazione a rampa (ISPP) di matrici di memoria NAND, studiando la dispersione statistica della tensione di soglia programmata. Mediante simulazioni numeriche di tipo Monte Carlo, la possibilità di aumentare l'accuratezza di programmazione grazie ad algoritmi ottimizzati con *double verify* (DV-ISPP) è stata investigata, evidenziando i benefici e gli inconvenienti di tali algoritmi.

Contents

CHAPTER 1

Introduction to Flash memory technology

In this chapter the fundamentals of Flash memory are presented, describing the memory cell structure, the array architectures and the device operation. Then, the major reliability issues of Flash technology are discussed, with a particular attention to the fundamental limits of the technology scaling, which will be carefully assessed in the next Chapters of this thesis. Finally, the future trends of Flash technology are considered, discussing possible evolutionary approaches which may be able to bring Flash into the Terabit regime and concluding with a glance at non-charge based technologies for innovative memory architectures.

1.1 Flash memories: a history of success

In the past decade, Flash technology emerged as the most successful non-volatile solid-state memory solution; in particular, its success was led by the impressive growth of floating gate NAND Flash technology, whose revenues have grown in a decade at a composite annual grow rate (CAGR) of almost 50%, while in the same period the total revenues of semiconductor memories have declined 8.7% (see Fig. 1.1) [1]. The success of Flash technology is closely linked to the expansion the consumer electronics market has experienced in the past years: on one hand, the constantly increasing demand of non-volatile storage capability of medium-capacity, with small dimensions, low power consumption, and high reliability has driven the Flash market explosion while, on the other hand, the availability of high-performance and low-cost NAND Flash memories has enabled new applications and products. As a result of this virtuous circle, nowadays Flash technology is widespread, providing memory storage to a wide selection of portable personal devices, such as smartphones, tablets, hand-held game consoles, e-book readers,

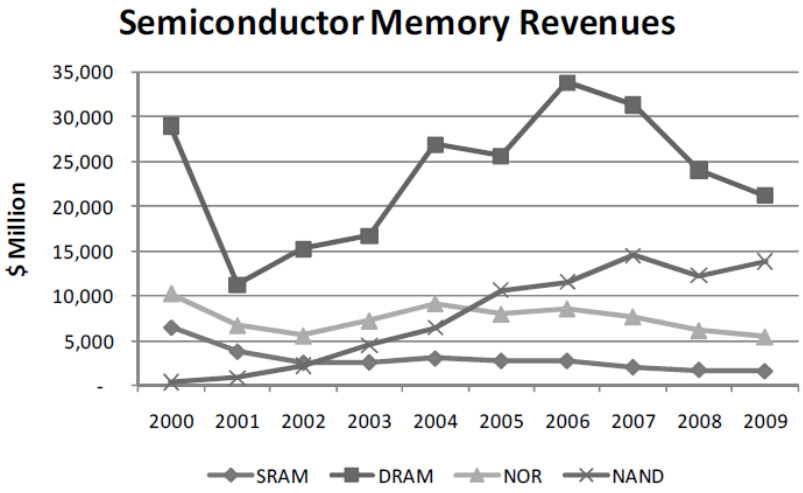
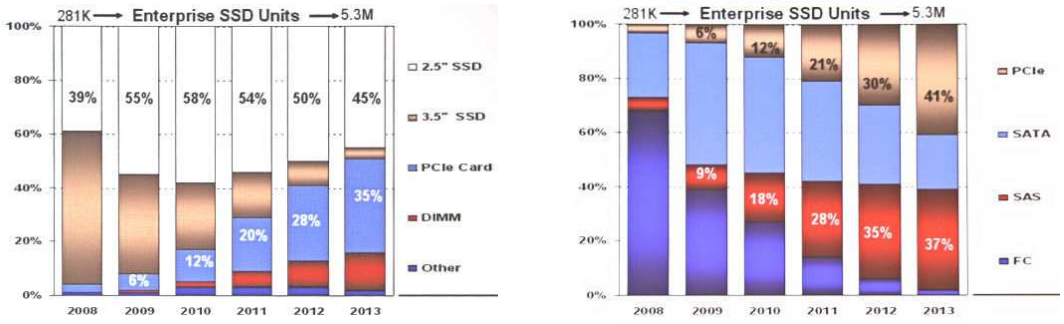


Figure 1.1: *Last decade semiconductor memory market (source: WSTS, Forward Insight, from [1]).*



(a) (b)

Figure 1.2: *SSD distribution by (a) form factor; and (b) interface. SSDs are moving from HDD replacement applications to new form factors and faster interfaces able to better exploits SSD superior performance in the enterprise market (e.g., in server applications) (source: Gartner).*

music player and digital cameras. Moreover, the introduction of high-capacity solid state drives (SSDs) is forecast to further enhance the success of Flash memories: even if the cost per GB is still higher than traditional hard disk drives (HDDs), SSDs have been replacing HDDs in high-end laptops and in the new product category of *ultrabooks* thanks to their higher performance, their lower power consumption and their lower susceptibility to mechanical damages. These features make SSDs suitable not only for consumer products, but also for enterprise solutions; Figs. 1.2(a)-1.2(b) forecast a growth of SSDs for enterprise-class storage systems, showing that they will move from HDD replacement applications to new form factors and faster interfaces, in order to address the needs of the server market. This suggests that Flash memories will play a key role also in the future development of internet-based *cloud* services.

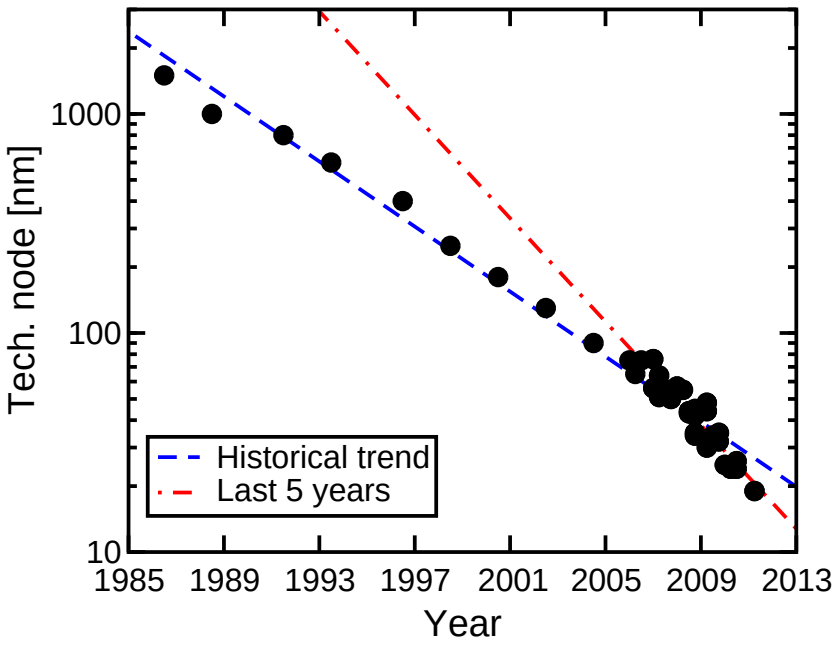


Figure 1.3: *Flash memory scaling time-line; the scaling trend is also shown comparing the historical one, that predicts the feature size to halve every 4 years, with the one of the last 5 years, during which the technology node halved every 2 years and a half (from [2]).*

The reason of this success can be found in the uninterrupted reduction of the memory cell feature size which, since the first proposal of Flash technology in 1987 [3], has led to a continuous increase of the integration density, making NAND Flash the chips with the highest number of integrated devices [4, 5] and proportionally reducing the cost per bit. Fig. 1.3 shows that the Flash feature size halved every 4 years, with an increase of the scaling rate in the last 5 years which is likely connected to the stronger competition in the Flash market (IM Flash Technology, a joint venture between Intel Corporation and Micron Technology was created in January 2006). This scaling trend has been enabled by several efforts in different fields, from the manufacturing technologies (e.g., improvements in lithography techniques, the introduction of innovative self-aligned technologies, the increase of wafer size, from 150 mm in 1987, to 300 mm in recent years) to the memory array architecture (the introduction of the NAND architecture allowed to obtain a cell area of $4F^2$, where F is the technology node), all aiming at the memory cost reduction [6].

This cell miniaturization process leads to unquestionable advantages in terms of cost reduction and integration density increase but it pushed the technology close to its physical limits, giving rise to a major drawback in terms of increased design complexity; as the cell size is reduced, in fact, not only the existent reliability issues get worse, setting even more stringent constraints on memory cell operation, but also new physical phenomena appear, potentially compromising the array functionality. At the decananometer scale, the discreteness of the matter and the charge granularity emerge as fundamental limitations and the closer packing of the cells into the array gives rise to undesired interaction between adjacent

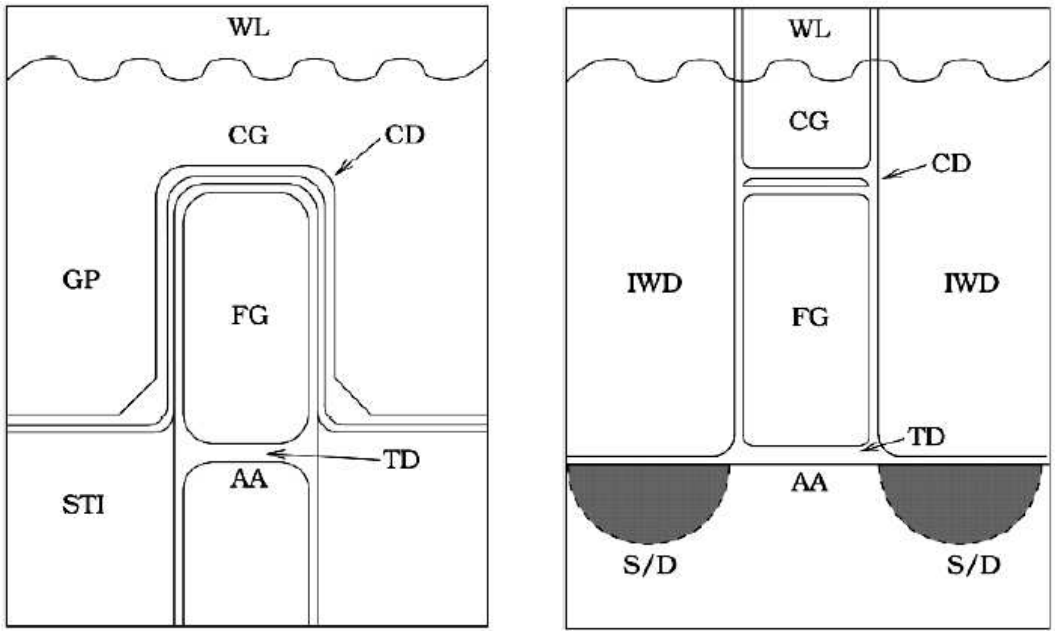


Figure 1.4: *Schematic cross sections of a floating-gate memory cell, along the direction (a) parallel to the word line and (b) parallel to the bit line (from [1]).*

cells, due to electrostatic interference [7]. This thesis work, in particular, aims at a detailed investigation of the emerging physical mechanisms limiting nanoscale Flash reliability.

Several solution have been proposed to manage these emerging reliability issues, including improvements in memory cell design, the introduction of new materials for the storage layer or the insulation dielectrics and advanced algorithms for the programming and reading operations; however, it still under debate if the conventional NAND technology will be able to scale down the 10 nm technology node. In this regards, two kind of approaches have been proposed to overcome the scaling limitations of Flash memories, the first of one relies on the evolution of the existing technology toward innovative architectures, such as three-dimensional integration of the NAND array. The second path, in turn, leads to completely new memory technologies, which are no longer based on the storage of electric charge and which can be arranged into new and more effective architectures. The following sections will briefly describe the conventional Flash memory cell, its working principles and the main scaling limitations, discussing the solutions proposed to overcome them in the frame of an evolutionary scenario, giving also a glance at the possible innovative, non-charge based solutions.

		NOR	NAND
Type of access		Random	Serial
Access time	Random:	$60 - 120$ ns	$60 - 120$ μs
	Page mode:	30 ns	$25 - 50$ ns
	Burst mode:	15 ns	—
Write speed	Random:	10 μs/byte	200 μs/byte
	Page mode:	—	200 μs/page (0.4 μs/byte)

Table 1.1: *Performance comparison between NOR and NAND Flash memories (source: Forward Insights).*

1.2 Flash Technology

1.2.1 Flash memory cell

The great success of the Flash memory, which nowadays is the leading non-volatile solid state memory technology, has been enabled by the versatility of the floating-gate (FG) transistor, which has shown to be an easily manufacturable and highly scalable memory cell; the realization process of a conventional Flash cell, in fact, is fully compatible with the CMOS process, only using standard materials and lithography. The FG transistor was initially proposed by D: Kahng and S. Sze at Bell Labs and has gone through several improvement and evolutions, becoming the Self-Aligned Shallow Trench Isolaction (SA-STI) cell which is commonly adopted in conventional Flash technologies and which is schematically depicted in Figs. 1.4(a)-1.4(b) [8, 9].

The FG transistor differs from the conventional MOSFET device for the presence of a polysilicon conductive layer interposed between the transistor active area (AA) and the control gate (CG) and which is completely surrounded by dielectric layers, hence the name of floating gate. By injecting/removing charge into/from the FG, the cell threshold voltage V_T can be modified, allowing the data storage. In so doing, the information is coded thanks to the charge stored in the FG, by a physical point of view and, thus, thanks to a controlled shift of the cell V_T value, from an electrical point of view; moreover, the possibility to achieve a fine V_T tuning allows to store more than one bit in a single cell, increasing the storage density, as it will discussed later in details. The electrical insulation provided by the dielectric layers enables the retention of the stored information (charge) for years, without any power supply. As shown in Figs. 1.4(a)-1.4(b), the insulation between the FG and the AA is provided by a tunnel oxide, or tunnel dielectric (TD) layer with thickness in the 7 to 8 nm range, while a coupling dielectric (CD), which in conventional devices is called inter-poly dielectric (IPD) and consists in an oxide-nitride-oxide (ONO) stack having an equivalent oxide thickness between 11 and 15 nm, isolates the FG from the highly-doped poly silicon word line [10–13]. Moreover, the separation between adjacent AAs is obtained thanks to the STI and the Inter Word line Dielectric (IWD) isolates the FGs from each others. In addition to that, the peculiar shape of CG, wrapping around FGs with its gate plugs (GP), as shown in Fig. 1.4(a), helps to achieve a better electrostatic shielding, as well as a higher electrostatic coupling between CG and FG.

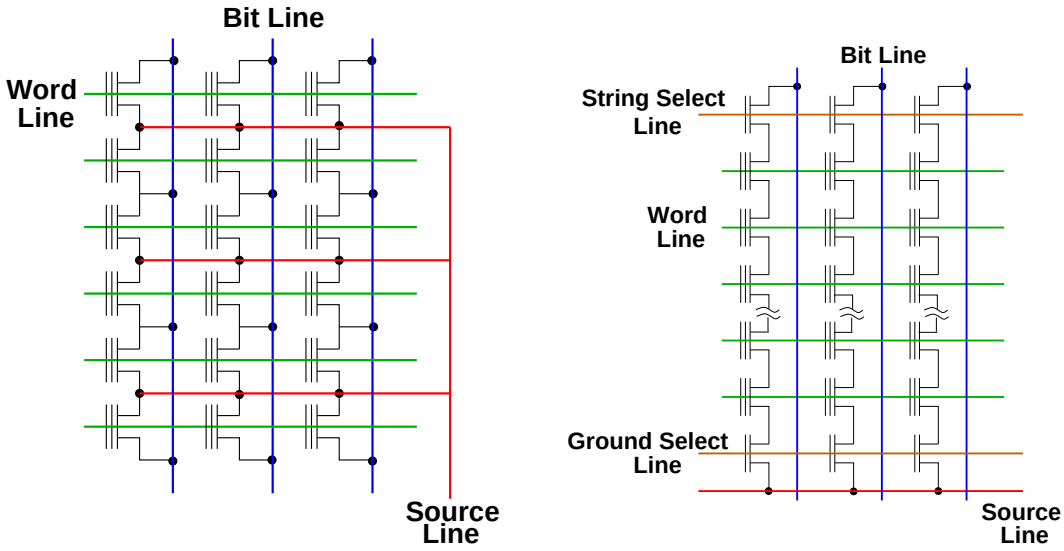
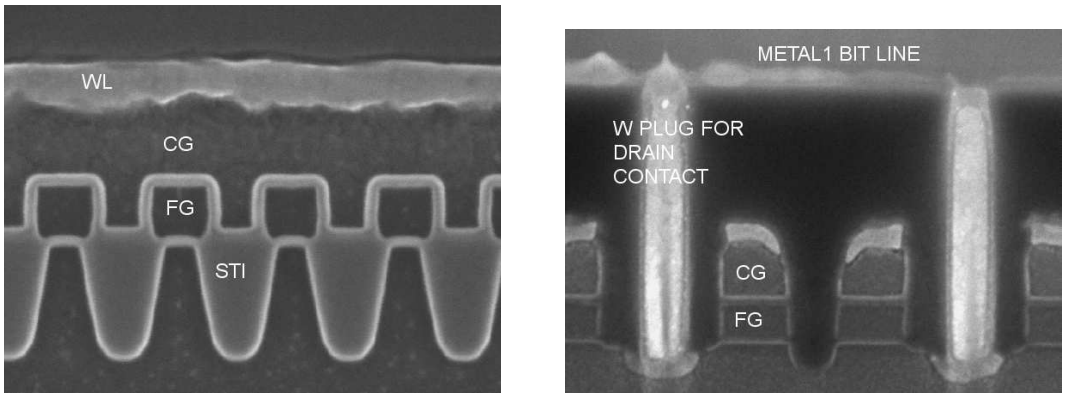


Figure 1.5: *Schematic representation of (a) NOR and (b) NAND arrays.*

1.2.2 Flash memory architectures

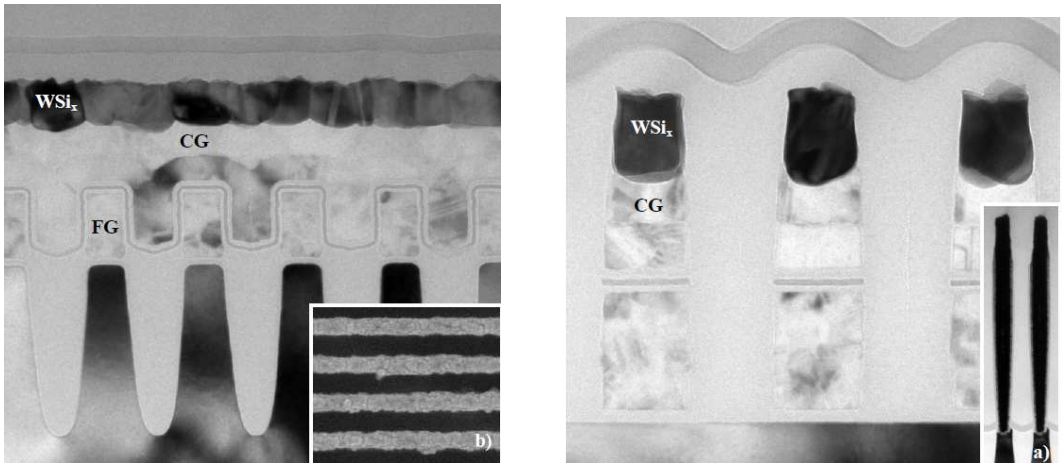
In order to achieve the desired memory capacity, a large number of cells can be closely packed into a two dimensional array. Fig. 1.5 shows the two main types of Flash architectures: NOR and NAND arrays, which are named after the resemblance to the pull-down networks in the corresponding CMOS logic gates. In the NOR array, cells are organized in a parallel architecture, aiming at a direct access of each cell, which is connected to its bit line through a dedicated drain contact. The cell direct access is the reason of the NOR superiority in the random access performance, as shown by the performance comparison between NOR and NAND in Table 1.1. In NAND array, on the contrary, memory cells are connected in series and organized into memory strings, trading off read/write random speed for an higher integration density and, thus, a lower cost per bit [16–18]. Figs 1.6(a)-1.6(b) show the cross sections of a NOR array along the bit line and the word line directions, while Figs 1.7(a)-1.7(b) refer to a NAND array: it is clear that the drain contact which is required for each NOR cell make the cell area occupation larger ($10F^2$, where F is the feature size) with respect to the NAND array, whose more compact layout allows to obtain a $4F^2$ cell area.

Flash memory cells can be programmed exploiting Channel Hot Electrons injection (CHE) or Folwer-Nordheim tunneling (FN) in order to inject electrons into the FG, thus increasing cell V_T by the desired amount [19]. NOR Flash exploits the former mechanisms (see 1.8(a)), which requires to apply a positive drain bias allowing the electrons to gain the energy required to be injected into the FG, thanks to the positive CG bias. NAND Flash, however, cannot use this mechanism due to the lack of drain contact and so they exploits FN tunneling from the channel to the FG (see 1.8(a)): this mechanism, on one hand, is slower than CHE, resulting in a slower single bit programming but, on the other hand,



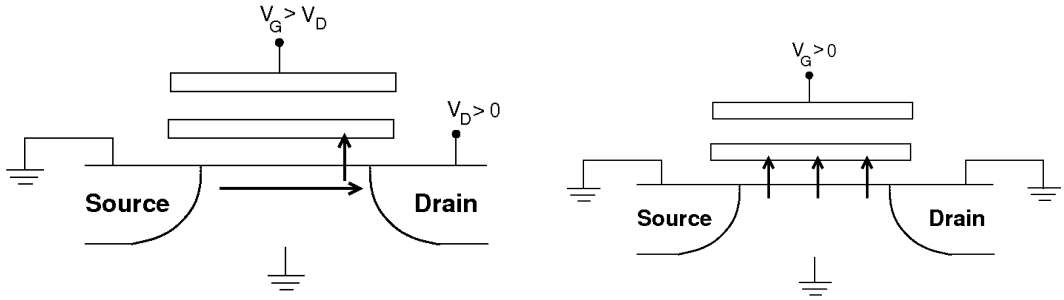
(a) (b)

Figure 1.6: *Cross sections of a 65 nm NOR Flash parallel to (a) the word line and (b) to the bit line direction (from [14]).*



(a) (b)

Figure 1.7: *Cross sections of a 8−Gb MLC NAND Flash parallel to (a) the word line and (b) to the bit line direction (from [15]).*



(a) (b)

Figure 1.8: *Schematic representation of (a) CHE and (b) FN programming.*

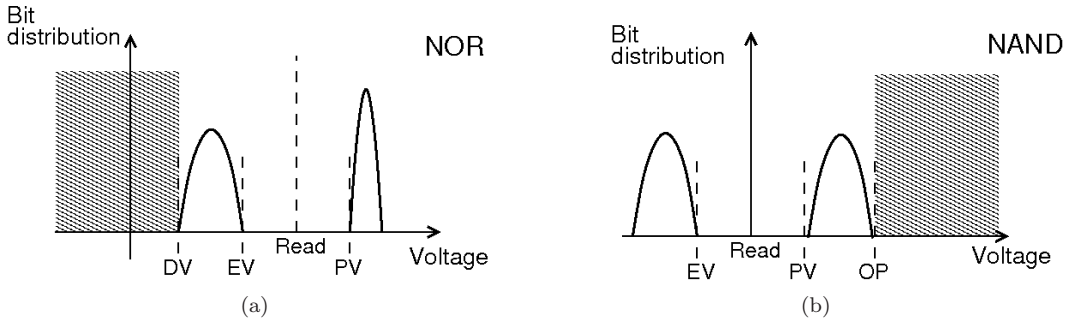
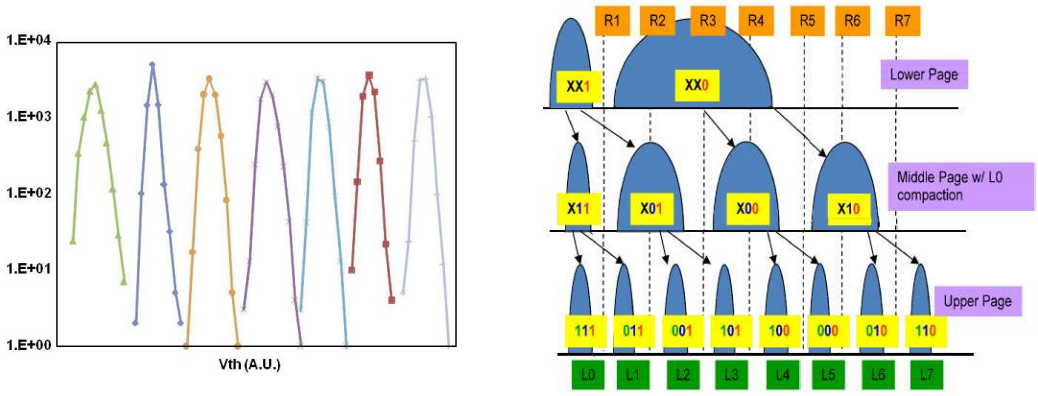


Figure 1.9: *Threshold voltage distribution for the programmed and erased cells in a (a) NOR and (b) NAND array architecture. Shaded areas must be avoided, as they give rise to incorrect readings. EV=Erase Verify level, PV=Program Verify level, OP=Over Programming level, DV=Depletion Verify level.*

it involves no current flow in the channel and so the lower programming current (flowing only through the gate) and the lower power consumption allow the parallel programming of several cells in the same array and largely enhances the overall write throughput with respect to NOR devices (see Table 1.1). In addition to that, NAND cell operation does not require the application of large biases to the drain, relieving the channel length scaling of the constraints given by drain-induced barrier lowering and punch-through. For this reason, the NAND cell not only is more easily scalable, but can also reach the minimum size obtainable with a given technology, with both active area length and width very close to the feature size, while the NOR cell has usually a channel length greater the the feature size, in order to mitigate the short channel effects. For the erase operation, instead, both NOR and NAND memories rely on FN tunneling from FG to the substrate, performed on a block basis. As a consequence of the different features and performance of these two architectures, NOR memories are optimized for performance code storage and execution, while NAND memories best address low cost mass storage applications, explaining the large difference in market revenues observed in Fig. 1.1.

To conclude this brief review of Flash operation, it should be pointed out that the differences between NOR and NAND architectures affect also the read operation and the V_T placement. In both cases, the reading mechanism consists in the application of a positive bias to the CG and to the bit line of the selected cell and in the sensing of the cell current, whose level gives a measure of the cell V_T and, hence, of the stored data. For the NOR cell, however, in order to guarantee a correct reading, all the cell must have their V_T greater than zero, avoiding a current leakage path when their CG is grounded. Fig. 1.9(a) schematically shows the V_T distributions in the programmed and in the erased state for a single level (SLC) NOR array: the erased V_T must be greater than the depletion verify level (DV) and lower than the erase verify level, while the programmed V_T must be only greater than the program verify level (PV), allowing to store one bit per cell and to determine the cell status by the comparison of cells V_T against one read level placed between EV and PV. In turn, the read operation for NAND cells is more



(a) (b)

Figure 1.10: *(a) Measured 8-level V_T distributions on a 20 nm NAND Flash memory (from [20]); and (b) schematic example of the programming algorithm adopted on a 25 nm TLC NAND Flash device: middle page programming and erase compaction are used in order to minimize impact of FG-FG coupling by reducing δV_T of neighboring cells (from [21]).*

complex, since it requires to sense the entire sting current: first of all, the memory string must be connected to the bit line and to the source line through the string select and the source select transistors (see Fig. 1.5) and all the unselected cells in the string must be turned on applying a V_{pass} bias to their CGs: due to this read scheme, on one hand cells V_T can be negative (the string is normally disconnected to the bit line and no current leakage can happen) but, on the other hand, it must not exceed the over programming level (OP), in order to turn on the unselected cells with the V_{pass} bias: Fig. 1.9(b) schematically depicted the V_T distributions for a SLC NAND device.

In order to increase the storage density, it is possible to increase the number of bit per cell, at the cost of a more complex programming algorithm required to obtain an accurate V_T placement (power supply and thus V_T window cannot be increased) and of a slower read operation (comparison against n read levels is required to determine the cell state in a n-bit per cell device). Nowadays, NOR and NAND multi level cell (MLC), storing 2 bits (4 V_T levels) and NAND triple-level cell (TLC), storing 3 bits/8 V_T levels, are widespread solutions: Fig. 1.10(a) shows the 7 V_T programmed distributions of a 20 nm TLC NAND Flash device [20]. Programming of MCL and TLC devices requires advanced programming schemes, like the one depicted in Fig. 1.10(b), which aims at an accurate 3 bit/cell V_T placement thanks to the minimization of the impact of FG-FG coupling obtained reducing ΔV_T of neighboring cells [21]. The physical constraints to V_T distribution width and the advanced programmed algorithms required in nanoscale MLC devices will be discussed in details in Chapter 6.

1.3 Reliability constraints to Flash operation

Conventional FG Flash memory cell has been continuously scaled down, up to the 25 nm technology node and beyond [5, 22]. The shrinking of cell size, how-

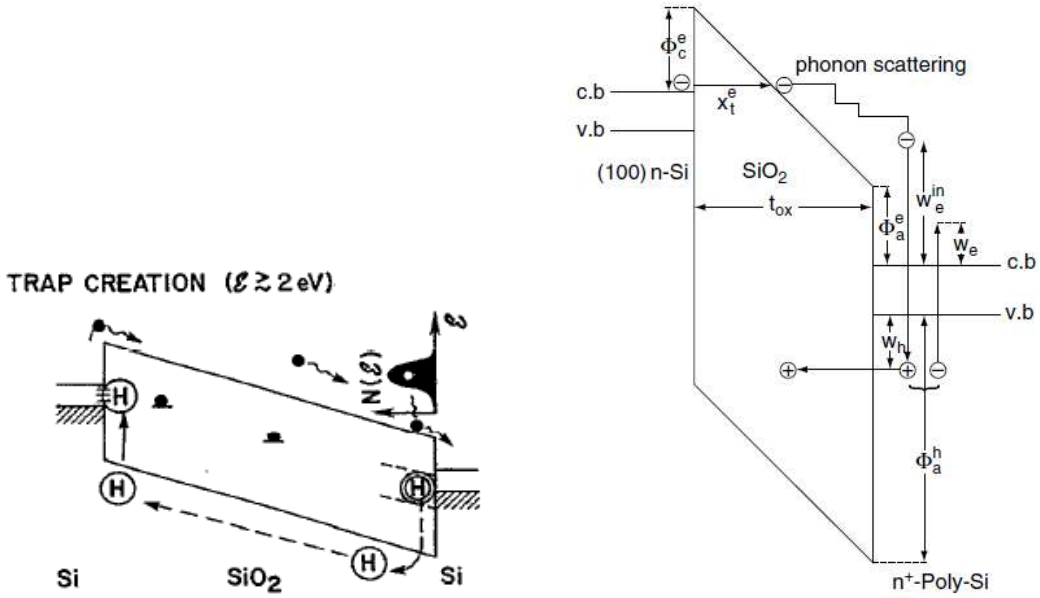
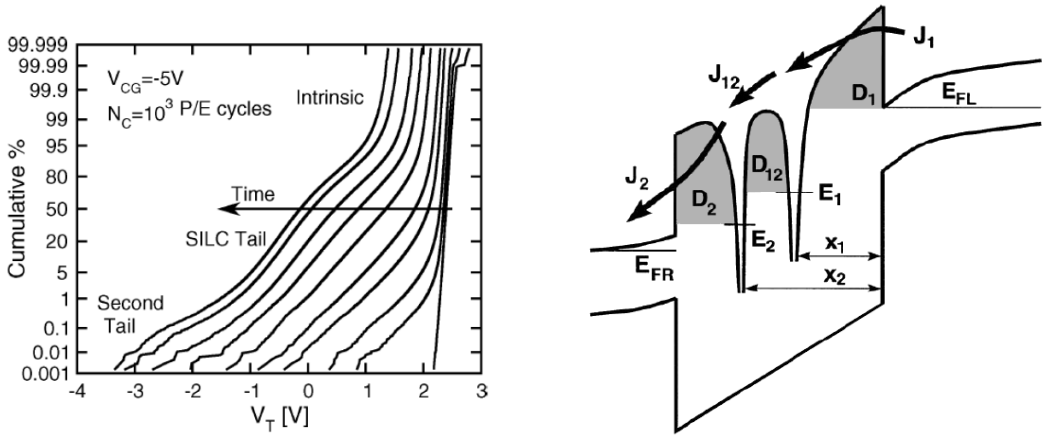


Figure 1.11: *Band diagrams of: (a) trap creation due to anode hydrogen release (from [23]); and (b) Anode Hole Injection (AHI) mechanism (from [24]).*

ever, results in larger impact of the existing reliability issues and forces to deal with new physical phenomena, related to the discrete nature of the matter and the charge flows. In addition to cell shrinking, also the cell packing into the array gets closer and closer, making the disturbs and the electrostatics interference between cells stronger. In the following paragraphs, a brief summary of the major reliability issues will be presented, while a detailed analysis of the emerging reliability constraints will be carried out throughout the entire thesis dissertation.

1.3.1 Cell-related reliability issues

Since the Flash memory cell relies on the dielectrics surrounding the FG in order to store the charge and provide the desired data retention, degradation and charge trapping which may occur bot in the tunnel oxide and in the IPD give rise to major issues for cell reliability and operation. High-field stress induced by FN tunneling during repeated program/erase (P/E) cycles, in fact, leads to the generations of defects in the tunnel oxide, due to charge trapping in the bulk oxide and interface state generations. Figs 1.11(a)-1.11(b) schematically show two possible oxide degradation mechanism: the first one consists in a trap creation near the cathode caused by mobile hydrogen release from sites near the anode, producing interface stares and oxide electron traps [23]; the second one is the Anode Hole Injection (AHI) mechanism, which induces interface states creation and charge trapping as well, as a consequence of impact ionization at the anode and injection of the secondary hole into the oxide. Stress-induced leakage current (SILC) [23, 26–53], fast moving and erratic bits [54], charge trapping/detrapping into/from the



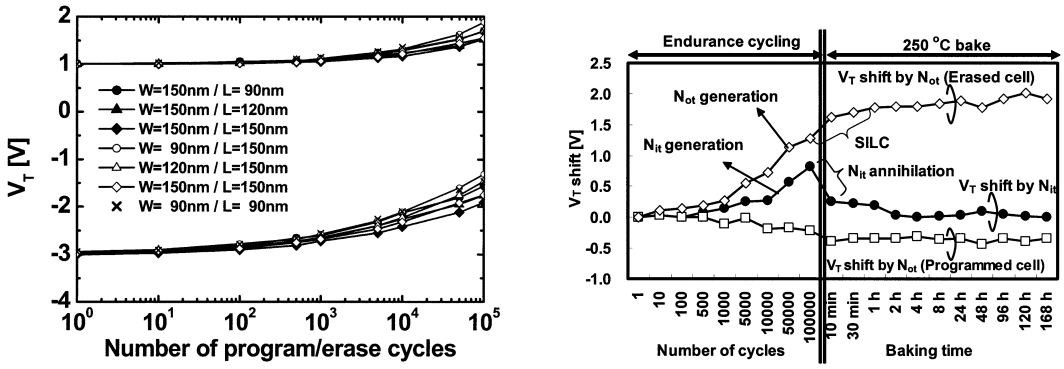
(a) (b)

Figure 1.12: *(a) Cumulative V_T distributions during an accelerated retention test with a negative bias applied at CG (from [25]); and (b) schematic representation of the 2TAT mechanism (from [26]).*

gate dielectrics and interface state generation/annealing [55–67], are well-known reliability issues for NAND Flash memories, related to oxide degradation and charge trapping phenomena.

The presence of trap defects into the oxide may give rise to a trap-assisted leakage path, resulting in a charge loss from FG due to Stress-Induced Leakage Current (SILC); Fig. 1.12(a) shows V_T distributions during a negative stress, revealing that, in addition to V_T loss be due to intrinsic charge loss from FG, a SILC tail and a second tail appear as a consequence of trap assisted tunneling (TAT) and 2TAT phenomena (a schematic band diagram representation of the latter one is given by Fig. 1.12(b)). Intrinsic charge loss and SILC set the main constrains to data retention of uncycled devices, limiting the tunnel oxide thickness scaling to $7 - 8$ nm [68]. Moreover, P/E cycling induced degradation limits the cell endurance, causing a significant V_T shift after about 10^5 P/E cycles and sets additional constraints to data retention due to charge detrapping from tunnel oxide and interface state annealing: Figs 1.13(a) and 1.13(b) show the P/E cycling characteristics and the bake V_T shift, decomposing the contributions due to SILC, charge detrapping and interface states creation/annealing. Reliability constraints due to post-cycling V_T instabilities will be discussed in details in Chapters 2, 3 and 4. In particular, charge trapping/detrapping phenomena emerged as a major issue for ultra-scaled Flash technologies since single electrons stored into the tunnel oxide result into more and more threshold voltage shift, due to the cell shrinking [5].

Besides these physical effects, decananometer technologies have been affected by new reliability issues due to emerging physical phenomena, among which random telegraph noise (RTN), fundamental variability sources due to the discrete nature of the substrate doping and the charge flow (discussed in Chapter 5) and cell-to-cell electrostatic interference appear as the most important to manage. Fig. 1.14(a) shows a typical drain current waveform of a Flash cell affected by



(a) (b)

Figure 1.13: *(a) Programming/erasing cycling characteristics of 16-cells string arrays of various cell sizes; and (b) evolution of threshold voltage shifts during programming/erasing cycling and data retention test, decomposed into oxide traps (N_{ot}) and interface traps (N_{it}) contribution (from [61]).*

RTN: the current exhibits discrete fluctuations between a high and a low current levels (corresponding to a low-V_T and a high-V_T state, respectively) as a consequence of an alternated capture/emission process of electrons into/from a trap close to the substrate Fermi level. RTN V_T fluctuations have a strong impact on the array functionality, resulting in fundamental source of V_T spread. Fig. 1.14(b) shows the cumulative distributions of ΔV_T between to consecutive read operation: the large number of cells in a high-density Flash array allows to investigate the statistical distribution of RTN amplitude, revealing that also very large V_T fluctuations (hence the definition of *giant* or *complex RTN* [71, 72]) can be detected, according to an exponential probability distribution. Moreover, the comparison between ΔV_T distributions obtained on different technology nodes reveals that the RTN amplitude is strongly increased by the cell shrinking; RTN fluctuations, thus, set an ultimate limit to the programming accuracy of Flash devices, adding a random widening to programmed V_T distributions and potentially compromising the cell verify operations required by the programming algorithms for an accurate V_T placement [73–75] For these reasons, RTN in Flash memory has been carefully investigated [69–74, 76–86], revealing that physical effects as the field intensifications at the active area corners [65, 70, 83], the atomistic nature of substrate doping [69, 87–93] and the spatial localization of trapped charges [69, 70, 78, 83, 94–96] must be taken into account in an accurate description of the RTN amplitude, especially in sub-100 nm technology nodes.

1.3.2 Array-related reliability issues

The cell scaling process results also in a more compact packing of the cells into the memory array, giving rise to new reliability issues due to cell-to-cell interaction. These array-related reliability issues include, on one hand, all the disturbs experienced during the array operations (program/read) by the memory cells which share the same terminals (word lines or bit lines); on the other hand, additional reliability constraints come from the electrostatic interference between adjacent

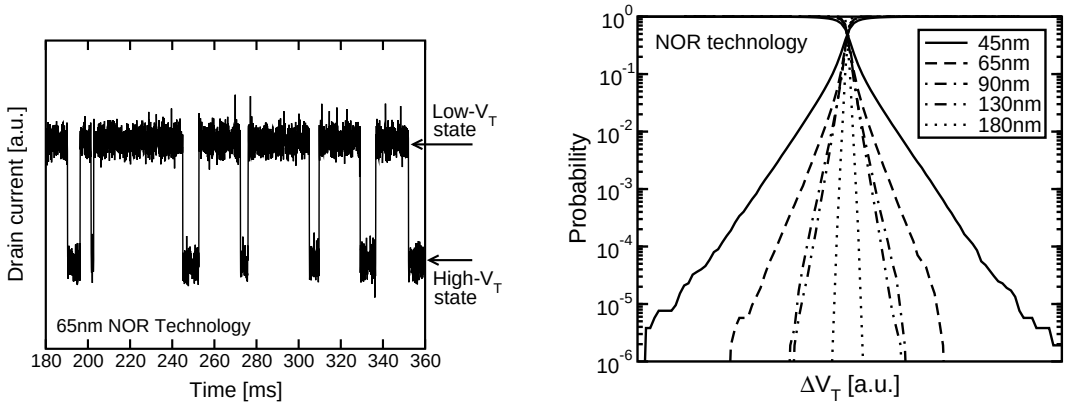


Figure 1.14: *(a) Single-trap RTN fluctuations measured on a selected cell out of a NOR Flash memory array in* 65 *nm technology (from [69]); and (b) cumulative distributions of RTN-induced* ΔV_T *for different NOR technologies ranging from* 180 *nm to* 45 *nm (from [70]).*

cells. All these mechanisms, however, result in an undesired change of the victim cell status due to the interaction with other cells in the array.

Among the former mechanisms, programming disturbs in NAND Flash arrays revealed to be strongly enhanced by the cell size reduction, due to short channel effects such as the gate-induced drain leakage (GIDL) and the drain-induced barrier lowering (DIBL). Fig. 1.15(a) shows the bias condition for NAND cell programming: in order to inhibit the programming of unselected cells, whose word lines are biased at the programming voltage V_{pgm}, a self-boosting method is used, raising the channel potential with the pass-voltage V_{pass} applied to the other cells in the string thanks to the capacitive coupling with the channel [99]. However, an excessively high pass-voltage may give rise to a V_{pass} disturbance to the cells in the same string of the programming cells, which is a soft programming mechanism by FN current with the V_pass voltage: thus, a trade-off must be obtained, choosing an intermediate V_pass value in order to avoid both V_{pgm} and V_{pass} disturbance [100, 101]. Novel kind of V_{pgm} disturbance, however, become a more sever issue with the technology node scaling, since new physical mechanisms can play a role during programming, depending on the cell biasing and the background pattern, as schematically depicted in Fig. 1.15(b): (1) GIDL at the source select transistor side may result in hot-carrier injection (HCI) into the adjacent FGs; (2) moreover, if the selected cell is programmed to the highest state (P3: $V_T \simeq 5$ V) an HCI effect by GIDL may happen also in the middle of the string; (3) finally, if the cell is scaled down to 32 nm and is programmed to P1 state ($V_T \simeq 1$ V), the HCI effect caused by DIBL, rather than GIDL, become the more prominent effect [98]. A solution to (1) is obtained avoiding to use the first WL for data storage, referred as *Dummy WL*, but mechanisms (2)-(3) cannot be avoided and thus V_T of the adjacent cells is increased due to HCI. Moreover, novel program disturb mechanisms, such as hot holes generation by FN electrons which are injected from channel/junction to the control gate (CG) along the isolation [102], conduction band distortion near S/D regions [103] or abnormal interference due

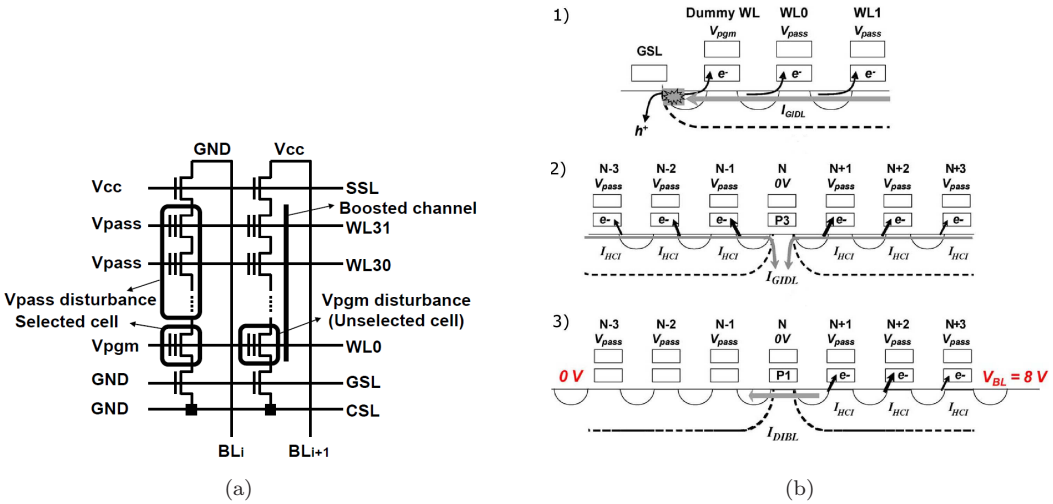
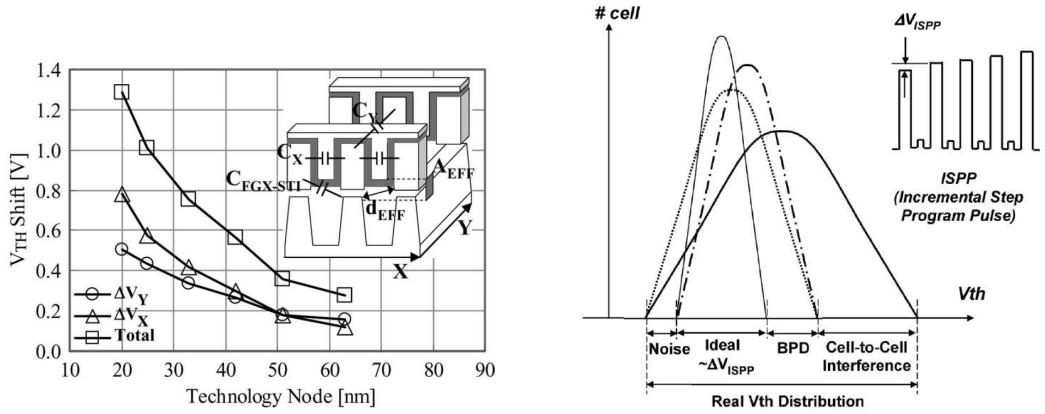


Figure 1.15: *(a) Bias condition for NAND cell programming (from [97]); and (b) a schematic summary of the different program disturb physical mechanisms: (1) GIDL at the source select transistor side; (2) GIDL for a cell programmed to P3 ($V_T \simeq 5$ V) in the middle of the string; (3) DIBL for a 32 nm cell programmed to P1 ($V_T \simeq 1$ V) (from [98]).*

to CG poly silicon depletion [104] have been reported for sub-30 nm technology nodes.

