POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione
RESEARCH DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

PERFORMANCE AND LIMITATIONS OF CHARGE-TRAP STORAGE FOR NON-VOLATILE MEMORY TECHNOLOGIES

Doctoral Dissertation of:
Alessandro MACONI

Advisor:
Prof. Alessandro S. Spinelli

Tutor:
Prof. Angelo Geraci

The Chair of the Doctoral Program:
Prof. Carlo Fiorini

2011 – XXIV

“***Forty-two***”
[The answer to the The Ultimate Question of Life, the Universe and Everything]

(D. Adams – The Hitchhiker's Guide to the Galaxy)

Preface

This Doctoral Dissertation concludes a period of more than four years that began in the summer 2007: since then I joined the *Dipartimento di Elettronica e Informazione* (DEI) of the *Politecnico di Milano*, first as a master thesis student and then, since January 2009, as a Ph.D. student in Information Technology. During both the master thesis and the doctoral studies I have been working on the modeling and characterization of electronic devices, focusing on the study of charge trap memories. I have to thank many people that I met during these years. The first acknowledgement is deserved to my supervisor, Prof. Alessandro Spinelli, and to the Assistant Professor Christian Monzio Compagnoni: both of them followed me during the last four years and supported my work with fruitful discussions. I also have to thank my colleague, Salvatore Amoroso, with which I have shared the work in the same field during both the thesis and the doctorate; along with him, I must thank all the people from *Micron* (former *Numonyx*), above all Aurelio Mauri and the whole TCAD group.

I thank all my colleagues and all the people that made a thesis in the *Nanolab*: in rigorous random order I thank Carmine, Simone L., Carlo, Niccolò, Ugo, Michele, Mattia, Prof. Ielmini, Andrea, Maria, Stefano, Simone B., Nicola, Giovanni, Pietro, Giuseppe, "my" students Gabriele and Evelyne, and all other people whom I worked with and I shared moments of fun and relax during the lunch breaks and after the work.

During the last year I also spent about six months in Leuven (Belgium), working in the memory group of *imec*: there I worked on the characterization and the modeling of the lateral charge migration in silicon nitride that is presented in the last chapter of this dissertation. I must thank a lot of people that I met there, first of all the manager of the group Jan Van Houdt, that gave me the opportunity to work in his group, then Antonio Arreghini, whom I worked with in strong collaboration: he helped me during the first period with all the bureaucracy and the duties I had to carry out, he explained me how to make the measurements, discussed with me the explanations of the results, the implementation of the simulator, and shared with me many dinners and beers. I also want to thank all the rest of the memory group and the people I met there, especially Koen and Silvia who hosted me when I was stuck in Belgium for Christmas due to the snow.

I also want to thank the rewievers of this thesis for the careful reading and the useful suggestions to improve it.

Finally I want to acknowledge all my friends and my family, that supported me and stood me when I was worried or nervous because of the work, in particular my mother, my grandmother, my brother and his future wife.

One last dedication to my father who, I am sure, would be proud of this work.

December MMXI,

Alessandro Maconi

Abstract

Semiconductor non-volatile memories have gained in the last decade an explosive success, thanks to the ever increasing market demand for portable consumer products requiring permanent data storage, such as digital cameras, MP3 players, removable cards, USB sticks, mobile phones, and, lately, solid state drives (SSD). The Flash technology, allowing high-density, small volume, and low-power devices, demonstrated the possibility to fulfill the requirements for these new applications to be developed. In particular, the NAND and the NOR Flash architectures represent today the winning solution for data and code storage, respectively. Floating-Gate (FG) Flash memories have been able so far to satisfy the market requirements, especially for the portable equipments, and to become the mainstream nonvolatile memory (NVM) technology. The increase of the integration density and the concurrent reduction of the producing cost are the basis of the growth of the storage semiconductor industry, as they allow not only to offer better memories at lower price, but also to open new market possibilities. Projecting into the next decade, though, there are several problems that must be faced to further scale the FG concept. Different approaches have been proposed to overcome these limitations, and one of the most promising seems to be the charge trap (CT) concept, such as SONOS and TANOS memories.