In addition to program disturbs, electrostatic interference emerged as one of the most serious limitation to decananometer NAND Flash operation, affecting the programming accuracy [13, 105–107]. Fig. 1.16(a) schematically depicts the main parasitic couplings between adjacent cells in the bit line (C_y) and world line direction (C_x) and the simulated V_T shift of an unselected cell when its first neighboring cells either in the $x-$ or in the $y-$direction are programmed from -5 V to $+5$ V, as a function of the technology node feature size, revealing the increased impact of electrostatic interference in scaled device. Besides this two major contributions, also a coupling in the diagonal xy-direction exists and additional emerging parasitic contributions may become non-negligible with a further feature size scaling, such as the direct electrostatic coupling between each cell active area and its neighboring floating gates in the word-line direction (schematically accounted by C_d in Fig. 1.16(a)) [13, 108]. C_x contribution is usually lower than C_y, due to the gate plugs wrapping the FG for a better insulation in the world line direction (see Fig. 1.4(a)). In order to quantify the impact of electrostatics interference on array operation, Fig 1.16(b) shows the programmed V_T distribution widening due to noise, background pattern dependences and cell-to-cell interference, revealing that the latter mechanisms results in the larger V_T displacement, strongly compromising the programming accuracy. Since a tight and accurate V_T placement is mandatory in MLC and TLC technologies, several efforts have been spent in programming algorithm design, in order to reduce the impact of electrostatic interference on the programmed V_T distributions [105, 106]. Fig. 1.17, for instance, schematically depicts the MSB Re-Pgm scheme: even pages cells are firstly programmed to temporary PV levels, lower than the PVs programming



(a) (b)

Figure 1.16: *(a) Simulation results of the V_T shift caused by cell-to-cell interference by changing the neighburing cell V_T from $-5$to 5 V, calculated for a decreasing cell size (from [13]); and (b) parasitic effects compromising programmed V_T distribution width of a NAND array: noise, background pattern dependences (BPD) and cell-to-cell interference (from [105]).*

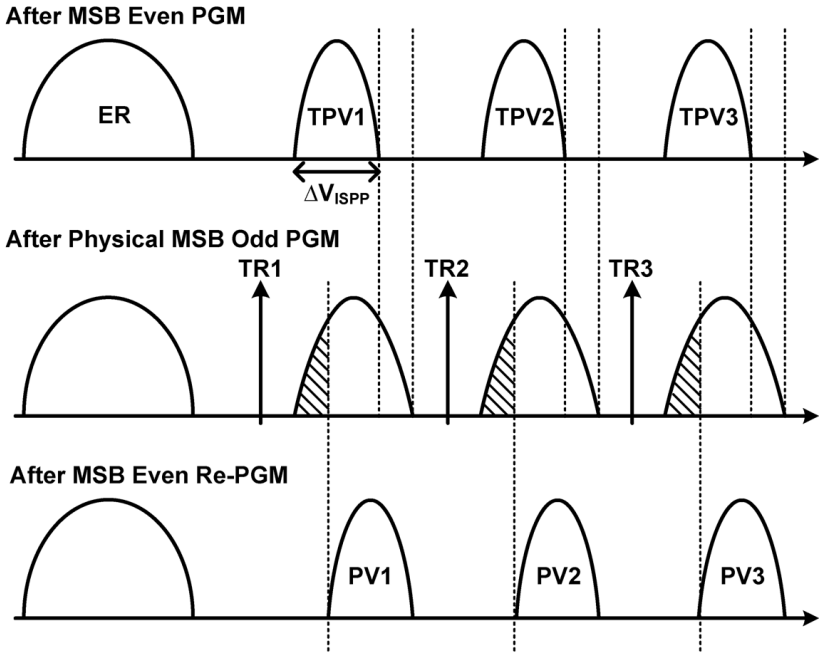
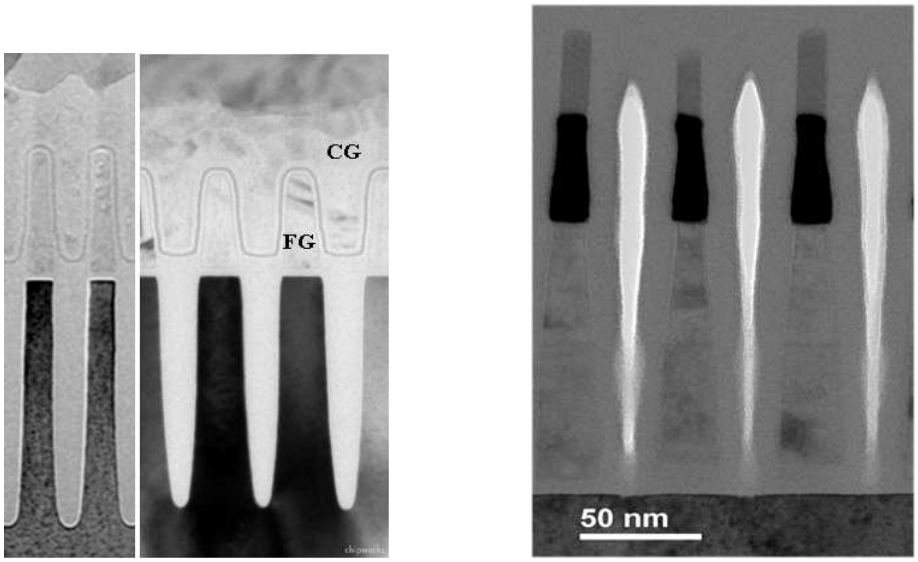


Figure 1.17: *V_T distributions of even pages during the MSB Re-PGM scheme, which aims at a cell-to-cell interference cancellation (from [106]).*



(a) (b)

Figure 1.18: *(a) Comparison between the cross section of (left) a* 27 *nm and (right) a* 35 *nm NAND technology by Samsung, highlighting the increase of FG aspect ratio with the node scaling (Chipworks, [106]); and (b) cross section of the IMFT* 25 *nm NAND technology, showing the air gap insulation between FGs in the direction parallel to the bit line(from [5]).*

algorithm aims at; then, even pages V_T distributions are enlarged upward due to adjacent cells electrostatic aggression after the odd pages programming to a random multi-level pattern and, so, a re-programming of even pages cells is required to bring them at the desired PV levels, cancelling the impact of the cell-to-cell interference during programming. Moreover, neighboring cells programming level should be considered: for this reason data randomization algorithms have been developed to minimize the V_T cell difference between the victim and the aggressor cells, thus reducing the interference [109]). All these solutions, however, have the main drawback of an increased algorithm complexity and a reduced programming speed.

1.4 Future trends for Flash technologies

1.4.1 Near the end of the roadmap?

High scalability of FG memory cell determined the widespread diffusion of Flash technology; however, the increasing technological complexity (e.g., the quad spacer patterning technology required at the 1X node [22]), the previously mentioned reliability issues and the fundamental limitation that Flash technology is facing in the decananometer regime, which will be extensively discussed throughout this thesis, bring into question the scalability of conventional Flash technologies beyond the 10 nm node [112, 113].

Among all the phenomena limiting the cell scaling, the main drawback of the FG technology appears to be its electrostatics: on one hand, a strong electrostatic

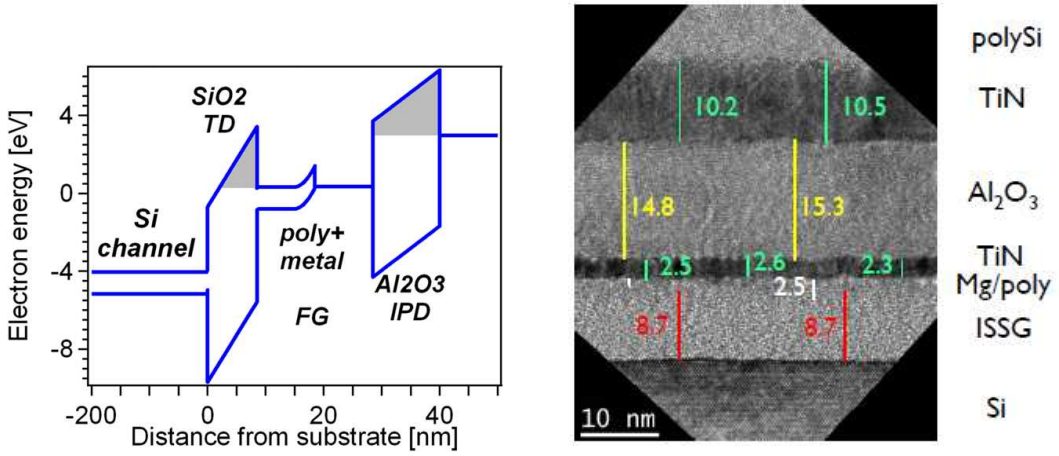


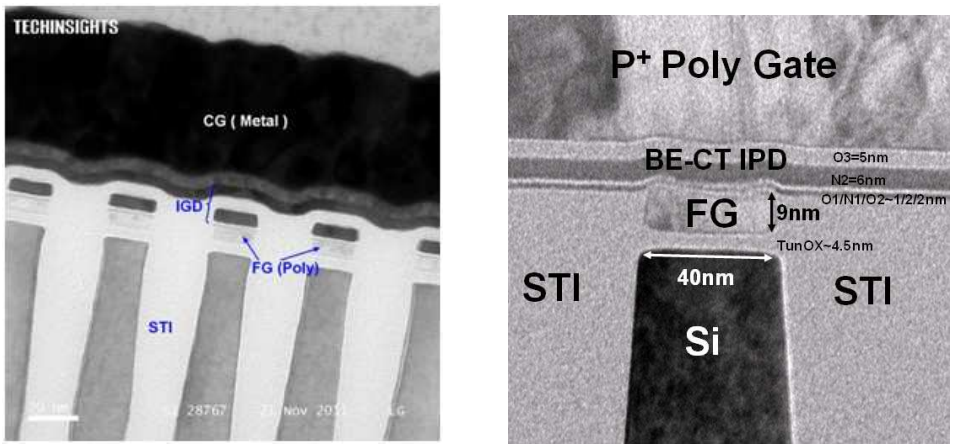
Figure 1.19: *The ultra-thin hybrid FG concept The FG is made of two layers, n-type polysilicon towards the tunnel oxide and p-type metal toward the IPD : (a) band diagram during erase (from [110]); and (b) TEM picture describing the different layers. FG has only a* 5 *nm thickness, consisting of approximately* 2.5 *nm Si and* 2.5 *nm TiN, and high-k IPD is adopted (from [111]).*

control of the FG potential is required for obtaining high programming efficiency and immunity from disturbs but, on the other hand, cell scaling increases the cell-to-cell interference. These two contrasting requirements lead to a trade-off in the FG geometry design: in fact, a decrease of FG height would be beneficial to cell-to-cell interference, while its increase, combined with a wrapped CG architecture, raise the CG-to-FG coupling and helps to get a better shielding in the WL direction. This trade-off has led to an almost constant FG height with the technology scaling: Fig. 1.18(a) shows a comparison between a 35 nm and 27 nm NAND cell of the same manufacturer (Samsung), highlighting the increase of the FG aspect ratio.

As a result of the electrostatic interference between cells, IPD and IWD thickness has become a limiting scaling factor and the FG spacing in 25 nm technology is already wider than its feature size, even if advanced solutions like air gap spacers (see Fig. 1.18(b)) are adopted to obtain a better insulation and to reduce the IWD thickness [5]. One possible solution for the wrapping CG geometry scaling may be the so-called *Ultimate FG* [112], consisting in an ultra-thin mono-crystalline FG with a thermally grown silicon oxide layer as IPD: in this case, the insulating layer thickness would be limited to $7 - 8$ nm, like the tunnel oxide, by SILC and FG charge loss; however, major issues may come from the integrity of this oxide and by the increase of technological complexity.

1.4.2 Planar FG geometry

A different approach to reduce the electrostatic interference consists in switching to a planar cell geometry, with a very thin FG, thus reducing the parasitic coupling with the neighboring cells, and replacing the conventional ONO IPD with a high-k



(a) (b)

Figure 1.20: *Two examples of planar FG geometry: (a) IMFT 20 nm technology, with poly silicon FG and high-k IPD (source: EETimes); and (b) Macronix planar FG cell with Barrier Engineered (BE) Charge Trapping (CT) IPD (from [114]).*

stack, to restore the CG-FG coupling. The main drawback of an ultra-thin FG geometry is the gate injection through IPD, due to the lower CG-FG coupling ratio leading to programming saturation and charge trapping in IPD: in order to overcome this issue, a hybrid-FG concept has been proposed in combination with a low-leakage high-k IPD: The FG is made of two layers, n-type poly silicon toward the tunnel oxide and p-type metal toward the IPD, whose higher work function suppress the tunneling through IPD [111, 112]. Figs. 1.19(a)-1.19(b) shows the band structure and the gate stack of a hybrid-FG device, showing a remarkable FG thickness of only 5 nm. However, this approach may be affected by process integration and material issues.

Despite the increased technological complexity, the planar approach is very promising for the 20 nm and beyond, as confirmed by Fig. 1.20(a), which shows the cross section of the first fully planar FG technology for mass production: IMFT's 20 nm NAND adopts a poly silicon thin FG with high-k IPD stack and a metal CG [115]. Fig. 1.20(b), in turn, presents a different approach to the planar FG geometry, proposed by Macronix [114]: in order to avoid the introduction of new materials in the cell structure, the high-k IPD is replaced by a charge-trapping device (a barrier engineered thin ONO), which should both trap the electrons, avoiding gate leakage current, and provide good reliability performances, due to the improved retention enabled by the optimized barrier.

1.4.3 Charge-trap memories

The solutions discussed so far try to mitigate Flash reliability issues by changing the cell geometry and the insulating dielectrics, but a different approach consists in the charge-trap memory concept, which replace the poly silicon FG with a thin trapping layer made of a dielectric material, such as silicon nitride (Si_3N_4), with a high traps density, as shown in Fig. reffig:CT. In so doing, the charge is no more

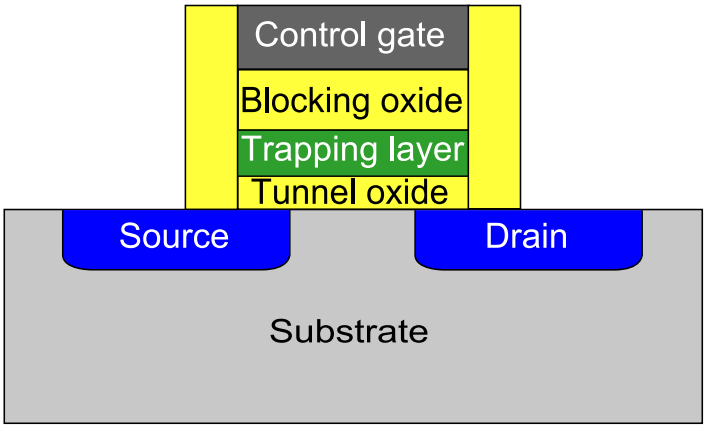
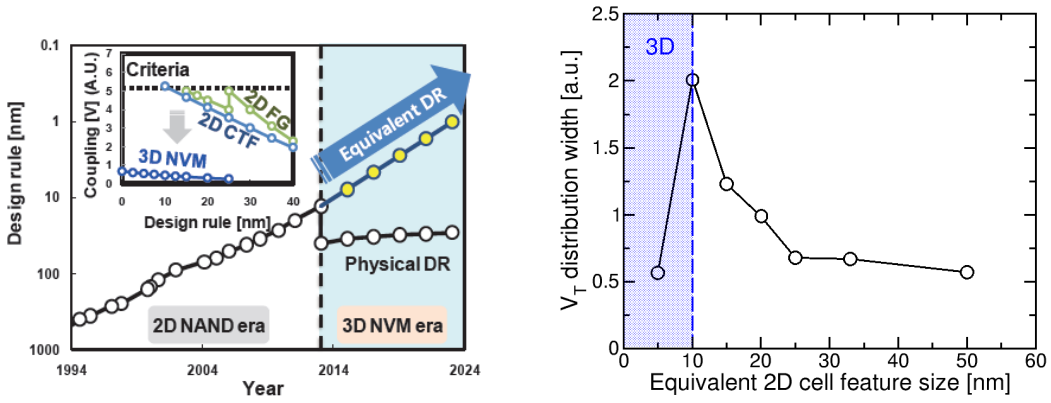


Figure 1.21: *Schematic view of a charge trap memory cell: the charge is stored in the trapping layer by applying the proper voltage to the control gate (CG); tunnel and blocking oxides are also shown (from [2]).*

stored in a conductive node, but it is trapped in localized defect sites, theoretically making the cell less vulnerable to leakage paths and local defects in the tunnel oxide (e.g., defects responsible for SILC). Moreover, the trapping layer is thick only few nm, providing all the beneficial effects in term of electrostatics interference given by a planar geometry [116]. One of the first CT concepts which was proposed is the Silicon-Oxide-Nitride-Oxide-Silicon c [117], where the name describes the composition of the gate stack: in particular, the silicon-nitride trapping layer is insulated by silicon oxide both from the silicon substrate and from the poly silicon CG. The use of silicon oxide as blocking oxide (the insulator between the trapping layer and the CG) give rise to several issues related to the gate leakage current which flows through it, leading to erase saturation and, thus, limiting the available programming window [118]. To overcome this limitations, several improvements have been proposed, first of all the adoption of a high-k blocking oxide, combined with a metal gate, such as in the TaN-Alumina-Nitride-Oxide-Silicon (TANOS) cell [119]. Charge-trap memories, however, failed to compete with FG Flash memories and to replace them in mass storage NAND products due to a wide range of issues, including program/erase efficiency and alumina trapping and leakage [120–122]; planar charge-trap memory will be unlikely developed in the future, since the planar FG has shown a more reliable operation.

Moreover, the planarization concept, even if it can relieve electrostatic interference, does not provide a long term solution to Flash scaling: planar cell will store a decreasing number of electrons (in the order of 10 electrons/bit in sub-20 nm technologies), facing fundamental limitations coming from the discrete nature of the charge and its variability effects both for FG cells (as discussed in Chapter 5) and for charge-trap technologies [123–126]. Possible solutions consist, on one hand, in increasing the integration density thanks to three-dimensional integration rather than due to cell scaling or, on the other hand, in moving to new memory concepts,



(a) (b)

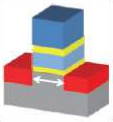
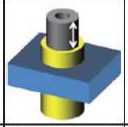
Figure 1.22: *(a) The scaling-up of design rule thanks to the transition to the 3D NVM era. The inset shows the parasitic coupling as a function of design rule for planar FG cells, planar charge-trap cells and 3D cells (from [113]); and (b) Monte Carlo simulation results showing the programmed V_T distribution width: 3D NAND geometry allow a tighter V_T placement, mainly due to lower parasitic coupling.*

which are not based on charge storage but rely on different physical phenomena: these approaches will be briefly discussed in the following Sections.

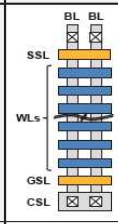
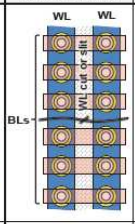
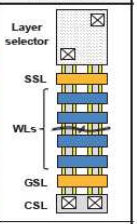
1.4.4 3D Flash approaches

The fundamental barriers to further scaling of Flash technology can be overcome thanks to the scaling-up of the physical design rule which can achieved with three-dimensional integration [113]; in so doing, the integration density can be increased exploiting the third dimension and the equivalent design rule can keeps its scaling trend, in terms of integration density increase (while the physical design rule can be even increased), bringing the Flash technology into the Terabit regime, as shown in Fig. 1.22(a). The inset of Fig. 1.22(a), moreover, compares the parasitic coupling in planar FG, planar CT and 3D devices, showing the great improvement achieved by 3D cells. In order to assess the improvements which can be achieved due to 3D architecture as a result of the lower parasitic couplings associated to a 3D NAND array, Monte Carlo simulations have been performed for different technology nodes: Fig. 1.22(b) shows that, on one hand, scaling of planar technology result in a strong widening of programmed distributions while, on the other hand, tighter V_T distributions can be obtained on 3D NAND arrays.

3D integration can be pursued through two different approaches, classified according to the channel and gate directions and hence named Vertical Channel (VC) and Vertical Gate (VG) scheme. Figs. 1.23(a)-1.23(b) summarize the main features of cell structures and NAND array architectures of both approaches, compared to the conventional 2D technology. In VC scheme, the channel is made by a poly silicon pillar, surrounded by the stacked word lines, while in the VG schemes the channels are stacked with subsequent deposition of storage and gate layers. From the 3D array viewpoint, VC NAND is less dependent from the lithography, compared to the 2D NAND, since the bit line and the word line

Dimension		2D	3D	
			Vertical Channel	Vertical Gate
Unit-cell				
Gate structure		Planar	Gate-All-Around	Dual -Gate
Unit-cell size		$4F^2$	$6F^2$	$4F^2$
Barriers for physical scaling		Lithography	Film thickness	Lithography & Film thickness
1x nm node	Cell swing	1 (ref.)	x ~ 0.75	x ~ 0.7
	On-cell current	1 (ref.)	x ~ 2.3	x ~ 1.2
Coupling & Direction		Strong WL + BL	Almost zero Vertical	Moderate BL + Vertical

(a)

Dimension	2D	3D	
		Vertical Channel	Vertical Gate
Plane view			
Program/Erase	FN	FN	FN
Major factors of program disturb	Channel coupling 1 (ref.)	NOP x~ 0.2	NOP & Vertical coupling x ~ 0.3
Required stacks for 1x nm node	-	16 ~ 64	8 ~ 32
Cell overhead	~ 20 %	Almost zero	35 ~ 70 %
Lithography dependence	Strong	Weak	Moderate

(b)

Figure 1.23: *Comparison between planar and 3D (a) cell structures and (b) array architectures (from [113]).*

pitches are determined by the film thickness; however channel and storage layer thickness is the main barrier to the scaling and determines a larger cell area ($6F^2$) than 2D NAND and VG NAND ($4F^2$). VG NAND, on the contrary, displays a moderate dependence on lithography, since the bit lines are patterned like in the 2D technology. Looking at the cell performance, the gate-all-around and the dual-gate geometries should determine an increase ($\sim$ 200%)of the on-current with respect to the planar case, limiting the subthreshold swing decrease due to the mobility degradation in the poly silicon channel ($\sim$ 70%). From an array point of view, the main advantage is given by the parasitic coupling reduction, which is almost cancelled in the VC NAND due to gate-all-around geometry, while is still present in the VG NAND; the array operation, however, is affected by the complexity of the 3D architectures, requiring a greater number of program disturbance (NOP) and inhibit conditions. The main issues in 3D architectures are the number of stacks required to achieve the desired density and the cell overhead needed for signal routing and cell selection trough different stacks. VC NAND is prone to suffer the first issue, since $\sim$ 64 layers to achieve a density comparable to the $1X$ node, while VG NAND needs a lower number of stacks but a larger number of transistor for layer selection [113].

Several 3D NAND Flash technology have been proposed in the past years to keep a trend of increasing bit density and reducing bit cost [129], most of them relying on a charge-trap (SONOS) memory cell, due its relatively easy integration in 3D structures: the most interesting examples of vertical channel technologies are given by the Pipe-shaped Bit-Cost Scalable concept (P-BICS), presented by Toshiba [127] and the Samsung's Terabit Cell Array Transistor (TCAT) [130], while several implementations of the vertical gate concept have been proposed by UCLA, Samsung [131–133] and by Macronyx [134], relying on a BE-SONOS device, in the latter case. The main drawbacks of these solutions, however, come from the adoption of charge-trap devices, in particular due to the charge spreading along the trapping layer, which cannot easily cut between different layers and thus creates additional leakage paths [135, 136]. In order to solve this is-

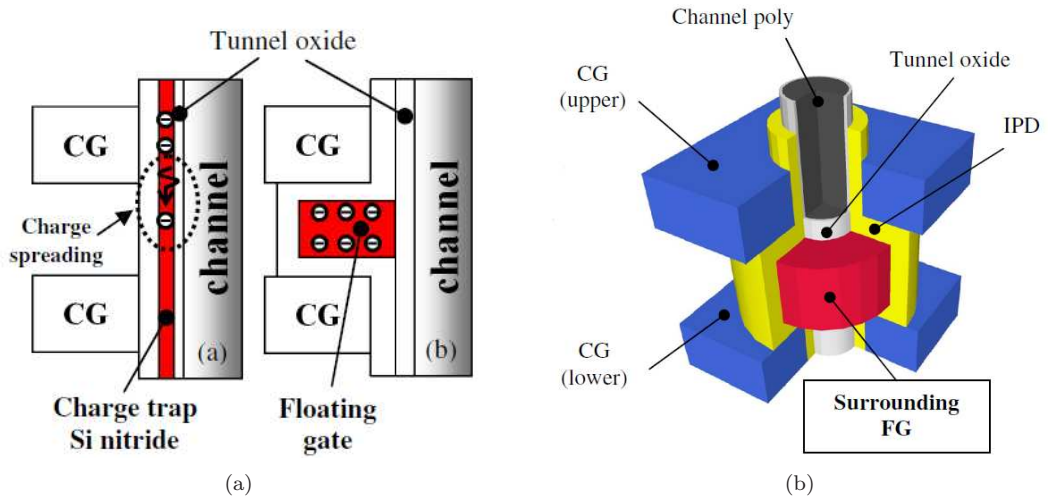


Figure 1.24: *Dual-gate surrounding FG cell proposed to solve lateral charge spreading in charge-trap 3D cells: (a) comparison with SONOS cell structure (BiCS [127]); and (b) bird's view of DC-SF NAND cell (from [128]).*

sue, Hynix has recently proposed a 3D dual CG NAND cell with surrounding FG (see Figs. 1.24(a),1.24(b)), opening the path to 3D integration also for FG based technologies [128].

1.4.5 Non-charge based memories

Besides these evolutions and improvements of the NAND Flash technology, a revolutionary approach, consisting in 2-terminals memory devices integrated in a 3D cross-point (or cross-bar) structure, may be considered for further scaling (see Fig. 1.25). In cross-point architectures the memory cells are no more arranged into NOR-like or NAND-like arrays, but they are placed at the intersection of mutually crossing perpendicular word lines and bit lines. This kind of layout in principle allows to realize the most compact cell area ($4F^2$) with a very simple process flow, using a minimal number of masks; the stacking of multiple layer, moreover, enable to further reduce cell occupation to $4F^2/n$, where n is the number of the layers. This memory concept, however, implies to move from the Flash technology to novel memories; in particular, since the storage node must have only 2 terminals, non-charge based technologies, which relies on the change of resistance of the active layer, appear the proper solution. Although cross-point architecture is very attractive from a scaling point of view, it suffers from an intrinsic weakness, since, depending on the memory pattern, a sneak current may flows through the memory elements close to the addressed one, compromising the current sensing and the stored bit reading. Thus, due to this sneak-path issue, cross-point architectures usually require a rectifying switch element for memory cell selection. The first cross-point architecture, for instance, consists in an antifuse, which realize a one-time-programmable (OTP) memory cell, addressed by a SiO_2 diode [138]. Among new proposed storage concepts based on the change of the resistance in the active material, which include magneto-resistive memory (MRAM) [139], phase-change

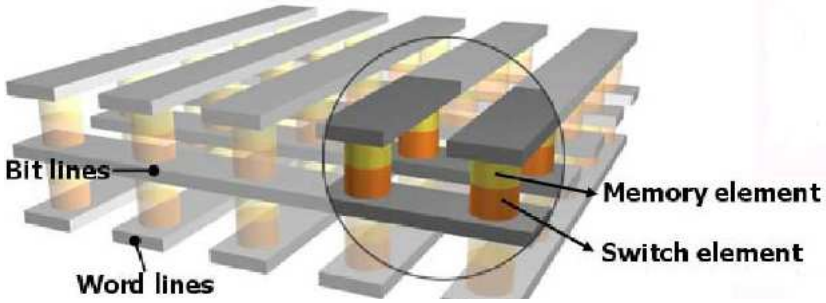
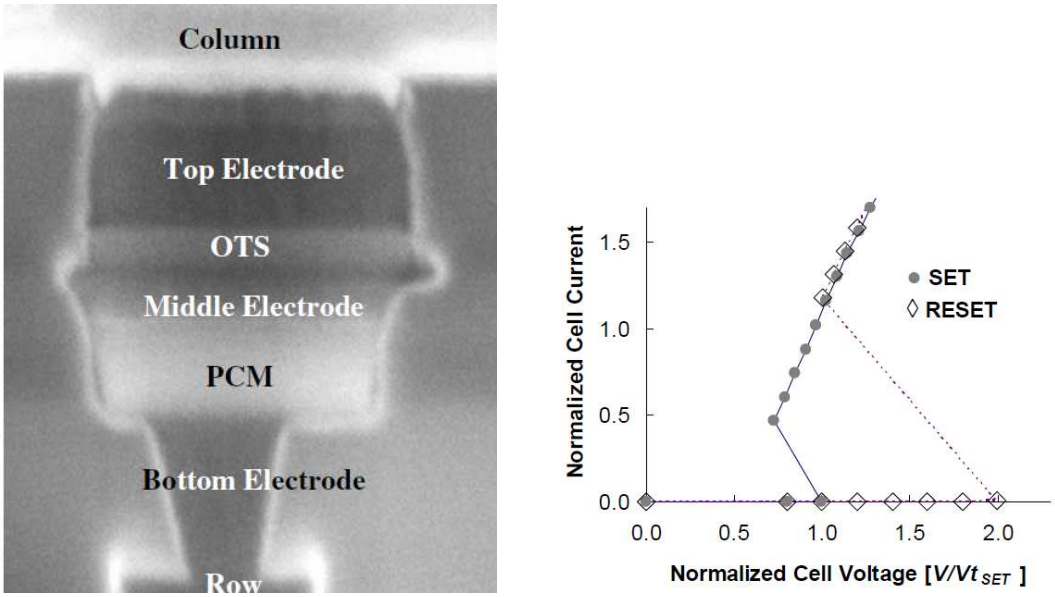


Figure 1.25: *The cross-point memory concept: the array is organized into mutually crossing perpendicular word lines and bit lines and the storage nodes are placed at the intersections. A novel memory technology is required for the storage node (e.g., a switch element for cell selection and a non-charge based memory element which can store data by changing is resistance). 3D stacking can be achieved to increase the integration density (from [137]).*

memory (PCM) [140] and resistive-switching memory (RRAM) [141, 142], the most advanced and promising ones for a cross-point integration are PCM and RRAM.

PCM technology, whose most advanced implementation for mass production is a planar 45-nm technology [144], is widely regarded as a Flash replacement, especially for NOR products, due to its higher speed and superior cycling capability, but it is not as widespread as the Flash one, because it is more expensive, requiring material that are not used in the standard CMOS process flow, and it is more power-hungry. PCM technology, however, emerged as a promising candidate for cross-point array integration: in the proposed implementations, the PCM cell can be selected by a conventional device, like a poly silicon diode [145], or by a chalcogenide-based not rectifying element: the Ovonic Threshold Switch (OTS) [146]. Figs. 1.26(a)-1.26(b) show the cross section of a stackable cross point phase change memory cell (PCMS), with OTS selector, and the $I - V$ characteristics of the cell, which is in the low resistive state (set state) when the chalcogenideis all crystalline, while the high resistive state (reset) is characterized by the presence of an amorphous active volume [143].

Another promising technology is the RRAM and several crossbar solutions employing oxide-based RRAM have been recently proposed [137]. Resistive switch between an high resistive state and a low resistive state can be achieved in some dielectric material thanks to a localized chemical transition, unlike PCM, whose switching relies on a bulk physical transition (i.e., the switch between crystalline and amorphous phase). Fig. 1.27 shows two possible resistive switching mechanisms in transition metal oxides, consisting in the formation/rupture of conductive filaments in an insulating matrix or in the switching at the interface between the electrode and the oxide [147]. Both the mechanisms are reversible, making possible to change the cell state, and the former one can be associated to unipolar or bipolar switching beahvior, while the latter usually leads to a bipolar beahvior, like in perovskite oxides. Cross-point integration of RRAM cells requires the use of a rectifying selection element, such as a diode, which must, on one hand, iso-



(a) (b)

Figure 1.26: *(a) SEM cross section of a stacked-PCM (PCMS) cell and (b) its $I-V$ characteristics. The cell is in the low resistive state (set state) when the chalcogenide is all crystalline, while the high resistive state (reset) is characterized by the presence of an amorphous active volume (from [143]).*

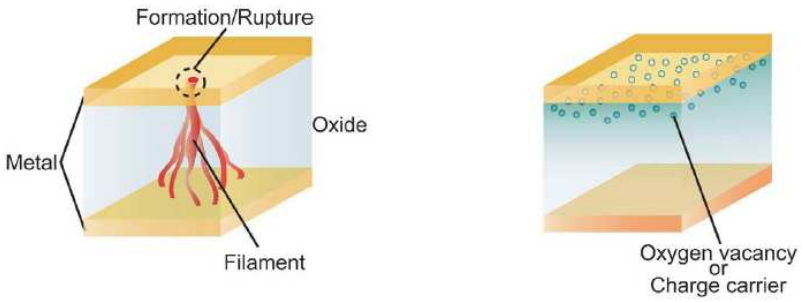


Figure 1.27: *Resistive switching mechanisms: (left) formation/rupture of conductive filaments in an insulating matrix; or (right) switching at the interface between the electrode and the oxide (from [147]).*

late unselected cell and, on the other hand, sustain the high reset current; these two requirements make the diode integration very difficult, since it must have a low leakage current, a high forward current and its size must be comparable with the cell size. For this reason, complementary switching in metal oxides is under investigation, in order to obtain diode-less crossbar devices [148].

However, cross point concepts shows several practical issue to be solved and feasibility and scaling proprieties of PCM- or RRAM-based cross point arrays should be still assessed; even if these approaches look promising for high density NAND replacement, floating gate Flash technology scaling and 3D integration nowadays is regarded as the main road for he near future memory evolution.

1.5 Description of the Ph.D. research activity

The research activity is focused on the physical mechanisms limiting the reliability of decananometer NOR and NAND Flash memories. Firstly, the attention is focused on oxide degradation and trapping/detrapping phenomena, which have been shown to set the ultimate limits to Flash endurance and data retention: threshold voltage instabilities after cycling have been investigated on NAND (Chapters 2-3) and NOR arrays (Chapter 4), highlighting their dependence not only on the bake but also on the cycling conditions, in terms of temperature and time. Experimental characterizations and data analyses assessed the validity of a universal model for damage recovery after distributed cycling, allowing the development of accelerated test schemes reproducing the real on-field usage of the devices. Moreover, the impact of short channel effects on string current in decananometer NAND Flash devices and the basic properties of damage creation and recovery (during cycling and bake, respectively) in nanoscale Flash devices have been also studied in detail in Chapter 3 thanks to a compact model for string current, providing a deeper insight into the time dynamics and activation energies of charge trapping/detrapping and interface states creation and annealing.

Then, among the emerging constraint to Flash reliability, a special interest has been devoted to fundamental variability sources and few electron phenomena. Few electron phenomena have been considered as a source of programming noise during incremental step pulse programming (ISPP) of NAND arrays, studying the statistical dispersion of the programmed threshold voltage distribution coming from the granular nature of the current flow to the floating gate. The impact of floating gate and control gate doping and geometry is investigated in Chapter 5, addressing the scaling trend of electron injection statistics and discussing possible cell design solution to mitigate its effects. Discrete electron emission from the floating gate and tunnel oxide, in fresh and cycled devices, respectively, has been also considered during data retention, investigating spread contributions to data retention coming from charge granularity and cell parameters fluctuations. Moreover, using Monte Carlo numerical simulations, the possibility to increase programming accuracy by means of double verify ISSP algorithms has been considered in Chapter 6, highlighting benefits and drawbacks of the algorithms.

The research activity provides a clear assessment of the fundamental limitations of Flash technology and, more in general, of the physical mechanisms every

future charge-based non-volatile technology should deal with in order to deliver a reliable operation. Moving from the physical understanding of the reliability constraints, the thesis discusses feasible solutions in order to extend the success of Flash technology to the decananometer technology nodes and develops the theoretical foundations of accelerated qualifications schemes for ultra-scaled technologies, in order to correctly reproduce the real on field usage of the devices.

CHAPTER 2

Cycling-Induced V_T Instabilities

This chapter presents a detailed investigation of the impact of cycling time and temperature on the threshold-voltage instability arising from damage recovery during data retention on nanoscale NAND Flash. Statistical results from the programmed state show that instabilities result, on average, in a threshold-voltage loss which increases logarithmically with the time elapsed since the end of cycling. Increasing the cycling time and temperature corresponds to an equivalent delay of the instant at which the first read operation on the array is performed. The delay is studied for a large variety of cycling and retention conditions, extracting the parameters required for a universal damage-recovery metric for NAND.

2.1 Introduction

Recovey of cycling–induced damage is a major source of threshold voltage (V_T) instability during data retention for deca-nanometer NAND Flash memories [60–62,65]. This instability represents the worst reliability issue coming from spurious charge trapping in the tunnel oxide and interface state creation during repeated program/erase (P/E) cycles [55, 57, 65, 66], arising from the possibility for the same oxide charge to detrap and for interface states to anneal out when cells have to preserve their data [56, 58–65]. Both charge detrapping and interface state annealing give rise to V_T displacements that are particularly critical for multi-level devices, where the higher number of bits per cell is obtained at the expense of reduced noise margins. The statistical nature of these displacements is well recognized [63, 64] and results from the stochastic fluctuation both of the number of defects per cell and of the impact of each single defect on cell V_T.

The former fluctuation is due to the very small device active area, resulting in such a low number of defects per cell that their Poissonian dispersion cannot be neglected. The fluctuation of the impact of each single defect on cell V_T is, instead, related to the possibility for the defect to be placed at different spatial positions over the cell active area in presence of percolative substrate conduction [66, 69, 75, 80, 96, 124]. Moreover, both positive and negative charges can be trapped in the tunnel oxide during cycling, giving V_T shifts of opposite polarities when detrapping takes place [57, 63]. As a result of the statistical dispersion of the V_T displacements, the array V_T distribution increases its spread as time elapses during data retention [63, 64]. In addition to that, the reliability issues coming from damage recovery are further worsened by the non-zero average value of the displacements, shifting the V_T distribution along a preferential direction.

The amount of cell damage contributing to V_T instabilities during data retention is the result not only of the number of P/E cycles previously performed on the array but also of the time delay between cycles and the cycling temperature [63, 64]. In fact, the possibility for partial damage recovery to take place during the time elapsing in between the cycles results into a lower amount of damage at the end of cycling when this is performed on a longer time interval. This means that lower V_T instabilities during data retention are expected when the cycling time is increased, as clearly shown for NOR Flash memories in [63,64]. As a consequence, characterization tests where P/E cycles are performed in quick succession to minimize the required experimental time, usually referred to as *fast–cycling* tests, provide only worst case results for the V_T instabilities during data retention. A more realistic test should, instead, reproduce the time distribution of P/E cycles that is reasonably expected in real device operation. To this aim, *distributed–cycling* experiments should be designed, trying to solve the trade-off between a low characterization time and a correct reproduction of the amount of damage at the end of cycling. In so doing, the increase of the cycling temperature to obtain in affordable experimental times the same damage recovery that should be obtained at the device working temperature on a much longer cycling timescale appears as the most practical solution [63, 64, 149].

This chapter presents a detailed investigation of the V_T instability determined by damage recovery during data retention on deca-nanometer NAND Flash, focusing on its basic phenomenology and on its dependence on cycling time and temperature. Extending the work we presented in [67] with more detailed discussions and additional results, we show that V_T instabilities from the programmed state result, on average, in a V_T loss which increases logarithmically with the time elapsed since the end of cycling. Electric field during data retention, cycling dose and probability level at which the shift of the array cumulative distribution is monitored are all important parameters for the slope of the logarithmic V_T loss. Changes in the cycling time and temperature correspond, instead, to an equivalent delay of the instant at which the first read operation on the array is performed. This delay is investigated for a large variety of cycling and retention conditions, extracting the value of the parameters needed to accurately model distributed–cycling effects.

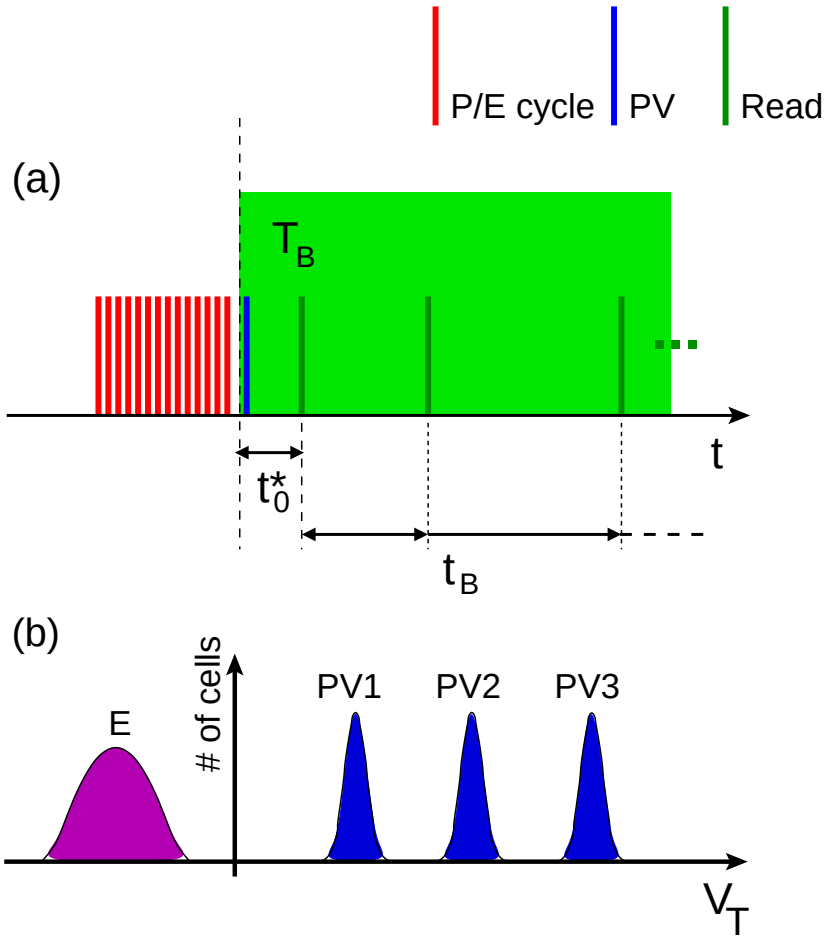
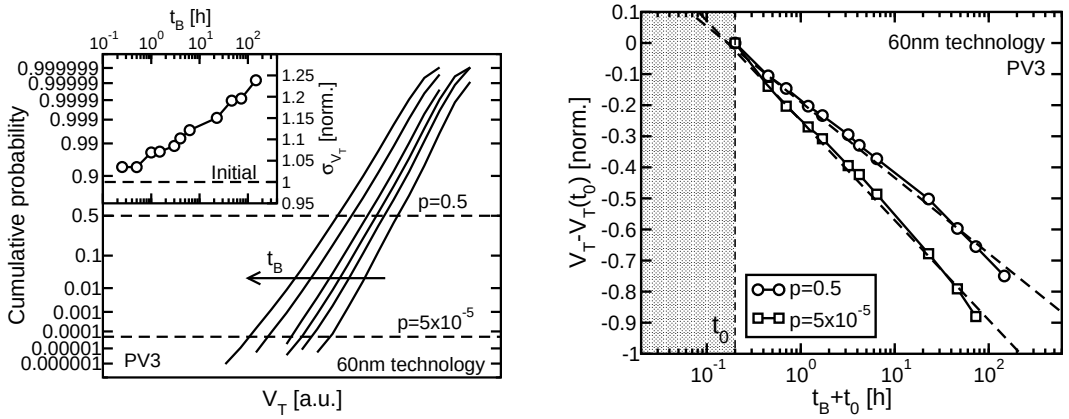


Figure 2.1: *Schematics for the fast–cycling experiment used to investigate the basic phenomenology of data retention V_T instabilities (a) and for V_T placement by PV in the multi–level device (b).*

2.2 Basic phenomenology

Fig. 2.1a schematically shows the simplest experimental procedure allowing the investigation of the basic phenomenology of V_T instabilities due to damage recovery during data retention. Referring to the most general case of multi-level NAND devices, a fast–cycling operation at room temperature (RT) is first performed on the NAND blocks, bringing all the cells from the erased to the highest programmed V_T level (namely, PV3, see Fig. 2.1b). The use of a fast–cycling test and of a solid cycling pattern between the extreme V_T levels aims at maximizing cell damage at the end of the P/E stress period, magnifying, in turn, V_T instabilities during the next data retention experiment. These instabilities are investigated for different electric fields in the tunnel oxide using NAND blocks selectively programmed with a solid pattern to the three high-V_T levels shown in Fig. 2.1b, *i.e.*, PV1, PV2 and PV3. This last program operation, achieved by a program-and-verify (PV) algorithm, immediately follows the end of cycling, as shown in Fig. 2.1a, and corresponds to the last high-field electrical stress to the array. After this operation,



(a) (b)

Figure 2.2: *V_T cumulative distribution at the first read operation after PV to the highest V_T level (PV3) and after increasing t_B up to 1 week (T_B= RT) (a), and corresponding V_T–loss transients at $p = 0.5$ and 5×10^{-5}, as a function of $t_0 + t_B$ (b). Data refer to a fast–cycling test on a block of our 60 nm test-chip. The inset in (a) shows the distribution standard deviation as a function of t_B.*

a bake experiment at temperature T_B is used to check V_T stability, reading the array V_T cumulative distribution at logarithmically-spaced times t_B since the first read operation, performed after a delay t_0 since the end of cycling. Note that in the case a T_B larger than RT is used for damage recovery acceleration, bakes are periodically interrupted and the devices cooled to RT for V_T reading. Moreover, in this case t_0 should be evaluated as the equivalent time at T_B that matches, from the damage recovery point of view, the time spent at RT between the end of cycling and the first read operation on the array (t_0^*), as assumed in Fig. 2.1a [64] and as will be better explained in Section 2.2.2.

2.2.1 V_T–loss dynamics

The experimental test of Fig. 2.1a was applied to 60 nm NAND test-chips, performing N_{cyc} = 10k P/E cycles and monitoring data retention at T_B = RT. Fast–cycling of single blocks required a cycling time $t_{cyc} \simeq 0.6$ h and a total time $t_0 \simeq 0.2$ h elapsed between the end of cycling and the first read operation on the array. Fig. 2.2(a) shows the V_T cumulative distribution at the first read after programming to PV3 and for increasing t_B up to 1 week. Note that the PV operation did not aim at minimizing the V_T distribution width at the beginning of the data retention experiment. In fact, in order to make random telegraph noise (RTN) negligible for the time evolution of the V_T distribution [80], an incremental step pulse programming (ISPP) algorithm with loose step amplitude was used [74, 150, 151], resulting in a nearly gaussian distribution with a rather large spread. This distribution has an average negative shift for increasing t_B as a result of damage recovery, increasing at the same time its standard deviation σ_{V_T} as shown in the inset of Fig. 2.2(a) [63, 64].

In order to gain the most complete picture of the damage recovery process from

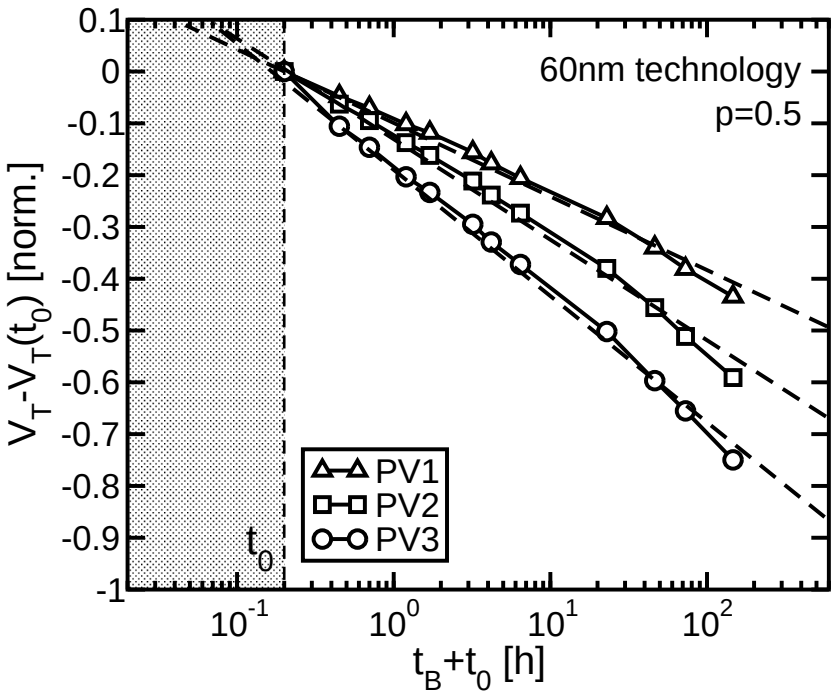


Figure 2.3: *V_T–loss transients ($p = 0.5$) during data retention at $T_B = RT$ after fast–cycling, for the three different PV levels schematically shown in Fig. 2.1b. The shadowed region corresponds to times lower than t_0, where V_T is not monitored by the experimental test.*

the cumulative distribution of Fig. 2.2(a), we investigated the V_T–loss transients at two different probability levels, *i.e.*, $p = 0.5$ and 5×10^{-5}. The former p value gives the average behavior of the array cells, while the latter allows the investigation of the shift of the distribution lower edge, including a contribution from the increase of σ_{V_T}. As a result, Fig. 2.2(b) shows that a faster V_T–loss appears for $p = 5 \times 10^{-5}$ than for $p = 0.5$. Moreover, note that for both the p values a logarithmic decrease of V_T clearly appears as a function of the total time elapsed since the end of cycling, *i.e.*, $t_0 + t_B$:

$$V_T(t_0 + t_B) - V_T(t_0) = \alpha \ln(t_0) - \alpha \ln(t_0 + t_B) \tag{2.1}$$

where α is a coefficient depending on p, in agreement with [58, 63, 64]. Fig. 2.3 shows that α depends also on the investigated PV level, increasing when moving from PV1 to PV2 to PV3 and revealing a strong dependence of damage recovery on the electric field in the tunnel oxide during data retention. As a general remark, note that the V_T–loss transients of Figs. 2.2(b)-2.3 have been normalized to the same arbitrary constant, allowing a direct quantitative analysis of their dependence on the main investigated parameters. The same normalization will be used in the rest of the paper when showing V_T–loss transients or when considering the V_T variation (ΔV_T) between the read operation at t_B and the first read operation on the array.

Despite deviations from the logarithmic behavior are expected when considering very short t_0, (2.1) gives a first expression able to quantitatively describe the

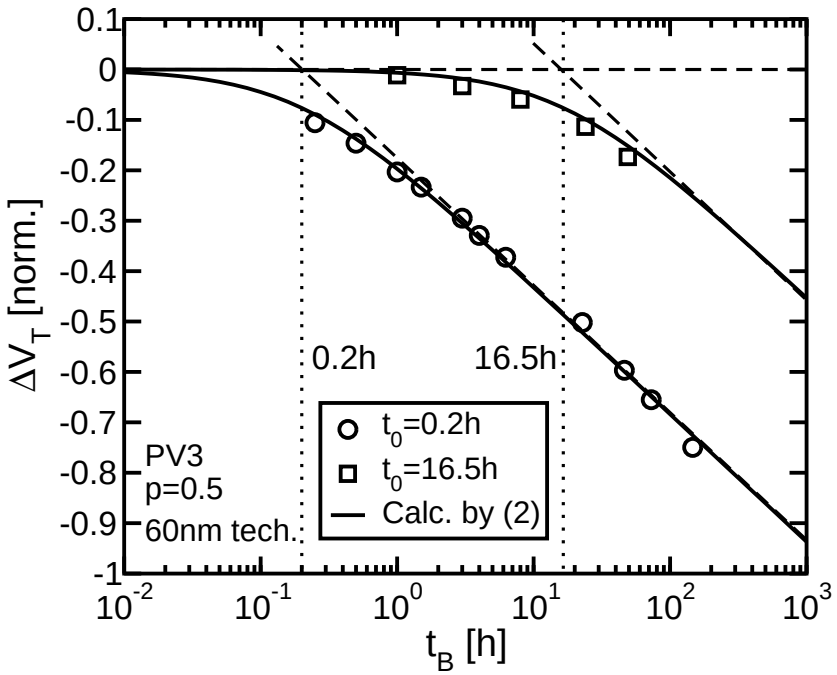


Figure 2.4: *ΔV_T transients as a function of t_B in the case of $t_0 = 0.2$ h and $t_0 = 16.5$ h ($p = 0.5$, PV3). Calculated results according to (2.2) are also shown (solid lines), using α extracted from the slope of the V_T–loss curves of Fig. 2.3. Dashed lines correspond to the asymptotic behaviors of ΔV_T.*

ΔV_T transients obtained from the experimental test of Fig. 2.1a:

$$\Delta V_T(t_B) = -\alpha \ln\left(1 + \frac{t_B}{t_0}\right) \tag{2.2}$$

Calculations according to (2.2) are shown in Fig. 2.4 and compared to experimental data (PV3, $p = 0.5$) in the case of $t_0 = 0.2$ h and $t_0 = 16.5$ h (the first read operation in Fig. 2.1a was postponed in this latter case), using the value of α extracted from the slope of the logarithmic V_T–loss of Fig. 2.3. Results highlight that t_0 represents a characteristic time for the ΔV_T transient, with $\Delta V_T \simeq 0$ for $t_B \ll t_0$ and $\Delta V_T \simeq -\alpha \ln(t_B/t_0)$ for $t_B \gg t_0$ (asymptotic behaviors of ΔV_T are shown as dashed lines in the figure). In addition to that, the increase of t_0 gives rise only to a horizontal shift of the ΔV_T curve along the logarithmic t_B axis, completely preserving its shape.

2.2.2 T_B effect on the ΔV_T transients

The possibility to speed up the damage–recovery process by thermal activation is widely exploited to reduce the experimental time required to investigate the reliability constraints given by V_T instabilities after cycling [64,149]. In particular, the increase of T_B is used with the aim of reproducing with short bake times the same V_T instabilities taking place at RT on a much longer timescale. The effect of T_B on the ΔV_T transients is shown in Fig. 2.5, as resulting from the fast–cycling tests of Fig. 2.1a (N_{cyc} = 10k P/E cycles, data retention from the PV3 level):

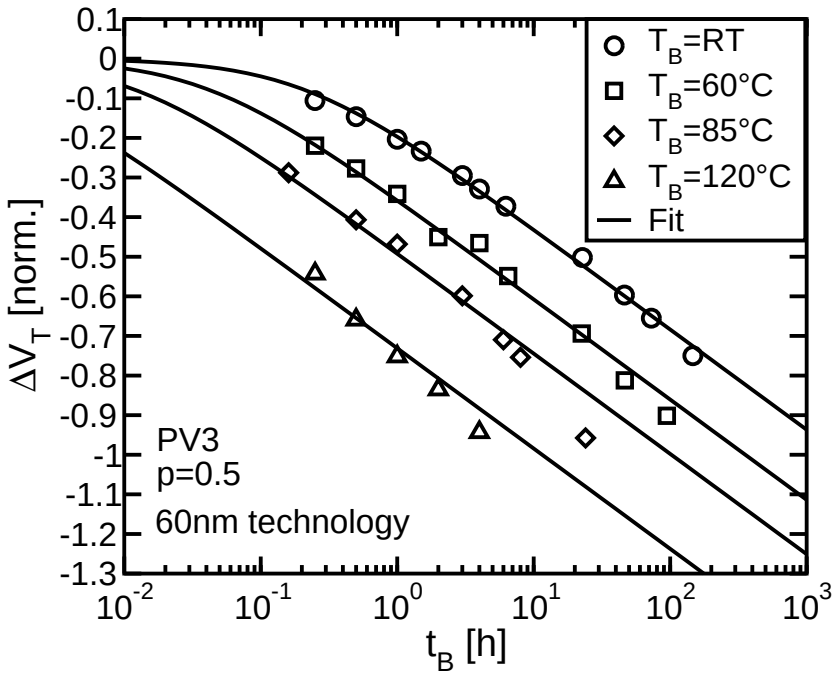


Figure 2.5: *ΔV_T transients as a function of t_B for different T_B ($p = 0.5$, PV3). A fitting of the experimental data according to (2.4) is also shown (solid lines), using α extracted from the slope of the V_T–loss curves of Fig. 2.3. From the fitting, t_0^* is extracted as a function of T_B.*

a leftward shift of the RT ΔV_T transient along the logarithmic t_B axis appears when T_B is increased, preserving the curve shape. In order to better highlight this behavior, Fig. 2.5 also shows a fit of the experimental data according to (2.2), using the same α extracted from the RT V_T–loss of Fig. 2.3 but reducing t_0 below 0.2 h when $T_B >$ RT are considered.

To practically manage the T_B effect on V_T instabilities, an Arrhenius law is usually assumed to describe the temperature dependence of the time needed to reach a selected ΔV_T, introducing, in so doing, an activation energy E_A for the damage–recovery process [64, 149]. Since the results of Fig. 2.5 highlight a rigid horizontal shift of the ΔV_T curves when T_B is changed, the Arrhenius relation can be directly used to define t_0^* as:

$$t_0^* = t_0 \cdot e^{E_A(1/kT_B - 1/kT_{RT})} \tag{2.3}$$

where kT_B and kT_{RT} are, respectively, the thermal energy at T_B and RT. According to this definition, t_0^* corresponds to the equivalent time at T_B that matches, from the damage–recovery point of view, the time t_0 spent at RT between the end of cycling and the first read operation. As a result, the ΔV_T transients at T_B can be described by replacing t_0 with t_0^* in (2.2):

$$\Delta V_T(t_B) = -\alpha \ln\left(1 + \frac{t_B}{t_0^*}\right) \tag{2.4}$$

where t_B is the physical time spent at T_B during data retention. Note, moreover, that the results on the V_T–loss dynamics of Section 2.2.1 can be generalized to the

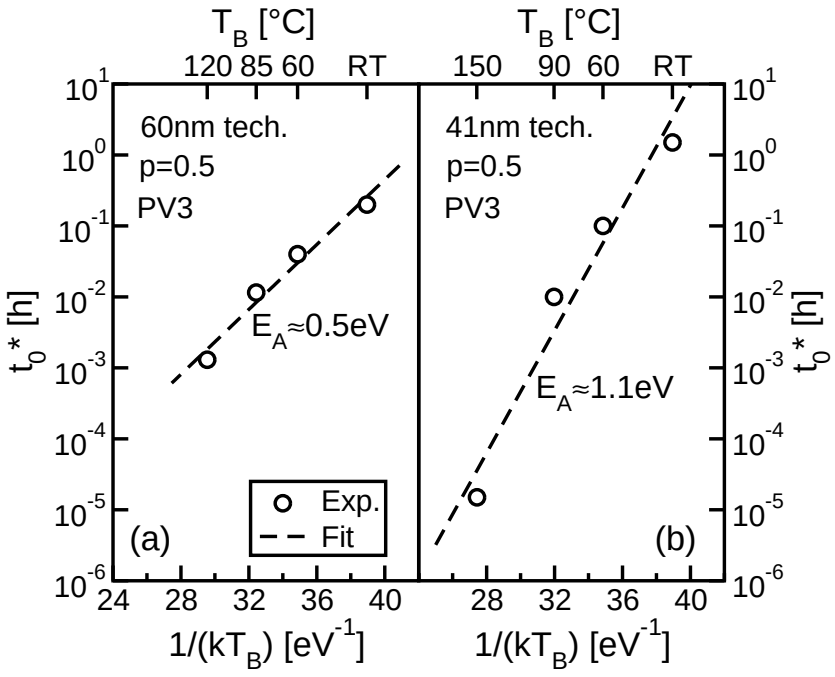


Figure 2.6: *Arrhenius plot for data retention: t_0^* is shown on a logarithmic scale as a function of $1/kT_B$. A linear fit of the experimental data is shown, allowing the evaluation of $E_A \simeq 0.5$ eV for our 60 nm test-chip (a) and $E_A \simeq 1.1$ eV for our 41 nm technology (b). Note that a different t_0 was required by the experiments on the 41 nm devices.*

case of an arbitrary $T_B \neq$ RT simply replacing t_0 with t_0^*, as implicitly assumed in the schematics for the experimental test of Fig. 2.1a.

Fig. 2.6a shows the Arrhenius plot obtained by reporting the extracted t_0^* on a logarithmic scale as a function of $1/kT_B$. From a linear fit of this graph, an activation energy $E_A \simeq 0.5$ eV can be obtained. This value is quite low if compared to the 1.1 eV that is usually reported for detrapping [63, 64, 149] and represents a unique feature of our 60 nm test-chip. In fact, all others NAND technologies we investigated, ranging from the 90 nm to the 32 nm node, displayed an E_A in the 1.0–1.2 eV range, as shown in Fig. 2.6b for our 41 nm device. The reason for the anomalous activation energy observed in the 60 nm test-chip has been traced back to an excessively thin tunnel oxide and to a non-optimized oxide/silicon interface, resulting into a significant contribution of direct tunneling on the V_T–loss, as discussed in [67].

2.2.3 Damage–creation and damage–recovery mechanisms

Fig. 2.7 shows that the slope α of the logarithmic data retention V_T–loss increases with the number N_{cyc} of P/E cycles previously performed on the NAND array. This is clearly the effect of a larger damage created in the cells as the cycling dose increases, resulting, in turn, in the possibility for larger V_T instabilities during data retention when partial damage recovery slowly takes place. The increase of cell damage with N_{cyc} is also responsible for the displacement of the high- and the low-V_T levels ($V_{T,E}$ and $V_{T,P}$, respectively) obtained by single-pulse program

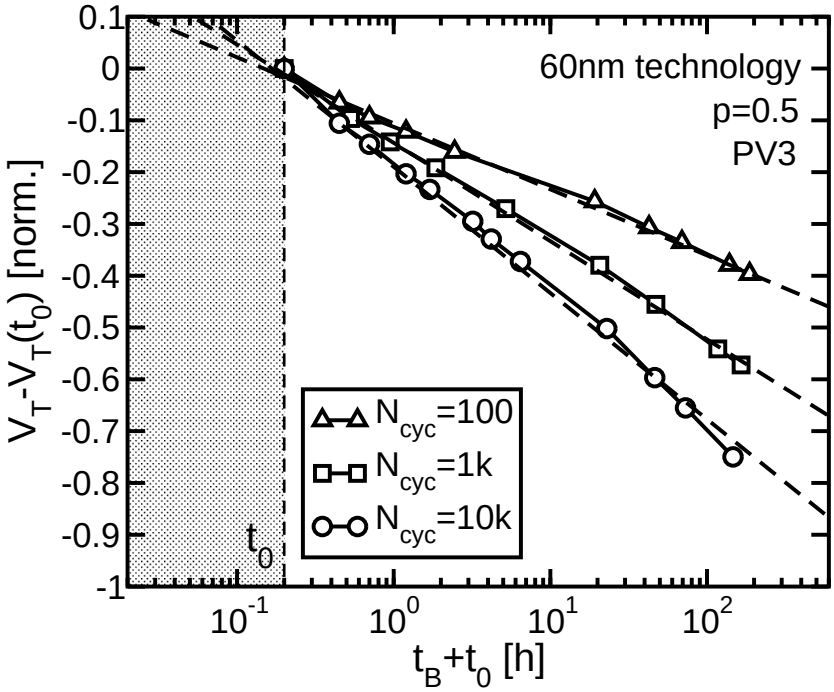


Figure 2.7: *V_T–loss transients ($p = 0.5$, PV3) during data retention at $T_B = RT$ after fast–cycling, for three different cycling doses. The shadowed region corresponds to times lower than t_0, where V_T is not monitored by the experimental test.*

and erase operations, as shown by the endurance curves of Fig. 2.8(a). In this figure, the average programmed V_T of a NAND block is compared with the results obtained from single cells out of analytical cell arrays (test element groups (TEGS), see [152]), showing a good agreement (note that the comparison cannot involve $V_{T,E}$, as this is not readable in NAND test-chips). The increase of $V_{T,E}$ and $V_{T,P}$ as cycling proceeds has been carefully investigated in [60,62,66], finding its source into charge trapping in the tunnel oxide and interface state generation at the substrate surface. In order to investigate the contribution of these two damage mechanisms on data retention V_T instabilities, we investigated the sub-threshold slope (STS) of the analytical cells of Fig. 2.8(a) as a function of N_{cyc}. Results are shown in Fig. 2.8(b), highlighting a different STS degradation for the programmed and erased cell state as cycling proceeds: STS starts growing since the very initial P/E cycles for the erased state, while a growth appears only when N_{cyc} exceeds 10^3 when referring to the programmed state. Differences between the STS of the programmed and erased cell arise from non-uniform substrate conduction and, in particular, from the possibility for different source-to-drain current paths when the cell has negative or positive V_T [65, 66, 153]. In these conditions, in fact, the same damage created in the cell in the form of interface states and trapped oxide charge may differently impact cell transcharacteristics, especially when considering that damage may be spatially localized as well [65, 66, 153]. From the results of Fig. 2.8(b), the STS degradation for the programmed cell state appears rather small up to 10^4 P/E cycles, revealing a small contribution

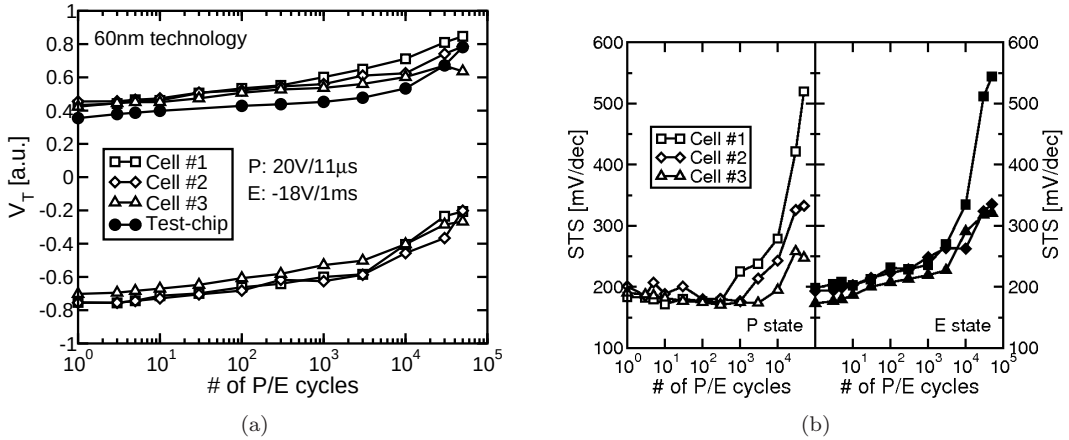


Figure 2.8: *Endurance results (a) for the 60 nm NAND test-chip (average V_T is considered) and for three analytical cells out of test-element groups of the same technology. Single pulse cycling with programming at 20 V, 11 µs and erasing at −18 V, 1 ms was used. STS as a function of cycling (b) for the three analytical cells shown in (a).*

of interface states to cell damage when the high-V_T state is considered. This, in turn, makes damage–recovery during data retention from the programmed state mainly a result of charge detrapping from the tunnel oxide. However, this could be not the case for very large N_{cyc} above 10^4, where the fast growth of cell STS resulting from Fig. 2.8(b) reveals a larger and larger impact of interface states on cell reliability.

2.3 Distributed–cycling results

Results presented in Section 2.2 clearly reveal that damage recovery takes place since the very beginning of the data retention period, giving rise to a logarithmic V_T–loss that has been observed for times as short as 0.2 h. As a direct consequence, the possibility for damage recovery to partially take place also during the time delays between subsequent P/E cycles should be carefully considered, leading to the requirement of correctly reproducing the cycling dynamics before the data retention period [63, 64]. In order for the V_T instability results to be quantitatively representative of device reliability, the fast–cycling test of Fig. 2.1a should be, therefore, replaced with the more general distributed–cycling test of Fig. 2.9a, where time delays between cycles have been introduced with the aim of reproducing the real cycling time of the memory device. For the sake of simplicity, a constant duration t_{wait} was assumed for these delays, resulting into $t_{cyc} = N_{cyc} \cdot t_{wait}$. Note, moreover, that all the results presented in this Section were obtained inserting the delays after cell programming, *i.e.*, when cells were in the PV3 state. In view of the strong field dependence of damage recovery shown in Fig. 2.3, this represents an additional important piece of information to be considered when quantitatively assessing the reliability constraints given by V_T instabilities. Finally, note that the possibility for the cycling temperature T_{cyc} to differ from the retention temperature T_B has been included in Fig. 2.9a.

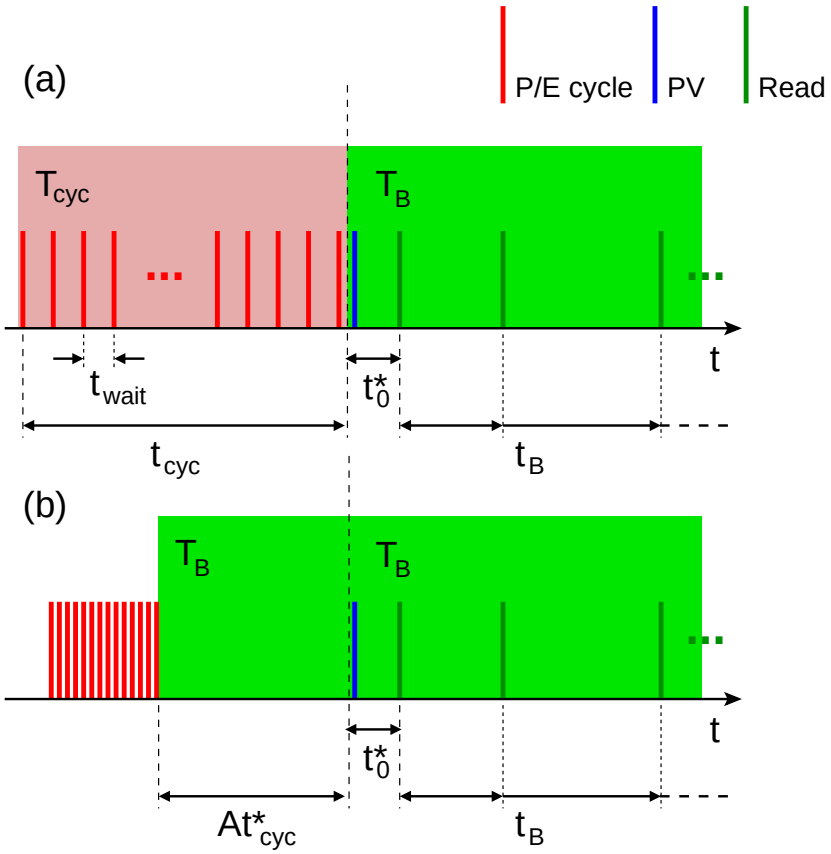


Figure 2.9: *Schematics for the experimental procedure used to investigate V_T instabilities due to damage recovery after cycling at different t_{cyc} and T_{cyc} (a) and equivalent model for distributed cycling (b).*

2.3.1 t_{cyc} and T_{cyc} effect on V_T instabilities

We applied the experimental test of Fig. 2.9a both to our 60 nm test-chips and to our 41 nm technology, applying N_{cyc} = 10k P/E cycles and then monitoring V_T instabilities at T_B= RT in the former case and at T_B = 60 °C in the latter. A $t_0 \simeq 1$ h was used in these experiments and different t_{cyc} and T_{cyc} were considered to investigate the effect of distributed–cycling on the V_T–loss during data retention. Fig. 2.10(a) shows that both longer t_{cyc} and higher T_{cyc} result into lower data retention V_T instabilities, confirming for NAND the results presented for NOR Flash in [63, 64]. In order to better highlight the impact of the cycling time and temperature on V_T instabilities, Fig. 2.10(b) shows that the ΔV_T transients obtained from the different distributed–cycling conditions overlap after a horizontal shift along the logarithmic t_B axis. As a consequence, the experimental ΔV_T transients can be reproduced by (2.2) using the same α extracted from fast–cycling experiments and replacing t_0 with an experimentally extracted time t_B^*:

$$\Delta V_T(t_B) = -\alpha \ln\left(1 + \frac{t_B}{t_B^*}\right) \tag{2.5}$$

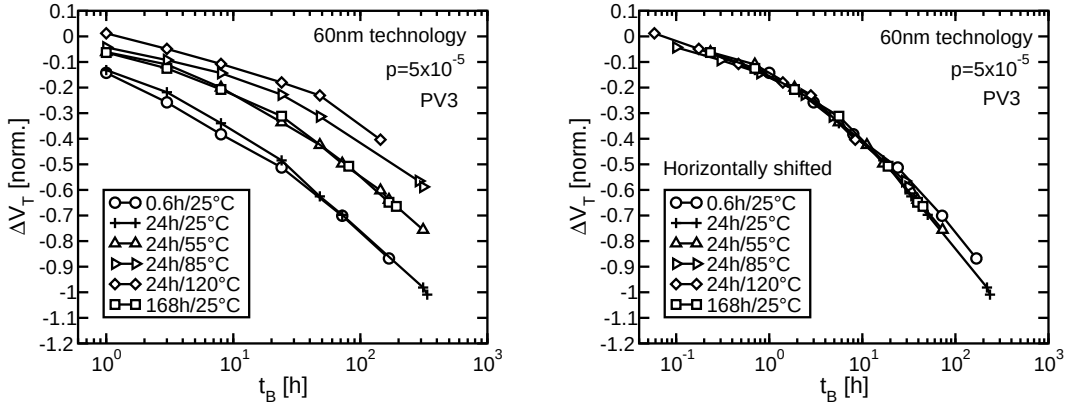


Figure 2.10: *ΔV_T transients (a) measured on our 60 nm test-hips as a function of t_B for different t_{cyc} and T_{cyc}. Bake temperature is T_B= RT. Transient can be horizontally shifted (b) to overlap with the fast–cycling ($t_{cyc} = 0.6$ h, $T_{cyc} = RT$) curve.*

Following a physical interpretation similar to that presented in [64], the effect of t_{cyc} and T_{cyc} on the ΔV_T transients can be understood referring again to Fig. 2.9a and making two main assumptions. The first one is that damage creation by P/E cycles depends neither on T_{cyc} nor on t_{wait}, which are, instead, parameters that strongly impact damage recovery during distributed cycling. For this hypothesis to hold, damage creation by P/E cycles should not depend on the damage already existent in the cells. The second assumption is that damage recovery occurring during the time delays in between the cycles can be reproduced by a single recovery phase of duration proportional to t_{cyc} at temperature T_{cyc} after damage has been created. Under these assumptions, Fig. 2.9a is equivalent, from the standpoint of cell damage present at the beginning of the data retention experiment, to the test of Fig. 2.9b. In this test, damage creation takes place during a fast–cycling experiment at RT, while partial damage recovery occurs in a single delay phase located between the end of cycling and the PV operation. Note that, using the results of Section 2.2.2, the temperature of this phase was changed from T_{cyc} to T_B by modifying the distributed–cycling duration according to the Arrhenius law:

$$t^*_{cyc} = t_{cyc} \cdot e^{E_A(1/kT_B - 1/kT_{cyc})} \tag{2.6}$$

In so doing, results for the experimental test depicted in Fig. 2.9b, and in turn of Fig. 2.9a, can be investigated from what shown in Section 2.2 for the test of Fig. 2.1a. In particular, the logarithmic V_T–loss taking place since the end of the fast–cycling experiment allows the quantification of ΔV_T as [64]:

$$\Delta V_T(t_B) = -\alpha \ln\left(1 + \frac{t_B}{t^*_0 + A t^*_{cyc}}\right) \tag{2.7}$$

and, by comparing (2.5) with (2.7), a direct expression for t^*_B can be obtained, resulting into:

$$t^*_B = t^*_0 + A t^*_{cyc} = t^*_0 + A t_{cyc} e^{E_A(1/kT_B - 1/kT_{cyc})} \tag{2.8}$$

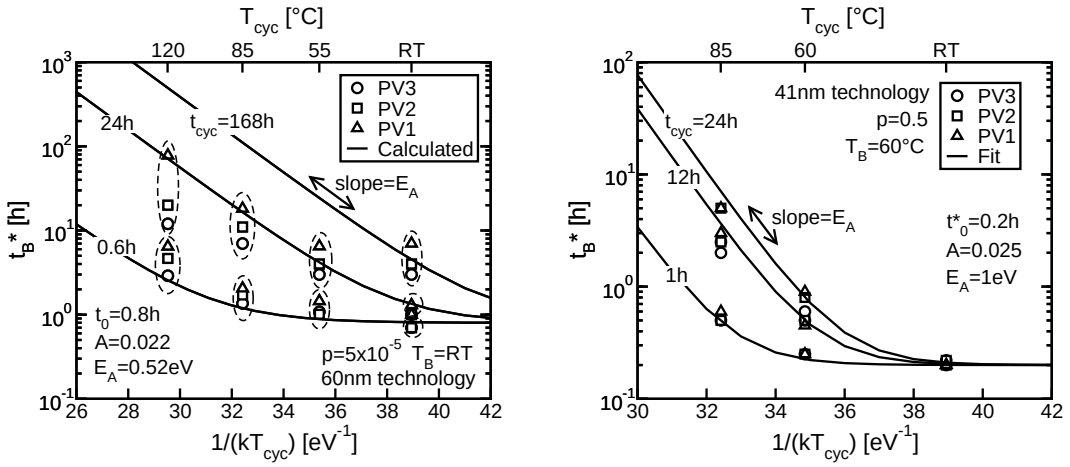


Figure 2.11: *Arrhenius plot for cycling: the t_B^* obtained from the fitting of the ΔV_T transients measured on our 60 nm (a) and 41 nm (b) test-chips for different t_{cyc} and T_{cyc} are shown as a function of $1/kT_{cyc}$.*

From (2.7) and (2.8), both the increase of t_{cyc} and of T_{cyc} are equivalent to a delay of the time at which the first read operation on the array is performed after a fast–cycling test, resulting into a delay of the fast–cycling ΔV_T transient along the logarithmic t_B axis, as observed in Fig. 2.10(a).

2.3.2 Arrhenius plot for cycling and UDM

Figs. 2.11(a)-2.11(b) show the t_B^* obtained from the fitting of the ΔV_T transients corresponding to different t_{cyc} and T_{cyc} on our 60 nm test-chips and on our 41 nm technology, respectively. Results are reported as a function of the reciprocal of the cycling temperature, obtaining what we defined as the *Arrhenius plot for cycling*. Note, in fact, that, similarly to Fig. 2.6, these graphs show on a logarithmic scale how the characteristic time of the data retention ΔV_T transients changes when the experimental temperature is modified. However, while Fig. 2.6 investigates the effect of the bake temperature T_B with constant T_{cyc}, Figs. 2.11(a)-2.11(b) consider the effect of the cycling temperature T_{cyc} with constant T_B. As a consequence, while a positive-slope fitting line results for the experimental points of Fig. 2.6, a negative slope is required in Figs. 2.11(a)-2.11(b) to fit the t_B^* dependence on $1/kT_{cyc}$ for fixed t_{cyc}. In fact, in this latter case, the increase of T_{cyc} determines a larger damage recovery directly during the distributed–cycling experiments, then reducing the damage contributing to V_T instabilities during data retention.

Experimental results for t_B^* reported in Figs. 2.11(a)-2.11(b) can actually be reproduced by (2.8), as shown by the solid lines in the figures. In so doing, besides validating the physical picture used in Section 2.3.1 to investigate distributed–cycling effects, the t_B^* dependence on t_{cyc} and T_{cyc} can better be highlighted. For fixed t_{cyc}, two temperature regimes exists for t_B^*: for high-T_{cyc}, t_B^* exponentially decreases when temperature is reduced, while for low-T_{cyc}, t_B^* saturates to a level nearly equal to t_0^*. The T_{cyc} value separating the two temperature regimes is a

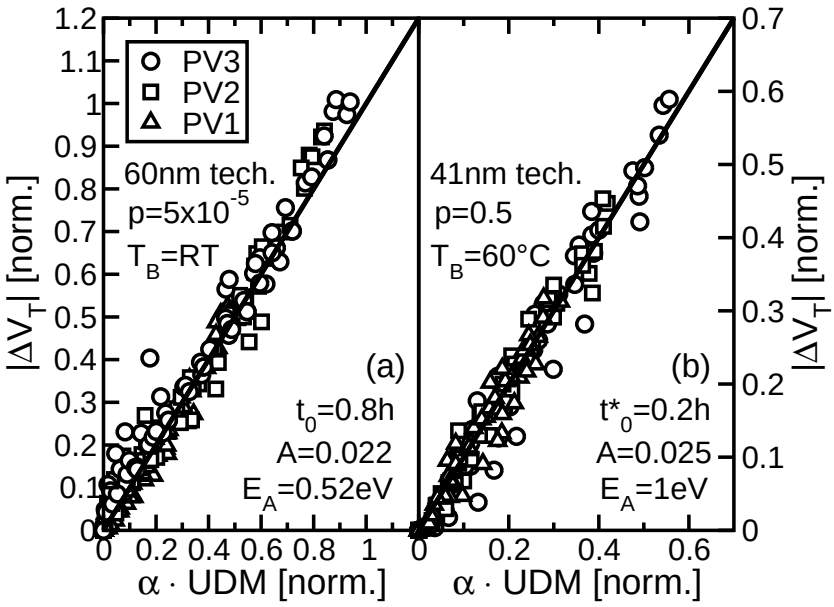


Figure 2.12: *ΔV_T transients from the experimental test of Fig. 2.9a reported as a function of $\alpha \cdot UDM$.*

function of t_{cyc}, being lower for longer t_{cyc}. As a consequence, the Arrhenius plot for cycling displays many branches of equal slope, each of them associated to a different t_{cyc} and departing from the saturation level t_0^* at different T_{cyc}.

From the fitting of the available experimental data in Figs. 2.11(a)-2.11(b), the parameters introduced in the previous Section to model the distributed–cycling effects can be extracted, obtaining $A = 0.022$ and $E_A \simeq 0.5$ eV for the 60 nm test-chips and $A = 0.025$ and $E_A \simeq 1$ eV for the 41 nm technology. The obtained E_A are in quite good agreement with the activation energies resulting from the Arrhenius plots for data retention of Fig. 2.6. With t_0, A and E_A known, the logarithmic term of (2.7) completely decribes the time dynamics of the ΔV_T transients, including all the times and temperatures involved in the experimental test, namely t_B, t_{cyc}, t_0, T_B and T_{cyc}. We considered, therefore, this term as a universal damage-recovery metric (UDM) [64]:

$$UDM = \ln\left(1 + \frac{t_B}{t_0^* + A t_{cyc}^*}\right) \tag{2.9}$$

Note that, instead, the dependence of ΔV_T on the explored PV level, p and N_{cyc} are included in the coefficient α that multiplies UDM in (2.7). To further check the validity of the extracted parameters and of (2.8) to describe the t_{cyc} and T_{cyc} dependence of the ΔV_T transients, Fig. 2.12 shows that all the experimental data collapse on a line of slope 1 when plotted as a function of α·UDM.

2.4 Conclusions

This chapter presented a detailed investigation of the cycling-induced V_T instabilities during data retention in NAND memories, considering both their basic

phenomenology and how they change when modifying t_{cyc} and T_{cyc}. Instabilities from the programmed state were shown to manifest themselves as a V_T–loss that logarithmically increases with the total time elapsed since the end of cycling. The slope of the logarithmic discharge was shown to depend on the electric field during data retention, the cycling dose and the probability level at which the array cumulative distribution is monitored. Changes in t_{cyc}, T_{cyc} and T_B correspond, instead, to an equivalent change of the instant at which the first read operation on the array is performed. Considering a large variety of t_{cyc} and T_{cyc}, the parameters needed to model the distributed–cycling effects on NAND were extracted, similarly to what done for NOR in [64].

CHAPTER 3

Compact modeling of NAND string current

This chapter presents a detailed compact-modeling investigation of the string current in decananometer NAND Flash arrays. This investigation allows, first of all, to highlight the role of velocity saturation, low-field mobility, and drain-induced barrier lowering on the string current versus read voltage characteristics. Results are validated on a 41−nm technology for different positions of the selected cell along the NAND string, different pass voltages, and different array background patterns. The effect of cycling on the string current is then investigated by means of post-cycling bake experiments, showing that the impact of charge trapping/detrapping and interface state generation/annealing varies as a function of the read current level.

3.1 Introduction

Aggressively shrinking cell active area while preserving an extremely compact layout, NAND Flash memories reached impressive integration densities in state-of-the-art technologies [5, 22]. The close packing of the memory cells in these technologies enhanced their interaction far beyond what comes from their series connection in the NAND string, complicating the investigation of single-cell operation and reliability [13, 85, 108, 152, 154, 155]. Modeling of the whole NAND string appears, therefore, mandatory for the quantitative analysis of many physical and electrical issues related to the array functionality, with compact modeling appearing as the most practical investigation approach [156–159].

This chapter presents a compact-modeling investigation of decananometer NAND Flash arrays, focused on the string current (I_S) sensed in READ conditions. The compact model allows, first of all, to analyze the effect of velocity saturation,

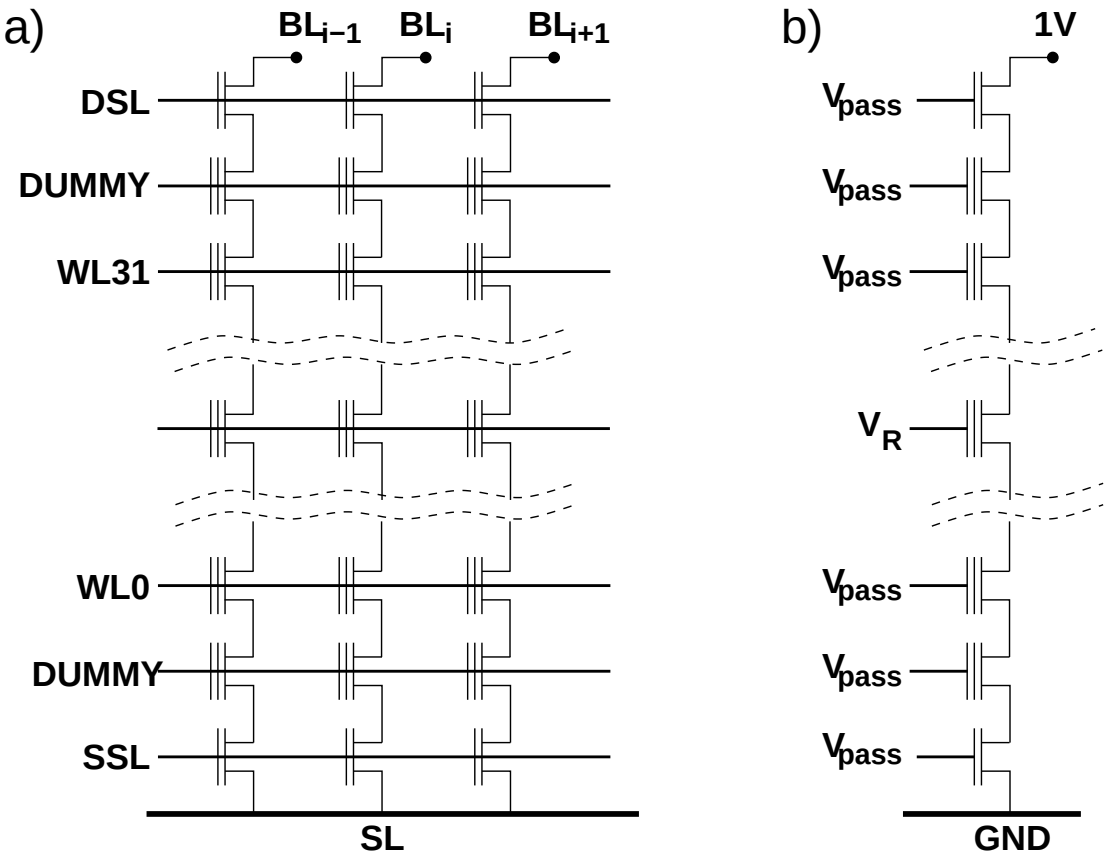


Figure 3.1: *Schematics for (a) cells connection in a NAND architecture and (b) bias applied to the string during sensing of I_S.*

low-field mobility, and drain-induced barrier lowering (DIBL) on the string current versus READ voltage ($I_S - I_R$) characteristics, which are here discussed for the first time. Referring to a 41−nm NAND technology, results are validated on experimental data for different positions of the selected cell along the NAND string, different pass voltages (V_{pass}), and different array background patterns. Postcycling bake experiments are then used to investigate the impact of charge trapping/detrapping and interface state generation/ annealing on I_S, showing that the former damage mechanism dominates at low READ current levels, while interface states come into play for I_S values that are close to the string saturation level via mobility degradation.

3.2 Compact modeling for the string current

3.2.1 Test structure

Fig. 3.1(a) shows a schematics for cell connection in our 41 nm NAND array, organized into strings of 32 memory cells and two dummy cells to minimize edge WL effects [97]. Each string can be addressed by two select transistors, driven by the drain-select line (DSL) and the source-select line (SSL), which connect the string to the bit-line (BL) and to the source line (SL), respectively. Fig. 3.1(b)

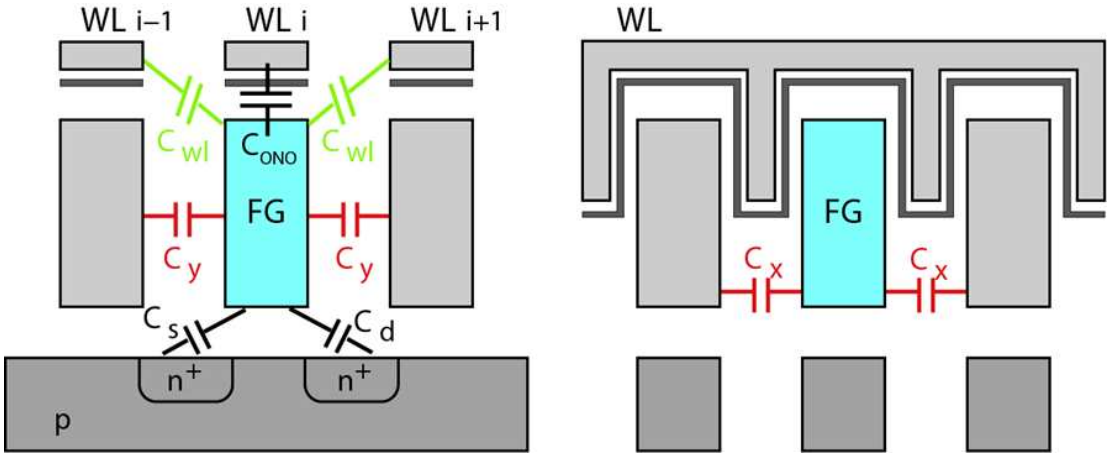


Figure 3.2: *Schematic cross section of a NAND array along (left) the bit line and (right) the WL directions, highlighting the main capacitive couplings of the floating gates.*

shows the bias conditions used in this paper for I_S investigation: a variable V_R is applied to the selected cell WL during READ, with not only all other WLs but also the DSL and SSL at V_{pass}. A $1-V$ bias was applied to the BL, with the SL grounded.