In Chapter 1 the working principle of the Flash memories is presented, explaining the reasons why this technology is facing ever increasing problems in the scaling process; then the possible alternatives are presented, dividing them in two big categories: (a) the evolutionary technologies, that slightly change the cell structure but maintain the same basic working principle and (b) new storing concepts, that completely change the approach to store informations. Among these the CT solution will be elaborated, explaining how it can solve some of the problems of the Flash cells, and also presenting the different possibilities, such as planar SONOS and TANOS, or 3D architectures, briefly discussing advantages and drawbacks of the different approaches.

Chapter 2 will focus on the modeling of planar CT memories: first a simple analytical model for the program operation of these devices will be presented, allowing to explain some of the fundamental differences between the FG and the CT memories. Afterwards a more accurate numerical model will be introduced, in order to address some of the peculiarities that cannot be described by the analytical one: the model is able to reproduce program and erase transients, and is tested against experimental data on SONOS devices. In order to understand the differences between the SONOS and TANOS memories, the impact of the introduction alumina layer in TANOS devices will be discussed, starting from experimental evidences and integrating the extracted properties in the model, allowing to reproduce experimental program, erase and retention transients on such devices.

In Chapter 3 a study of the Incremental Step Pulse Programming (ISPP) is presented, with experiments on large area SONOS and TANOS capacitors, and using a 3D model on deca-nanometer devices. The analysis made on large area SONOS capacitors allows to better understand the physical differences between these devices and the FG Flash, mainly highlighting and explaining the decrease in the trapping efficiency, and then extending the characterization to TANOS devices, pointing out the role of the alumina layer. Finally, an analysis on ultra-scaled devices will be presented, revealing a further decrease of the programming efficiency, that is caused by the fringing field.

Chapter 4 will be dedicated to the modeling of cylindrical CT memories: a physics-based analytical model will be presented, obtaining the electrostatic solution and studying the curvature effect impact on the tunneling current and on the transient dynamics. The model will then be tuned against experimental data, and a parametric analysis of the gate-all-around CT cells will be presented.

At the end, in Chapter 5 lateral charge migration in the nitride layer will be studied: in 3D structures the nitride is not cut at the borders of each cell, and this can lead to worse retention transient. First, experimental data on planar SONOS cells will be presented, explaining how the results can be interpreted in terms of lateral diffusion of the charge out of the active area of the cell. Then a 2D model to simulate retention transients is developed and tested against the experimental data, highlighting the need of this diffusion process to reproduce retention transients obtained on cells with the nitride layer continuing beyond the active area. The model is then extended to cylindrical geometries, and an analysis of 3D structures is carried on, allowing to understand the impact of the lateral charge migration on these devices. In conclusion, disturbs to neighboring cells and impact of the lateral migrated charge on the string resistance are also evaluated in detail.

Riassunto

Il mercato delle memorie non volatili a semiconduttore ha avuto una crescita esponenziale nell'ultimo decennio, grazie alla crescente domanda di prodotti portatili che necessitino di memorizzazione permanente di dati, come per esempio fotocamere digitali, lettori MP3, schede di memoria, chiavette USB, telefoni cellulari e, ultimamente, dischi fissi a stato solido. La tecnologia Flash che permette di ottenere dispositivi ad alta densità, di piccole dimensioni e basso consumo, ha dimostrato la possibilità di rispondere alle esigenze di queste applicazioni e di permetterne lo sviluppo. In particolare le architetture NAND e NOR Flash rappresentano oggi la soluzione vincente per la memorizzazione di dati e codice rispettivamente. Le memorie Flash a Floating-Gate sono state in grado finora di soddisfare le richieste del mercato, specialmente per i dispositivi portatili, diventando così la principale tecnologia di memorizzazione non volatile. L'aumento della densità di integrazione e la contemporanea riduzione dei costi di produzione sono alla base della crescita del mercato delle memorie a stato solido, perché permettono non solo di offrire dispositivi di memoria migliori a prezzi più bassi, ma anche di aprire nuove possibili applicazioni. Proiettando però questa tendenza nel prossimo decennio, ci sono diverse limitazioni che devono essere affrontate per scalare ulteriormente la cella a Floating-Gate. Diverse tecnologie alternative sono state proposte per superare queste limitazioni, e una delle più promettenti sembra essere la tecnologia a trappole discrete usata, per esempio, nelle memorie SONOS e TANOS.

Nel Capitolo 1 il principio di funzionamento delle memorie Flash è presentato, spiegando le ragioni per cui questa tecnologia sta affrontando problemi crescenti nel processo di scaling; quindi sono presentate le possibili tecnologie alternative, dividendole in due grosse categorie: (a) tecnologie evolutive, che modificano leggermente la struttura della cella, ma mantengono lo stesso principio di funzionamento, e (b) nuove tecnologie di memorizzazione, che cambiano completamente l'approccio alla memorizzazione delle informazioni. Delle diverse tecnologie, sarà analizzata la soluzione rappresentata dalle memorie a trappole discrete,

illustrando i motivi per cui può risolvere alcuni dei problemi delle celle Flash e presentando le diverse alternative che cadono in questa categoria, come per esempio celle SONOS e TANOS planari e le architetture 3D, discutendo brevemente vantaggi e svantaggi delle diverse alternative.

Il Capitolo 2 si concentra invece sulla modellistica delle memorie planari a trappole discrete: inizialmente viene presentato un semplice modello analitico per la programmazione, che permette di spiegare alcune delle differenze fondamentali fra le memorie a Floating-Gate e quelle a trappole discrete. Quindi è sviluppato un più accurato modello numerico, in modo da poter comprendere alcune delle peculiarità non spiegabili con il modello analitico: il modello numerico è in grado di riprodurre transitori di programmazione e cancellazione, ed è testato su dati sperimentali di dispositivi SONOS. Per comprendere le differenze fra memorie SONOS e TANOS, viene poi studiato l'impatto dell'introduzione dello strato di allumina nei dispositivi TANOS, partendo da evidenze sperimentali e integrando le proprietà così estratte nel modello e permettendo quindi di riprodurre transitori di programmazione, cancellazione e ritenzione sperimentali su questo tipo di dispositivi.

Nel Capitolo 3, è presentato uno studio della programmazione ISPP (Incremental Step Pulse Programming – programmazione a impulsi crescenti), tramite caratterizzazione sperimentale di condensatori di grande area SONOS e TANOS e, con l'ausilio di un simulatore 3D, tramite simulazioni di dispositivi deca-nanometrici. L'analisi effettuata su dispositivi SONOS di grande area, permette di comprendere meglio le differenze fisiche fra questi dispositivi e le celle Flash a Floating-Gate, mettendo in evidenza e spiegando il calo nell'efficienza di intrappolamento, ed estendendo quindi la caratterizzazione a dispositivi TANOS, mostrando il ruolo dell'allumina in questo tipo di programmazione. Infine sarà presentata un'analisi su dispositivi ultra scalati, che mostra un ulteriore calo dell'efficienza di programmazione dovuto all'aumento del campo elettrico ai bordi della cella.

Il Capitolo 4 sarà dedicato alla modellizzazione di dispositivi a trappole discrete cilindrici: sarà presentato un modelo analitico, ottenendo così una soluzione elettrostatica e studiando l'impatto della curvatura sulla corrente di tunneling e sulle dinamiche dei transitori di programmazione e cancellazione. Il modello è poi confrontato con dati sperimentali, permettendo infine un'analisi parametrica delle celle gate-all-around.