3.2.2 Electrostatics

Fig. 3.2 shows a schematic cross section of the NAND array along the BL and WL directions, highlighting the main capacitive couplings of the floating gates. Due to the close packing of the array cells, parasitic couplings between adjacent floating gates have a non-negligible impact on their potential, which for i-th cell along the string (V_{FG}^i), was computed from

$$\begin{aligned} V_{FG}^i = {} & \alpha_G V_{WL}^i + \alpha_S V_S^i + \alpha_D V_D^i + \frac{Q_{FG}^i}{C_{TOT}} + \\ & + \alpha_x \left(V_{FG}^{i,L} + V_{FG}^{i,R} \right) + \alpha_y \left(V_{FG}^{i+1} + V_{FG}^{i-1} \right) + \\ & + \alpha_{WL} \left(V_{WL}^{i+1} + V_{WL}^{i-1} \right) \end{aligned} \tag{3.1}$$

The coupling coefficients α_G, α_S and α_D in (3.1) are given by the capacitance between the floating gate and its WL, source and drain, respectively, divided by the total floating gate capacitance C_{TOT}. Similarly, α_x and α_y are the ratios between C_x and C_y (see Fig. 3.2) and C_{TOT}, accounting for inter-floating-gate coupling in the WL direction (with adjacent FGs at potentials $V_{FG}^{i,L/R}$) and in the BL direction (with adjacent FGs at potentials $V_{FG}^{i\pm1}$). The parameter α_{WL} accounts for the direct parasitic coupling between a floating gate and its adjacent WLs (at potentials $V_{WL}^{i\pm1}$). All the previous coupling coefficients were evaluated by TCAD simulations of a 3×3 cell structure.

The relation between V_{FG}^i and surface potentials ψ_s^i for the i-th cell is given

by [156, 159]:

$$V_{FG}^{i} = V_{FB} + \psi_s^i + \frac{\sqrt{2\epsilon_{\mathrm{Si}} q N_A (V_S^i + \psi_s^i)}}{C_{ox}} - \theta V_{DS}^i \tag{3.2}$$

where V_{FB} is the flatband voltage, C_{ox} the tunnel oxide capacitance, ϵ_{Si} the silicon dielectric constant, q is the electron charge, N_A is the substrate doping and V_{DS}^i is the drain-source voltage of the i-th cell. The fitting parameter θ models the DIBL effect and also partially accounts for channel-length modulation effects. The electrostatic threshold voltage $V_T^{FG,i}$ of the equivalent transistor, corresponding to the floating-gate bias bringing the cell into the strong-inversion regime, can be calculated from (3.2) assuming $\psi_s = 2\psi_B$, ψ_B being the distance of the Fermi level in the bulk and the V_S^i term allowing the body effect to be correctly included in our model.

3.2.3 Conduction and mobility model

The drain current of the i-th cell along the NAND string was calculated according to conventional MOSFET models for the strong-inversion ($I_{D,SI}^i$) and subthreshold regimes ($I_{D,sub}^i$) [160]:

$$I_{D,SI}^i = -W v_d(y) Q_{inv}^i \left(V_{FG}^i, V_T^{FG,i}, V^i(y), V_S^i\right) \tag{3.3}$$

$$\begin{aligned} I_{sub}^i =& \mu_{eff} \frac{W}{L} \sqrt{\frac{\epsilon_{Si} q N_a}{2\psi_s^i}} \left(\frac{kT}{q}\right)^2 \left(\frac{n_i}{N_a}\right)^2 \\ & \exp\left[\frac{q\left(\psi_s^i - V_S^i\right)}{kT}\right] \left(1 - \exp\left[\frac{-qV_{DS}^i}{kT}\right]\right) \end{aligned} \tag{3.4}$$

where W and L are the cell width and length, $v_d(y)$ and $V^i(y)$ are the electron velocity and the electron quasi-Fermi potential in the source-to-drain (y) direction, Q_{inv}^i is the electron charge in the channel, n_i is the intrinsic carrier concentration, and kT is the thermal energy. The total drain current was modeled as.

$$I_D^i = I_{D,SI}^i + I_{D,sub}^i \tag{3.5}$$

provided that $I_{D,sub}^i$ is limited in the strong-inversion regime as done in [156] to account for the pinning of surface potential.

To take into account v_d saturation at high longitudinal electric fields E_y, the following well-known relationship was used [160]:

$$v_d = \frac{\mu_{eff}}{\left[1 + \left(E_y \frac{\mu_{eff}}{v_{sat}}\right)^2\right]^{1/2}} E_y \tag{3.6}$$

where $v_{sat} = 8x10^6$ cm s^{-1} is the electron saturation velocity in MOSFET channels [160]. The dependence of the low-field electron mobility μ_{eff} in (3.6) on the effective normal electric field E_{eff} was described according to the empirical model [161, 162]:

$$\mu_{eff} = \frac{\mu_0}{1 + \left(\frac{E_{eff}}{E_0}\right)^{\nu}} \tag{3.7}$$

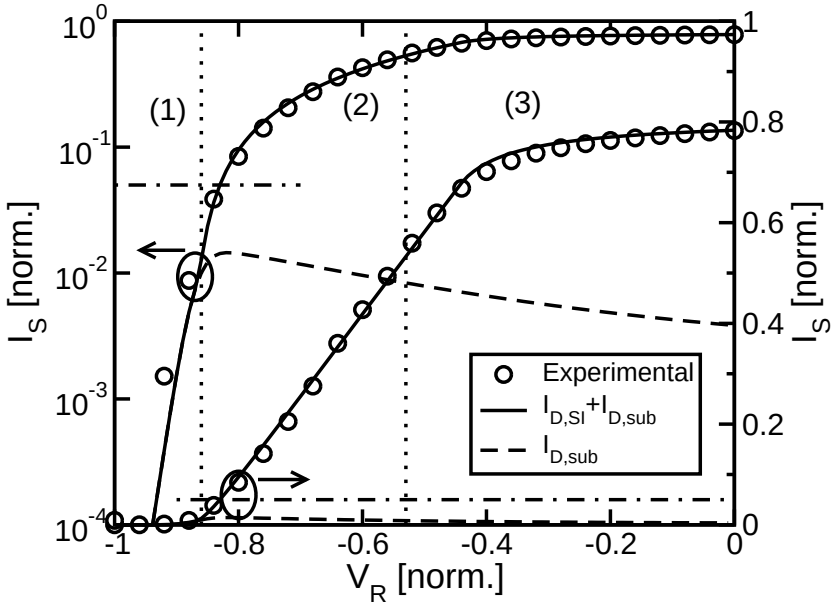


Figure 3.3: *(Symbols) Experimental $I_S - V_R$ curve compared to (lines) model calculations, plotted on linear and semilogarithmic scales. The conduction regimes of the selected cell are highlighted: (1) subthreshold; (2) velocity saturation; and (3) ohmic.*

where μ_0, E_0 and ν were treated as fitting parameters. the validity if (3.7) is limited to E_{eff} values high enough to neglect the Coulomb scattering contribution to mobility.

3.2.4 Simulation scheme for the NAND string

For fixed V_{pass}, V_R and Q^i_{FG}, string electrostatics and conduction can be determined by requiring $I^i_{DS} = I_{string}$, since the current is constant along the string. This allows to determine not only I_S but also V^i_{FG}, V^i_S and V^i_D. The select transistors current was calculated with the same model used for the cell, directly applying V_{pass} as the floating-gate bias of the devices and using their respective W and L. Particular attention was devoted to parameter calibration (θ, μ_0, E_0 , ν, Q_{FG}) for the selected cell via comparison against experimental data, as discussed in the next sections.

3.3 Experimental results

3.3.1 Parameter extraction

Fig. 3.3 shows the experimental $I_S - V_R$ curve obtained by sweeping aV_R on WL15, with $V_{pass} = 8.5$ V and all the cells in the erased state (all the voltage axes and all the current axes have been normalized to the same constant throughout this chapter). The compact model of Section 3.2 can nicely reproduce the experimental data both on the linear and on the logarithmic current scales, allowing a detailed analysis of the curve. At low V_R (region 1 in Fig. 3.3), the selected cell on WL15 is in the subthreshold regime, increasing its current exponentially until region 2

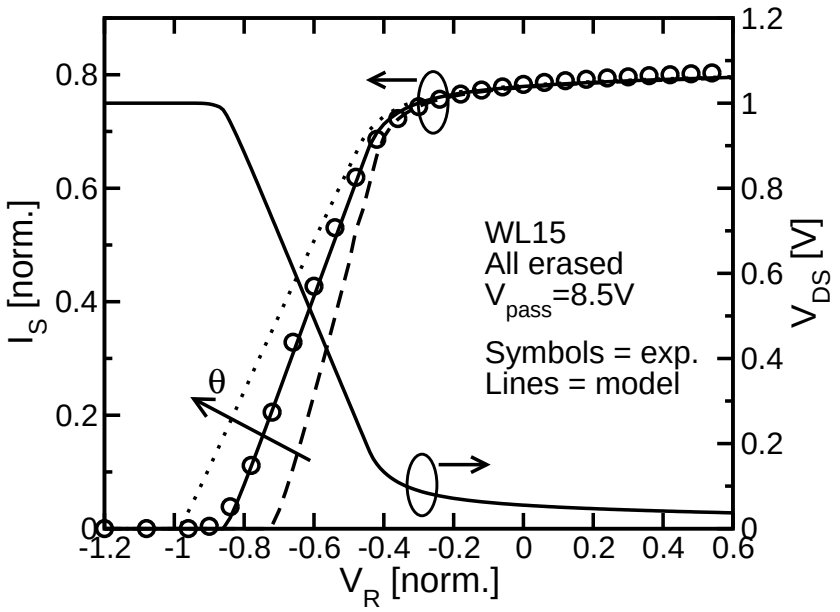


Figure 3.4: *Experimental data and model calculations for different values of the DIBL parameter θ. Calculated V_{DS} of the selected cell is also shown.*

is reached, determining the onset of strong inversion. In this latter region, the selected cell works in velocity saturation conditions, displaying a nearly linear increase of I_S with V_R. Finally, for high V_R, the cells move to the ohmic regime (region 3), where I_S rapidly reaches a saturation value I_S^{sat} determined by the channel resistance of the unselected cells and select transistors, which are always in the ohmic regime. The different contributions to I_S in (3.5) are also shown, revealing that the weak-inversion current has already become negligible at the current levels (dashed-dotted lines in the figure, corresponding to a normalized $I_{read} = 0.05$) at which cell threshold voltage (V_T) is typically extracted in NAND devices.

The comparison with experimental data allows calibrating the fitting parameters of our model. For example, Fig. 3.4 shows three $I_S - V_R$ curves computed with different values of the DIBL coefficient θ, showing its nonnegligible impact on both V_T and string transconductance g_m. To explain this result, V_{DS} for the selected cell is also shown in the figure: when the selected cell is in the subthreshold regime, V_{DS} equals the applied BL bias since there is a negligible voltage drop on the passing cells; however, this drop increases as the selected cell enters the velocity saturation and then the ohmic regime, reducing V_{DS} as I_S increases. This means that the DIBL correction to the electrostatic V_T^{FG} depends on I_S, hence affecting the string g_m.

Fig. 3.5 shows the low-field mobility μ_{eff} as a function of Q_{inv} as resulting from our fit of the experimental data on our 41 nm NAND strings ($\mu_0 = 450$ cm^2/Vs, $E_0 = 0.45$ MV/cm, $\nu = 3$). The extracted μ_{eff} is significantly lower than what was reported for long channel transistor [161], but is in good agreement with results obtained on transistors of similar gate length [163, 164]. This confirms the detrimental effect of L scaling on mobility already highlighted on MOSFETs in

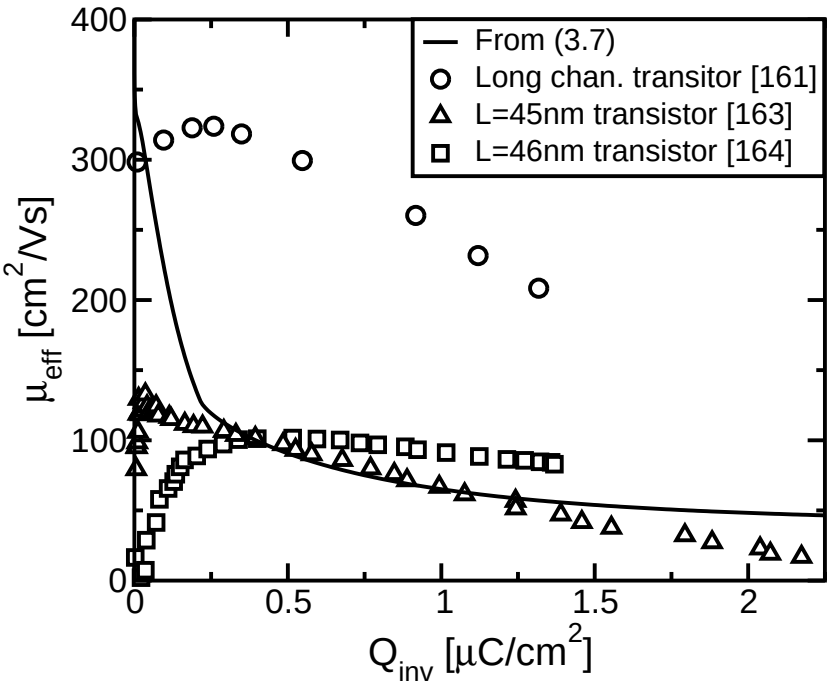


Figure 3.5: *Electron mobility for our* 41 *nm memory cells as obtained from (3.7), compared with experimental data for long- and short-channel transistors.*

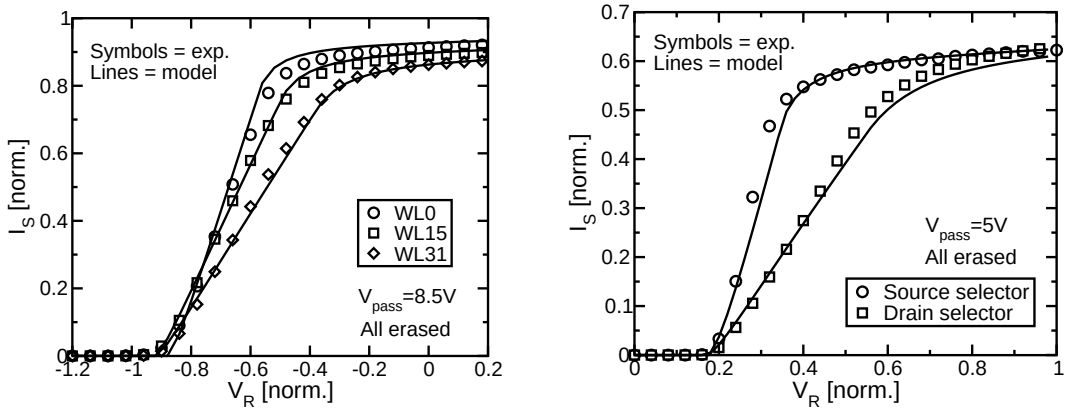
the sub-100 nm devices.

3.3.2 Model validation

An additional assessment of the model validity was carried out against experimental data obtained modifying the WL of the selected cell, V_{pass}, and the array background pattern. In all these cases, fitting parameters were kept constant and only the floating-gate charge of the selected cell was adjusted to take into account slight variations due to the program operation or cell variability. Fig. 3.6(a) shows $I_S - V_R$ curves for different selected WLs along the NAND string (all the cells are in the erased state and $V_{pass} = 8.5$ V. Neither significant horizontal shifts of the curve nor changes in I_S^{sat} can be observed, although a strong variation in g_m is detected: moving from WL0 to WL31, in fact, the number of unselected cells at the source side and, thus, the equivalent source resistance R_S of the selected cell increases, determining a g_m degradation [85]. This effect is far more evident in Fig.3.6(b), where the $I_S - V_R$ curves of the select transistors reveal the strongest variation in g_m, due to the minimum and the maximum R_S in the string.

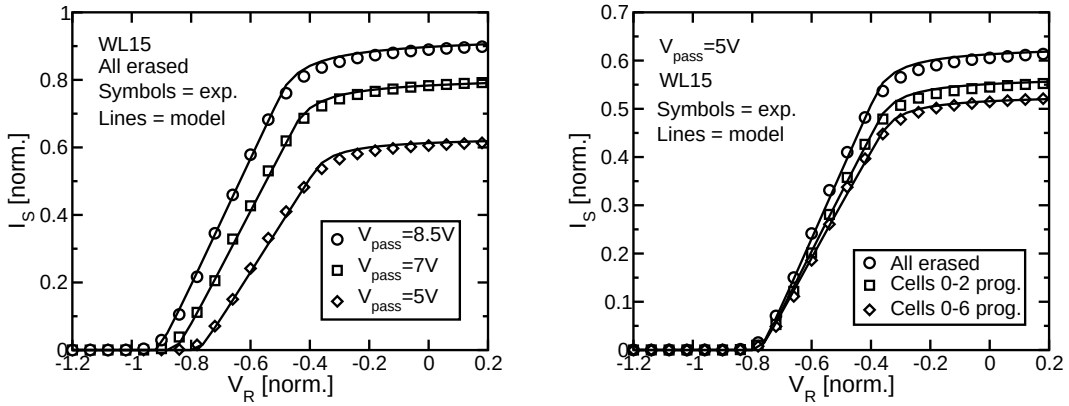
Fig. 3.7(a) shows the impact of V_{pass} on the $I_S - V_S$ curve when WL15 is selected. When V_{pass} increases from 5 V to 8.5 V, I_S^{sat} and g_m increase due to the reduction of the channel resistance of the unselected cells. In addition to that, a shift of the $I_S - V_R$ characteristics toward negative V_R can be clearly seen; this is the result of parasitic coupling between the selected cell floating-gate and the adjacent WLs ($WL_{i\pm1}$ in Fig. 3.2) via C_{WL} and C_y.

Finally, Fig. 3.7(b) shows the $I_S - V_R$ curves obtained for WL15 selected when



(a) (b)

Figure 3.6: *$I_S - V_R$ curves for (a) cell at different WLs in the string and for (b) source and drain select transistors. All the other cells in the string are in the erased state and (a) $V_{pass} = 8.5$ V, (b) $V_{pass} = 5$ V.*



(a) (b)

Figure 3.7: *$I_S - V_R$ curves for WL15 selected and: (a) different V_{pass}'s (all cells in the string are in the erased state); (b) different background patterns of the string (all cells erased, cells from WL0 to WL2 programmed and cells from WL0 to WL6 programmed). $V_{pass} = 5$ V.*

an increasing number of cells along the string are programmed. The increase of the channel resistance of the unselected cells that are moved to high-V_T degrades both I_S^{sat} and g_m. No horizontal shift of the curve appears, instead, from the figure, as the cells that are moved to high-V_T are not adjacent to the selected one, therefore preventing parasitic interference effects among their floating-gates. These effects clearly appear, instead, in Fig. 3.8, where the $I_S - V_R$ characteristics are shown in the case cells belonging to the string at the left and at the right of the selected one are in the erased state and in the programmed state: a horizontal shift of the curve, with no degradation of g_m and I_S^{sat}, appears in this case as a result of the change in the array background pattern, due to the parasitic capacitances C_x leading to a change in the selected floating-gate potential (in deeply-scaled technologies, even a direct coupling capacitance between adjacent floating-gates

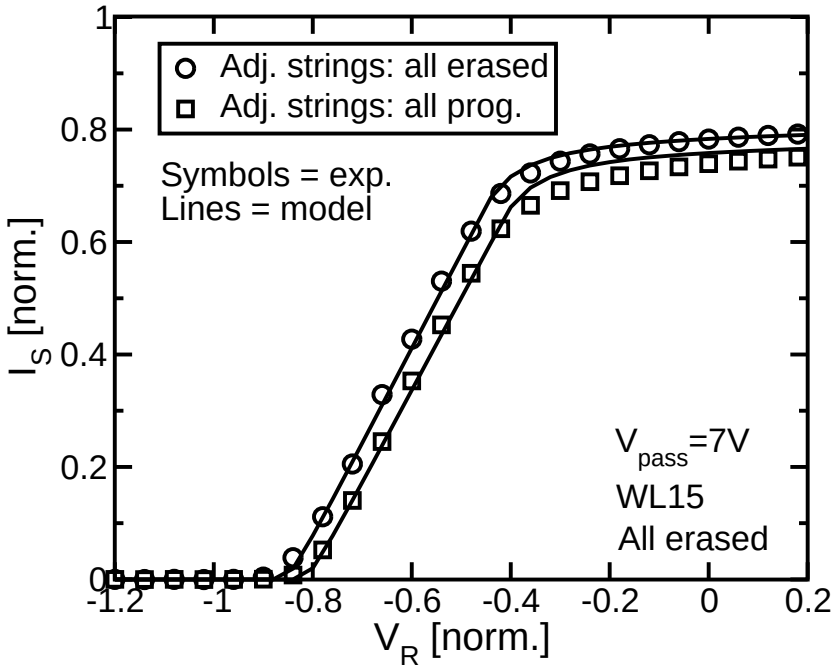


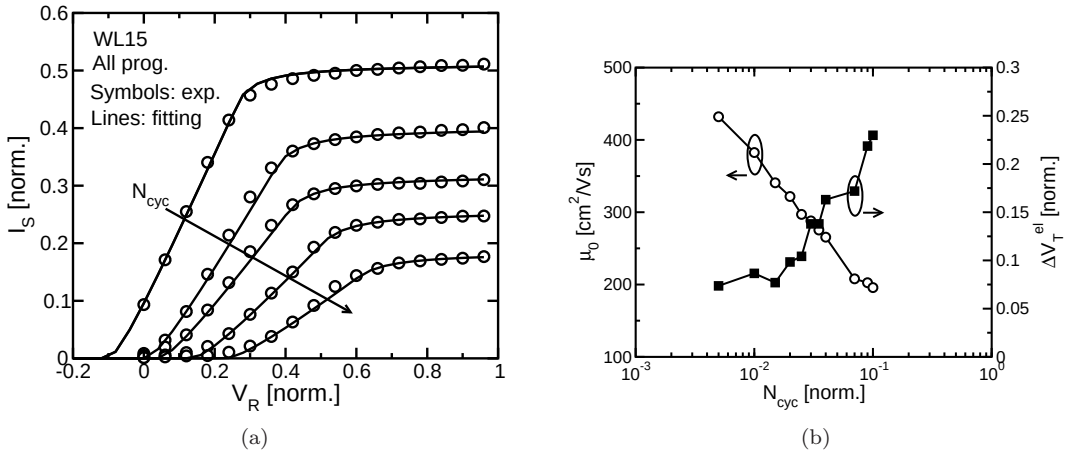
Figure 3.8: *$I_S - V_R$ curves for WL15 when all the cells of the adjacent strings are in the erased or in the programmed state. All the cells in the selected string are erased, and $V_{pass} = 7$ V*

and the selected string active area may contribute to this shift; see [154]).

3.4 Reliability analysis

The NAND string compact model was employed for an investigation of V_T instabilities in cycled devices. V_T instabilities are a major concern for array reliability [58, 60, 62–65, 67, 165] and mainly arising from charge trapping/detrapping into/from the tunnel oxide and interface state generation/annealing. Fig. 3.9(a) shows that cycling has three main effects on the experimental $I_S - V_R$ curve: a reduction of the transconductance and saturation level and a positive shift of the curve [60,62]. These effects are the result of cycling-induced charge trapping in the tunnel oxide and interface states generation, impacting V_T^{FG}, the efficiency of the electron transfer to/from the floating gate during program/erase and degrading carrier mobility in the channel. To discriminate between all these, we first fitted the results of Fig. 3.9(a), resulting in the μ_0 and V_T^{FG} dependence on the number of cycles N_{cyc} shown in Fig. 3.9(b); all other parameters remained constant.

For cycling-induced V_T instabilities investigation, postcycling bakes at different temperatures T_B's were performed on programmed arrays, collecting $I_S - V_R$ curve of the selected cell after increasing bake intervals t_B, as shown in Fig. 3.10. A negative shift of the curve and a recovery of g_m and I_S^{sat} are observed as t_B elapses, these effects being more pronounced for higher T_B and revealing that the underlying physical phenomena are thermally activated. Experimental data were fitted by the NAND string compact model by changing μ_0, to take into account mobility degradation/recovery dynamics, and introducing a term ΔV_T^{FG}, generi-



(a) (b)

Figure 3.9: *(a) $I_S - V_R$ curve for a cell at word-line* 15 *for different cycling conditions (all cells are in the programmed state, $V_{pass} = 8$ V); and (b) μ_0 and V_T^{el} as a function of the number of program/erase cycles, extracted from $I_S - V_R$ fitting.*

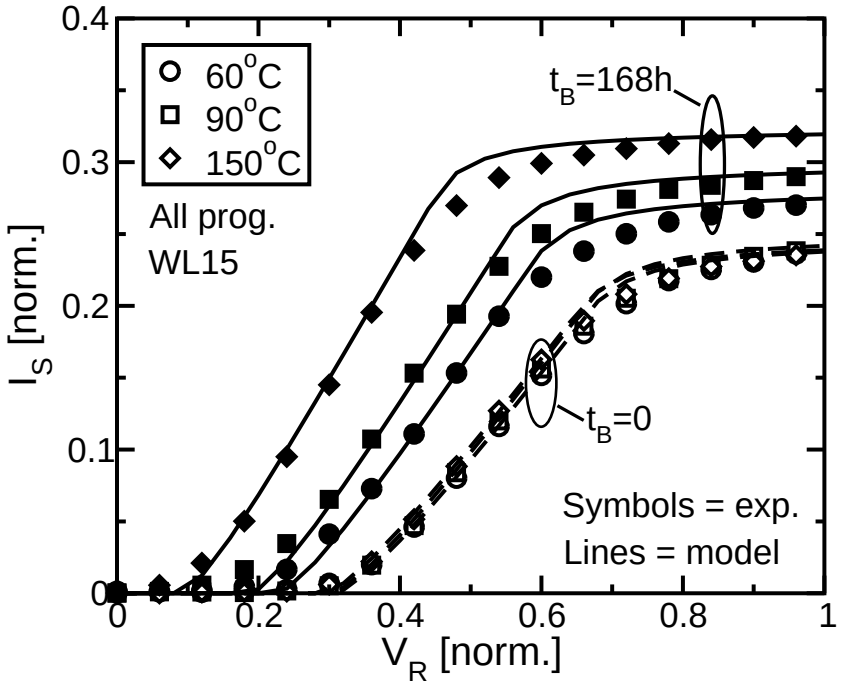
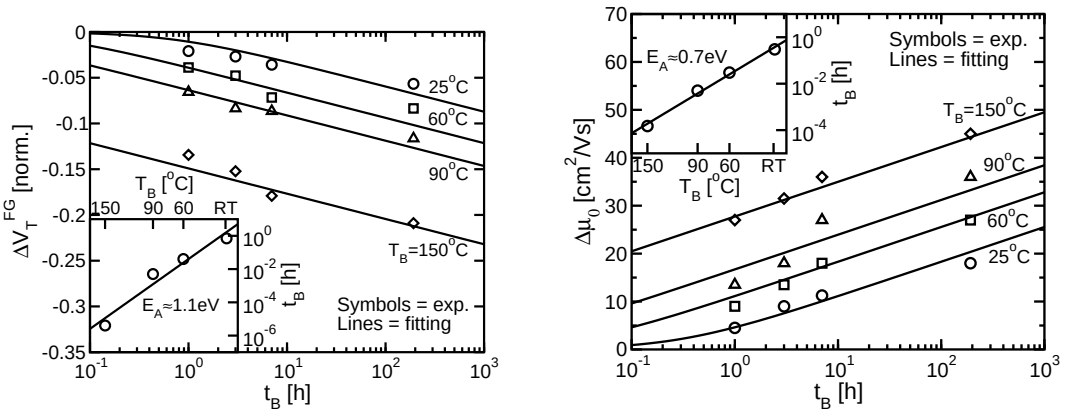


Figure 3.10: *$I_S - V_R$ curves for WL15 selected measured on a cycled device before and after several* 168 *h bakes performed at different T_B's. All the unselected cells are programmed, with $V_{pass} = 8$ V.*



(a) (b)

Figure 3.11: *(a) ΔV_T^{FG} and (b) μ_0 obtained by fitting experimental data during postcycling bakes for T_B ranging from room temperature to 150° C. The corresponding Arrhenius plot is shown in the inset.*

cally accounting for a net charge loss from the tunnel oxide. Their dependence on t_B and T_B is given in Figs. 3.11(a) and 3.11(b): both exhibit a logarithmic dependence on t_B and activation energies $E_A \simeq 1.1$ and 0.7 eV, respectively (see the Arrhenius plots in the insets). $E_A \simeq 1.1$ eV is the typical value reported for charge detrapping [63, 64, 67, 165], revealing that this damage mechanism is the main responsible for the extracted ΔV_T^{FG}. The lower activation energy obtained for mobility recovery suggests, instead, that μ_0 variations are mainly due to interface state generation/annealing (note that, in principle, the μ_0 transients in Fig. 3.11(b) may be quantitatively affected by charge detrapping phenomena over the source/drain regions, changing the parasitic resistance of these regions during the bake experiments; these phenomena should, however, display an activation energy of 1.1 eV, similar to charge detrapping over the channel). For interface state annealing, in fact, $E_A \simeq 0.52$ eV was reported in sub-90 nm Flash technologies [166], while a value of 0.55 eV can be extracted from the data in [62], relative to the annealing of interface state density (N_{it}) extracted from the subthreshold slope on MOSFET's (see Fig. 3.12 and inset).

The impact of mobility recovery due to interface state annealing on postcycling instability of constant-current V_T was investigated monitoring the V_T loss during bakes at the normalized current levels $I_R = 0.05$ and for $I_R = 0.15$. The V_T transients are shown in Fig. 3.13, where $E_A \simeq 1.1$ eV is obtained for $I_{read} = 0.05$ (see the Arrhenius plot in the inset), confirming that the dominant failure mechanism at low string current levels is charge detrapping from tunnel oxide [67, 165]; for $I_R = 0.15$, instead, a lower activation energy $E_A \simeq 0.9$ eV is extracted, revealing that mobility recovery affects V_T when this is extracted at high I_R, due to the increase of the g_m of both the selected and the unselected cells. It should be pointed out that $I_R = 0.15$ can be considered an unrealistic value for typical NAND operation but is representative of a condition that can be reached when I_S^{sat} is reduced as a consequence of V_{pass} lowering, programming to higher V_T of unselected cells or increasing the number of cells in the string.

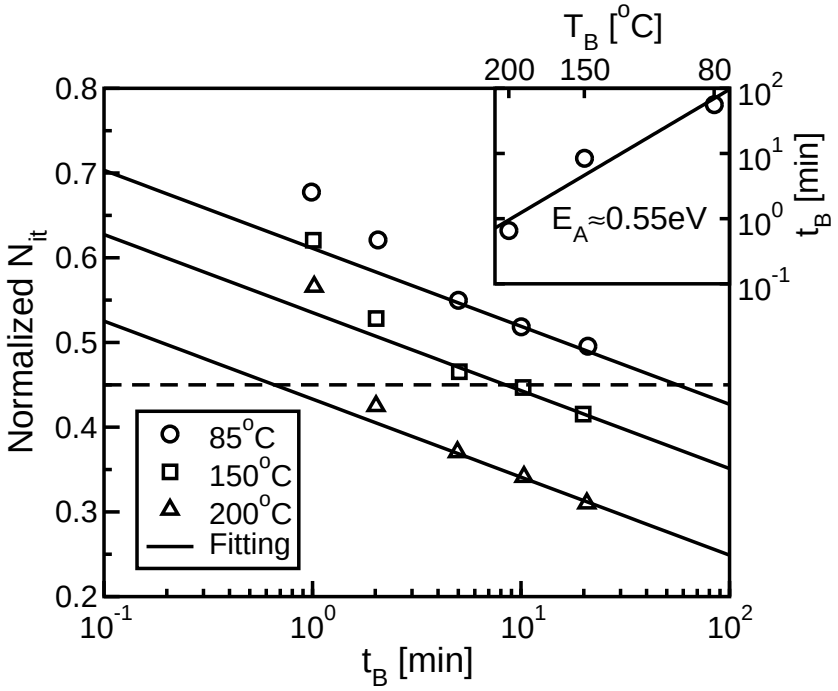


Figure 3.12: *Interface annihilation results measured on MOSFETs [62] at different temperatures, and (see inset) corresponding Arrhenius plot.*

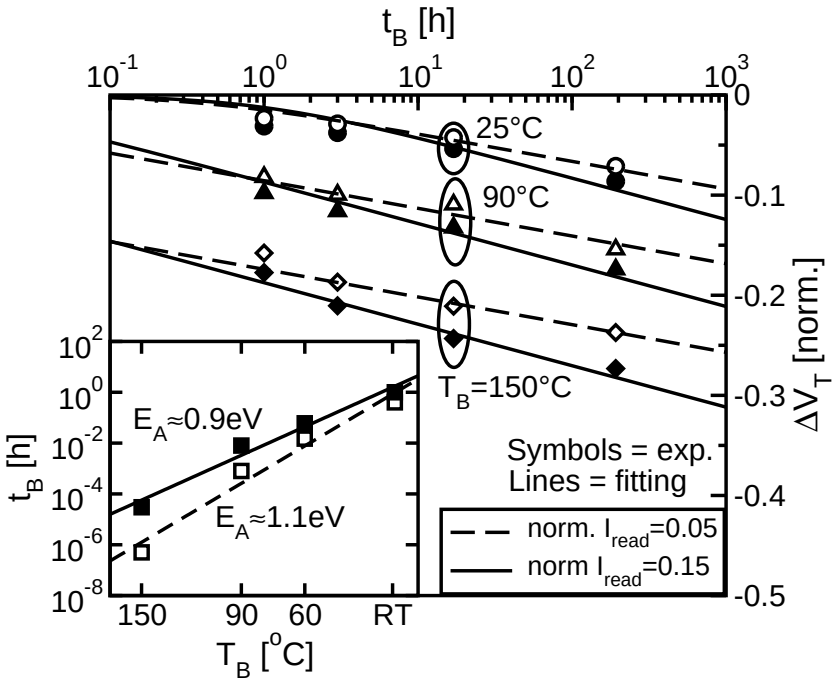


Figure 3.13: *ΔV_T transients during postcycling bakes at various T_B's for WL15 selected, evaluated at normalized $I_R = 0.05$ and $I_R = 0.15$. The unselected cells are programmed and biased with $V_{pass} = 8$ V. The inset shows the Arrhenius plot for data retention.*

3.5 Conclusions

This chapter presented a detailed compact-modeling investigation of the $I_S - V_R$ characteristics in decananometer NAND strings Flash arrays. The role of velocity saturation, low-field mobility and DIBL on I_S were discussed, validating the results on experimental data for a 41−nm technology. Then, the compact model was used to analyze the contribution of charge trapping/detrapping and interface states on I_S and on cycling-induced V_T instabilities, revealing that the latter comes into play when cell V_T is monitored at high I_R, due to carrier mobility degradation. Results show that V_T instabilities are increased and their activation energy is lowered with respect to the usual 1.1 eV value given by charge detrapping whenever the saturation value of I_{string} moves too close to I_{read}.

CHAPTER 4

Distributed-cycling schemes for NOR Flash arrays

This chapter investigates the validity of distributed-cycling schemes on scaled Flash memory technologies. These schemes rely on the possibility to emulate on-field device operation by increasing the cycling temperature according to an Arrhenius law, but the assessment of the activation energy that has to be used on scaled technologies requires a careful control of the experimental tests, preventing spurious second-order effects to emerge. In particular, long gate-stresses required to gather the array threshold voltage (V_T) map are shown to give rise to parasitic V_T-drifts, which add to the V_T-loss coming from damage recovery during post-cycling bake. When the superposition of the two phenomena is taken into account, the effectiveness of the conventional qualification schemes relying on a 1.1 eV activation energy is fully confirmed at the 45 nm NOR node.

4.1 Introduction

Floating-gate devices are the most widely used non-volatile semiconductor memories, thanks to their high density and good performance, in terms of throughput and data retention [5, 167, 168]. The demand for ever higher storage densities has led to a continuous cell scaling and to the development of multi-level (MLC) technologies [169–171], leading to an increased sensitivity to spurious charges trapped in the dielectrics (e.g., a single charge trapped in the tunnel oxide can give a V_T-shift of 100 mV in state-of-the-art NAND cells [5]). For this reason, V_T instabilities during post-cycling bakes emerged as one of the major reliability concerns for decananometer Flash technologies [63, 64, 67, 165] and great efforts have been spent to develop adequate qualification tests [172–174]. These instabilities are due to charge trapping/detrapping into/from the tunnel oxide and interface states an-

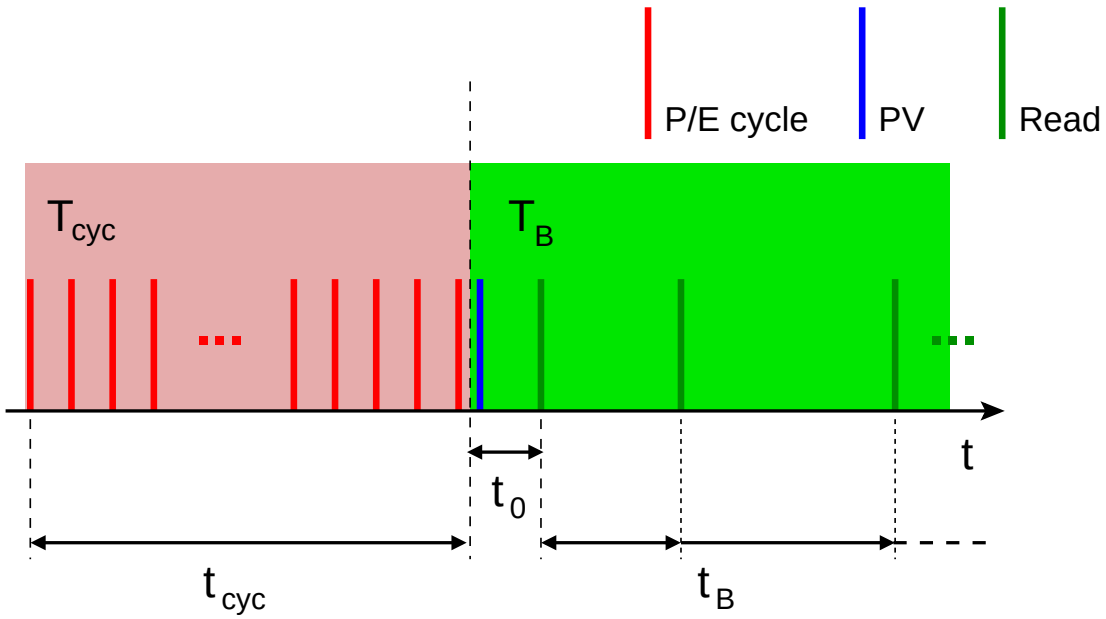


Figure 4.1: *Schematics for the experimental procedure used to investigate V_T-loss during post-cycling bakes.*

nealing [55–67, 165], representing a partial recovery of the previously generated cycling-induced cell damage.

In order to correctly reproduce the real on-field usage of the devices, distributed-cycling schemes have been developed for the characterization of post-cycling V_T-instabilities [63, 64, 67, 165, 172–174]. These schemes aims at taking into account the partial damage recovery that takes place in the time interval between the P/E cycles, which plays a non-negligible role in the real on-field operation of the device. To save experimental time, this can be obtained by increasing the cycling temperature according to an Arrhenius law whose activation energy corresponds to that of the damage recovery processes, typically $E_a = 1.1$ eV [63, 64, 67, 165, 172–174]. In so doing very long cycling experiments can be equivalently reproduced in short time scales, making the qualification tests more practical.

This chapter presents a detailed assessment of the validity of distributed-cycling schemes on the 45 nm NOR technology node, showing that the extraction of the activation energy for damage recovery requires a careful control of experimental tests. Monitoring the average V_T-loss of a cycled memory array at different bake temperatures, can be shown, in fact, that an apparent activation energy $E_{aa} \simeq 0.8$ eV, lower than the typical 1.1 eV value, is obtained. Comparing data retention on fresh and cycled devices, this is explained in terms of an additional contribution to V_T instabilities coming from the long gate-stress times present in the adopted experimental procedure, activating spurious charge displacements in the oxide/nitride/oxide (ONO) stack. Provided that all phenomena are correctly taken into account, the validity of distributed-cycling schemes is fully confirmed at the 45 nm NOR node, with $E_a = 1.1$eV. Finally, the parameter values re-

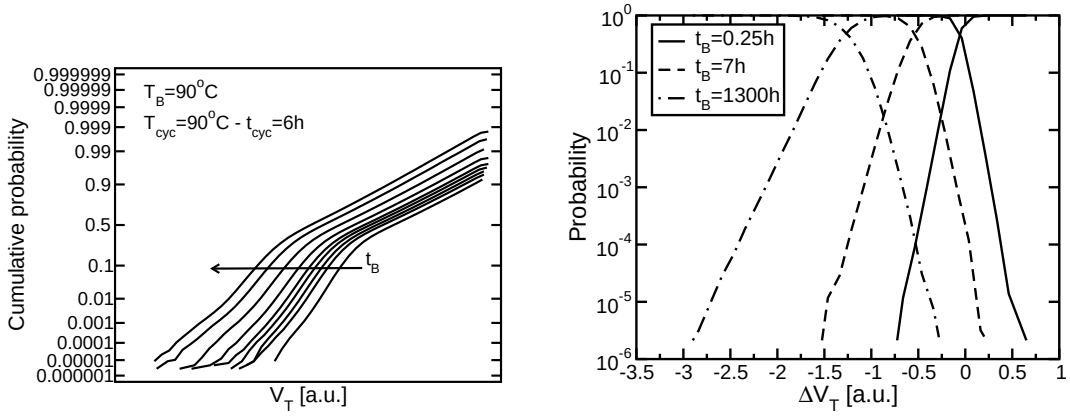


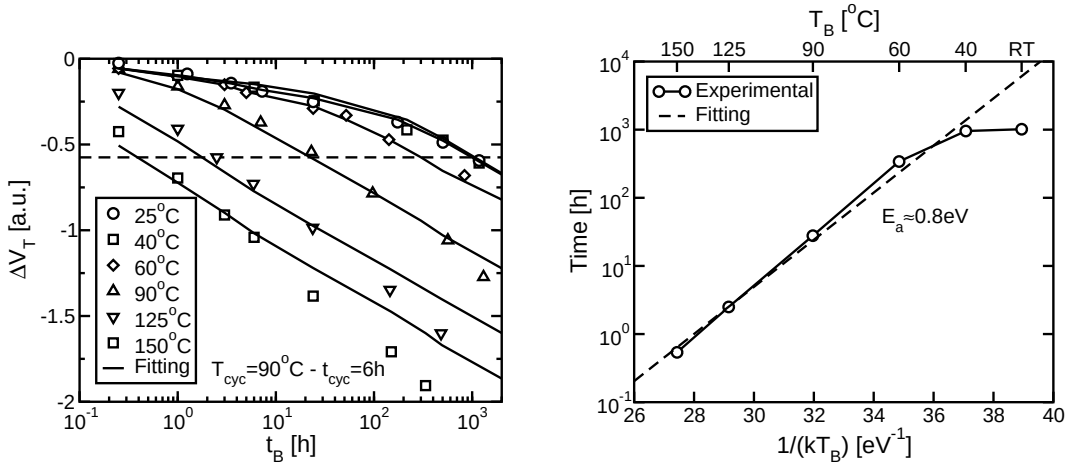
Figure 4.2: *(a) Array V_T cumulative distribution for increasing bake times t_B at T_B=90^{o}C, after cycling with T_{cyc}=90^o C, t_{cyc}=6 h; and (b) ΔV_T cumulative distributions F and $1-F$ calculated between V_T maps acquired at increasing bake times t_B and the V_T map at the beginning of the bake ($t_B = 0$ h).*

quired for a universal damage-recovery metric (UDM) for the 45 nm NOR node are extracted.

4.2 Apparent activation energy for damage recovery

The experimental procedure adopted to investigate V_T instabilities during post-cycling bakes on 45 nm NOR test chips is schematically depicted in Fig. 4.1 [67] and consists of three phases: (1) a stress phase of duration t_{cyc}, during which N_{cyc} P/E cycles are performed on a memory block at a constant temperature T_{cyc}; (2) a single program-and-verify (PV) operation required to bring the array to the desired V_T level [75, 150, 175]; (3) a bake phase performed at a constant temperature T_B (which can be different from T_{cyc}) during which the V_T map is periodically collected at increasing bake times t_B (in Fig. 4.1, t_0 represents the delay between the end of cycling and the first map acquisition). Both the PV operation and V_T map acquisitions are performed at room temperature (RT). T_{cyc} ranging from RT up to 100^{o}C and T_B from RT up to 150^{o}C were explored. As a final remark, in order to maximize the electrical stress, a uniform cycling pattern was used, moving the cells from the erased to the highest programmed level available in the multilevel device under test.