Infine, nel capitolo 5 verrà studiata la migrazione laterale di carica nel nitruro: nei dispositivi 3D lo strato di nitruro non è tagliato ai bordi di ogni cella e questo può provocare un peggioramento dei transitori di ritenzione. Inizialmente saranno presentati dati sperimentali su celle SONOS planari, spiegando come i risultati ottenuti possano essere interpretati in termini di diffusione laterale di carica al di fuori dell'area attiva della cella. Quindi sarà sviluppato un modello 2D per simulare i transitori di ritenzione e sarà utilizzato per riprodurre i dati sperimentali, mettendo in evidenza la necessità di considerare questo processo per seguire l'andamento dei dati sperimentali ottenuti su celle in cui il nitruro continua al di fuori dell'area attiva. Il modello verrà infine esteso a geometrie cilindriche, e verrà presentata un'analisi delle strutture 3D, permettendo di comprendere l'impatto della migrazione laterale di carica su questi dispositivi. Infine saranno valutati nel dettaglio i disturbi sulle celle adiacenti e l'impatto della migrazione laterale di carica sulla resistenza di stringa.

Contents

Chapter 1

Introduction to non-volatile memory technologies

In this chapter various solid state non-volatile memory technologies will be presented. Flash technology is the one dominating the market nowadays, but it is facing always increasing problems with the scaling of the dimensions of the cell: for this reason alternatives must be considered, such as charge-trap devices, that slightly change the memory stack in order to overcome some of the problems of the Foating Gate Flash technology.

1.1 Non Volatile Memory: An Introduction

The growing demand of portable devices, such as smartphones, tablets, handheld game consoles, and e-book readers, brings with it an always increasing demand of non-volatile storage capability with small dimensions, low power consumption, and high reliability: Flash memories can answer to all these demands with increasing capacity and decreasing cost per unit capacity as the scaling continues. The scaling of the Flash memories has continued for about 25 years leading to a market that, according to iSuppli, exceeded $26 billions in 2010. This enormous success was essentially driven by Moore's Law, that lead to dramatic reductions in the feature size for memory over the past few decades. In Fig. 1.1 is shown the timeline for the technology node used for the Flash memory,

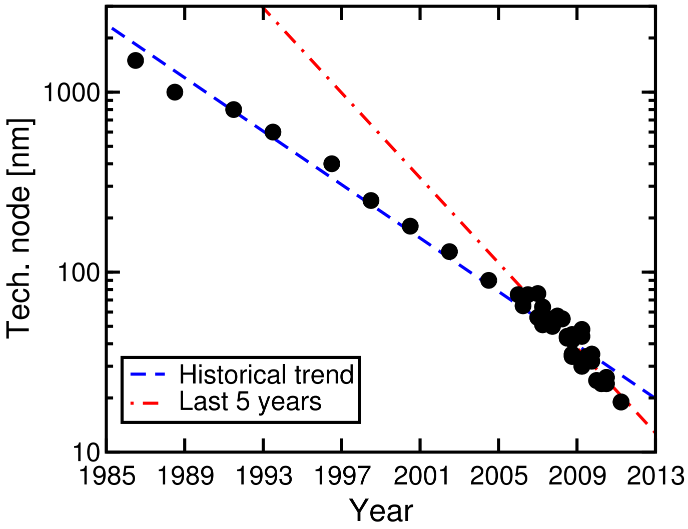


Figure 1.1: Flash memory scaling timeline; the scaling trend is also shown comparing the historical one, that predicts the feature size to halve every 4 years, with the one of the last 5 years, during which the technology node halved every 2 years and a half.

starting from the 1.5 μm cell built in the mid 80's, to the first cell under 20 nm in 2011. From the figure it is evident that in the past few years the scaling trend is accelerating: a fitting of the historical scaling trend gives that the feature size has been halved about every 4 years, while the fitting of the last 5 years only leads to a technology node halved in only 2 years and a half. This continuous reduction in feature size lead to higher density of integration and ultimately to a large reduction of the memory price, that enabled creation of new markets, driving an ever increasing demand for more memory bits that, in turn, largely repaid the efforts devoted for the manufacturing of memory chips with increased performance and functionality, in a sort of virtuous circle. Thus, despite their higher cost per bit with respect to magnetic hard disk drives, semiconductor memories resulted the winning solution in all the consumer products requiring light weight, low size, low power consumption and high reliability. The advances in the lithography technology were the essential driving force of this scaling trend, but the reduction of the unit cost also benefited from innovative self-aligned technologies, from the introduction of the NAND architecture that minimized the cell area for a given technology node, from the introduction of the multi-level cell technology that allowed storing more than one bit per cell, and from the increased wafer size, from 150 mm in 1987, to 300 mm in recent