Fig. 4.2(a) shows the array V_T cumulative distribution during a T_B=90^{o}C bake after 3k cycles with T_{cyc} = 90^{o}C, t_{cyc} = 6 h. All the cells were previously programmed to the highest V_T level of the multi-level chip in order to maximize the V_T instability observed during bake [165]. A negative shift of the distribution appears as the bake time elapses due to partial damage recovery, with an increase of the distribution spread [63, 64, 67, 165]. It should be pointed out that the adopted experimental procedure allows to determine not only the V_T cumulative distribution, but also the entire array V_T map, giving the possibility to perform a wider range of data analyses. Fig. 4.2(b), for instance, shows the ΔV_T cumulative dis-



(a) (b)

Figure 4.3: *(a) Average ΔV_T of cycled arrays during bakes at different T_B, after the same cycling experiment (T_{cyc}=90^{o}C, t_{cyc}=6 h); and (b) Arrhenius plot obtained from the time needed to reach a selected V_T-loss (dashed line in (a))*

tributions obtained from the same data of Fig. 4.2(a) when making a cell-to-cell comparison of V_T at time t_B and at time 0: the ΔV_T distribution shifts toward negative values, consistently with Fig. 4.2(a), with clear exponential tails due to RTN [69, 71, 73, 78, 80, 82] enlarging as time proceeds. Despite this effect will not be discussed here, it appears quite clear that the acquisition of the array V_T map allows a more complete reliability analysis.

In Fig. 4.3(a), the ΔV_T transients at a probability level $p = 50\%$ are reported as a function of t_B for different T_B from RT to 150^{o}C, but for the same cycling conditions (3k cycles, T_{cyc} = 90^{o}C, t_{cyc} = 6 h). The time required to reach a selected V_T-loss (dashed line in the figure) reduces for higher T_B, leading to the Arrhenius plot for data retention of Fig. 4.3(b). An apparent activation energy $E_{aa} \simeq 0.8$ eV appears in the 40^{o}C - 150^{o}C temperature range, which is lower than the typical 1.1 eV assumed by standard distributed-cycling tests [63, 64, 67, 165, 172–174] and questioning their validity for state-of-the-art NOR technologies.

4.3 Parasitic gate-stress

In order to further check the validity of the activation energy evaluated in Fig. 4.3(b), the impact of the periodic acquisition of the array V_T map on the results has been carefully investigated. In the adopted experimental set-up the V_T measurement of each cell in the array requires a staircase voltage waveform to bias the selected cell word-line (WL), with the bit-line (BL) at constant bias and WLs and BLs of unselected cells grounded. The V_T of the selected cell is defined as the minimum WL bias resulting in a cell drain current higher than a read current level I_{read}. To obtain the desired resolution for the V_T sensing, a staircase with a very tight step, equal to 20 mV was used: this means that the acquisition of the V_T map causes very long gate-stresses, unlikely encountered during on-field array operation. In

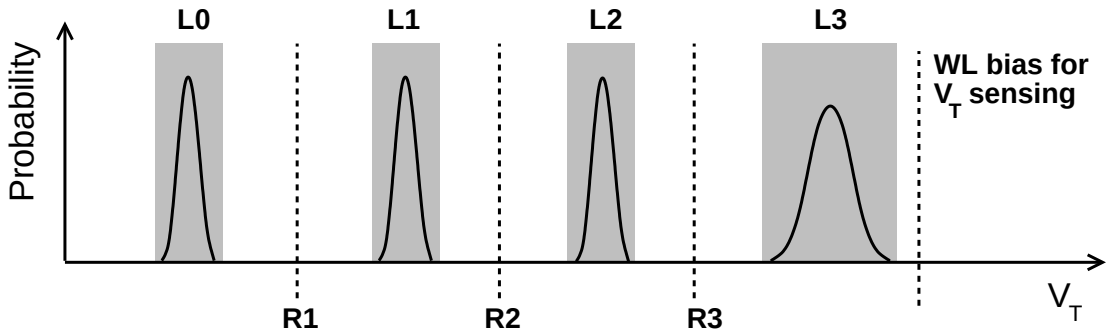


Figure 4.4: *Schematics of cell V_T distribution in the multi-level cell: 4 different V_T levels (L0 - L1) allow to store 2bit/cell. V_T comparison against 3 read levels R1 - R3 are required to discriminate the cell state.*

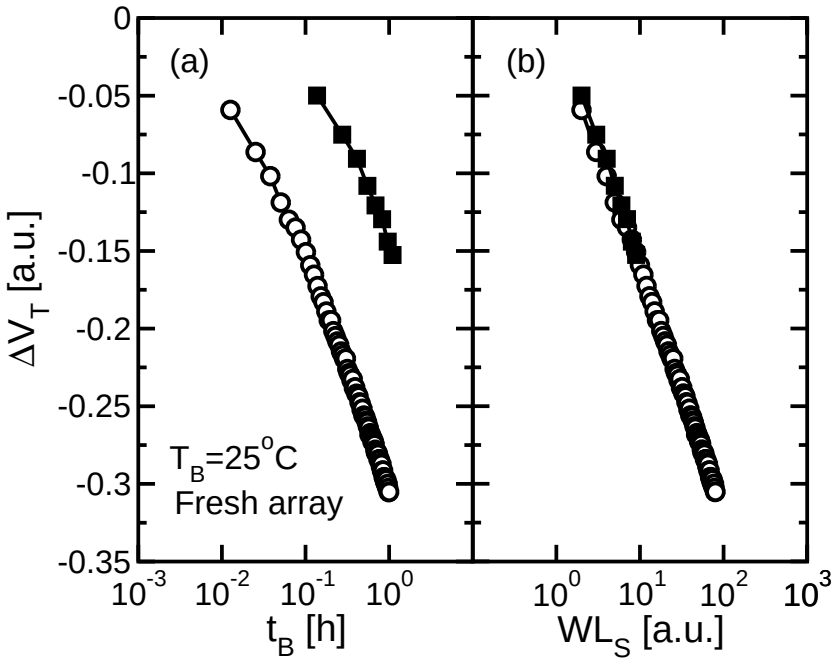
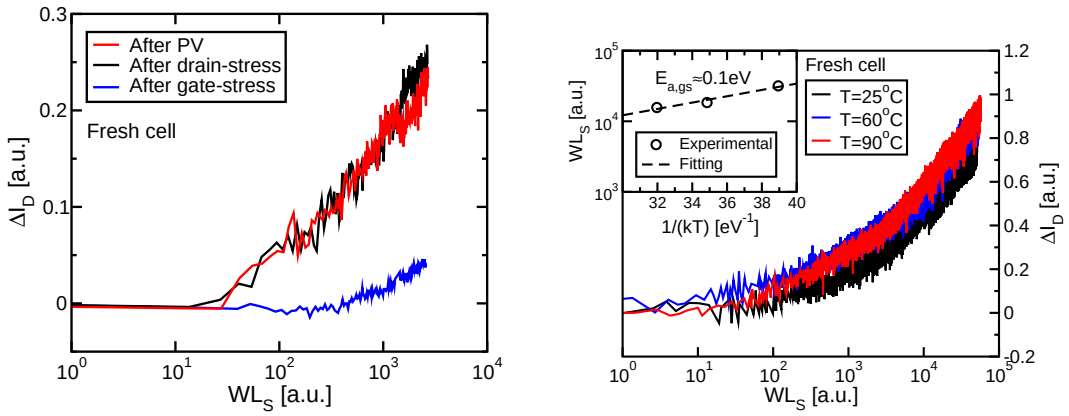


Figure 4.5: *Average ΔV_T transients on fresh arrays (no cycling prior to the bake) for many (circles) and few (squares) V_T map acquisitions during bake, as a function of bake time t_B (a) and WL stress time WL_S (b).*



(a) (b)

Figure 4.6: *(a) ΔI_D of a typical cell: dc current sensing was performed just after the PV operation, after a drain-stress or after a gate-stress; and (b) ΔI_D on a typical fresh cell during dc current sensing, for different temperatures (inset shows the $E_{a,gs}$ extraction).*

addition to this, and most important, it should be noted that during on-field operation a cell in the highest V_T level is not turned-on during read, while this happens in the experimental test: this is schematically shown in Fig. 4.4, where the V_T distributions of a 2bit/cell MLC technology are shown: during on-field operation, cell V_T is compared against 3 read levels R1 - R3 but, in the experimental procedure, a WL stress higher than R3 is required to measure V_T of cells belonging to L3. Thus, a careful investigation of the impact of stress field during V_T acquisition is mandatory [176–180].

Disturbs arising from the V_T map acquisition were evaluated by performing bake tests on fresh arrays, simply omitting the cycling phase in the experimental procedure depicted in Fig. 4.1. Fig. 4.5a shows the ΔV_T transients obtained from two different bake experiments at RT: in the first case, V_T maps were collected continuously at the highest rate allowed by the experimental set-up in a time $t_B = 1$ h; in the second case only 1/10 V_T maps were collected at logarithmically-spaced intervals up to $t_B = 1$ h: in both cases long gate-stresses required for the V_T map acquisition revealed themselves as a source of parasitic V_T-drifts. This confirms that gate-stresses required to gather the V_T maps and not bakes are the cause of the observed parasitic V_T-drift. This appears more clearly when comparing the ΔV_T transients as a function of the WL stress time WL_S (Fig. 4.5b), where the transients collapse on the same curve.

In order to further investigate the dynamics of the gate-stress induced parasitic V_T-drifts, Fig. 4.6(a) shows the results of a dc sensing of I_D on a single array cell. Results display an increase of I_D as WL_S elapses, in agreement with the V_T reduction of Fig. 4.5. Moreover, the I_D transient does not change when a drain-stress is applied after the PV operation and prior to the I_D sensing, while large reductions of it appear when a gate-stress is inserted. This result confirms that the V_T-drift of Fig. 4.5 and the ΔI_D of Fig.4.6(a) arise from the long gate-stress required to gather the V_T map and not from the drain bias. In addition, Fig. 4.6(b) shows that the I_D-drift during dc current sensing is rather temperature

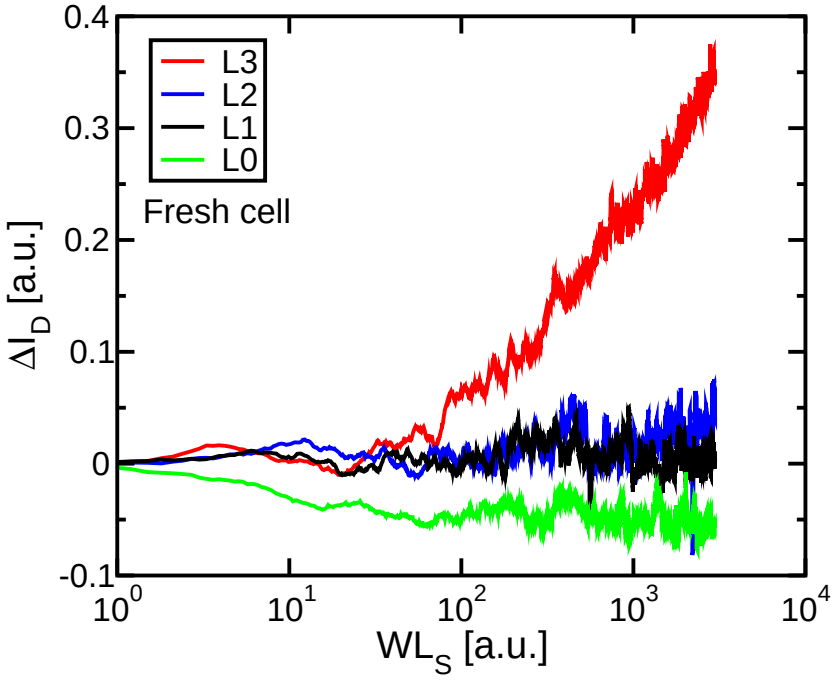


Figure 4.7: *ΔI_D on a typical fresh cell, programmed to the levels L0 - L3. The initial value of I_D is almost the same in all the measurements.*

independent in the range from RT to $T = 90^oC$, resulting in a quite low activation energy $E_{a,gs} \simeq 0.1$ eV (see the Arrhenius plot in the inset).

In order to investigate the dependence of the V_T drift on the program level, Fig. 4.7 shows the ΔI_D transients obtained on a typical cell programmed to the levels L0 - L3, changing the WL bias to obtain the same initial I_D: a strong dependence on the program level is observed, revealing that the highest I_D instabilities are detected when the cell is on the L3 level.

From the results shown in Figs. from 4.6(a) to 4.7, the observed V_T-drifts on fresh arrays can be attributed to small charge displacements in the ONO interpoly dielectric during positive gate-stresses. In particular, the nitride layer of the ONO stack may act as a trapping layer for electrons during cell operation, increasing cell V_T [180, 181]. During positive gate stresses, these trapped electrons can slightly move inside the nitride, getting closer to the control gate and, thus, leading to a negative V_T-shift, as previously reported for 90 nm NOR technologies [182]. Note that Fig. 4.7 excludes the possibility that charge displacements take place in the tunnel oxide because, at the beginning of I_D sensing, the electric field in the tunnel oxide is the same for the L0 - L3 curves.

4.4 Activation energy assessment

In order to assess the activation energy of post-cycling damage recovery and the validity of conventional distributed-cycling schemes, the impact of the parasitic gate-stress discussed in Section 4.3 for fresh devices should be evaluated also on cycled devices. Fig. 4.8 compares the average V_T transient of fresh and cycled

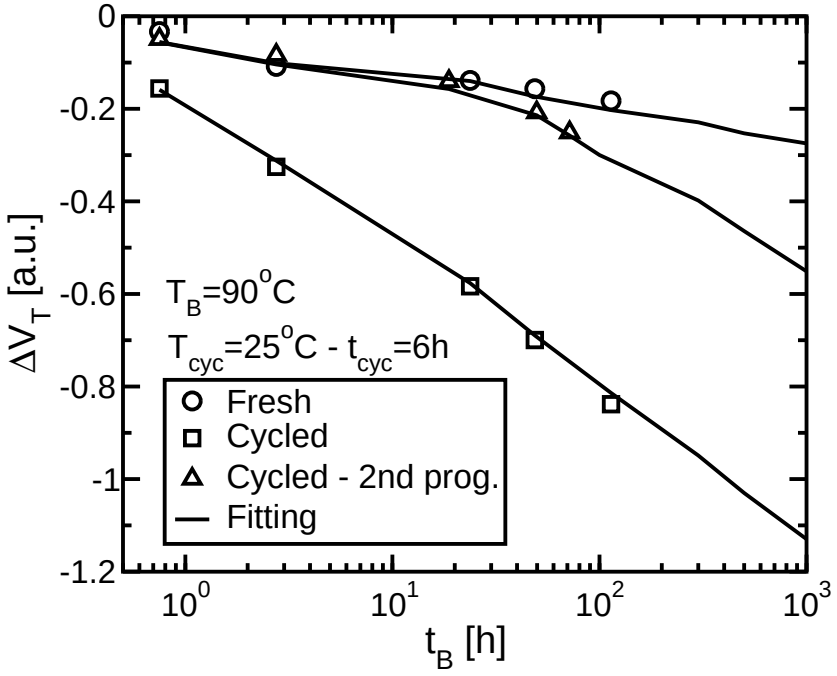


Figure 4.8: *Average ΔV_T transients during bake (T_B = 90^{o}C) on fresh and cycled arrays (T_{cyc}=90^{o}C, t_{cyc}=6 h). Results obtained from a second bake experiment after erase and reprogramming of the cycled array are also shown.*

arrays: a large increase of the V_T-loss appears after cycling with respect to the fresh array, due to partial damage recovery during bake. To explore the cycling dependence of the parasitic V_T-drift, after 100 h bake at T_B = 90^{o}C, the array was erased and reprogrammed and a second bake was performed. Results show that the second V_T transient on the cycled array follows that of the fresh sample up to $t_B \simeq$ 100 h, revealing, first of all, the negligible contribution of damage-recovery to the second 100 h V_T-loss when a first 100 h bake has been already performed. Moreover, this result shows that the parasitic gate-stress induced V_T-drift remains substantially unaltered after cycling, leading to the conclusion that the gate induced drift measured on fresh samples simply adds to the V_T-loss coming from damage recovery during post-cycling bake, compromising the E_a extraction due to its negligible temperature dependence.

According to Figs. 4.5-4.7, a logarithmic dependence of the parasitic V_T-drift (ΔV_T) on WL_S can be assumed:

$$|\Delta V_T| = \alpha_S \ln\left(1 + \frac{WL_S}{WL_S^*}\right) \tag{4.1}$$

where α_S gives the magnitude of the logarithmic decrease of V_T due to gate-stress. In turn, post-cycling damage-recovery during bake features a logarithmic V_T-loss as a function of t_B [64, 67, 165], that is:

$$|\Delta V_T| = \alpha \ln\left(1 + \frac{t_B}{t_0^*}\right) \tag{4.2}$$

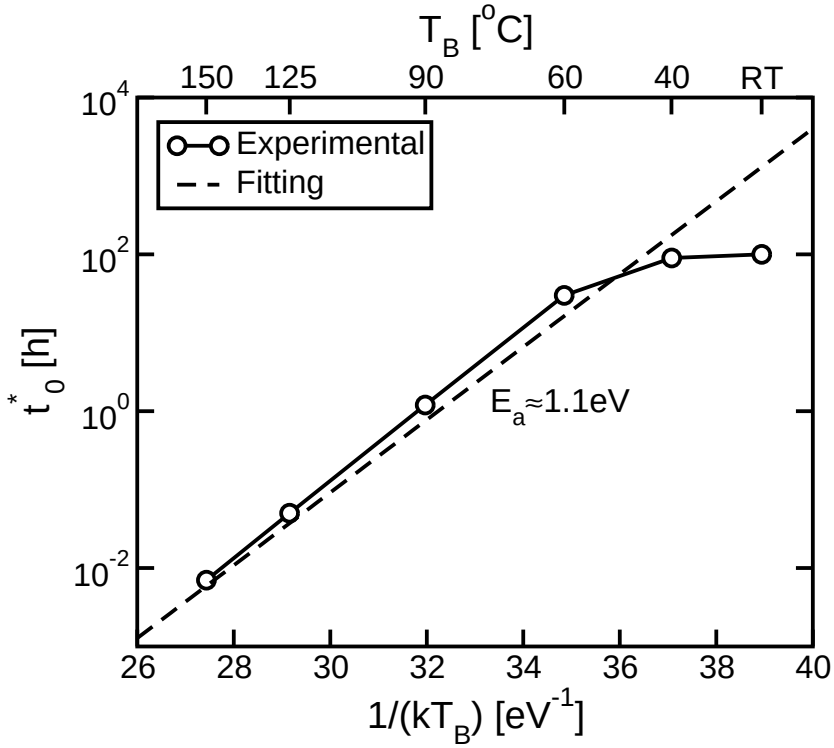


Figure 4.9: *Arrhenius plot for the t_0^* obtained from the fitting of the ΔV_T transients of Fig. 4.3(a) according to (4.3).*

where α gives the slope of the post-cycling V_T-loss as a function of the logarithm of bake time and t_0^* represents the effective time elapsed at the bake temperature T_B between the end of the cycling and the first V_T map acquisition [67,165]. Once the V_T-drift of fresh devices was fitted according to (4.1) in Fig. 4.8, the post-cycling ΔV_T transients could be fitted as the sum of the parasitic V_T-drift and the post-cycling V_T instabilities:

$$|\Delta V_T| = \alpha \ln\left(1 + \frac{t_B}{t_0^*}\right) + \alpha_S \ln\left(1 + \frac{WL_S}{WL_S^*}\right) \tag{4.3}$$

where the same α_S and WL_S^* are used for both fresh and cycled devices. From this last fitting, α and t_0^* could be extracted. Note that the obtained parameters allow a good fitting also of the second bake transient in Fig. 4.8: since the array was reprogrammed after a first bake of duration $t_{B,1} = 100$ h at $T_B = 90^o$C, this can be straightforwardly obtained adding 100 h to t_0^*.

Eq. (4.3) represents a powerful formula to investigate the activation energy issue raised by Figs. 4.3(a)-4.3(b): since the parasitic V_T-drift has been shown to display a negligible dependence on bake temperature, the ΔV_T transients of Fig. 4.3(a) can be reproduced by (4.3) changing only t_0^* for the different T_B curves and keeping for α and α_S the values extracted from Fig. 4.8. Doing so, the dependence of damage-recovery on bake temperature can be analyzed reporting the resulting t_0^* as a function of $1/kT_B$, as done in Fig. 4.9. Results in this figure refer only to the damage recovery process and display the typical activation energy $E_a \simeq 1.1$ eV

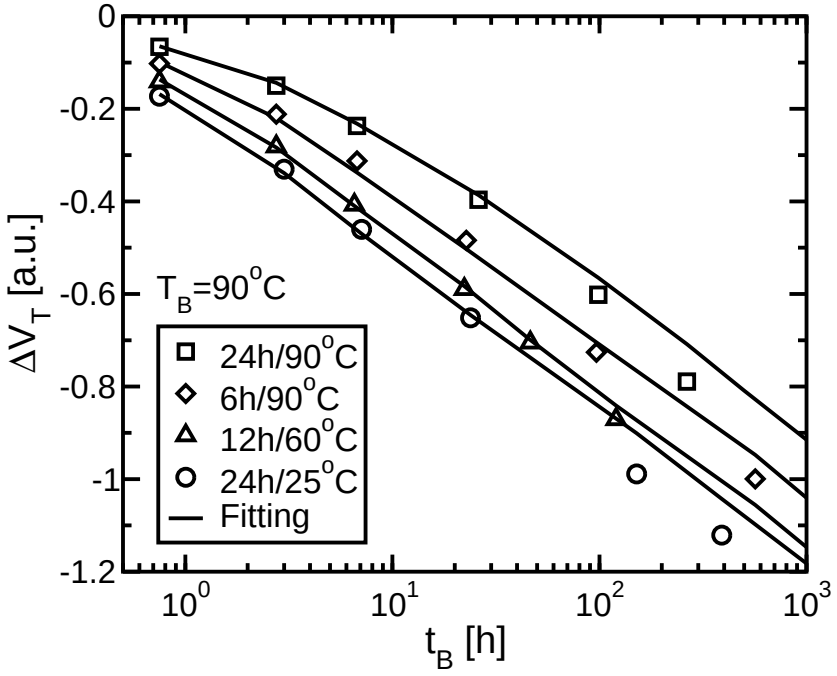


Figure 4.10: *ΔV_T transients at $T_B = 90^oC$ after distributed-cycling with different t_{cyc} and T_{cyc}.*

observed in all previous NOR and NAND technology nodes [63, 64, 67, 165]. This result leads to the conclusion that the apparent activation energy $E_{aa} \simeq 0.8$ eV observed in Fig. 4.3(b) is a spurious effect of the parasitic V_T-drift due to gate-stresses during the V_T map acquisition and not to the damage recovery dynamics.

Finally, provided that the parasitic V_T-drift arising from gate-stress is correctly taken into account, the validity of conventional distributed-cycling schemes was assessed on the 45 nm NOR technology. To this aim, $N_{cyc} = 3$k P/E cycles were performed on the array for different t_{cyc} from 6 h to 24 h, with T_{cyc} from RT to 100^{o}C. The same bake temperature $T_B = 90^o$C was used in all the experiments in order to evaluate only the effect of distributed-cycling on the V_T instability during bake. Fig. 4.10 shows some of the ΔV_T transients obtained for different t_{cyc} and T_{cyc}: it can be observed that all the curves follow the same logarithmic trend and that they are only delayed (i.e., rigidly shifted on the log time axis) according to the cycling conditions. Thus, results can be fitted by (4.3) changing only t_0^* in agreement with [165], where the following formula was proposed for t_0^* :

$$t_0^* = t_0 + At_{cyc} \cdot e^{E_a(1/kT_B - 1/kT_{cyc})} \tag{4.4}$$

Fig. 4.11(a) shows that this expression for t_0^* can correctly reproduce the experimental data, allowing the extraction of $E_a \simeq 1.1$ eV. This activation energy matches that from bake experiments of Fig. 4.9, confirming the validity of standard distributed-cycling tests at the 45 nm NOR node. Finally, the inset shows t_0^* as a function of t_{cyc} for $T_{cyc} = T_B$: in this case (4.4) reduces to $t_0^* = t_0 + At_{cyc}$, allowing the evaluation of A [64, 67, 165].

Fig. 4.11(b) shows that a Universal Damage-recovery Metric (UDM) [64, 67,

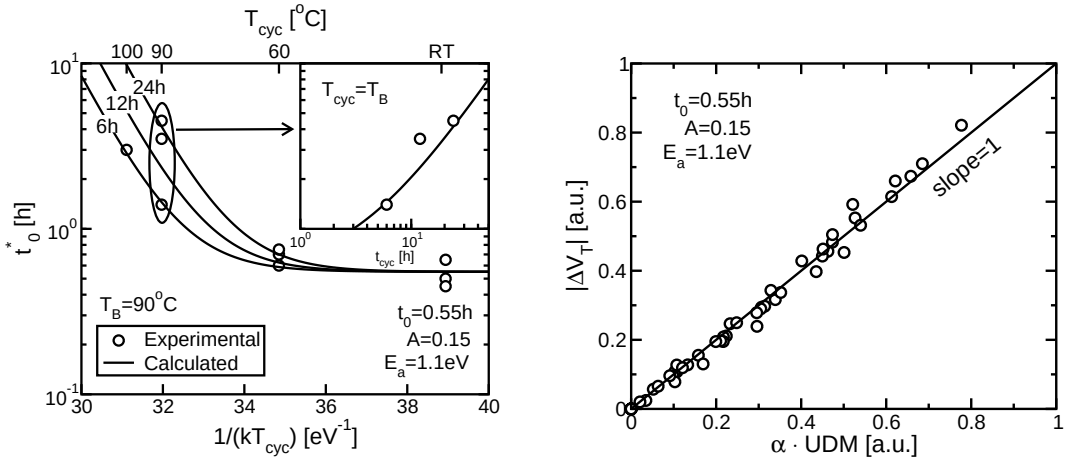


Figure 4.11: *(a) Arrhenius plot for distributed-cycling: the t_0^* resulting from the fitting of the ΔV_T transients of Fig. 4.10 are shown as a function of $1/kT_{cyc}$; and (b) ΔV_T transients obtained subtracting the parasitic V_T-drift given by (4.1) from experimental data as a function of $\alpha \cdot UDM$. Data for all the t_{cyc} and T_{cyc} reported in Fig. 4.11(a) are shown.*

165], using the parameter values obtained from Fig. 4.11(a), holds also for 45 nm NOR, if the parasitic V_T-drift is taken into account. Defining:

$$UDM = \ln\left(1 + \frac{t_B}{t_0^*}\right) \tag{4.5}$$

(4.2) can be written as $|\Delta V_T| = \alpha \cdot UDM$, leading all the ΔV_T transient from the distributed-cycling experiments to lay on the same linear relation with slope 1, in Fig. 4.11(b), when reported as a function of $\alpha \cdot UDM$.

4.5 Conclusions

This chapter presented a detailed assessment of the validity of conventional distributed-cycling schemes relying on the 1.1 eV activation energy for damage recovery at the 45 nm NOR node. This has required a careful investigation of the experimental tests, since parasitic contributions to the V_T-drift have been shown to appear due to the long gate-stresses required to gather the V_T map of the memory array. This spurious V_T instability can be observed both on fresh and cycled devices, adding, in the latter case, to the V_T-loss coming from post-cycling damage recovery. If this additional source of V_T instability, which shows negligible dependence on temperature and cycling, is not correctly taken into account, an apparent activation energy $E_{aa} < 1.1$ eV is extracted from conventional test schemes. Moreover, the validity of UDM is compromised as a consequence of the temperature and cycling independence of the parasitic V_T-drift. However, when the superposition of the two phenomena is taken into account, the validity of distributed-cycling schemes and UDM are fully confirmed. Finally, it was shown that parasitic V_T-drift effects does not compromise neither the normal on-field usage of the memory device nor

the coverage of JEDEC JESD47's HTDR and LTDR qualification tests [173,174], since they come into play only when longer stress times and higher voltages then normal on-field operation are adopted, as in case the adopted experimental procedure.

CHAPTER 5

Fundamental variability sources in Flash memories

This chapter discusses the impact of fundamental variability sources on sub-100 nm NAND Flash memories operation; in particular, the atomistic nature of the substrate doping and the discrete nature of the electrons flow into the floating-gate have been previously reported as fundamental phenomena, affecting the statistical dispersion of neutral cell threshold voltage and programming accuracy, respectively. This chapter investigates the impact of control-gate and floating-gate doping and geometry on the electron-injection spread of nanoscale NAND Flash memories, addressing the electron-injection spread scaling trend. Then, the attention is focused on data retention, presenting a comparison of the dispersion contribution due to the neutral threshold-voltage spread and the electron-emission statistics from the floating gate, evaluating their effect on the data retention transients of a memory array. Finally, the first array-level experimental observation of post-cycling discrete electron detrapping from the tunnel oxide in sub-30nm NAND Flash arrays is given, providing an accurate stochastic model of cycled array data retention.

5.1 Introduction

Variability effects are becoming more and more important for NAND Flash memories, affecting the uniformity and reproducibility of device characteristics and operations as cell dimensions scale down [107, 151, 183, 184]. In this respect, two main classes of variability sources can be identified: the first one is related to cell-to-cell parameter variations, due to process tolerances or fundamental reasons [107]; the second one accounts for the statistics of few-electron phenomena and it is related to the discrete electron transfer to/from the floating gate (FG)

during cell operation [151, 183, 184]. In particular, the statistical injection of electrons into the floating gate (FG) has been recently shown to set the ultimate limit to the accuracy of the incremental step pulse programming (ISPP) algorithm for deca-nanometer NAND Flash memories [151, 184–186]. Scaling of device dimensions results, in fact, into a continuous decrease of the FG charge controlling cell state, making the fundamental fluctuation of the number of electrons transferred to the FG during ISPP more and more important for the programmed threshold voltage (V_T) distribution of the array. While this latter variability source appears only when the charge state of the FG is changed, fluctuations in cell parameters also result into a spread of the neutral threshold voltage ($V_{T,0}$) of the cells in the memory array [107].

This chapter firstly investigate the impact of CG and FG design on the electron-injection spread (EIS) of deca-nanometer NAND Flash memories, not only in terms of their geometry but also of their doping, which is becoming more and more important for device operation [187]; then, a comparison of cell-to-cell parameter variations and of electron-emission statistics (EES) on the retention characteristics of nanoscale NAND Flash memories is presented. Referring to fresh cells programmed to the same threshold voltage (V_T) level, both the $V_{T,0}$ spread and EES are shown to represent a source of statistical variability for data retention, resulting into a broadening of the array V_T distribution as retention time elapses. A quantitative assessment of the two contributions clearly shows, however, that the $V_{T,0}$ spread, and therefore cell-to-cell parameter variations, dominates over the EES spread for data retention variability of nanoscale NAND Flash memories. Finally, a statistical analysis of post-cycling data retention is presented, showing by means of experimental and modeling analysis that discrete electron detrapping from tunnel oxide rules data retention on sub-30 nm NAND Flash arrays and that detrapping statistics is a major post-cycling spread source on aggressively scaled arrays.

5.2 Fundamental variability sources

5.2.1 Neutral cell threshold voltage spread

As a consequence of the high integration density which can be achieved in nanoscale Flash memory devices, variability has emerged as a major issue for array operation and reliability. The continuous shrinking of the cell size, up to decananometer scale, has been largely increasing the variability sources (e.g., due to cell-to-cell parameter variations), setting new constraints to array functionality. In the case of Flash memories, an accurate control of cell V_T is required in order to guarantee the desired functionality: the variability sources result, first of all, in a statistical variation of the neutral cell threshold voltage ($V_{T,0}$) [107]; Fig. 5.1(a) shows the $V_{T,0}$ spread as a function of the technology node, highlighting how the variability effects get more severe with the technology scaling. The major variability sources that contributes to $V_{T,0}$ spread are individually shown in Fig. 5.1(b): besides process-induced tolerances (e.g., cell length/width fluctuations) and $V_{T,0}$ variations due to oxide traps fluctuations (OTF), random dopant fluctuations (RDF) are predicted to play a major role in sub-30 nm technologies.

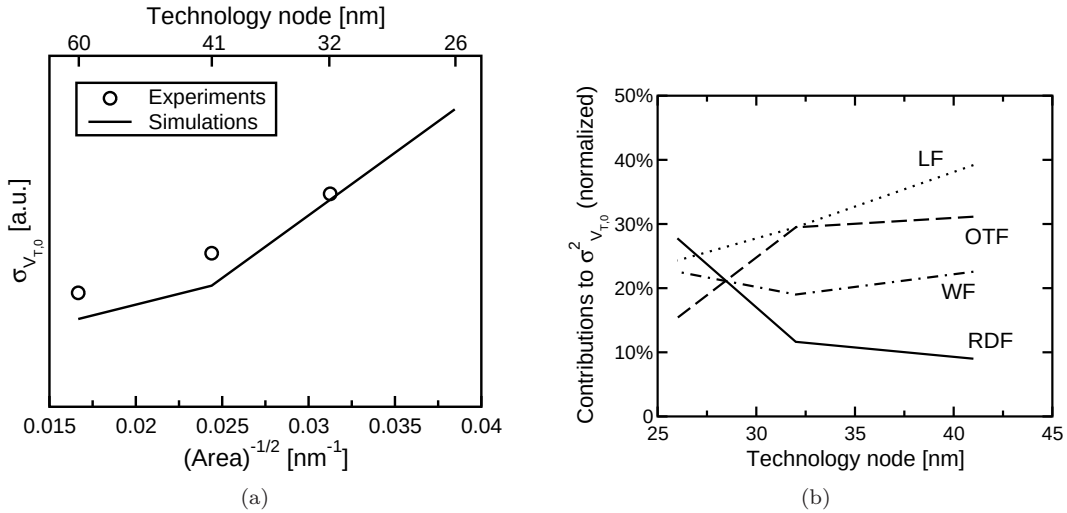


Figure 5.1: *(a) $V_{T,0}$ spread and (b) its most important contributions as a function of technology node. LF: L fluctuations; WF: W fluctuations; RDF = random dopant fluctuations, OTF: oxide trap fluctuations (from [107]).*

The randomness both of the number of dopants and of their positions in the channel region, in fact, has been reported as a fundamental V_T variability source for decananometer MOSFET devices [88, 90]: Fig. 5.2 shows the simulated potential distributions and equi-concentration contours for a $50X50$ nm MOSFET, showing how the randomness of dopants induces fluctuations of the channel potential, thus affecting the V_T of the device [90]. Since RDF is a physical phenomena related to the discrete nature of the matter, it is a fundamental and unavoidable variability source which becomes more and more severe with device scaling, as clearly shown in Fig. 5.1(b).

This increase of $V_{T,0}$ spread with scaling has a non-negligible impact on memory array functionality: a very accurate V_T placement is required, especially in multi-level cell (MLC) and triple-level cell (TLC) Flash technologies, and thus programming algorithms should be designed in order to cancel out cell-to-cell $V_{T,0}$ variability and to achieve the desired programming accuracy [169–171,188]. When a tight V_T placement is achieved, however,$V_{T,0}$ spread results in a statistical dispersion of the tunnel oxide electric field of programmed cells, strongly affecting data retention transients, as discussed in Section 5.4.

5.2.2 Electron injection statistics (EIS) during programming

NAND Flash programming is performed injecting electrons from the substrate into the FG thanks to Fowler-Nordheim tunneling mechanism, applying a large positive bias to the CG, while the other terminals are grounded. However, in order to obtain a narrow programmed V_T distribution, reducing the impact of cell-to-cell parameter fluctuation (such as $V_{T,0}$ spread), programming algorithms more complex than just applying a constant CG bias are adopted: Fig. 5.3(a) shows an example of the CG waveform and the resulting V_T programming transient of a memory cell during the incremental step pulse programming algorithm

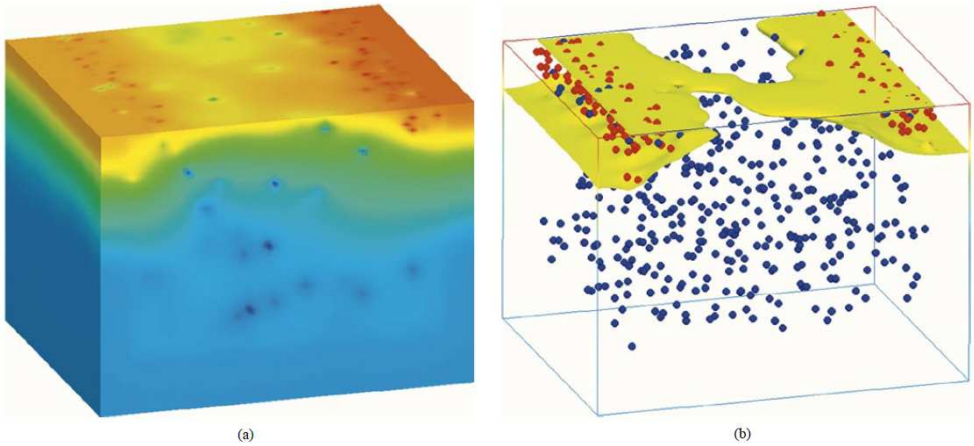


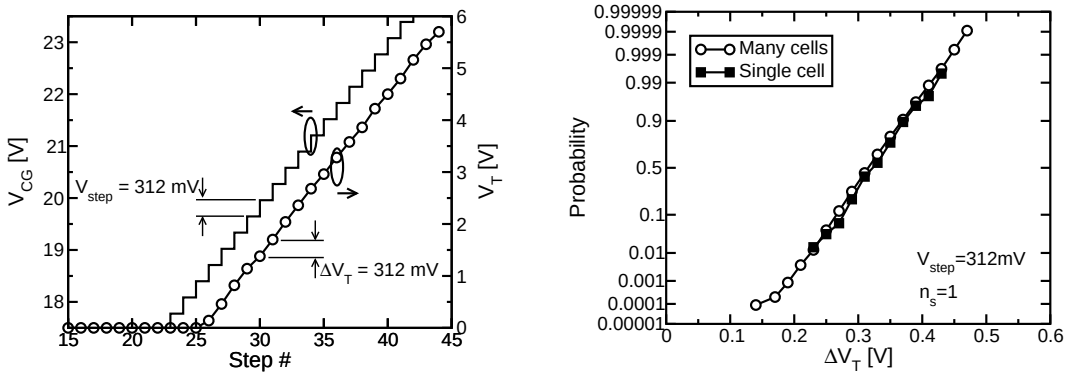
Figure 5.2: *Simulation domain in typical Density Gradient simulations of a* $50X50$ *nm MOSFET. (a) Potential distribution indicating also the positions of the individual dopants.(b) One equi-concentration contour (from [90]).*

(ISSP) [150, 151]. The periodic application of small programming steps of the same amplitude V_{step} and duration τ_s aims at bringing the cell in a stationary condition, in which the threshold voltage shift ΔV_T equals the step amplitude, thus allowing an accurate V_T placement. After each programming step, cell V_T is compared against one or more program verify levels (PVs) and the programming is inhibited when the desired level is reached: more details of single and double verify ISSP algorithms will be discussed in Chapter 6.

During ISPP programming, a constant average Fowler-Nordheim tunneling current equals to $\overline{I_G} = V_{step} \cdot C_{pp}/\tau_s$ is reached during each programming step (where C_{pp} is the CG to FG capacitance), which corresponds to the injection into the FG of an average number of electrons per step $\overline{\Delta n} = V_{step} \cdot C_{pp}/q$. This means that in the example of Fig. 5.3(a), assuming $C_{pp} \simeq 30$ aF, on average only 50 electrons are injected into the FG during each programming step, making the statistical fluctuations of the number of electrons a non-negligible variability source. Assuming a Poisson statistics for the electron injection process (which means that electron injection events are assumed to be independent on from each other), the spread of the cell ΔV_T per step can be evaluated as:

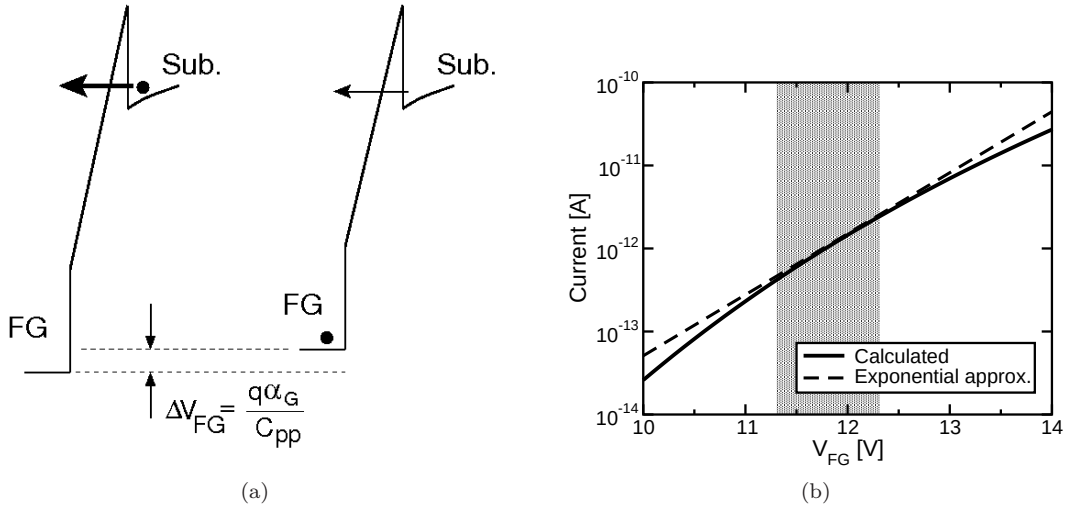
$$\sigma_{\Delta V_T} = \frac{q}{C_{pp}}\sqrt{\overline{\Delta n}} = \sqrt{\frac{q}{C_{pp}}\overline{\Delta V_T}}, \tag{5.1}$$

where $\overline{\Delta V_T} = q\overline{\Delta n}/C_{pp}$. This equation leads to predict a relative spread $\sigma_{\Delta V_T}/\overline{\Delta V_T} = 13\%$ for the example of Fig. 5.3(a). Fig. 5.3(b) shows the cumulative ΔV_T due to programming pulse confirming that a Gaussian spread can be detected (in a normal probability plot a Gaussian distribution is represented by a straight line). In decananometer Flash technologies, thus, as the number of electrons required to obtain $\Delta V_T = 100$ mV is continuously decreasing due to C_{pp} scaling [5], the discrete flow of electrons transferred into the FG become a major variability source, setting a fundamental limit to the programming accuracy, as discussed in Chapter 6 with more details.



(a) (b)

Figure 5.3: *(a) Example of the V_{CG} waveform applied during ISPP of NAND cells and resulting V_T transient on a 60 nm technology device; and (b) cumulative distribution function of the $\Delta V_{T,s}$ resulting from a single step of the ISPP algorithm of (a) (from [151]).*



(a) (b)

Figure 5.4: *(a) Schematics for the conduction band profile of a NAND cell during Fowler-Nordheim programming, before and after the storage of a single electron into the floating gate; and (b) I_G-V_{FG} characteristics of a 60 nm NAND cell and corresponding exponential approximation around the stationary working point of a typical ISPP algorithm (from [184]).*

The evaluation of electron injection spread previously presented is based on the assumption that EIS follows a Poisson statistics; a deeper insight into the electron transfer from the substrate into the floating gate, however, reveals that this is intrinsically a sub-Possonian process. Each single electron injection event, in fact, modify the FG potential and reduces the electric field in the tunnel oxide, introducing a negative feedback in the electron injection process, as schematically depicted in the band diagrams of Fig. 5.4(a).

A detailed analysis of electron injected process was presented in [184] and further discussed in [7]: if the Fowler-Nordheim tunneling current is approximated around the ISSP stationary working point as:

$$I_G = I_0 exp\left[\frac{\gamma}{\alpha_G} V_{FG}\right], \tag{5.2}$$

where γ/α_G represents the slope of the I_G curve in the semilogarithmic plot and I_0 is a constant, a field feedback factor can be introduced as

$$f = e^{q\gamma/C_{pp}}. \tag{5.3}$$

This means that the tunneling current I_G is scaled by a factor f (which is always larger than 1) after each electron injection event; taking into account this effect, is it possible to replace (5.1) with the following analytical expression for $\sigma_{\Delta V_T}$:

$$\sigma_{\Delta V_T} = \frac{q}{C_{pp}}\sigma_{\Delta n} = \sqrt{\frac{q}{\gamma C_{pp}}\left(1 - e^{-\gamma\overline{\Delta V_T}}\right)}. \tag{5.4}$$

Fig. 5.5 shows that the model can correctly reproduce the experimental data for $\sigma_{\Delta V_T}$ as a function of $\overline{\Delta V_T}$. Experimental data reveals that the Poisson assumption of (5.1) holds for low values of $\overline{\Delta V_T}$, since the feedback effect can be neglected in this regime; however, when $\overline{\Delta V_T}$ increases, a greater number of electron is injected during each programming step, making the feedback stronger and, thus, resulting in a saturation of $\sigma_{\Delta V_T}$ due to the sub-Poissonian nature of the injection process.

Moreover, a detailed Monte Carlo model which carefully reproduces the electron injection process during ISPP, removing all the approximation (such as the exponential approximation of I_G), will be discussed in Chapter 6, in order to investigate the ultimate limitations that the discrete nature of charge flow sets to programming algorithms accuracy.

5.3 Impact of CG and FG design on the EIS

5.3.1 Polysilicon doping and C_{pp} bias dependence

EIS analysis presented in the previous Section, however, neglects several second order effects, such as the effects due to FG doping and geometry. Doping of the CG and FG polysilicon will have a larger and larger impact on NAND cell operation as device dimensions are continuously scaled down, due to dopant activation issues and more relevant depletion effects [187]. During ISPP, a polysilicon depletion layer appears both in the CG and at the tunnel-oxide side of the FG, while electron accumulation is present at the interpoly side of the FG. In order to investigate

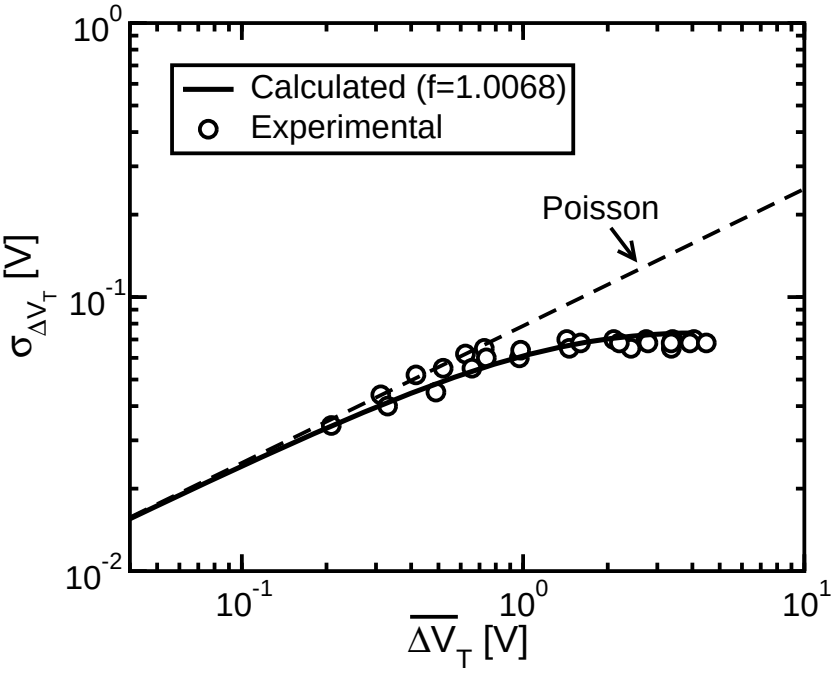


Figure 5.5: *Calculated and experimental $\sigma_{\Delta V_T}$ for 60 nm NAND cells, as a function of $\overline{\Delta V_T}$. (from [184]).*

these effects and their impact on the EIS, accurate TCAD simulations of cell electrostatics are mandatory.

Fig. 5.6 shows the cross-section of a deca-nanometer NAND cell along the word-line (WL) direction, as appearing from a TEM image and as reproduced by TCAD tools. The correct reproduction of 3-D cell geometry in TCAD simulations is mandatory to account for non-uniform cell electrostatics and fringing fields. In addition, the correct estimation of FG and CG polysilicon doping is of fundamental importance for the accurate description of cell performance during programming, due to non-negligible depletion effects. To this aim, polysilicon dopings for TCAD devices were extracted from process simulations and confirmed by capacitance-voltage measurements on capacitors having the same gate stack and polysilicon widths similar to the investigated cell technologies. As a result of CG polysilicon depletion, Fig. 5.7(a) shows a reduction of the small-signal C_{pp} when the voltage between FG and CG (V_{pp}) is increased, leading to a different C_{pp} value in read ($C_{pp,R}$, for $V_{pp} < 2$ V) and program ($C_{pp,P}$, for $V_{pp} > 5$ V) conditions. As a consequence, the effect of a single electron stored in the FG on cell electrostatics is different during the read and program operations. To include this effect into the EIS calculation, the following expression for the standard deviation of the number of electrons Δn injected into the FG during ISPP should be considered [184]:

$$\sigma_{\Delta n} = \sqrt{\frac{1 - f^{-\overline{\Delta n}}}{\ln f}} \tag{5.5}$$

where $\overline{\Delta n}$ is the average number of transferred electrons and f is the feedback

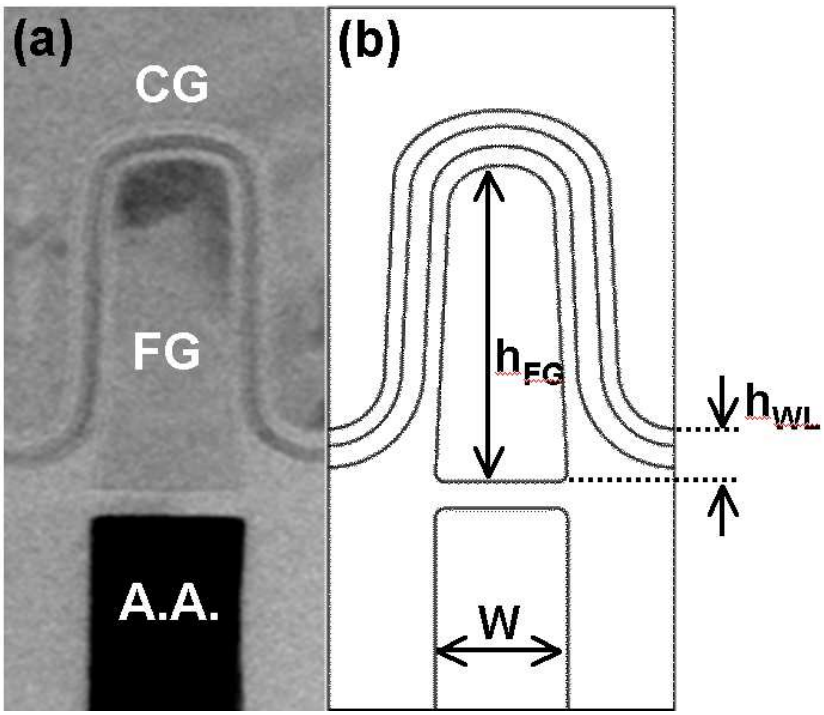
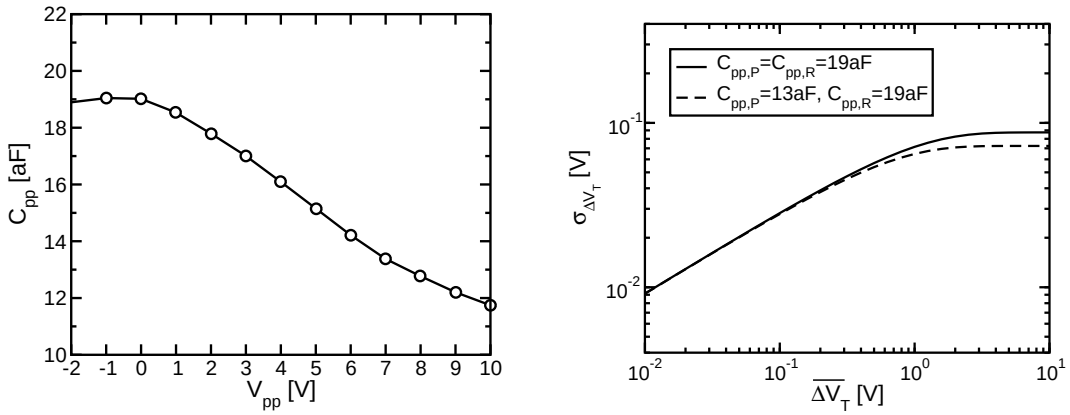


Figure 5.6: *Cross-section of a deca-nanometer (sub-60nm) NAND cell along the WL direction, as appearing from a TEM image (a) and as reproduced by TCAD tools (b).*



(a) (b)

Figure 5.7: *(a) TCAD simulation results for the C_{pp} of a deca-nanometer NAND cell as a function of V_{pp}; and (b) calculated $\sigma_{\Delta V_T}$ as a function of $\overline{\Delta V_T}$, considering and neglecting the reduction of C_{pp} at large bias.*

factor giving the reduction of the programming tunneling current after each electron storage in the FG. Assuming an exponential approximation for the tunneling current-FG voltage characteristics, f is constant and equal to [184]:

$$f = e^{q\gamma/C_{pp,P}} \tag{5.6}$$

where q is the electronic charge and γ is proportional to the slope of the tunneling current characterisitics around the stationary working point of the ISPP algorithm, as discussed in detail in [184]. Note that in (5.6), the C_{pp} value in program conditions has been considered, as the change of cell electrostatics when storing one electron in the FG during ISPP is required. Cell ΔV_T resulting from the storage of Δn electrons in the FG is, instead, given by $C_{pp,R}$ according to the simple relation:

$$\Delta V_T = \frac{q}{C_{pp,R}} \Delta n \tag{5.7}$$

Merging (5.7) with (5.5) and (5.6) it can be obtained:

$$\sigma_{\Delta V_T} = \sqrt{\frac{q}{\gamma C_{pp,R}} \cdot \frac{C_{pp,P}}{C_{pp,R}} \cdot \left(1 - e^{-\gamma \frac{C_{pp,R}}{C_{pp,P}} \overline{\Delta V_T}}\right)} \tag{5.8}$$

which gives the ΔV_T spread during ISPP in the general case where C_{pp} has a significant change between program and read conditions. Assuming $C_{pp,R} = 19$ aF and $C_{pp,P} = 13$ aF, Fig. 5.7(b) shows that the effect of $C_{pp,P} \neq C_{pp,R}$ appears only near the saturation of the $\sigma_{\Delta V_T}$ curve, while in the low-$\overline{\Delta V_T}$ regime, including the V_s values of interest for ISPP, the curve does not change with respect to the case of $C_{pp,P} = C_{pp,R} = 19$ aF. This result reveals that EIS is mainly dominated by the C_{pp} value in read conditions, as directly obtained from (5.8) under the hypothesis of small $\overline{\Delta V_T}$, giving:

$$\sigma_{\Delta V_T} \simeq \sqrt{\frac{q}{C_{pp,R}} \overline{\Delta V_T}} \tag{5.9}$$

From this result, a scaling analysis of $\sigma_{\Delta V_T}$ can be straightforwardly developed neglecting any polysilicon-depletion effect in the CG and FG, as will be done in the next Section. Note, however, that polysilicon depletion remains of fundamantal importance for the ISPP speed, due to the voltage drop lost on the depletion layers.

5.3.2 Polysilicon geometry and C_{pp} scaling

Fig. 5.6 shows the typical FG and CG geometry adopted for all the NAND technologies under investigation below the 60 nm node. The width of the FG polysilicon equals the active area (A.A.) width at its tunnel-oxide side and slightly decreases moving from bottom to top. The CG polysilicon wraps around the FG, up to a distance h_{WL} from the FG lower surface. A rough estimation of C_{pp} for this structure can be straightforwardly obtained with a three parallel-plate capacitors approximation, *i.e.* neglecting the curvature effects at the top surface of the FG polysilicon:

$$C_{pp} = \frac{\epsilon_{ox}}{EOT_{ono}} \left[WL + 2L\left(h_{FG} - h_{WL}\right)\right] \tag{5.10}$$

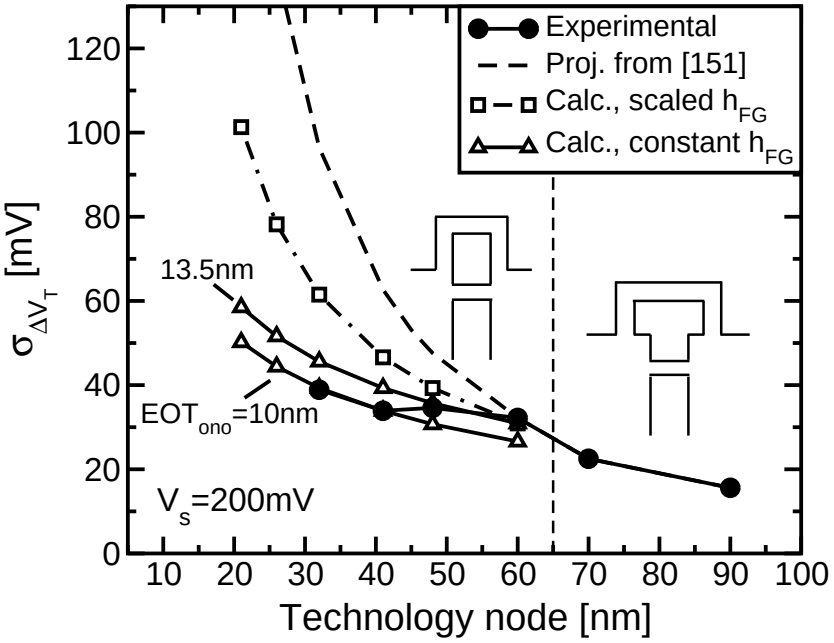


Figure 5.8: *Calculated $\sigma_{\Delta V_T}$ for $\overline{\Delta V_T} = 200$ mV, assuming constant h_{FG} (with $EOT_{ono} = 13.5$ and 10 nm) and h_{FG} decreasing proportionally to F ($EOT_{ono} = 13.5$ nm). Experimental results are also shown, together with the first projections reported in [151].*

where ϵ_{ox} is the oxide dielectric constant, EOT_{ono} is the equivalent oxide thickness of the oxide-nitride-oxide (ONO) interpoly dielectric and h_{FG} is the height of the FG polysilicon. It is possible to use (5.10) for a first order investigation of the C_{pp} and $\sigma_{\Delta V_T}$ evolution with the technology node feature size F, under the simplifying assumptions of $W = L = F$, constant $EOT_{ono} = 13.5$ nm and $h_{WL} = 12$ nm. To this aim, two possible scaling scenarios for the FG height can be considered: assuming $h_{FG} = 80$ nm for the 60 nm node, the case of constant h_{FG} and of h_{FG} decreasing proportionally to F was considered. Fig. 5.8 shows the resulting $\sigma_{\Delta V_T}$ from (5.9) when $V_s = 200$ mV: due to the lower reduction of cell C_{pp}, a lower growth of $\sigma_{\Delta V_T}$ appears with cell scaling when h_{FG} is constant than when h_{FG} is reduced proportionally to F. In particular, Fig. 5.8 shows that experimental results down to the 32 nm node have actually followed the trend given by constant h_{FG}, with the small displacements from the calculated results due to a small reduction of EOT_{ono} as scaling proceeds and to the extremely simplified C_{pp} calculation given by (5.10). The effect of a reduction of EOT_{ono} from 13.5 nm to 10 nm, predicted in [10] for sub-32 nm technologies, is shown in Fig. 5.8 for the case of constant h_{FG}, highlighting that the consequent increase of cell C_{pp} contributes to the lowering of $\sigma_{\Delta V_T}$. Finally, from the $\sigma_{\Delta V_T}$ values of Fig. 5.8, the EIS effect on the programmed V_T distribution in presence of a program-verify level can be easily estimated, as done in [186]. This results into a programmed V_T distribution width at 10^{-5} equal to 280mV for the 60nm node, and large values for more scaled cells.

As a last remark, note that Fig. 5.8 also shows the projections for $\sigma_{\Delta V_T}$ first presented in [151], which appear now to have over-estimated the EIS growth. This was due to a strong change in the FG design since the 60 nm node. First, for

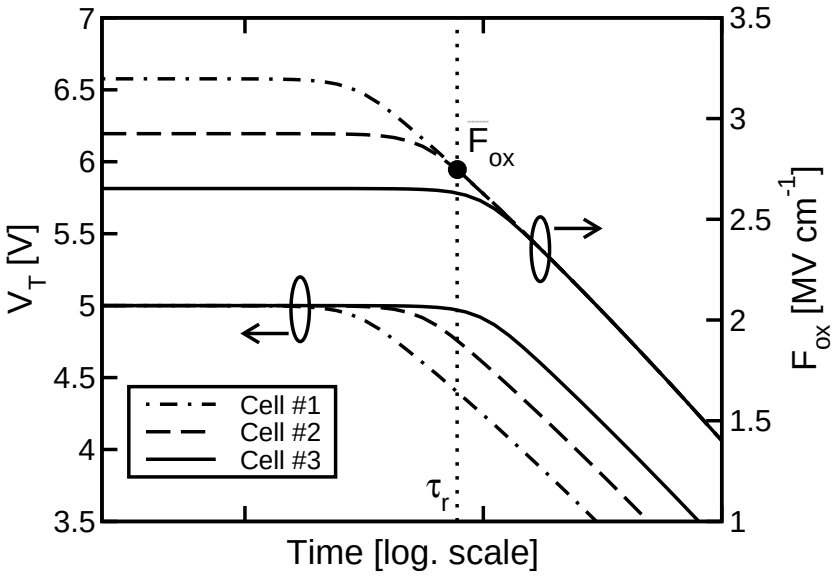


Figure 5.9: *Calculated retention transient for three cells having different $V_{T,0}$.*

$F > 60$ nm, an increase of the FG width was present at the top of the FG (see the inset of Fig. 5.8) to increase the CG to FG capacitive coupling ratio α_G. However, to reduce the gate pitch, the FG structure evolved into the one reported in Fig. 5.6 below the 60 nm node. Moreover, to preserve α_G and the program/erase performance, C_{pp} did not scale with cell area, with h_{FG} that remained nearly constant. Despite this scaling solution has increased the reliability issues coming from cell electrostatic interference [189], it has strongly limited few electron phenomena, such as EIS and retention-time dispersion [190].

5.4 Impact of $V_{T,0}$ spread and EES on data retention

Fundamental variability sources, like $V_{T,0}$ spread and charge granularity, can play a role also in data retention of nanoscale Flash arrays. In the following sections, a detailed investigation of $V_{T,0}$ spread and electron emission statistics (EES) impact on data retention will be presented, showing that both these variability sources result into a V_T distribution broadening as retention time elapses.

5.4.1 Effect of the $V_{T,0}$ spread on data retention

A Gaussian $V_{T,0}$ distribution with zero mean value and standard deviation $\sigma_{V_{T,0}}$ was assumed. To easily quantify this value as a function of cell parameters, only the variability contribution of atomistic doping was considered, given by [90]:

$$\sigma_{V_{T,0}} = 3.19 \times 10^{-8} \frac{t_{ox} N_A^{0.4}}{\alpha_G \sqrt{WL}} \tag{5.11}$$

where t_{ox} is the tunnel-oxide thickness, N_A is the substrate doping, and W and L are the cell width and length, respectively. Note that to adapt the formula presented in [90] for MOS transistors to the case of FG devices, the control-gate

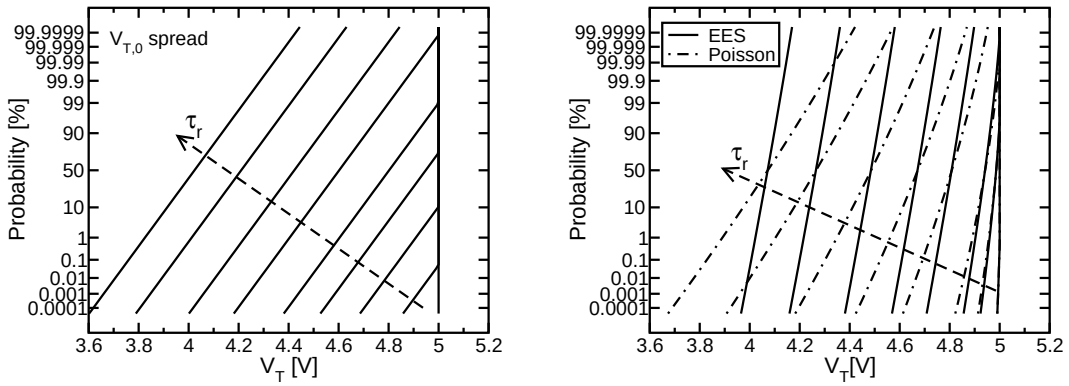


Figure 5.10: *(a) Calculated array V_T distribution for increasing τ_r, assuming the cells initially placed at the same $V_{T,P} = 5$ V and including only the $V_{T,0}$ effect; and (b) same as in Fig. 5.10(a), but considering identical cells with $V_{T,0} = 0$ and including only the effect of EES. Results from a poissonian statistics are also shown for the same average V_T loss.*

(CG) to FG capacitive coupling ratio α_G was introduced in (5.11). Assuming $W = L = 60$ nm, $t_{ox} = 8$ nm, $N_A = 5 \times 10^{17}$ cm^{-3} and $\alpha_G = 0.6$, the previous expression predicts $\sigma_{V_{T,0}} = 85$ mV, which must be viewed as a lower bound for the real $V_{T,0}$ spread of a 60 nm NAND array. In fact, despite atomistic doping is one of the main variability sources for $V_{T,0}$, many other contributions exist, *e.g.* due to process-induced tolerances [107].

The effect of the $V_{T,0}$ spread on data retention was investigated referring to a fresh device at low temperature, where the V_T shift from the programmed state only comes from the FG charge loss through the tunnel oxide via direct quantum-mechanical tunneling, neglecting all the possible detrapping and defect-assisted charge loss processes limiting data retention after cycling. An ideal V_T distribution at the beginning of the data retention experiment was assumed, with all the cells placed exactly at the same programmed V_T level, $V_{T,P} = 5$ V. In these conditions, as a result of the $V_{T,0}$ spread, cells initially experience a different electric field in the tunnel oxide, F_{ox}, given by:

$$F_{ox} = \frac{\alpha_G}{t_{ox}} \left(V_T - V_{T,0} + V_{FB} \right), \tag{5.12}$$

where the substrate and polysilicon voltage drops has been neglected and $V_{FB} \simeq -1.1$ V is the cell flat-band voltage, whose weak dependence from substrate doping was also neglected. Once F_{ox} is known, the tunneling current I_{ox} discharging the FG can be calculated [191] and the V_T retention transient can be evaluated knowing that $dV_T/dt = -I_{ox}/C_{pp}$, where C_{pp} is the CG to FG capacitance. Fig. 5.9 shows the calculated results for three cells having different $V_{T,0}$: a clear converging behavior is featured by F_{ox}, but not by V_T. Rather, cells start losing charge at different times, when the main F_{ox} discharge characteristics catches their initial oxide field. For example, at a given data retention time τ_r in Fig. 5.9, cells having initial $F_{ox} < \overline{F_{ox}}$ (*e.g.*, cell # 3 in the figure) still retain their initial V_T value, while the ones with initial $F_{ox} > \overline{F_{ox}}$ (cells #1 and #2) give rise to a low-V_T tail

in the array distribution, having spread $\sigma_{V_{T,0}}$.

From this result, the cumulative V_T distribution can be easily obtained for each τ_r, and is shown in Fig. 5.10(a). The distribution is initially a step-like function centered at $V_{T,P}$, *i.e.* a vertical line in a normal-scale plot. As τ_r increases, a tail of slope equal to $\sigma_{V_{T,0}}$ appears at a probability level given by the fraction of cells having their initial $F_{ox} > \overline{F_{ox}}$. Eventually, all the cells will lose charge and the V_T distribution becomes a shifted replica of the intrinsic one.

5.4.2 EES effect on data retention

In the case of a negligible $V_{T,0}$ spread, a broadening of the V_T distribution during data retention is still expected due to the EES, representing a fundamental variability source related to the granular nature of I_{ox} and not to any physical difference among the cells [183]. In order to investigate the EES impact on data retention, identical cells having $V_{T,0} = 0$ were considered, all initially placed at $V_{T,P} = 5$ V. By (5.12), a unique initial F_{ox} can be defined for all the cells, corresponding to an initial tunneling current through the tunnel oxide $I_{ox,1}$. From this value, the average time required for the emission of the first electron from the FG during data retention can be easily calculated as $\tau_1 = q/I_{ox,1}$. After one electron has been emitted from the FG, a reduction of F_{ox} equal to $\Delta F_{ox} = q\alpha_G/(t_{ox}C_{pp})$ takes place, in turn reducing I_{ox} and increasing the average time required for the next electron emission. This means that the electron emission process is intrinsically sub-poissonian, with the average time $\tau_n = q/I_{ox,n}$ required for the emission of the n^{th} electron from the FG increasing with n, due to the reduction of the tunneling current $I_{ox,n}$ when $n-1$ electrons have already been emitted. The probability $P_n(\tau_r)$ to have n electron emission events during the retention time τ_r can be obtained from the following relation [184, 192]:

$$[P_0(\tau_r) \quad P_1(\tau_r) \quad P_2(\tau_r) \ldots] = [1 \quad 0 \quad 0 \ldots]\, e^{\mathbf{F}\tau_r}, \tag{5.13}$$

where $\mathbf{F}$ is a matrix of the following form:

$$\mathbf{F} = \begin{bmatrix} -1/\tau_1 & 1/\tau_1 & 0 & 0 & 0 & \ldots \\ 0 & -1/\tau_2 & 1/\tau_2 & 0 & 0 & \ldots \\ 0 & 0 & -1/\tau_3 & 1/\tau_3 & 0 & \ldots \\ 0 & 0 & 0 & -1/\tau_4 & 1/\tau_4 & \ldots \\ 0 & 0 & 0 & 0 & -1/\tau_5 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The probabilities $P_n(\tau_r)$ given by (5.13) correspond to the probability for a cell to decrease its V_T by an amount qn/C_{pp} in a time τ_r and can be used to calculate the V_T distribution during data retention. Results are shown in Fig. 5.10(b) (solid lines), displaying an increase of the distribution width in the first stages of the V_T loss. However, a saturation of the distribution spread is clearly evident for long τ_r, as discussed in [151,184]. Dashed lines in the figure show, instead, results obtained according to a Poisson distribution for the electron emission process, reported for

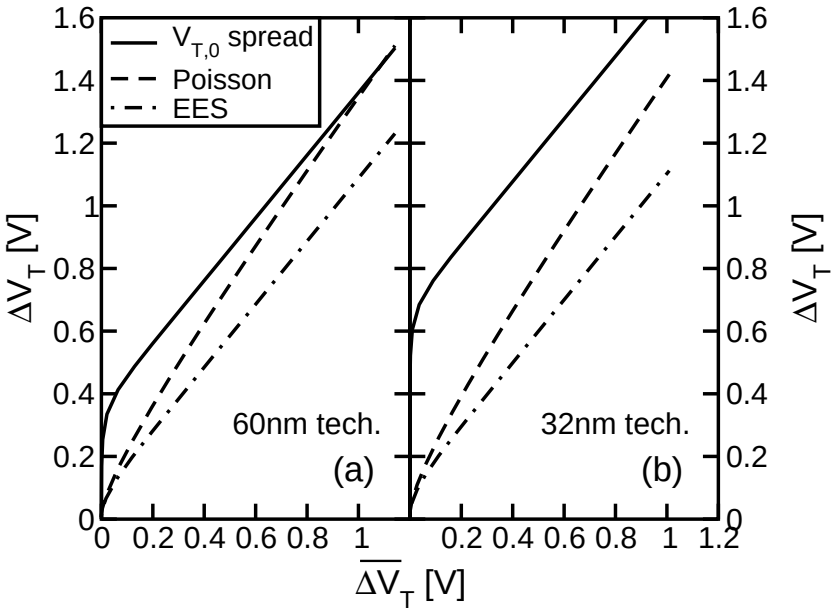


Figure 5.11: *Calculated edge V_T loss as a function of the average loss for a* 60 *(a) and a* 32 *nm (b) NAND technology.*

the same average V_T loss. Clearly, such predictions largely over-estimate the distribution spread when large V_T losses are reached.

To compare the two effects previously discussed, Fig. 5.11a shows the calculated edge V_T loss at a probability level equal to 10^{-5} (ΔV_T) as a function of the average array V_T loss ($\overline{\Delta V_T}$), for the same τ_r. This representation is preferred to a plot of the rms value of the V_T distribution at τ_r due to the asymmetrical shape of the distributions of Fig. 5.10(a). Results are shown including either the $V_{T,0}$ spread or the EES contribution to data retention dispersion. It can be seen that the larger edge V_T loss is given by the $V_{T,0}$ spread, revealing that cell-to-cell parameter variations represent the dominant variability source for data retention (similar results were also obtained starting from different $V_{T,P}$ distributions, having a non-zero width as a result of a non-ideal program operation). This is even more true considering that only the atomistic doping contribution to $\sigma_{V_{T,0}}$ has been here taken into account and that the V_T transients in Fig. 5.10(a) have been calculated by (5.12) neglecting any statistical fluctuation in α_G, t_{ox} and in the tunneling current I_{ox} (due, for instance, to fluctuations in the curvature of the floating-gate edge). All these fluctuations represent additional cell-to-cell variability sources to be accounted for a comprehensive investigation of the statistical dispersion of data retention. Note also that if a poissonian approximation is (incorrectly) used for the EES (dashed line), its contribution to the spread of the V_T distribution in Fig. 5.10(b) continuously increases, eventually exceeding the intrinsic variability. Though this conclusion is wrong, it should be pointed out that the cross-over takes place at quite large ΔV_T, much larger than acceptable values for multi-level NAND devices.

Finally, Fig. 5.11b shows that these same conclusions are valid for a 32 nm

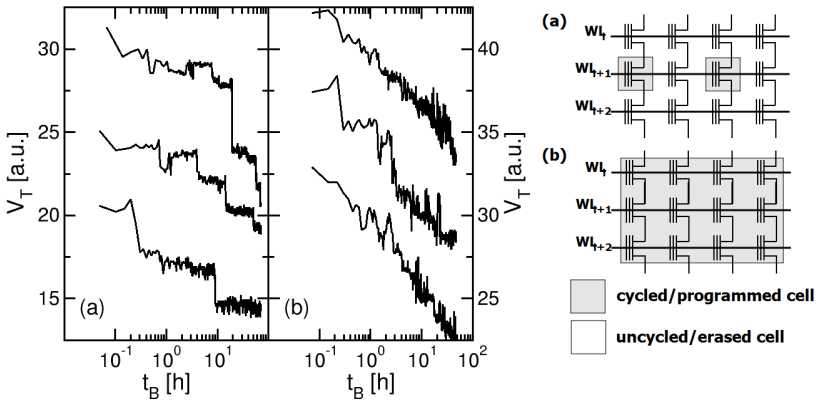


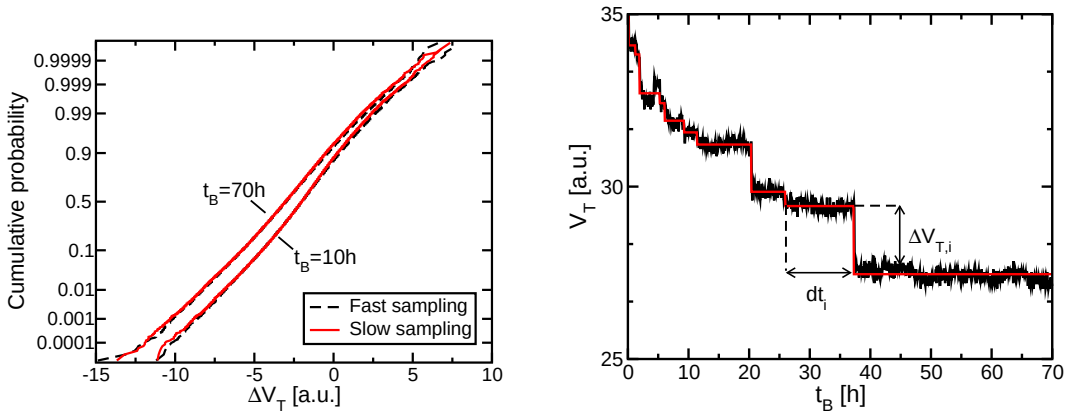
Figure 5.12: *V_T transients of some cells when (a) a single page or (b) all the cells are subjected to cycling and bake.*

NAND technology, due to the interplay between the $\sigma_{V_{T,0}}$ increase in inverse proportion to the square root of the cell area in (5.11) and the increase of the standard deviation of the EES in Fig. 5.10(b), in inverse proportion with the square root of C_{pp} [184]. Since C_{pp} is usually not scaled proportionally to the cell area in modern NAND technologies, the EES impact on data retention is expected to remain negligible with respect to $V_{T,0}$ variability.

As a last remark, note that the same analysis presented here for the direct-tunneling charge-loss can be easily extended to the case of defect-assisted tunneling,by changing the considered I_{ox}-F_{ox} characteristics used to perform the calculations leading to Figs. 5.10(a) and 5.10(b).

5.5 Statistical analysis of discrete detrapping events

The analysis presented in the previous section accounts for data retention variability sources on fresh arrays; charge detrapping from the tunnel oxide of cycled cells, however, represents a historic reliability limitation for Flash memories [58, 63], leading to data retention limiting threshold-voltage (V_T) instabilities, as discussed in Chapters 2-4. Notwithstanding a deep investigation on the topic, all experimental evidence reported thus far has shown continuous V_T transients [60, 63], where the expected discrete nature of the detrapping process is not readily apparent due averaging effects. Discrete V_T transients have been shown but only for the case of electron emission from the floating gate in uncycled, single test devices having an ultra-thin tunnel oxide [183]. This section demonstrate the first array-level experimental observation of post-cycling discrete electron detrapping from the tunnel oxide in sub-30nm NAND Flash arrays, investigating the statistical nature of the mechanism, including its deviation from the usually assumed Poisson statistics. Finally, this section shows that such experimental data can be used to reliably and accurately model the post cycled data retention performance of arrays. These results provide important insights into the fundamental scaling challenges of aggressively scaled NAND Flash technologies, where the impact of single electrons and defects becomes increasingly important, and pave the way for



(a) (b)

Figure 5.13: *(a) ΔV_T cumulative distributions after $t_B = 10$ and 70 h for fast sampling (1 read every $\simeq 10$ min) and slow sampling (1 read every $\simeq 1$ h); and (b) example of $\Delta V_{T,i}$ and dt_i extraction from a detrapping V_T transient (fast sampling).*

further analyses of NAND Flash reliability, based on the individual monitoring of the behavior of single electrons and defects.

V_T instabilities were investigated during post-cycling bake experiments at 100^{o}C on a single page of a sub-30 nm NAND Flash array. To resolve the discrete nature of the detrapping process, a special care was paid on minimizing parasitic contributions coming from damage recovery in the other cells in the array, which can affect the measurement of the V_T transient via interference effects and series resistance recovery, as discussed in Chapter 3. To this aim, the program operations were first applied to the selected page only, limiting the electrical stress of the unselected cells to the block erase operation only. Then, to minimize V_T instabilities in the unselected cells during bake, only the selected page was brought to the programmed state after cycling, leaving all the other pages erased. Fig. 5.12a shows the resulting V_T transients for a few cells of the selected page during bake, clearly revealing discrete V_T-loss events as time elapses, corresponding to single-electron detrapping from the tunnel oxide. For comparison, Fig. 5.12b shows V_T transients from a second experiment where all the cells in the array were subjected to both program and erase operations during cycling and all the pages were brought to the programmed state prior to bake. In this (more common) case, a more complex situation appears, where the detrapping events are masked by damage recovery in the unselected cells, resulting in a nearly-continuous behavior.

Fig. 5.12a reveals that a V_T sampling time in the tens-of-minutes timescale was used to accurately resolve the time dynamics of the detrapping events. This choice is a consequence of cycling dose and bake temperature and was met by directly performing V_T reads at the bake temperature, thus preventing any time uncertainty coming from wafer cooling and warming during the experiment. However, this rather short sampling time has the drawback of giving rise to a very frequent read stress on the array, which might activate parasitic effects contributing to V_T instabilities (see Chapter 4). In order to exclude this possibility, the cumulative distribution of the cell-to-cell V_T shift (ΔV_T) after 10 and 70 h of bake

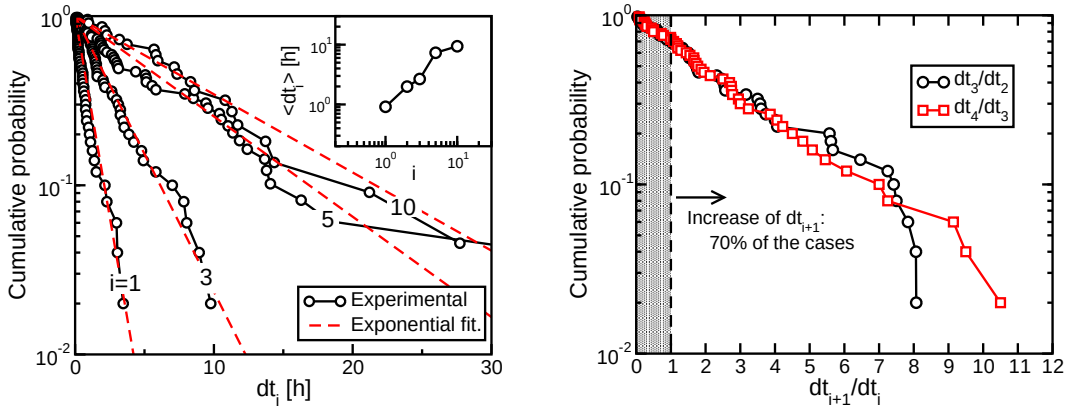


Figure 5.14: *(a) dt_i cumulative distributions for $i = 1, 3, 5, 10$; the inset shows $\langle dt_i \rangle$ as a function of i; and (b) cumulative probability of the ratio dt_{i+1}/dt_i: in more than 70% of the cases the ratio is > 1, i.e., dt_{i+1} is longer than dt_i.*

obtained from the previous experiment is compared with the one obtained using a longer sampling period of 1 hour (Fig. 5.13(a)): the distributions are the same, confirming that the observed V_T transients are only due to the detrapping process.

The main statistical features of the discrete detrapping process were investigated on a population of 50 cells out of the high-ΔV_T tail of the distribution in Fig. 5.13(a). For each cell, the V_T transient was "squared" as shown in Fig. 5.13(b), identifying the time delay dt_i needed for the i^{th} detrapping event to occur and the related V_T drop ($\Delta V_{T,i}$). Fig. 5.14(a) shows that the resulting dt_i statistics can be approximated by an exponential distribution whose average value $\langle dt_i \rangle$ increases with i (see inset). This means that a detrapping event reduces the probability for the next to occur, as confirmed in Fig. 5.14(b), where the statistical distribution of the ratio dt_{i+1}/dt_i highlights that in the vast majority of cells (nearly 70%) dt_i increases with i. This increase may originate from: 1) a reduction in the number of electrons available for detrapping (decrease in the supply); 2) a reduction in the tunnel-oxide electric field affecting the detrapping rate; 3) activation of deeper traps (movement of the emission front away from the interface).

Results in Figs. 5.14(a)-5.14(b) reveal that the electron detrapping process is *not* Poissonian. However, if the number of detrapping events (n_d) occurring in a time t_B is small, a negligible error is committed by adopting a Poisson distribution, as shown in Fig. 5.15(a). This appears also from the standard deviation vs. mean value (σ_{n_d} vs. $\langle n_d \rangle$) relation for n_d in Fig. 5.15(b), which can be approximated by a square-root dependence typical of a Poisson process. However, a departure of σ_{n_d} from this dependence is expected at high $\langle n_d \rangle$ due to the true sub-Poissonian nature of the detrapping process appearing in Figs. 5.14(a)-5.14(b). This deviation is shown in Fig. 5.15(b) by calculation results obtained assuming a sub-Poissonian emission process with the $\langle dt_i \rangle$ increase extracted from the inset of Fig. 5.14(a). Calculations were carried out in the same way as described in the previous section (see (5.13).

To complete the statistical analysis of the discrete detrapping process, Fig. 5.16(a)

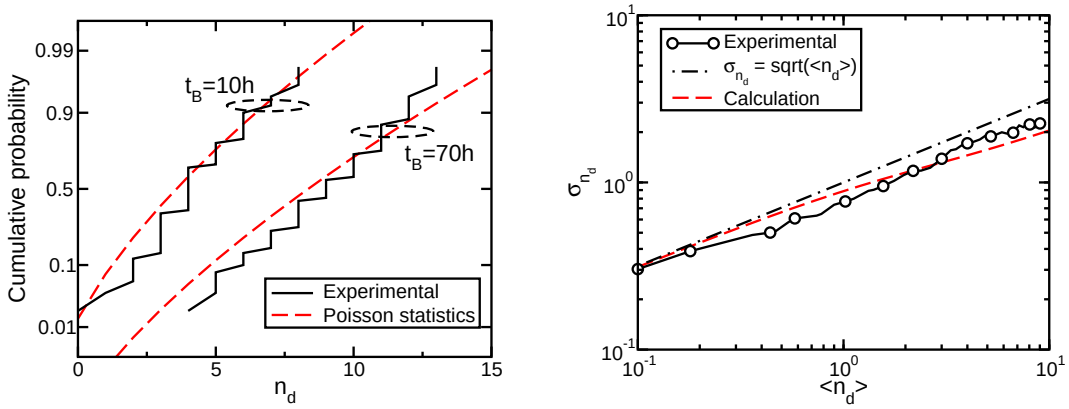


Figure 5.15: *(a) Number of detrapping events n_d after $t_B = 10$ and 70 h, compared with calculations according to Poisson statistics; and (b) experimental σ_{n_d} vs. $\langle n_d \rangle$ relation, compared with a Poisson approximation and more refined calculations.*

shows that the $\Delta V_{T,i}$ statistics displays an exponential distribution, as usually observed for localized point charges randomly placed over the channel of decananometer MOS devices, where percolative source-to-drain conduction occurs [69]. Using this latter piece of information and assuming a Poisson statistics for electron detrapping, the ΔV_T distribution at fixed t_B can be reproduced as shown in Fig. 5.16(b). However, electron detrapping toward the channel can only provide negative ΔV_T. Random telegraph noise (RTN), which allows for the possibility of a cell V_T to randomly increase between two successive read operations, is included in the model to account for the positive ΔV_T shift observed in the experimental data. In conclusion, this section reports the first experimental evidences of discrete electron emission giving rise to V_T detrapping transients in Flash memories, investigating its statistical properties and showing that a suitable model accounting for detrapping and RTN can reproduce the statistical retention data.

5.6 Conclusions

This chapter presented a careful analysis of the fundamental sources of variability which set the ultimate limits in nanoscale Flash memories operation. Besides cell-to-cell parameter variations and process tolerances, due to the technology scaling, the discrete nature of both the substrate doping and the charge flow into/from the FG emerges as a fundamental limitation to memory functionality. In particular, neutral threshold voltage spread, on one hand, sets additional burdens to programming algorithms; on the other hand, $V_{T,0}$ spread was shown to strongly affect retention transients of the cells in the array. Charge granularity, in turn, was demonstrated to set the ultimate limit to programming accuracy, thus requiring an accurate cell design in order to minimize its impact. Limitations to program accuracy due to EIS and programming algorithms optimizations, however, will be addressed in details in Chapter 6.

In this respect, a careful investigation of the impact of CG and FG design on the

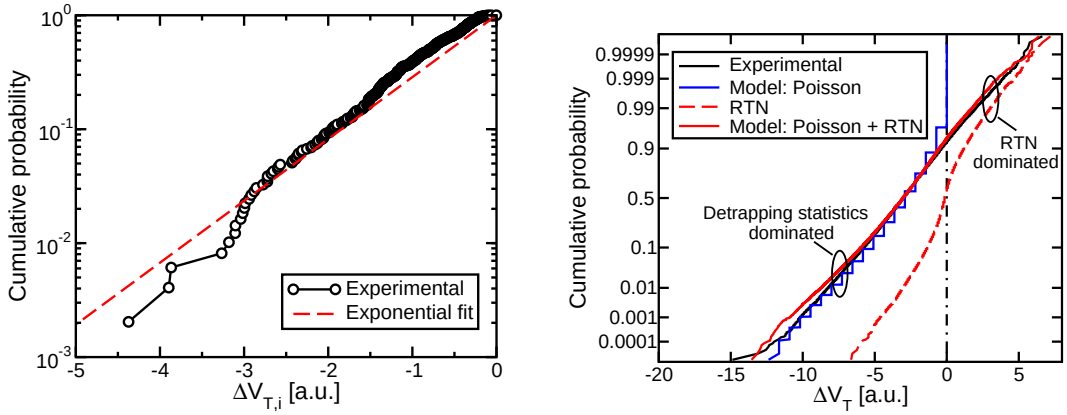


Figure 5.16: *(a) $\Delta V_{T,i}$ cumulative distribution; and (b) RTN and post-bake ΔV_T cumulative distributions compared with model results, including only Poisson statistics on the number of emission events or Poisson statistics and RTN fluctuations.*

EIS of deca-nanometer NAND Flash memories was carried out. CG polysilicon doping was shown to rule the reduction of C_{pp} when moving from the read to the program conditions which, however, barely impacts EIS. In fact, in the low V_s regime commonly used by the ISPP algorithm, EIS was shown to depend only on $C_{pp,R}$. Finally, the scaling trend of C_{pp} and EIS was addressed, discussing the evolution of the FG polysilicon in terms of geometry and dimensions.

Then, a detailed investigation of the statistical dispersion of data retention of nanoscale NAND Flash memories due to $V_{T,0}$ spread and EES has been presented. Both the previous sources of statistical dispersion result into a broadening of the array V_T distribution as retention time elapses, but the larger variability in data retention appears from the $V_{T,0}$ spread, and not from EES. These results reveal that cell-to-cell variability and not EES will represent the major issue for data retention dispersion in future nanoscale NAND technologies.

Finally, results showed in the Section 5.5 report that careful array measurements facilitate the observation of discrete electron emission from the tunnel oxide with sufficient detail to enable the development of an accurate stochastic model of cycled array data retention. Such a model has important practical application in its predictive ability for sub-30nm NAND Flash arrays.

CHAPTER 6

Programming accuracy of ISPP algorithms

This chapter presents a detailed investigation of the performance of a double-verify algorithm for accurate programming of deca-nanometer NAND Flash memories. In order to minimize the programmed threshold-voltage distribution width in presence of discrete and statistical electron injection, a weakened programming step is applied to cells if their threshold voltage falls between a low- and a high-program-verify level during incremental step pulse programming. Clear improvements are shown with respect to the single-verify case, with minimal burdens on programming time and complexity.

6.1 Introduction

Accurate programming of NAND Flash memories is usually obtained by the incremental step pulse programming (ISPP) algorithm [150], consisting in the application to cell control-gate (with grounded bulk and channel) of short programming pulses of equal duration τ_s and increasing amplitude. This algorithm allows very tight threshold voltage (V_T) distributions to be obtained when a constant increase V_s is given to the control-gate pulses, leading to an average V_T variation per step ($\Delta V_{T,s}$) rapidly converging to V_s [150,151,184,193]. In this case, inserting a verify operation after each pulse and stopping the algorithm when cell V_T exceeds the desired program-verify (PV) level, a maximum width of the programmed V_T distribution equal to V_s should theoretically be obtained, regardless of the width of the neutral V_T distribution. However, this result was shown to be compromised by the electron injection statistics (EIS), introducing a statistical spread in $\Delta V_{T,s}$ and allowing the cells to be displaced from the PV level more than V_s [5,151,184,186]. The severe scaling of the NAND technology, and in particular of the control-gate

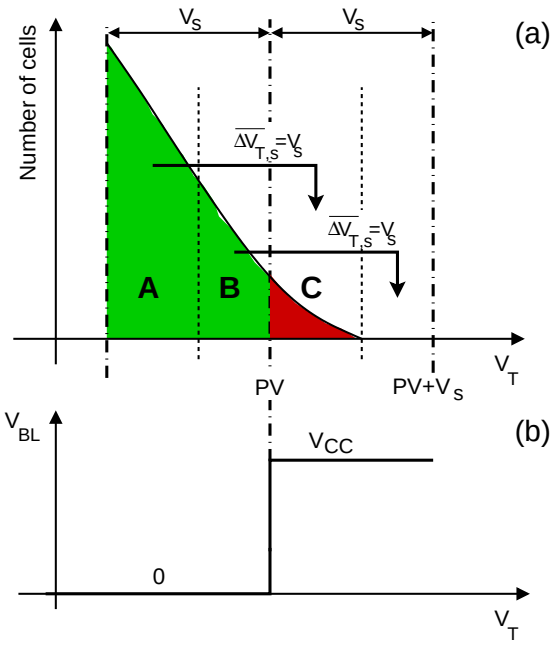


Figure 6.1: *Schematic for the single-verify algorithm (a) and bit-line bias applied during the next programming pulse as a function of V_T (b): the bit-line is kept to ground for cells below the PV level (region* **A** *and region* **B** *of the V_T axis), while V_{BL} is raised to V_{CC} for program inhibit when cells overcome the PV level (region* **C***).*

to floating-gate capacitance (C_{pp}) [194], has in fact increased the impact of single electrons stored in the floating gate on cell V_T, reducing, in turn, the number of electrons to be transferred to accomplish the program operation [5, 189, 190]. As a consequence, the statistical process ruling the granular electron injection into the floating gate during programming has become a major source of dispersion for the final cell V_T [5, 151, 184, 186]. The possibility to overcome the single-verify (SV) accuracy limitations by means of more complex double-verify (DV) algorithms [195, 196] has never been clearly assessed so far.

This chapter investigates an ISPP-based DV algorithm, considering its ability to tighten the V_T distribution in presence of EIS. The algorithm compares cell V_T with two PV levels, namely a high-PV (HPV), used to determine the end of the program operation, and a low-PV (LPV) level: in the case cell V_T falls between LPV and HPV, a positive bit-line bias (V_{BL}) is applied to the selected string to reduce the V_T growth when the next ISPP pulse is applied. In so doing, the programmed V_T distribution can be tighter than V_s, trading-off the programming speed with a better programming accuracy. In order to correctly evaluate the algorithm performance, the programmed V_T distribution width is studied by means of Monte Carlo (MC) simulations for the electron injection process, therefore carefully accounting for the EIS spread. Results show clear improvements in the programmed V_T distribution width, with minimal burdens on the programming time and complexity.

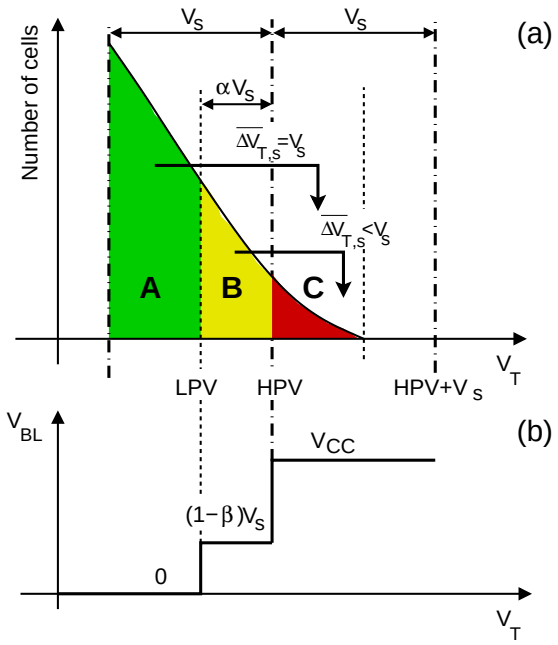


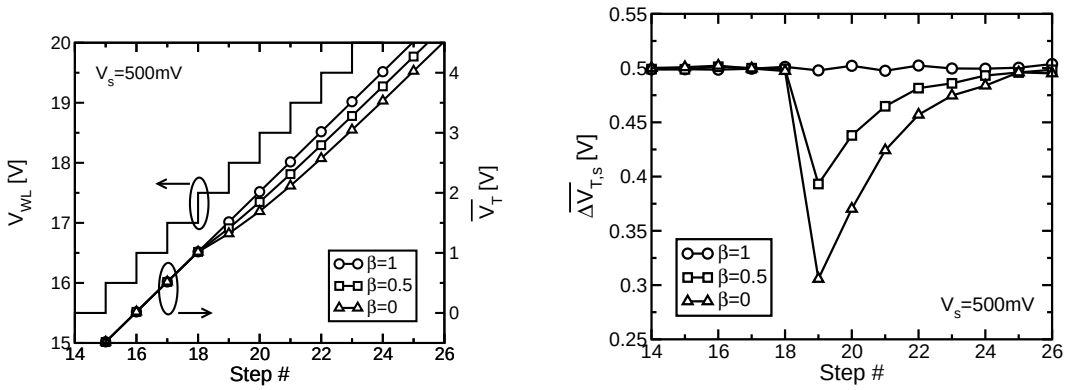
Figure 6.2: *Schematic for the double-verify algorithm investigated in this work (a) and bit-line bias applied during the next programming pulse as a function of V_T (b): the bit-line is kept to ground for cells that fall in region* **A** *of the V_T axis, while $V_{BL} = (1-\beta)V_s$ is applied if cell V_T is intermediate between the LPV and HPV levels (region* **B***). When cells overcome the HPV level (region* **C***), V_{BL} is raised to V_{CC} for program inhibit.*

6.2 Double-verify ISPP algorithm

6.2.1 Algorithm description

The basic features of SV ISPP of NAND arrays are schematically shown in Fig. 6.1: when a sequence of programming pulses whose amplitude has a constant increase V_s is applied to the selected word-line, cells below the PV level (namely, **A** and **B** in the figure) display an average V_T increase per step $\overline{\Delta V_{T,s}} = V_s$, while cells having $V_T >$ PV (namely, cells **C**) preserve their V_T state thanks to the inhibit bit-line voltage $V_{BL} = V_{CC}$ applied for channel boosting [99]. In so doing, the width of the programmed V_T distribution is mainly set by cells **B**, *i.e.* those that are closer to the PV level before overcoming it, on average, by V_s.

The basic idea for a DV ISPP algorithm [195, 196] is to reduce the V_T shift of the cells that come in close proximity to the PV level. To this aim, the conventional ISPP algorithm is modified as schematically depicted in Fig. 6.2: cell V_T is compared against two PV levels, namely HPV and LPV=HPV-αV_s, with $0 < \alpha < 1$, and three possible values of the bit-line bias are applied at the next programming pulse: (1) $V_{BL} = 0$ V for cells with $V_T <$ LPV (region **A** of the V_T axis in Fig. 6.2), (2) $V_{BL} = (1-\beta)V_s$ for cells with LPV $< V_T <$ HPV (region **B**) and (3) $V_{BL} = V_{CC}$ for cells with $V_T >$ HPV (region **C**). Cells in region **A** are considered sufficiently far from the end of programming, obtained when V_T overcomes the HPV level, to withstand an average increase of their V_T equal to V_s. As a consequence, their bit-line is kept to ground when the next word-line pulse



(a) (b)

Figure 6.3: *(a) Word-line bias applied during ISPP of the NAND cells (V_s = 500 mV). The resulting average V_T evolution from 10^3 MC simulations is also shown for the case $\beta = 1$ (conventional ISPP algorithm), $\beta = 0.5$ and $\beta = 0$, starting applying the bit-line bias at step* 18*; and (b) average $\Delta V_{T,s}$ as a function of the programming step. At step* 18 *the bit-line bias is applied, giving rise to a clear reduction of $\overline{\Delta V_{T,s}}$ in the next steps with respect to V_s.*

is applied as in the standard ISPP algorithm. Cells whose V_T is above the LPV but below the HPV level (region **B**) are instead close to the end of programming: a positive bit-line bias $V_{BL} = (1-\beta)V_s$ (with $\beta \leq 1$) is applied in this case during all the following programming pulses. The bit-line bias applied to cells **B** aims at reducing the increase of their tunnel-oxide electric field during the next pulses and should, however, be sufficiently low to keep the string-select transistor ON during program. In so doing, these cells will experience an average V_T increase that is lower than V_s, therefore limiting the maximum displacement they can have from the HPV level when they overcome it. Finally, cells having V_T beyond the HPV level (region **C**) have reached the end of their programming transient and no further V_T increase is needed. Therefore, their bit-line starts to be biased at V_{CC} to inhibit the effect of the following word-line pulses.

6.2.2 Effect of a bit-line bias on the ISPP transients

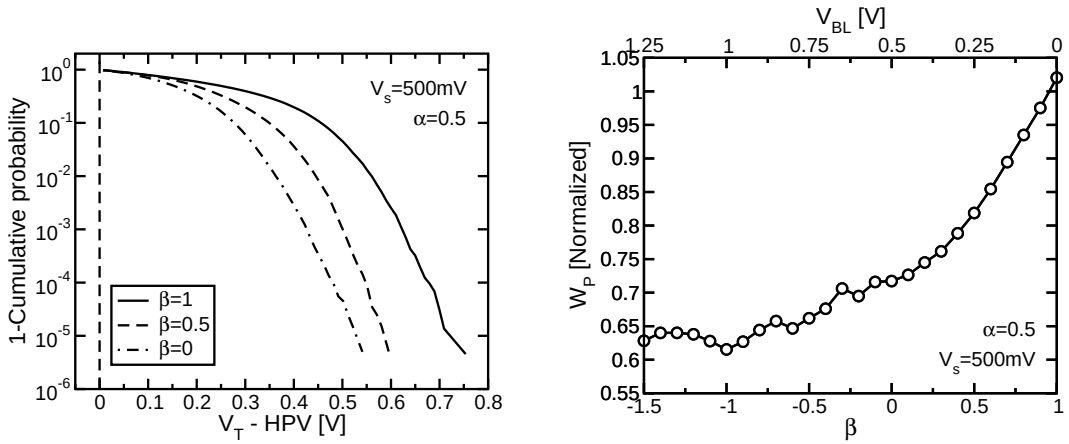
Figs. 6.3(a)-6.3(b) display the average behavior out of 10^3 MC simulations for the V_T transient during ISPP, in the case of β = 1, 0.5 and 0. The MC model, originally presented and validated against experimental data in [151] for SV ISPP, reproduces the discrete electron injection during programming, correctly describing the tunnel-oxide field variations due to (1) electron storage in the FG, (2) word-line bias increase and, in the case of the DV algorithm, (3) bit-line bias increase (more details on the simulation procedure will be given in the next Section). The same neutral V_T was initially assumed for the cells. Fig. 6.3(a) shows the staircase waveform applied to the selected word-line (V_s = 500 mV) and the simulated average V_T evolution in the case $\beta = 1$ (circles), clearly displaying $\overline{\Delta V_{T,s}} = V_s$. Note, in fact, that $\beta = 1$ represents the reference programming transient, with no bit-line bias applied, and should be compared to the $\beta = 0.5$ and $\beta = 0$ cases, corresponding to V_{BL} = 250 mV and 500 mV. The bit-line bias was

always applied since step 18 in the figure: in the following steps, the V_T transients for $\beta = 0.5$ and $\beta = 0$ depart from the reference transient, first showing a slower programming for some steps and then growing parallel to the $\beta = 1$ curve, with only a horizontal shift depending on the selected β value. This is also clearly highlighted in Fig. 6.3(b), where the average $\Delta V_{T,s}$ abruptly moves from V_s to a quite lower value after step 18. The reduction of $\overline{\Delta V_{T,s}}$ is due to the application of the bit-line bias, increasing the channel potential and decreasing the tunnel oxide field. This is, however, only a transient effect, with $\overline{\Delta V_{T,s}}$ that increases and converges again to V_s after some programming steps. This is due to the constant increase V_s of the word-line pulses, leading to the recovery of the stationary working point for the tunneling current [184].

6.2.3 Algorithm design parameters

The reduction of $\overline{\Delta V_{T,s}}$ immediately after the application of the bit-line bias in Fig. 6.3(b) can be exploited to tighten the programmed V_T distribution. In fact, if the cells overcome the HPV level during the transient phase during which $\overline{\Delta V_{T,s}} < V_s$, their maximum displacement from this level at the end of programming can be reduced. This can be obtained when the HPV is not too far from the LPV level, determining the initial step for the application of V_{BL}. Otherwise, if the HPV level is not overcome during the transient V_T growth following the application of V_{BL}, the same width of the V_T distribution of the conventional ISPP algorithm is obtained, due to the convergence of $\overline{\Delta V_{T,s}}$ to V_s. In addition to that, note that the achievable reduction of $\overline{\Delta V_{T,s}}$ in Fig. 6.3(b) depends on β, with lower β giving a more abrupt decrease of $\overline{\Delta V_{T,s}}$ and the possibility for tighter V_T distributions to be obtained. All these considerations suggest that the programmed V_T distribution width obtained by the DV algorithm can be optimized by a careful selection of both α and β. For this optimization, the maximum number of pulses required by cells receiving the bit-line bias to overcome the HPV level is also a very important parameter. In fact, if very low $\overline{\Delta V_{T,s}}$ are obtained but the HPV is too far from the LPV level, a large number of programming pulses may be required by the cells to reach the end of the program operation, compromising the programming speed. All these points should be carefully considered for a proper design of the DV programming scheme.

As a final remark, note that the optimization of the algorithm in terms of α and β requires the width of the programmed V_T distribution to be investigated including all the sources of statistical spread that may compromise the algorithm accuracy. Spread sources may impact either the verify operation or the charge transfer to the floating gate during the programming pulses: precision of V_T sensing, stability of the PV levels and RTN [69, 71, 73, 74, 80, 92, 186] are among the main accuracy constraints of the former group, while the latter includes EIS and erratic behaviors [151,184,197,198]. In the next Sections, attention will be focused only to EIS, as this is related to device physics and not to the sensing circuitry, and, differently from erratic phenomena, impacts the programmed V_T distribution at high probability levels [199, 200].



(a) (b)

Figure 6.4: *(a) Simulated cumulative distributions of V_T obtained after ISPP programming with $V_s = 500\ mV$, in the case of the conventional SV algorithm ($\beta = 1$) and of two implementations of the DV algorithm ($\beta = 0.5$ and $\beta = 0$); and (b) simulated V_T distribution width at a probability level of 10^{-4} as a function of β for $V_s = 500\ mV$ and $\alpha = 0.5$. Results are normalized to the distribution width obtained from the SV algorithm.*

6.3 Programming accuracy in presence of EIS

6.3.1 Simulation methodology

In order to investigate the programming accuracy of the DV ISPP algorithm, a 32 nm NAND technology was taken as a reference, simulating the ISPP operation from the erased cell state in a MC fashion, taking into account EIS limitations. Each MC run consisted in the extraction of initial cell V_T from a gaussian distribution with average value equal to the erased level and standard deviation corresponding to the dispersion of neutral cell V_T of the technology [107]. Then the electron transfer process to the floating gate was simulated extracting the time delay between one electron injection to the next from an exponential distribution with average value q/I_t, where I_t is the tunneling current through the tunnel oxide, which is a function of the floating gate potential V_{FG}. In order to carefully reproduce the electron injection process during ISPP, V_{FG} and I_t were updated both after each electron storage in the floating gate and at the beginning of each programming pulse, when the word-line and, in some cases, the bit-line bias is modified [151]. In so doing, simulations account for the non-homogeneous nature of the Poisson process ruling electron injection from the substrate to the floating gate [184, 201], and can be used to extract the EIS contribution to programming dispersion for arbitrary V_s. In order to correctly reproduce the DV algorithm, V_{FG} is calculated taking into account the applied bit-line bias since the beginning of the first programming step at which cell V_T becomes larger than the LPV level and the application of the programming pulses is stopped when V_T overcomes the HPV level.

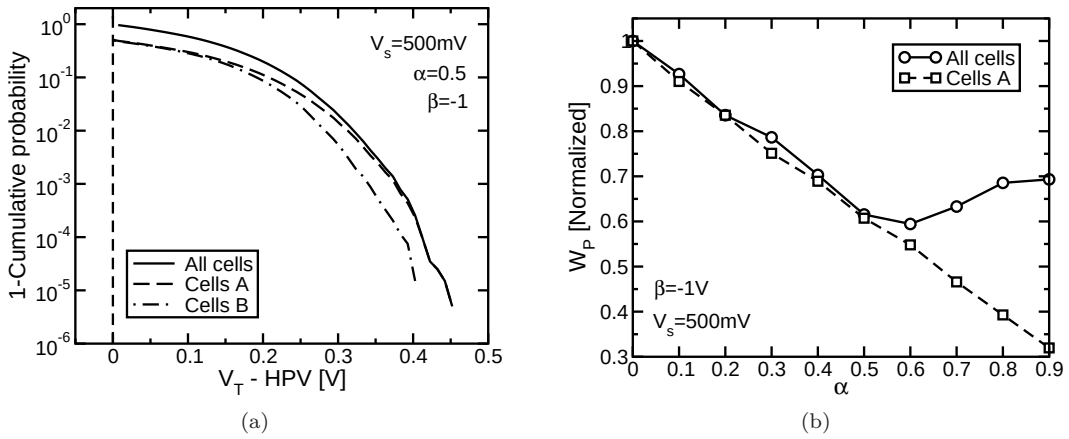


Figure 6.5: *(a) Simulated cumulative distribution of V_T programmed by the DV algorithm in the case $V_s = 500$ mV, $\alpha = 0.5$ and $\beta = -1$. Results including all the cells, only cells* **A** *or only cells* **B** *of Fig. 6.2 are shown; and (b) simulated V_T distribution width at a probability level of 10^{-4} as a function of α for $V_s = 500$ mV and $\beta = -1$. Results are normalized to the distribution width obtained from the SV algorithm.*

6.3.2 Simulation results

Fig. 6.4(a) shows the simulation results for the V_T distribution obtained after ISPP with $V_s = 500$ mV and $\alpha = 0.5$, assuming $\beta = 1$, 0.5 and 0. The former case gives the reference curve corresponding to the SV algorithm, showing the EIS impact making the distribution width larger than V_s, while the other two cases represent possible implementations of the DV algorithm. Results confirm that the algorithm presented in Section 6.2 can actually narrow the programmed V_T distribution with respect to the conventional SV programming scheme, pointing out also that the improvements strictly depend on the value of β. In order to discuss the results more quantitatively, the V_T distribution width (W_P) was extracted at a reference probability level of 10^{-4}. This level was chosen to have reliable results from the 10^5 MC simulations used to obtain the distributions of Fig. 6.4(a), nevertheless not being too much higher than the ECC level for NAND Flash. For $V_s = 500$ mV and $\alpha = 0.5$, Fig. 6.4(b) shows that a strong reduction of W_P with respect to what obtained from the SV algorithm can be obtained, revealing a 35% decrease when β is reduced from 1 to -1, with a clear saturation of W_P for $\beta < -1$.

The reduction of W_P with β in Fig. 6.4(b) reveals that the V_T distribution width after program is determined by cells **B** of Fig. 6.2, receiving the bit-line bias during the last programming pulses of the ISPP algorithm. The reduction of β from 1 ($V_{BL} = 0$) to -1 ($V_{BL} = 2V_s$) makes, in fact, cells **B** reduce their $\overline{\Delta V_{T,s}}$ in the steps immediately following the application of V_{BL}, as shown in Fig. 6.3(b). This, in turn, allows cells **B** to lower their maximum displacement from the HPV level at the end of programming, compacting the V_T distribution when cells **A** have a lower displacement from the HPV level. In this way, the best achievable accuracy is then obtained when the statistical distribution of cells **B** over the HPV level becomes narrower than the statistical distribution of cells **A**. These cells, in fact, do not receive any bit-line bias during programming and their final

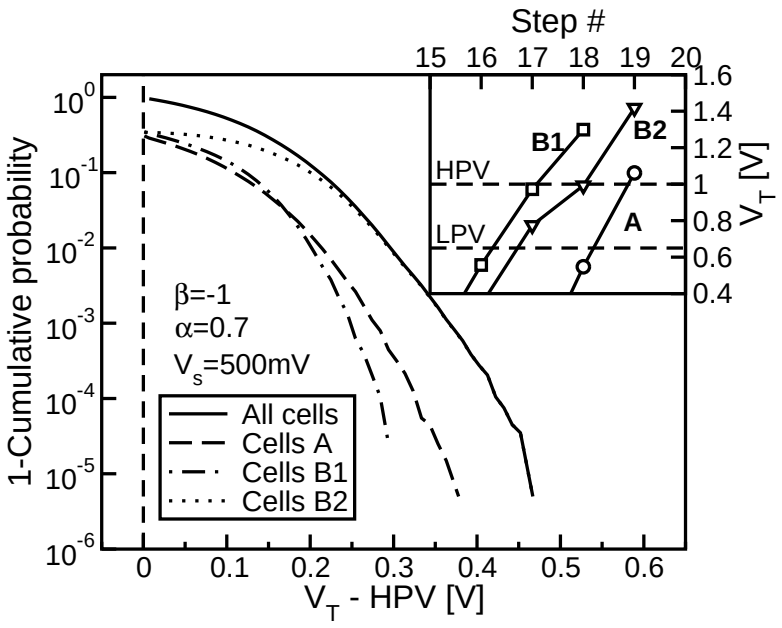


Figure 6.6: *Simulated cumulative distribution of V_T programmed by the DV algorithm in the case $V_s = 500$ mV, $\alpha = 0.7$ and $\beta = -1$. Cells* **B** *are shown considering those requiring only 1 (***B1***) or more than 2 (***B2***) steps to overcome HPV since V_{BL} is applied (see inset for an example).*

V_T distribution does not depend on β. Fig. 6.5(a) shows that for $V_s = 500$ mV and $\alpha = 0.5$ this condition is obtained when $\beta = -1$, as for this value the high-V_T tail of the programmed distribution is given by cells **A**, with cells **B** being closer to the HPV. This result means that for $V_s = 500$ mV and $\alpha = 0.5$, $\beta = -1$ represents the optimum conditions for the reduction of the programmed V_T distribution width, as confirmed in Fig. 6.4(b) by the saturation of W_P for β values lower than -1.

For $\beta = -1$, Fig. 6.5(b) shows that W_P is limited by and therefore decreases with the programmed V_T distribution of cells **A** for α ranging from 0 to 0.5, while for larger values of α, W_P is limited by cells **B** and grows for $\alpha > 0.6$. This is due to the larger number of cells of group **B** requiring more than two steps to overcome the HPV level after the application of V_{BL} (namely, **B2**) for larger α: as resulting from Fig. 6.3(b), these cells display a $\overline{\Delta V_{T,s}}$ at their last programming pulse that is larger than that of cells completing program in one pulse (**B1**) and, therefore, are the main limitation to W_P for large α (see Fig. 6.6).

6.3.3 Algorithm optimization

In order to explore the possibility for a further reduction of W_P when changing α from the 0.5 case addressed in Fig. 6.4(b), Fig. 6.7(a) shows a contour plot for the simulated W_P as a function of α and β. Results show that the parameters $\alpha = 0.5$ and $\beta = -1$ are near the optimum value for the DV algorithm in the case $V_s = 500$ mV, with only a slightly better W_P obtained when α approaches 0.6. In the optimal conditions, Fig. 6.7(a) reveals that a reduction nearly equal to 40% can be obtained from the DV with respect to the SV algorithm. Note

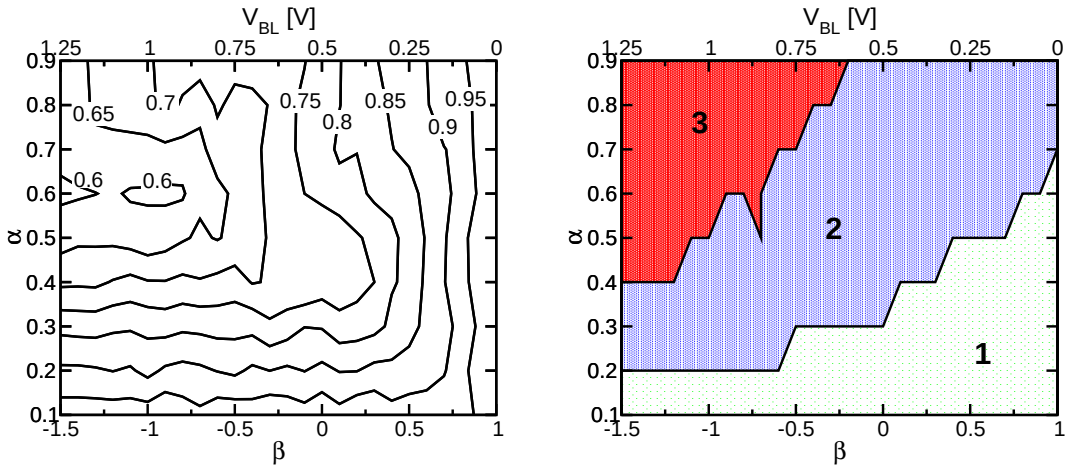


Figure 6.7: *Contour plot for: (a) the simulated W_P at a probability level of 10^{-4}; and (b) the maximum number of steps required by cells* **B** *to overcome the HPV level since the application of the bit-line bias; as a function of α and β in the case $V_s = 500$ mV. W_P levels are normalized to the distribution width obtained from the SV algorithm. The bit-line bias $V_{BL} = (1-\beta)V_s$ applied to cells* **B** *(see Fig. 6.2) is quoted on the upper axis. Results have been extracted from a statistics of 10^5 MC simulations.*

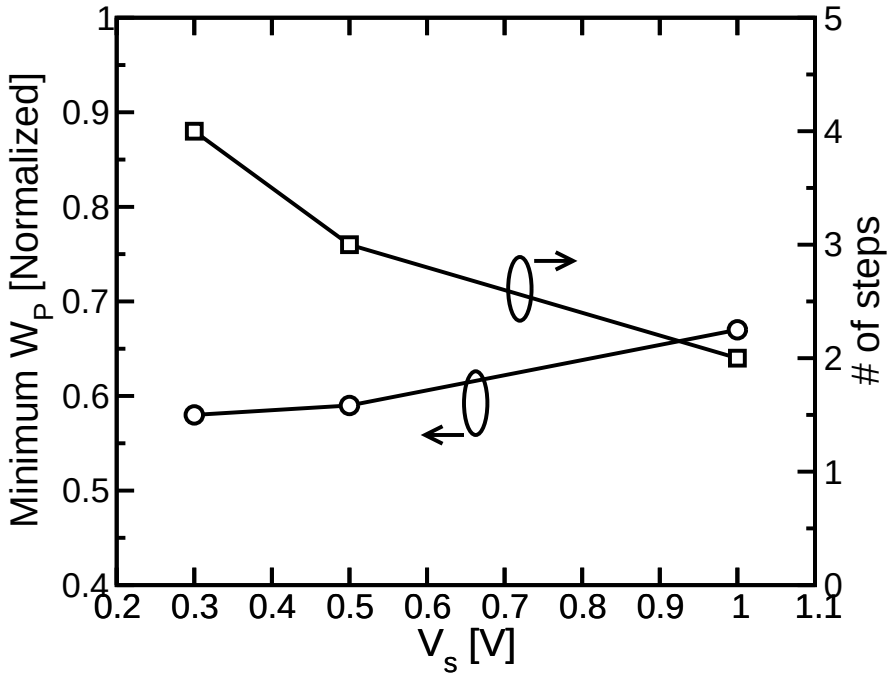


Figure 6.8: *Minimum W_P (normalized) achievable by the DV algortihm as a function of V_s, for $0 < \alpha < 1$ and limiting β to require a bit-line bias lower than 1 V. The maximum number of steps required by cells* **B** *to complete the program operation since the application of the bit-line bias is also shown.*

that this large improvement in W_P does not require a significant increase in the number of steps needed to complete the program operation. Fig. 6.7(b) shows, in fact, the maximum number of steps required by cells **B** to overcome the HPV level since the application of the bit-line bias, for $V_s = 500$ mV. Results have been extracted from a statistics of 10^5 MC simulations, considering the worst case cell for each value of α and β. From Fig. 6.7(b), in the case $\alpha = 0.5$ and $\beta = -1$, a maximum number of 3 steps are required for a cell to overcome the HPV level since the application of V_{BL}. This does not represent a critical delay of the programming speed when considering that a larger dispersion of the number of programming pulses is determined by the statistical spread of neutral cell V_T. This is is also confirmed by Fig. 6.8, showing the minimum W_P and the corresponding maximum number of steps required by cells **B** to complete program as a function of V_s, keeping the additional constraint of $V_{BL} < 1$ V in order to ensure that the string-select transistor is ON during program.

6.4 Conclusions

This chapter presented a detailed investigation of the accuracy of a DV algorithm for deca-nanometer NAND Flash memories. In order to account for EIS, optimization of the algorithm was studied by means of MC simulations for the electron injection process, showing that quite large V_T distribution narrowing can be obtained (*e.g.* for $V_s = 500$ mV and the optimal conditions $\alpha = 0.6$ and $\beta = -1$, corresponding to $V_{BL} = 1$ V applied to cells **B**, a narrowing nearly equal to 40% is obtained). These results are of fundamental importance for future NAND Flash technologies, especially for multi-level memory devices.

Summary of results

The research activity carried out during the Ph.D. program was focused on the reliability constraints of decananometer Flash memories and resulted in a deep understanding from a physical standpoint of the emerging failure mechanisms which affect Flash operation and set the ultimate limits to the technology scaling. Experimental activities and modeling efforts, including Monte Carlo modeling of programming algorithms, analytical modeling of electron injection/emission statistics, analysis of post-cycling threshold voltage instabilities and compact modeling of NAND string current, aimed at the characterization of a wide range of physical mechanisms, such as few electron phenomena, variability sources of cell parameters, tunnel oxide and IPD degradation due to electron trapping/detrapping and interface state creation/annealing, electrostatic interference between adjacent cells, impact of short channel effects and mobility degradation on NAND string current. The careful analysis of these mechanisms and the assessment of their impact on Flash reliability allowed to determine the fundamental scaling limitations and to propose feasible solutions for the future technology nodes, consisting in cell and algorithm design and optimization. Experimental and modeling activities, moreover, were devoted to the design of characterization techniques for ultra-scaled Flash technologies: accelerated testing schemes for the evaluation of distributed cycling effects and threshold voltage instabilities during data retention, in fact, emerged as a challenging and demanding task for Flash qualification and the impact of this research activity was recognized by both the semiconductor industry and the scientific community, since the main results achieved in this field were cited by JEDEC documents and awarded at the 2012 IEEE International Physics Reliability Symposium (see the List of Publications for more details).

In order to assess the constraints to post-cycling data retention, the attention was focused to cycling-induced damage creation and post-cycling threshold voltage instabilities, which were investigated under a wide range of experimental conditions. This analysis led to the assessment and validation of a universal damage-recovery model, allowing the development of accelerated qualification schemes which accurately reproduce the real on-field usage of the memory device.

Spurious effects which may emerge during the testing procedure itself, compromising the activation energy evaluation and the correct interpretation of distributed-cycling qualification schemes were also addressed, gaining a deep understanding of qualification schemes for nanoscale Flash devices. From a physical standpoint, damage creation and recovery dynamics and mechanisms were addressed via compact modeling of NAND string current. This approach allowed to evaluate the impact of electrostatics interference and short channel effect on NAND operation and to identify the major failure oxide degradation and recovery mechanisms, consisting in the charge trapping / detrapping in the oxide and in the interface state creation / annealing. The analysis of threshold voltage instabilities revealed that the latter mechanism comes into play when cell threshold voltage is monitored at high read current, due to carrier mobility degradation. Results show that threshold voltage instabilities are increased and their activation energy is lowered with respect to the usual 1.1 eV value given by charge detrapping whenever the saturation value of the string current moves too close to read current, thus, providing useful hints for the design of the read scheme and the choice of the operating voltages of the memory array.

Then, the investigation of few electron phenomena highlighted that cell state is controlled by a low number of electrons, which progressively decreases with size scaling, and, thus, the granularity of the charge flux into the floating gate emerged as a source of program noise, setting the ultimate limit to programming accuracy. In this regards, a careful investigation of the impact of CG and FG design on the electron injection statistics (EIS) of decananometer NAND Flash memories was carried out, discussing the EIS scaling trend and the evolution of the FG design in terms of geometry and dimensions. Moreover, a detailed investigation of the accuracy of programming algorithms was performed by means of Monte Carlo simulations for the electron injection process, showing that narrower programmed distribution can be obtained with an optimized double-verify ISPP algorithm, with respect to a standard single-verify ISPP scheme, with minimal burdens on algorithm complexity and programming speed. These results are of fundamental importance for future NAND Flash technologies, especially for multi-level memory devices. Then, the impact of few electron phenomena also on data retention was taken into account, investigating the statistical dispersion of data retention of nanoscale NAND Flash memories due to neutral threshold voltage spread and electron emission statistics (EES): both the phenomena results in a distribution widening during retention, but the modeling results reveal that cell-to-cell variability and not EES will represent the major issue for data retention dispersion of fresh devices in future nanoscale technologies. Finally, single electron detrapping from tunnel oxide during post-cycling retention was experimentally demonstrated in the sub-30 nm regime and a statistical model was developed, showing that post-cycling data retention variability is ruled by detrapping statistics in aggressively scaled Flash technologies.

In conclusion, this research activity provided the physical understanding, the modeling tools and the characterization techniques required to investigate the reliability of the state-of-the-art Flash memory technologies; the most of the emerging mechanisms, moreover, will likely affect also the future generations of charge-based

memory technologies, including planar and 3D Flash architectures, and, thus, this thesis provides the reference against novel memories should be tested, in order to assess the improvements with respect to the conventional Flash technology.

Bibliography

[1] R. Micheloni, L. Crippa, and A. Marelli, *Inside NAND Flash Memories*. Springer, 2010.

[2] A. Maconi, *Performance and Limitations of Charge-Trap Storage for Non-Volatile Memory Technologies*. PhD thesis, Politecnico di Milano - Doctoral Program in Information Technology, 2011.

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

[4] K. Kim, "Future memory technology: challenges and opportunities," in *VLSI-TSA Tech. Dig.*, pp. 5–9, 2008.

[5] K. Prall and K. Parat, "25 nm 64 Gb MLC NAND technology and scaling challenges," in *IEDM Tech. Dig.*, pp. 102–105, 2010.

[6] S. Lai, "Non-volatile memory technology: the quest for ever lower cost," in *IEDM Tech. Dig.*, pp. 1–6, 2008.

[7] C. M. Compagnoni, A. S. Spinelli, A. L. Lacaita, A. Ghetti, and A. Visconti, *NONVOLATILE MEMORIES -Materials, Devices and Applications*, ch. Emerging Constraints on NAND Flash Memory Reliability. American Scientific Publishers, 2012.

[8] D. Kahng and S. Sze, "A floating-gate and its application to memory devices," in *The Bell System Technical Journal*, pp. 1288–1295, 1967.

[9] S. Aritome, S. Satoh, T. Maruyama, H. Watanabe, S. Shuto, G. Hemink, R. Shirota, S. Watanabe, and F. Masuoka, "A 0.67μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V-only 256 Mbit NAND EEPROMs," in *IEDM Tech. Dig.*, pp. 61–64, 1994.

[10] "International technology roadmap for semiconductors," 2009.

[11] T. Kamigaichi, F. Arai, H. Nitsuta, M. Endo, K. Nishihara, T. Murata, H. Takekida, T. Izumi, K. Uchida, T. Maruyama, I. Kawabata, Y. Suyama, A. Sato, K. Ueno, H. Takeshita, Y. Joko, S. Watanabe, Y. Liu, H. Meguro, A. Kajita, Y. Ozawa, and T. Watanabe, "Floating gate super multi level NAND Flash memory technology for 30 nm and beyond," in *IEDM Tech. Dig.*, pp. 827–830, 2008.

[12] H. Nitta, T. Kamigaichi, F. Arai, T. Futatsuyama, M. Endo, N. Nishihara, T. Murata, H. Takekida, T. Izumi, K. Uchida, T. Maruyama, I. Kawabata, Y. Suyama, A. Sato, K. Ueno, H. Takeshita, Y. Joko, S. Watanabe, Y. Liu, H. Meguro, A. Kajita, Y. Ozawa, Y. Takeuchi, T. Hara, and T. Watanabe, "Three bits per cell floating gate NAND Flash memory technology for 30 nm and beyond," in *Proc. IRPS*, pp. 307–310, 2009.

[13] M. Park, K. Kim, J.-H. Park, and J.-H. Choi, "Direct field effect of neighboring cell transistor on cell-to-cell interference of NAND Flash cell arrays," *IEEE Electron Device Lett.*, vol. 30, pp. 174–177, Feb. 2009.

[14] G. Servalli, D. Brazzelli, E. Camerlenghi, G. Capetti, S. Costantini, C. Cupeta, D. DeSimone, A. Ghetti, T. Ghilardi, P. Gulli, M. Mariani, A. Pavan, and R. Somaschini, "A 65 nm NOR Flash technology with 0.042 μm^2 cell size for high performance multilevel application," in *IEDM Tech. Dig.*, pp. 869–872, 2005.

[15] D. James, "Nano-Scale Flash in the Mid-Decade," in *ASMC 2007. IEEE/SEMI*, pp. 371–376, 2007.

[16] K. Takeuchi, Y. Kameda, S. Fujimura, H. Otake, K. Hosono, H. Shiga, Y. Watanabe, T. Futatsuyama, Y. Shindo, M. Kojima, M. Iwai, M. Shirakawa, M. Ichige, K. Hatakeyama, S. Tanaka, T. Kamei, J.-Y. Fu, A. Cernea, Y. Li, M. Higashitani, G. Hemink, S. Sato, K. Oowada, S.-C. Lee, N. Hayashida, J. Wan, J. Lutze, S. Tsao, M. Mofidi, K. Sakurai, N. Tokiwa, H. Waki, Y. Nozawa, K. Kanazawa, and S. Ohshima, "A 56-nm CMOS 99-mm^2 8-Gb multi-level NAND Flash memory with 10-MB/s program throughput," *IEEE J. Solid-State Circuits*, vol. 42, pp. 219–232, Jan. 2007.

[17] D. Nobunaga, E. Abedifard, F. Roohparvar, J. Lee, E. Yu, A. Vahidimowlavi, M. Abraham, S. Talreja, E. Sundaram, R. Rozman, L. Vu, C. L. Chen, U. Chandrasekhar, R. Bains, V. Viajedor, W. Mak, M. Choi, D. Udeshi, M. Luo, S. Qureshi, J. Tsai, F. Jaffin, Y. Liu, and M. Mancinelli, "A 50 nm 8 Gb NAND Flash memory with 100 MB/s program throughput and 200 MB/s DDR interface," in *Proc. ISSCC*, pp. 426–427, 2008.

[18] C. Trinh, N. Shibata, T. Nakano, M. Ogawa, J. Sato, Y. Takeyama, K. Isobe, B. Le, F. Moogat, N. Mokhlesi, K. Kozazai, P. Hong, T. Kamei, K. Iwasa, J. Nakai, T. Shimizu, M. Honma, S. Sakai, T. Kawaai, S. Hoshi, J. Yuh, C. Hsu, T. Tseng, J. Li, J. Hu, M. Liu, S. Khalid, J. Chen, M. Watanabe, H. Lin, J. Yang, K. McKay, K. Nguyen, T. Pham, Y. Matsuda, K. Nakamura, K. Kanebako, S. Yoshikawa, W. Igarashi, A. Inoue, T. Takahashi, Y. Komatsu, C. Suzuki, K. Kanazawa, M. Higashitani, S. Lee, T. Murai, K. Nguyen, J. Lan, S. Huynh, M. Murin, M. Shlick, M. Lasser, R. Cernea, M. Mofidi, K. Schuegraf, and K. Quader, "A 5.6 MB/s 64 Gb 4 b/cell NAND Flash memory in 43 nm CMOS," in *Proc. ISSCC*, pp. 246–247, 2009.

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

[20] K.-T. Park, O. Kwon, S. Yoon, M.-H. Choi, I.-M. Kim, B.-G. Kim, M.-S. Kim, Y.-H. Choi, S.-H. Shin, Y. Song, J.-Y. Park, J.-E. Lee, C.-G. Eun, H.-C. Lee, H.-J. Kim, J.-H. Lee, J.-Y. Kim, T.-M. Kweon, H.-J. Yoon, T. Kim, D.-K. Shim, J. Sel, J.-Y. Shin, P. Kwak, J.-M. Han, K.-S. Kim, S. Lee, Y.-H. Lim, and T.-S. Jung, "A 7MB/s 64Gb 3-bit/cell DDR NAND Flash memory in 20nm-node technology," in *ISSCC Tech. Dig.*, pp. 212–213, 2011.

[21] M. Goldman, K. Pangal, G. Naso, and A. Goda, "25nm 64Gb 130mm^2 3bpc NAND Flash Memory," in *Proc. IMW*, pp. 1–4, 2011.

[22] J. Hwang, J. Seo, Y. Lee, S. Park, J. Leem, J. Kim, T. Hong, S. Jeong, K. Lee, H. Heo, H. Lee, P. Jang, K. Park, M. Lee, S. Baik, J. Kim, H. Kkang, M. Jang, J. Lee, G. Cho, J. Lee, B. Lee, H. Jang, S. Park, J. Kim, S. Lee, S. Aritome, S. Hong, and S. Park, "A middle-1x nm NAND Flash memory cell (M1X-NAND) with highly manufacturable integration technologies," in *IEDM Tech. Dig.*, pp. 199–202, 2011.

[23] D. J. DiMaria and E. Cartier, "Mechanism for stress-induced leakage currents in thin silicon dioxide films," *J. Appl. Phys.*, vol. 78, pp. 3883–3894, 1995.

[24] P. Samanta, "Hole trapping due to anode hole injection in thin tunnel gate oxides in memory devices under Fowler-Nordheim stress," *Applied Physics Letters*, vol. 75, pp. 2966–2968, Nov. 1999.

[25] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and M. J. van Duuren, "Defect generation statistics in thin gate oxides," *IEEE Trans. Electron Devices*, vol. 51, pp. 1288–1295, Aug. 2004.

[26] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "A new two-trap tunneling model for the anomalous SILC in flash memories," *Microelectron. Eng.*, vol. 59, pp. 189–195, 2001.

[27] J. Maserjian and N. Zamani, "Observation of positively charged state generation near the Si/SiO_2 interface during Fowler-Nordheim tunneling," *J. Vac. Sci. Technol.*, vol. 20, pp. 743–746, 1982.

[28] D. A. Baglee and M. C. Smayling, "The effects of write/erase cycling on data loss in EEPROMs," in *IEDM Tech. Dig.*, pp. 624–626, 1985.

[29] P. Olivo, T. N. Nguyen, and B. Riccò, "High-field-induced degradation in ultra-thin SiO_2 films," *IEEE Trans. Electron Devices*, vol. 35, pp. 2259–2267, Dec. 1988.

[30] K. Naruke, S. Taguchi, and M. Wada, "Stress induced leakage current limiting to scale down EEPROM tunnel oxide thickness," in *IEDM Tech. Dig.*, pp. 424–427, 1988.

[31] R. Rofan and C. Hu, "Stress-induced oxide leakage," *IEEE Electron Device Lett.*, vol. 12, pp. 632–634, Nov. 1991.

[32] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," in *IEDM Tech. Dig.*, pp. 139–142, 1992.

[33] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of Flash memory cells," *Proc. IEEE*, vol. 81, pp. 776–788, May 1993.

[34] P. P. Apte and K. C. Saraswat, "Correlation of trap generation to charge-to-breakdown (Q_{bd}): A physical-damage model for of dielectric breakdown," *IEEE Trans. Electron Devices*, vol. 41, pp. 1595–1602, Sept. 1994.

[35] H. Satake and A. Toriumi, "Common origin for stress-induced leakage current and electron trap generation in SiO_2," *Appl. Phys. Lett.*, vol. 67, pp. 3489–3490, 1995.

[36] S. Yamada, K. Amemiya, T. Yamane, H. Hazama, and K. Hashimoto, "Non-uniform current flow through thin oxide after Fowler-Nordheim current stress," in *Proc. IRPS*, pp. 108–112, 1996.

[37] M. Kimura and T. Ohmi, "Conduction mechanism and origin of stress-induced leakage current in thin silicon dioxide films," *J. Appl. Phys.*, vol. 80, pp. 6360–6369, 1996.

[38] N. Matsukawa, S. Yamada, K. Amemiya, and H. Hazama, "A hot hole-induced low-level leakage current in thin silicon dioxide films," *IEEE Trans. Electron Devices*, vol. 43, pp. 1924–1929, Nov. 1996.

[39] G. J. Hemink, K. Shimizu, S. Aritome, and R. Shirota, "Trapped hole enhanced stress induced leakage current in NAND EEPROM tunnel oxides," in *Proc. IRPS*, pp. 117–121, 1996.

[40] J. D. Blauwe, J. Van Houdt, D. Wellekens, R. Degraeve, P. Roussel, L. Haseslagh, L. Defern, G. Groeseneken, and H. E. Maes, "A new quantitative model to predict SILC-related disturb characteristics in Flash E^2PROM devices," in *IEDM Tech. Dig.*, pp. 343–346, 1996.

[41] E. F. Runnion, S. M. Gladstone, R. S. Scott, D. J. Dumin, L. Lie, and J. C. Mitros, "Thickness dependence of stress-induced leakage currents in silicon oxide," *IEEE Trans. Electron Devices*, vol. 44, pp. 993–1001, 1997.

[42] K. Sakakibara, N. Ajika, M. Hatanaka, H. Miyoshi, and A. Yasuoka, "Identification of stress-induced leakage current components and the corresponding trap models in SiO_2 films," *IEEE Trans. Electron Devices*, vol. 44, pp. 2267–2273, June 1997.

[43] K. Sakakibara, N. Ajika, K. Eikyu, K. Ishikawa, and H. Miyoshi, "A quantitative analysis of time-decay reproducible stress-induced leakage current in SiO_2 films," *IEEE Trans. Electron Devices*, vol. 44, pp. 1002–1007, June 1997.

[44] S. Satoh, G. Hemink, K. Hatakeyama, and S. Aritome, "Stress-induced leakage current of tunnel oxide derived from Flash memory read-disturb characteristics," *IEEE Trans. Electron Devices*, vol. 45, pp. 482–486, Feb. 1998.

[45] J. D. Blauwe, J. Van Houdt, D. Wellekens, G. Groeseneken, and H. E. Maes, "SILC-related effects in Flash E^2PROM's–Part I: A quantitative model for steady-state SILC," *IEEE Trans. Electron Devices*, vol. 45, pp. 1745–1750, Aug. 1998.

[46] J. D. Blauwe, J. Van Houdt, D. Wellekens, G. Groeseneken, and H. E. Maes, "SILC-related effects in Flash E^2PROM's–Part II: Prediction of steady-state SILC-related disturb characteristics," *IEEE Trans. Electron Devices*, vol. 45, pp. 1751–1760, Aug. 1998.

[47] G. B. Alers, B. E. Weir, M. A. Alam, G. L. Timp, and T. Sorch, "Trap assisted tunneling as a mechanism of degradation and noise," in *Proc. IRPS*, pp. 76–79, 1998.

[48] S. Takagi, M. Takayanagi, and A. Toriumi, "Experimental examination of physical model for direct tunneling current in unstressed/stressed ultrathin gate oxides," in *IEDM Tech. Dig.*, pp. 461–464, 1999.

[49] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC based on electron and hole tunneling – Part I: Transient effects," *IEEE Trans. Electron Devices*, vol. 47, pp. 1258–1265, June 2000.

[50] D. Ielmini, A. S. Spinelli, M. A. Rigamonti, and A. L. Lacaita, "Modeling of SILC based on electron and hole tunneling – Part II: Steady-state," *IEEE Trans. Electron Devices*, vol. 47, pp. 1266–1272, June 2000.

[51] F. Schuler, R. Degraeve, P. Hendrickx, and D. Wellekens, "Physical description of anomalous charge loss in floating gate based NVM's and identification of its dominant parameter," in *Proc. IRPS*, pp. 26–33, 2002.

[52] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "A statistical model for SILC in Flash memories," *IEEE Trans. Electron Devices*, vol. 49, pp. 1955–1961, 2002.

[53] P. Cappelletti, R. Bez, A. Modelli, and A. Visconti, "What we have learned on Flash memory reliability in the last ten years," in *IEDM Tech. Dig.*, pp. 489–492, 2004.

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

[55] S. Yamada, Y. Hiura, T. Yamane, K. Amemiya, Y. Ohshima, and K. Yoshikawa, "Degradation mechanism of Flash EEPROM programming after program/erase cycles," in *IEDM Tech. Dig.*, pp. 23–26, 1993.

[56] M. Kato, N. Miyamoto, H. Kume, A. Satoh, T. Adachi, M. Ushiyama, and K. Kimura, "Read-disturb degradation mechanism due to electron trapping in the tunnel oxide for low-voltage flash memories," in *IEDM Tech. Dig.*, pp. 45–48, 1994.

[57] Y.-B. Park and D. K. Schroder, "Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a Flash EEPROM," *IEEE Trans. Electron Devices*, vol. 45, pp. 1361–1368, June 1998.

[58] R. Yamada, Y. Mori, Y. Okuyama, J. Yugami, T. Nishimoto, and H. Kume, "Analysis of detrap current due to oxide traps to improve flash memory retention," in *Proc. IRPS*, pp. 200–204, 2000.

[59] R. Yamada, T. Sekiguchi, Y. Okuyama, J. Yugami, and H. Kume, "A novel analysis method of threshold voltage shift due to detrap in a multi-level Flash memory," in *Symp. VLSI Tech. Dig.*, pp. 115–116, 2001.

[60] J.-D. Lee, J.-H. Choi, D. Park, and K. Kim, "Degradation of tunnel oxide by FN current stress and its effects on data retention characteristics of 90 nm NAND Flash memory cells," in *Proc. IRPS*, pp. 497–501, 2003.

[61] J.-D. Lee, J.-H. Choi, D. Park, and K. Kim, "Data retention characteristics of sub-100 nm NAND Flash memory cells," *IEEE Electron Device Lett.*, vol. 24, pp. 748–750, 2003.

[62] J.-D. Lee, J.-H. Choi, D. Park, and K. Kim, "Effects of interface trap generation and annihilation on the data retention characteristics of Flash memory cells," *IEEE Trans. Device and Materials Reliab.*, vol. 4, pp. 110–117, Mar. 2004.

[63] 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 Trans. Device and Materials Reliab.*, vol. 4, pp. 335–344, Sep. 2004.

[64] N. Mielke, H. Belgal, A. Fazio, Q. Meng, and N. Righos, "Recovery effects in the distributed cycling of Flash memories," in *Proc. IRPS*, pp. 29–35, 2006.

[65] M. Park, K. Suh, K. Kim, S.-H. Hur, K. Kim, and W.-S. Lee, "The effect of trapped charge distributions on data retention characteristics of NAND Flash memory cells," *IEEE Electron Device Lett.*, vol. 28, pp. 750–752, Aug. 2007.

[66] A. Fayrushin, K. Seol, J. Na, S. Hur, J. Choi, and K. Kim, "The new program/erase cycling degradation mechanism of NAND Flash memory devices," in *IEDM Tech. Dig.*, pp. 823–826, 2009.

[67] C. Monzio Compagnoni, C. Miccoli, R. Mottadelli, S. Beltrami, M. Ghidotti, A. L. Lacaita, A. S. Spinelli, and A. Visconti, "Investigation of the threshold voltage instability after distributed cycling in nanoscale NAND Flash memory arrays," in *Proc. IRPS*, pp. 604–610, 2010.

[68] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "Statistical modeling of reliability and scaling projections for Flash memories," in *IEDM Tech. Dig.*, pp. 703–706, 2001.

[69] A. Ghetti, C. Monzio Compagnoni, A. S. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer Flash memories," *IEEE Trans. Electron Devices*, vol. 56, pp. 1746–1752, Aug. 2009.

[70] A. Ghetti, C. Monzio Compagnoni, F. Biancardi, A. L. Lacaita, S. Beltrami, L. Chiavarone, A. S. Spinelli, and A. Visconti, "Scaling trends for random telegraph noise in deca-nanometer Flash memories," in *IEDM Tech. Dig.*, pp. 835–838, 2008.

[71] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, "The impact of random telegraph signals on the scaling of multilevel Flash memories," in *Symp. VLSI Circ. Dig.*, pp. 140–141, 2006.

[72] P. Fantini, A. Ghetti, A. Marinoni, G. Ghidini, A. Visconti, and A. Marmiroli, "Giant random telegraph signals in nanoscale floating-gate devices," *IEEE Electron Device Lett.*, vol. 28, pp. 1114–1116, Dec. 2007.

[73] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, "Random telegraph signal in Flash memory: its impact on scaling of multilevel Flash memory beyond the 90-nm node," *IEEE J. Solid-State Circuits*, vol. 42, pp. 1362–1369, 2007.

[74] C. Monzio Compagnoni, M. Ghidotti, A. L. Lacaita, A. S. Spinelli, and A. Visconti, "Random telegraph noise effect on the programmed threshold-voltage distribution of Flash memories," *IEEE Electron Device Lett.*, vol. 30, pp. 984–986, Sep. 2009.

[75] C. Monzio Compagnoni, L. Chiavarone, M. Calabrese, M. Ghidotti, A. L. Lacaita, A. S. Spinelli, and A. Visconti, "Fundamental limitations to the width of the programmed V_T distribution of NOR Flash memories," *IEEE Trans. Electron Devices*, vol. 57, pp. 1761–1767, Aug. 2010.

[76] R. Gusmeroli, C. Monzio Compagnoni, A. Riva, A. S. Spinelli, A. L. Lacaita, M. Bonanomi, and A. Visconti, "Defects spectroscopy in SiO_2 by statistical random telegraph noise analysis," in *IEDM Tech. Dig.*, pp. 483–486, 2006.

[77] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate Flash memory," in *IEDM Tech. Dig.*, pp. 491–494, 2006.

[78] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random telegraph noise in Flash memories - model and technology scaling," in *IEDM Tech. Dig.*, pp. 169–172, 2007.

[79] H. Miki, T. Osabe, N. Tega, A. Kotabe, H. Kurata, K. Tokami, Y. Ikeda, S. Kamohara, and R. Yamada, "Quantitative analysis of random telegraph signals as fluctuations of threshold voltages in scaled Flash memory cells," in *Proc. IRPS*, pp. 29–35, 2007.

[80] C. Monzio Compagnoni, R. Gusmeroli, A. S. Spinelli, A. L. Lacaita, M. Bonanomi, and A. Visconti, "Statistical model for random telegraph noise in Flash memories," *IEEE Trans. Electron Devices*, vol. 55, pp. 388–395, Jan. 2008.

[81] C. Monzio Compagnoni, R. Gusmeroli, A. S. Spinelli, and A. Visconti, "RTN V_T instability from the stationary trap-filling condition: an analytical spectroscopic investigation," *IEEE Trans. Electron Devices*, vol. 55, pp. 655–661, Feb. 2008.

[82] C. Monzio Compagnoni, A. S. Spinelli, S. Beltrami, M. Bonanomi, and A. Visconti, "Cycling effect on the random telegraph noise instabilities of NOR and NAND Flash arrays," *IEEE Electron Device Lett.*, vol. 29, pp. 941–943, Aug. 2008.

[83] A. Ghetti, M. Bonanomi, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita, and A. Visconti, "Physical modeling of single-trap RTS statistical distribution in Flash memories," in *Proc. IRPS*, pp. 610–615, 2008.

[84] A. S. Spinelli, C. Monzio Compagnoni, R. Gusmeroli, M. Ghidotti, and A. Visconti, "Investigation of the random telegraph noise instability in scaled Flash memory arrays," *Jpn. J. Appl. Phys.*, vol. 47, pp. 2598–2601, 2008.

[85] S.-M. Joe, J.-H. Yi, S.-K. Park, H.-I. Kwon, and J.-H. Lee, "Position-dependent threshold-voltage variation by random telegraph noise in NAND Flash memory strings," *IEEE Electron Device Lett.*, vol. 31, pp. 635–637, July 2010.

[86] S.-M. Joe, J.-H. Yi, S.-K. Park, H. Shin, B.-G. Park, Y. J. Park, and J.-H. Lee, "Threshold voltage fluctuation by random telegraph noise in floatin gate NAND Flash memory string," *IEEE Trans. Electron Devices*, vol. 58, pp. 67–73, Jan. 2011.

[87] H.-S. Wong and Y. Taur, "Three-dimensional "atomistic" simulation of discrete random dopant distribution effects in sub-0.1 μm MOSFET's," in *IEDM Tech. Dig.*, pp. 705–708, 1993.

[88] A. Asenov, A. R. Brown, J. H. Davies, and S. Saini, "Hierarchical approach to "atomistic" 3-D MOSFET simulation," *IEEE Trans. Comput.-Aided Design*, vol. 18, pp. 1558–1565, Nov. 1999.

[89] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, "On discrete random dopant modelling in drift-diffusion simulations: physical meaning of "atomistic" dopants," *Microelectron. Reliab.*, vol. 42, pp. 189–199, 2002.

[90] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, pp. 1837–1852, Sep. 2003.

[91] G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional Nano-MOSFETs," *IEEE Trans. Electron Devices*, vol. 52, pp. 3063–3070, May 2006.

[92] K. Sonoda, K. Ishikawa, T. Eimori, and O. Tsuchiya, "Discrete dopant effects on statistical variation of random telegraph signal magnitude," *IEEE Trans. Electron Devices*, vol. 54, pp. 1918–1925, Aug. 2007.

[93] M. F. Bukhori, S. Roy, and A. Asenov, "Statistical aspects of reliability in bulk MOSFETs with multiple defect states and random discrete dopants," *Microelectron. Reliab.*, vol. 48, pp. 1549–1552, Sep. 2008.

[94] H. H. Mueller and M. Schulz, "Random telegraph signal: An atomic probe of the local current in field-effect transistors," *J. Appl. Phys.*, vol. 83, pp. 1734–1741, 1998.

[95] A. Asenov, R. Balasubramaniam, A. R. Brown, J. H. Davies, and S. Saini, "Random telegraph signal amplitudes in sub 100 nm (decanano) MOSFETs: a 3d "atomistic" simulation study," in *IEDM Tech. Dig.*, pp. 279–282, 2000.

[96] A. Asenov, R. Balasubramaniam, A. R. Brown, and J. H. Davies, "RTS amplitudes in decananometer MOSFETs: 3-D simulation study," *IEEE Trans. Electron Devices*, vol. 50, pp. 839–845, Mar. 2003.

[97] J.-D. Lee, C.-K. Lee, M.-W. Lee, H.-S. Kim, K.-C. Park, and W.-S. Lee, "A new programming disturbance phenomenon in NAND Flash memory by source/drain hot-electrons generated by GIDL current," in *Proc. Non-Volatile Semiconductor Memory Workshop*, pp. 31–33, 2006.

[98] M. Kang, W. Hahn, I. H. Park, J. Park, Y. Song, H. Lee, C. Eun, S. Ju, K. Choi, Y. Lim, S. Jang, S. Cho, B.-G. Park, and H. Shin, "DIBL-Induced program disturb characteristics in 32-nm NAND Flash memory array," *IEEE Trans. Electron Devices*, vol. 58, pp. 3626–3629, Oct. 2011.

[99] K.-D. Suh, B.-H. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.-S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, and H.-K. Lim, "A 3.3 V 32 Mb NAND Flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, 1995.

[100] S. Satoh, H. Hagiwara, T. Tanzawa, K. Takeuchi, and R. Shirota, "A novel isolation-scaling technology for NAND EEPROMs with the minimized program disturbance," in *IEDM Tech. Dig.*, pp. 291–294, 1997.

[101] J.-D. Choi, S.-S. Cho, Y.-S. Yim, J.-D. Lee, H.-S. Kim, K.-J. Joo, S.-H. Hur, H.-S. Im, J. kim, J.-W. Lee, K.-I.Seo, M.-S. Kang, K.-H. Kim, J.-L. Nam, K.-C. Park, and M.-Y. Lee, "Highly manufacturable 1Gb NAND Flash using 0.12 μm process technology," in *IEDM Tech. Dig.*, pp. 25–28, 2001.

[102] S. Seo, H. Kim, S. Park, S. Lee, S. Aritome, and S. Hong, "Novel negative V_T shift program disturb phenomena in $2x \sim 3x$ nm nand flash memory cells," in *Proc. IRPS*, pp. 6B.2.1–6B.2.4, 2011.

[103] B. Cho, C. Lee, K. Seol, S. Hur, J. Choi, J. Choi, and C. Chung, "A new cell-to-cell interference induced by conduction band distortion near S/D region in scaled NAND Flash memories," in *Proc. IMW*, pp. 1–4, 2011.

[104] E. Kwon, D. Oh, B. Lee, J. hyong Yi, S. Kim, G. Cho, S. Park, and J. Choi, "An abnormal floating gate interference and a low program performance in 2y nm NAND Flash devices," in *SISPAD Conf. Proc.*, pp. 207–210, 2011.

[105] K.-T. Park, M. Kang, D. Kim, S.-W. Hwang, B. Y. Choi, Y.-T. Lee, C. Kim, and K. Kim, "A zeroing cell-to-cell interference page architecture with temporary LSB storing and parallel MSB program scheme for MLC NAND Flash memories," *IEEE J. Solid-State Circuits*, vol. 43, pp. 919–928, Apr. 2008.

[106] C.-H. Lee, S.-K. Sung, D. Jang, S. Lee, S. Choi, J. Kim, S. Park, M. Song, H.-C. Baek, E. Ahn, J. Shin, K. Shin, K. Min, S.-S. Cho, C.-J. Kang, J. Choi, K. Kim, J.-H. Choi, K.-D. Suh, and T.-S. Jung, "A highly manufacturable integration technology for 27 nm 2 and 3 bit/cell NAND Flash memory," in *IEDM Tech. Dig.*, pp. 98–101, 2010.

[107] A. Spessot, A. Calderoni, P. Fantini, A. S. Spinelli, C. Monzio Compagnoni, F. Farina, A. L. Lacaita, and A. Marmiroli, "Variability effects on the V_T distribution of nanoscale NAND Flash memories," in *Proc. IRPS*, pp. 970–974, 2010.

[108] A. Ghetti, L. Bortesi, and L. Vendrame, "3D simulation study of gate coupling and gate cross-interference in advanced floating gate non-volatile memories," *Solid-State Electron.*, vol. 49, pp. 1805–1812, 2005.

[109] C. Lee, S.-K. Lee, S. Ahn, J. Lee, W. Park, Y. Cho, C. Jang, C. Yang, S. Chung, I.-S. Yun, B. Joo, B. Jeong, J. Kim, J. Kwon, H. Jin, Y. Noh, J. Ha, M. Sung, D. Choi, S. Kim, J. Choi, T. Jeon, H. Park, J.-S. Yang, and Y.-H. Koh, "A 32-Gb MLC NAND Flash memory with Vth endurance enhancing schemes in 32 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 46, pp. 97–106, Jan. 2011.

[110] P. Blomme, M. Rosmeulen, A. Cacciato, M. Kostermans, C. Vrancke, S. V. Aerde, T. Schram, I. Debusschere, M. Jurczak, and J. V. Houdt, "Novel dual layer floating gate structure as enabler of fully planar flash memory," in *Symp. VLSI TechnologyTech. Dig*, pp. 129–130, 2010.

[111] D. Wellekens, P. Blomme, M. Rosmeulen, T. Schram, A. Cacciato, I. Debusschere, J. Van Houdt, and S. Van Aerde, "An ultra-thin hybrid floating gate concept for Sub-20nm NAND Flash technologies," in *Proc. IMW*, pp. 1–4, 2011.

[112] J. V. Houdt, "Charge-based nonvolatile memory: Near the end of the roadmap?," *Current Applied Physics*, vol. 11, no. 2, Supplement, pp. e21 – e24, 2011.

[113] J. Choi and K. S. Seol, "3D approaches for non-volatile memory," in *Symp. VLSI Tech. Dig.*, pp. 178–179, 2011.

[114] H.-T. Lue, Y.-H. Hsiao, K.-Y. Hsieh, S.-Y. Wang, T. Yang, K.-C. Chen, and C.-Y. Lu, "Scaling feasibility study of planar thin floating gate (FG) NAND Flash devices and size effect challenges beyond 20nm," in *IEDM Tech. Dig.*, pp. 9.2.1–9.2.4, 2011.

[115] Y.-M. Kwon, "Delving deep into Micron and Intel's 20-nm 64-Gbit MLC NAND flash memory." www.eetimes.com, 2012.

[116] H.-T. Lue, T.-H. Hsu, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "Study of incremental step pulse programming ISPP and STI edge effect of BE-SONOS NAND Flash," in *Proc. IRPS*, pp. 693–694, 2008.

[117] M. White, D. Adams, and J. Bu, "On the go with SONOS," *IEEE Circuits and Devices Magazine*, no. 16, pp. 22–31, 2000.

[118] B. De Salvo, C. Gerardi, R. van Schaijk, S. Lombardo, D. Corso, C. Plantamura, S. Serafino, G. Ammendola, M. van Duuren, P. Goarin, W. Yuet Mei, K. van der Jeugd, T. Baron, M. Gely, P. Mur, and S. Deleonibus, "Performance and reliability features of advanced nonvolatile memories based on discrete traps (silicon nanocrystals, SONOS)," *IEEE Trans. Device and Materials Reliab.*, vol. 4, pp. 377–389, Sep. 2004.

[119] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K. Kim, "A novel SONOS structure of SiO_2/SiN/Al_2O_3 with TaN metal gate for multi-giga bit flash memories," in *IEDM Tech. Dig.*, pp. 613–616, 2003.

[120] A. Mauri, C. Monzio Compagnoni, S. Amoroso, A. Maconi, F. Cattaneo, A. Benvenuti, A. S. Spinelli, and A. L. Lacaita, "A new physics-based model for TANOS memories program/erase," in *IEDM Tech. Dig.*, pp. 555–558, 2008.

[121] C. Monzio Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi, and A. S. Spinelli, "Physical modeling for programming of TANOS memories in the Fowler-Nordheim regime," *IEEE Trans. Electron Devices*, vol. 56, pp. 2008–2015, Sep. 2009.

[122] S. M. Amoroso, A. Mauri, N. Galbiati, C. Scozzari, E. Mascellino, E. Camozzi, A. Rangoni, T. Ghilardi, A. Grossi, P. Tessariol, C. Monzio Compagnoni, A. Maconi, A. L. Lacaita, A. S. Spinelli, and G. Ghidini, "Reliability constraints for TANOS memories due to alumina trapping and leakage," in *Proc. IRPS*, pp. 966–969, 2010.

[123] A. Mauri, C. Monzio Compagnoni, S. M. Amoroso, A. Maconi, A. Ghetti, A. S. Spinelli, and A. L. Lacaita, "Comprehensive investigation of statistical effects in nitride memories – Part I: Physics-based modeling," *IEEE Trans. Electron Devices*, vol. 57, pp. 2116–2123, Sep. 2010.

[124] C. Monzio Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi, E. Greco, A. S. Spinelli, and A. L. Lacaita, "Comprehensive investigation of statistical effects in nitride memories – Part II: Scaling analysis and impact on device performance," *IEEE Trans. Electron Devices*, vol. 57, pp. 2124–2131, Sep. 2010.

[125] S. M. Amoroso, A. Maconi, A. Mauri, C. Monzio Compagnoni, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional simulation of charge-trap memory programming – Part I: Average behavior," *IEEE Trans. Electron Devices*, vol. 58, pp. 1864–1871, July 2011.

[126] A. Maconi, S. M. Amoroso, C. Monzio Compagnoni, A. Mauri, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional simulation of charge-trap memory programming – Part II: Variability," *IEEE Trans. Electron Devices*, vol. 58, pp. 1872–1878, July 2011.

[127] R. Katsumata, M. Kito, Y. Fukuzumi, M. Kido, H. Tanaka, Y. Komori, M. Ishiduki, J. Matsunami, T. Fujiwara, Y. Nagata, L. Zhang, Y. Iwata, R. Kirisawa, H. Aochi, and A. Nitayama, "Pipe-shaped BiCS Flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices," in *Symp. VLSI Tech. Dig.*, pp. 136–137, 2009.

[128] S. Whang, K. Lee, D. Shin, B. Kim, M. Kim, J. bin, J. Han, S. Kim, B. Lee, Y. Jung, S. Cho, C. Shin, H. Yoo, S. Choi, K. Honag, S. Aritome, S. Park, and S. Hong, "Novel 3-dimensional dual control-gate with surrounding floating-gate (DC-SF) NAND flash cell for 1 Tb file storage applications," in *IEDM Tech. Dig.*, pp. 668–671, 2010.

[129] Y.-H. Hsiao, H.-T. Lue, T.-H. Hsu, K.-Y. Hsieh, and C.-Y. Lu, "A critical examination of 3D stackable NAND Flash memory architectures by simulation study of the scaling capability," in *Proc. IMW*, pp. 142–145, 2010.

[130] J. Jang, H.-S. Kim, W. Cho, H. Cho, J. Kim, S. Shim, Y. Jang, J.-H. Jeong, B.-K. Son, D. W. Kim, K. Kim, J.-J. Shim, J. S. Lim, K.-H. Kim, S. Y. Yi, J.-Y. Lim, D. Chung, H.-C. Moon, S. Hwang, J.-W. Lee, Y.-H. Son, U.-I. Chung, and W.-S. Lee, "Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND Flash memory," in *Symp. VLSI Tech. Dig.*, pp. 192–193, 2009.

[131] J. Kim, A. Hong, M. Ogawa, S. Ma, E. Song, Y.-S. Lin, J. Han, U.-I. Chung, and K. Wang, "Novel 3-D structure for ultra high density flash memory with VRAT (Vertical-Recess-Array-Transistor) and PIPE (Planarized Integration on the same PlanE)," in *Symp. VLSI Tech. Dig.*, pp. 122–123, 2008.

[132] J. Kim, A. Hong, S. M. Kim, E. Song, J. H. Park, J. Han, S. Choi, D. Jang, J. T. Moon, and K. Wang, "Novel Vertical-Stacked-Array-Transistor (VSAT) for ultra-high-density and cost-effective nand flash memory devices and SSD (Solid State Drive)," in *Symp. VLSI Tech. Dig.*, pp. 186–187, 2009.

[133] W. Kim, S. Choi, J. Sung, T. Lee, C. Park, H. Ko, J. Jung, I. Yoo, and Y. Park, "Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage," in *Symp. VLSI Tech. Dig.*, pp. 188–189, 2009.

[134] H.-T. Lue, T.-H. Hsu, Y.-H. Hsiao, S. Hong, M. Wu, F. Hsu, N. Lien, S.-Y. Wang, J.-Y. Hsieh, L.-W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, and C.-Y. Lu, "A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND Flash using junction-free buried channel BE-SONOS device," in *Symp. VLSI Tech. Dig.*, pp. 131–132, 2010.

[135] C. Kang, J. Choi, J. Sim, C. Lee, Y. Shin, J. Park, J. Sel, S. Jeon, Y. Park, and K. Kim, "Effects of lateral charge spreading on the reliability of TANOS (TaN/AlO/SiN/Oxide/Si) NAND Flash memory," in *Proc. IRPS*, pp. 167–170, 2007.

[136] A. Maconi, A. Arreghini, C. Compagnoni, G. Van den bosch, A. Spinelli, J. Van Houdt, and A. Lacaita, "Impact of lateral charge migration on the retention performance of planar and 3D SONOS devices," in *Proc. ESSDERC*, pp. 195–198, 2011.

[137] M.-J. Lee, Y. Park, B.-S. Kang, S.-E. Ahn, C. Lee, K. Kim, W. Xianyu, G. Stefanovich, J.-H. Lee, S.-J. Chung, Y.-H. Kim, C.-S. Lee, J.-B. Park, and I.-K. Yoo, "2-stack 1D-1R cross-point structure with oxide diodes as switch elements for high density resistance RAM applications," in *IEDM Tech. Dig.*, pp. 771–774, 2007.

[138] M. Johnson, A. Al-Shamma, D. Bosch, M. Crowley, M. Farmwald, L. Fasoli, A. Ilkbahar, B. Kleveland, T. Lee, T. yi Liu, Q. Nguyen, R. Scheuerlein, K. So, and T. Thorp, "512-Mb PROM with a three-dimensional array of diode/antifuse memory cells," *IEEE J. Solid-State Circuits*, vol. 38, pp. 1920–1928, nov. 2003.

[139] S. Tehrani, J. M. Slaughter, M. Deherrera, B. N. Engel, N. D. Rizzo, J. Salter, M. Durlam, R. W. Dave, J. Janesky, B. Butcher, K. Smith, and G. Grynkewich, "Magnetoresistive random access memory using magnetic tunnel junctions," in *Proceedings of the IEEE*, pp. 703–714, 2003.

[140] S. Lai, "Current status of the phase change memory and its future," in *IEDM Tech. Dig*, pp. 255–258, 2003.

[141] I. Baek, D. Kim, M. Lee, H.-J. Kim, E. Yim, M. Lee, J. Lee, S. Ahn, S. Seo, J. Lee, J. Park, Y. Cha, S. Park, H. Kim, I. Yoo, U.-I. Chung, J. Moon, and B. Ryu, "Multi-layer cross-point binary oxide resistive memory (OxRRAM) for post-NAND storage application," in *IEDM Tech. Dig.*, pp. 750–753, 2005.

[142] M. Kozicki, M. Park, and M. Mitkova, "Nanoscale memory elements based on solid-state electrolytes," *Nanotechnology, IEEE Transactions on*, vol. 4, pp. 331–338, May 2005.

[143] D. Kau, S. Tang, I. Karpov, R. Dodge, B. Klehn, J. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, S. Lee, T. Langtry, K. wei Chang, C. Papagianni, J. Lee, J. Hirst, S. Erra, E. Flores, N. Righos, H. Castro, and G. Spadini, "A stackable cross point phase change memory," in *IEDM Tech. Dig.*, pp. 1–4, 2009.

[144] G. Servalli, "A 45nm generation phase change memory technology," in *IEDM Tech. Dig.*, pp. 1–4, 2009.

[145] Y. Sasago, M. Kinoshita, T. Morikawa, K. Kurotsuchi, S. Hanzawa, T. Mine, A. Shima, Y. Fujisaki, H. Kume, H. Moriya, N. Takaura, and K. Torii, "Cross-point phase change memory with $4f^2$ cell size driven by low-contact-resistivity poly-Si diode," in *Symp. VLSI Tech. Dig.*, pp. 24–25, 2009.

[146] S. R. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," *Phys. Rev. Lett.*, vol. 21, pp. 1450–1453, 1968.

[147] A. Sawa, "Resistive switching in transition metal oxides," *Materials Today*, vol. 11, no. 6, pp. 28 – 36, 2008.

[148] F. Nardi, S. Balatti, S. Larentis, and D. Ielmini, "Complementary switching in metal oxides: Toward diode-less crossbar RRAMs," in *IEDM Tech. Dig.*, pp. 31.1.1–31.1.4, 2011.

[149] "JEDEC Standard JEP122E: Failure mechanisms and models for semiconductor devices," tech. rep., JEDEC Solid State Technology Association, March 2009.

[150] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome, and R. Shirota, "Fast and accurate programming method for multi-level NAND EEPROMs," in *Symp. VLSI Tech. Dig.*, pp. 129–130, 1995.

[151] C. Monzio Compagnoni, A. S. Spinelli, R. Gusmeroli, S. Beltrami, A. Ghetti, and A. Visconti, "Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics," *IEEE Trans. Electron Devices*, vol. 55, pp. 2695–2702, Oct. 2008.

[152] C. Monzio Compagnoni, A. Ghetti, M. Ghidotti, A. S. Spinelli, and A. Visconti, "Data retention and program/erase sensitivity to the array background pattern in deca-nanometer NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 57, pp. 321–327, Jan. 2010.

[153] M. Park, E. Ahn, E. Cho, K. Kim, and W.-S. Lee, "The effect of negative V_{TH} of NAND Flash memory cells on data retention characteristics," *IEEE Electron Device Lett.*, vol. 30, pp. 155–157, Feb. 2009.

[154] J.-D. Lee, S.-H. Hur, and J.-D. Choi, "Effects of floating-gate interference on NAND Flash memory cell operation," *IEEE Electron Device Lett.*, vol. 23, pp. 264–266, May 2002.

[155] Y. S. Kim, D. J. Lee, C. K. Lee, H. K. Choi, S. S. Kim, J. H. Song, D. H. Song, J.-H. Choi, K.-D. Suh, and C. Chung, "New scaling limitation of the floating gate cell in NAND Flash memory," in *Proc. IRPS*, pp. 599–603, 2010.

[156] B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. 22, pp. 558–566, Aug. 1987.

[157] L. Larcher, A. Padovani, P. Pavan, P. Fantini, A. Calderoni, A. Mauri, and A. Benvenuti, "Modeling NAND Flash memories for IC design," *IEEE Electron Device Lett.*, vol. 29, pp. 1152–1154, Oct. 2008.

[158] A. Spessot, C. Monzio Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, "Compact modeling of variability effects in nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 58, pp. 2302–2309, Aug. 2011.

[159] W. Liu, X. Jin, J. Chen, M.-C. Jeng, Z. Liu, Y. Cheng, K. Chen, M. Chan, K. Hui, J. Huang, R. Tu, P. K. Ko, and C. Hu, "BSIM3v3.2 MOSFET model users' manual," Tech. Rep. UCB/ERL M98/51, EECS Department, University of California, Berkeley, 1998.

[160] Y. T. Taur and T. H. Ning, *Fundamentals of modern VLSI devices*. Cambridge University Press, 1998.

[161] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFET's: Part I-effects of substrate impurity concentration," *IEEE Trans. Electron Devices*, vol. 41, pp. 2357–2362, Dec. 1994.

[162] K. Chen, H. C. Wann, J. Dunster, P. K. Ko, C. Hu, and M. Yoshida, "MOSFET carrier mobility model based on gate oxide thickness threshold and gate voltages," *Solid-State Electron.*, vol. 39, pp. 1515–1518, Oct. 1996.

[163] A. Cros, K. Romanjek, D. Fleury, S. Harrison, R. Cerutti, P. Coronel, B. Dumont, A. Pouydebasque, R. Wacquez, B. Duriez, R. Gwoziecki, F. Boeuf, H. Brut, G. Ghibaudo, and T. Skotnicki, "Unexpected mobility degradation for very short devices : A new challenge for CMOS scaling," in *IEDM Tech. Dig.*, pp. 1–4, Dec. 2006.

[164] G. Bidal, D. Fleury, G. Ghibaudo, F. Boeuf, and T. Skotnicki, "Guidelines for MOSFET device optimization accounting for L-dependet mobility degradation," in *IEEE 2009 Silicon Nanoelecronics Workshop*, pp. 5–6, June 2009.

[165] C. Miccoli, C. Monzio Compagnoni, S. Beltrami, A. S. Spinelli, and A. Visconti, "Threshold-voltage instability due to damage recovery in nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 58, pp. 2406–2414, Aug. 2011.

[166] H. Yang, H. Kim, S.-I. Park, J. Kim, S.-H. Lee, J.-K. Choi, D. Hwang, C. Kim, M. Park, K.-H. Lee, Y.-K. Park, J. K. Shin, and J.-T. Jeong-Taek Kong, "Reliability issues and models of sub-90nm NAND Flash memory cells," in *Proc. ICSICT*, pp. 760–762, Oct. 2006.

[167] R. Fastow, R. Banerjee, P. Bjeletich, A. Brand, H. Chao, J. Gorman, X. Guo, J. B. Heng, N. Koenigsfeld, S. Ma, A. Masad, S. Soss, and B. J. Woo, "A 45 nm NOR Flash technology with self-aligned contacts and 0.024 μm^2 cell size for multi-level applications," in *VLSI-TSA Tech. Dig.*, pp. 81–82, 2008.

[168] A. Fazio, "Future directions of non-volatile memory in compute applications," in *IEDM Tech. Dig.*, pp. 641–644, 2009.

[169] M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf, D. Leak, K. Loe, D. Mills, P. Ruby, R. Rozman, S. Sweha, S. Talreja, and K. Wojciechowski, "A multilevel-cell 32Mb Flash memory," in *Proc. ISSCC*, pp. 132–133, 1995.

[170] B. Riccò, G. Torelli, M. Lanzoni, A. Manstretta, H. E. Maes, D. Montanari, and A. Modelli, "Nonvolatile multilevel memories for digital applications," *Proc. IEEE*, vol. 86, pp. 2399–2421, Dec. 1998.

[171] A. Modelli, R. Bez, and A. Visconti, "Multi-level Flash memory technology," in *Proc. SSDM Conf.*, pp. 516–517, 2001.

[172] "JEDEC JEP122G: Failure mechanisms and models for semiconductor devices," tech. rep., JEDEC Solid State Technology Association.

[173] "JEDEC Standard JESD47H: Stress-test-driven qualification of integrated circuit." JEDEC Solid State Technology Association.

[174] "JEDEC Standard JESD22-A117C: Electrically erasable programmable ROM (EEPROM) program/erase endurance and data retention stress tes." JEDEC Solid State Technology Association.

[175] C. Calligaro, A. Manstretta, A. Modelli, and G. Torelli, "Technological and design constraints for multilevel Flash memories," in *Proc.* 3^{rd} *IEEE Int. Conf. on Electronics, Circuits and Systems*, pp. 1005–1008, 1996.

[176] A. Brand, K. Wu, S. Pan, and D. Chin, "Novel read-disturb failure mechanism induced by FLASH cycling," in *Proc. IRPS*, pp. 127–132, 1993.

[177] C. Dunn, C. Kaya, T. Lewis, T. Strauss, J. Schreck, P. Hefley, M. Middennorf, and T. San, "Flash EEPROM disturb mechanism," in *Proc. IRPS*, pp. 299–308, 1994.

[178] T. Wada, "Acceleration method for gate-disturb degradation on embedded flash EEPROM," *Microelectron. Reliab.*, vol. 40, pp. 1279–1283, 2000.

[179] Y.-H. Lee, N. Mielke, W. McMahon, Y.-L. R. Lu, Q. Meng, and L. Jiang, "Drain read disturb assessment of NOR Flash memory," in *VLSI-TSA Tech. Dig.*, pp. 83–84, 2008.

[180] Y.-H. Lee, W. McMahon, Y.-L. R. Lu, J.-Y. J. Tewg, and S. T. Ma, "On the scaling of Flash cell spacer for gate disturb and charge retention otimization," *IEEE Trans. Electron Devices*, vol. 56, pp. 1959–1965, Sep. 2009.

[181] S. Mori, Y. Y. Araki, M. Sato, H. Meguro, H. Tsunoda, E. Kamiya, K. Yoshikawa, N. Arai, and E. Sakagami, "Thickness scaling limitation factors of ONO interpoly dielectric for nonvolatile memory devices," *IEEE Trans. Electron Devices*, vol. 43, pp. 47–53, Jan. 1996.

[182] G. Tao, S. Nath, C. Ouvrard, H. Chauveau, D. Dormans, and R. Verhaar, "Experimental study of charge displacement in nitride layer and its effect on threshold voltage instability of advanced Flash memory devices," in *Proc. IPFA*, pp. 76–80, July 2007.

[183] G. Molas, D. Deleruyelle, B. De Salvo, G. Ghibaudo, M. Gely, L. Perniola, D. Lafond, and S. Deleonibus, "Degradation of floating-gate memory reliability by few electron phenomena," *IEEE Trans. Electron Devices*, vol. 53, pp. 2610–2619, 2006.

[184] C. Monzio Compagnoni, R. Gusmeroli, A. S. Spinelli, and A. Visconti, "Analytical model for the electron-injection statistics during programming of nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 55, pp. 3192–3199, Nov. 2008.

[185] C. Monzio Compagnoni, A. S. Spinelli, R. Gusmeroli, A. L. Lacaita, S. Beltrami, A. Ghetti, and A. Visconti, "First evidence for injection statistics accuracy limitations in NAND Flash constant-current Fowler-Nordheim programming," in *IEDM Tech. Dig.*, pp. 165–168, 2007.

[186] C. Friederich, J. Hayek, A. Kux, T. Muller, N. Chan, G. Kobernik, M. Specht, D. Richter, and D. Schmitt-Landsiedel, "Novel model for cell-system interaction MCSI in NAND Flash," in *IEDM Tech. Dig.*, pp. 831–834, 2008.

[187] A. Spessot, C. Monzio Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, "Effect of floating-gate polysilicon depletion on the erase efficiency of NAND Flash memories," *IEEE Electron Device Lett.*, vol. 31, pp. 647–649, July 2010.

[188] D. W. Lee, S. Cho, B. W. Kang, S. Park, B. Park, M. K. Cho, K.-O. Ahn, Y. S. Yang, and S. W. Park, "The operation algorithm for improving the reliability of TLC (Triple Level Cell) NAND Flash characteristics," in *Proc. IMW*, pp. 1–2, 2011.

[189] Y. Shin, "Non-volatile memory technologies for beyond 2010," in *Symp. VLSI Tech. Dig.*, pp. 156–159, 2005.

[190] G. Molas, D. Deleruyelle, B. De Salvo, G. Ghibaudo, M. Gely, S. Jacob, D. Lafond, and S. Deleonibus, "Impact of few electron phenomena on floating-gate memory reliability," in *IEDM Tech. Dig.*, pp. 877–880, 2004.

[191] P. Palestri, N. Barin, D. Brunel, C. Busseret, A. Campera, P. A. Childs, F. Driussi, C. Fiegna, G. Fiori, R. Gusmeroli, G. Iannaccone, M. Karner, H. Kosina, A. L. Lacaita, E. Langer, B. Majkusiak, C. Monzio Compagnoni, A. Poncet, E. Sangiorgi, L. Selmi, A. S. Spinelli, and J. Walczak, "Comparison of modeling approaches for the capacitance-voltage and current-voltage characteristics of advanced gate stacks," *IEEE Trans. Electron Devices*, vol. 54, pp. 106–114, Jan. 2007.

[192] M. J. Faddy, "Extended Poisson process modelling and analysis of count data," *Biometrical Journal*, vol. 39, pp. 431–440, 1997.

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

[194] C. Monzio Compagnoni, C. Miccoli, A. L. Lacaita, A. Marmiroli, A. S. Spinelli, and A. Visconti, "Impact of control-gate and floating-gate design on the electron-injection spread of decananometer NAND Flash memories," *IEEE Electron Device Lett.*, vol. 31, pp. 1196–1198, Nov. 2010.

[195] T. Tanaka and J. Chan, "Non-volatile semiconductor memory device adapted to store a multi-valued data in a single memory cell." U.S. patent 6 643 188 B2, November 4 2003.

[196] V. Moschiano, G. Santin, T. Vali, and M. Rossini, "Non-volatile multilevel memory cell programming." U.S. Patent 7 692 971 B2, April 6 2010.

[197] T. C. Ong, A. Fazio, N. Mielke, S. Pan, N. Righos, G. Atwood, and S. Lai, "Erratic erase in ETOX^{TM} Flash memor array," in *Symp. VLSI Tech. Dig.*, pp. 83–84, 1993.

[198] A. Chimenton, P. Pellati, and P. Olivo, "Analysis of erratic bits in Flash memories," *IEEE Trans. Device and Materials Reliab.*, vol. 4, pp. 179–184, Dec. 2001.

[199] S. Gregori, A. Cabrini, O. Khouri, and G. Torelli, "On-chip error correcting techniques for new-generation Flash memories," *Proc. IEEE*, vol. 91, pp. 602–616, Apr. 2003.

[200] N. Mielke, T. Marquart, N. Wu, J. Kessenich, H. Belgal, E. Schares, F. Trivedi, E. Goodness, and L. Nevill, "Bit error rate in nand flash memories," in *Proc. IRPS*, pp. 9–19, 2008.

[201] P. A. W. Lewis and G. S. Shedler, "Simulation methods for Poisson processes in nonstationary systems," in *Proc. of the* 10^{th} *conference on Winter simulation - Volume 1*, pp. 155–163, 1978.

List of publications

1. C. Monzio Compagnoni, C. Miccoli, A. L. Lacaita, A. Marmiroli, A. S. Spinelli, A. Visconti. "Impact of the Control-Gate and Floating-Gate Design on the Electron-Injection Spread of Decananometer NAND Flash Memories". IEEE Electron Device Letters, vol. 31, no. 11, pp. 1196–1198, Nov. 2010
2. C. Miccoli, C. Monzio Compagnoni, S. M. Amoroso, A. Spessot, P. Fantini, A. Visocnti, A. S. Spinelli. "Impact of Neutral Threshold-Voltage Spread and Electron-Emission Statistics on Data Retention of Nanoscale NAND Flash". IEEE Electron Device Letters, vol. 31, no. 11, pp. 1202–1204, Nov. 2010
3. C. Monzio Compagnoni, C. Miccoli, R. Mottadelli, S. Beltrami, M. Ghidotti, A. L. Lacaita, A. S. Spinelli, A. Visconti. "Investigation of the Threshold Voltage Instability after Distributed Cycling in Nanoscale NAND Flash Memory Arrays". IRPS 2010, pp. 604–610
4. C. Miccoli, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita. "Investigation of the Programming Accuracy of a Double-Verify ISPP Algorithm for Nansoscale NAND Flash Memories". IRPS 2011, pp. MY.5.1–MY.5.6
5. C. Miccoli, C. Monzio Compagnoni, S. Beltrami, A. S. Spinelli, A. Visconti. "Threshold-Voltage Instability Due to Damage Recovery in Nanoscale NAND Flash Memories". IEEE Transactions on Electron Devices, vol. 58, no. 8, pp. 2406–2414, Aug. 2011
6. G. M. Paolucci, C. Miccoli, C. Monzio Compagnoni, L. Crespi, A. S. Spinelli, A. L. Lacaita. "Investigation of Cycling-Induced V_T Instabilities in NAND Flash Cells via Compact Modeling". IMW 2012
7. G. M. Paolucci, C. Miccoli, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita. "String Current in Decananometer NAND Flash Arrays: A Compact-Modeling Investigaion". IEEE Transactions on Electron Devices, vol. 59, no. 9, pp. 2331–2337, Sept. 2012
8. C. Miccoli, C. Monzio Compagnoni, L. Chiavarone, S. Beltrami, A. L. Lacaita, A. S. Spinelli, A. Visconti. "Assessment of Distributed-cycling Schemes on 45nm NOR Flash Memory Arrays". IRPS 2012, pp. 2A.1.1–2A.1.7
9. C. Miccoli, J. Barber, C. Monzio Compagnoni, G. M. Paolucci, J. Kessenich, A. L. Lacaita, A. S. Spinelli, R. J. Koval, A. Goda. "Resolving Discrete Emission Events: a New Perspective for Detrapping Investigation in NAND Flash Memories". Accepted for IRPS 2013

Awards and recognitions

[3] is cited by: "JEDEC JEP122C: Failure Mechanisms and Models for Semiconductor Devices", tech. rep., JEDEC Solid State Technology Association, October 2011

[8] was awarded the Best Student Paper Award at the 2012 IEEE International Reliability Physics Symposium (IRPS), which took place from 15th to 19th of April in Anaheim (CA) - USA