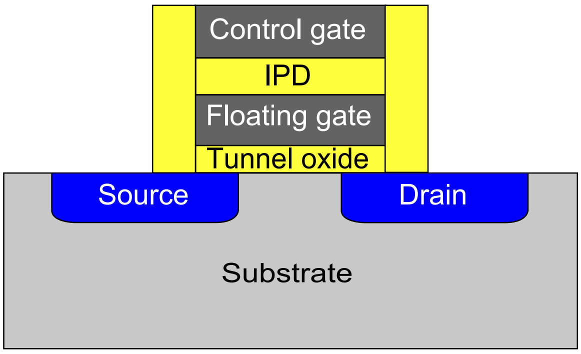
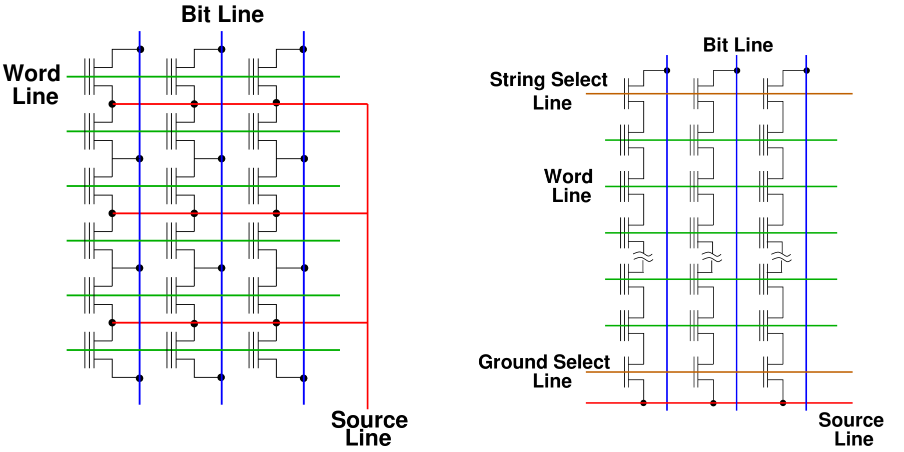


Figure 1.2: Schematic view of a Flash memory cell: the charge is stored in the floating gate (FG) by applying the proper voltage to the control gate (CG); tunnel oxide and interpoly dielectric (IPD) are also shown.

years [1]. However, ensuring that storage capacities continue to increase and that cost per bit continues to fall will require something more than the scaling of feature size for next technology nodes. Further scaling of traditional Flash cell is in fact mined by fundamental physical and technological constraints. In response to this challenge, several emerging memory architectures have been proposed in the last decade as possible alternatives to Flash memory. Such technologies may be classified into two big categories, i.e., (a) evolutionary memories, that essentially rely on the continuation of the existing ones, and (b) completely new storing concepts, no more charge-based. In the following sections, the traditional Flash memory cell, its working principle, and the main scaling limitations will be briefly described, then discussing both the evolutionary scenario, with a particular attention to the charge-trap memories, and the non charge-based solutions.

1.2 Flash Technology

The Flash technology, first proposed in 1979 [2], has become the dominating technology in the field of solid state non-volatile memories. It is based on the use of the Floating Gate (FG) transistor, a principle scheme of which is reported in Fig. 1.2. The FG cell is not different from a MOS transistor, except for the presence of a conductive layer inside the gate oxide of the transistor itself; this layer is completely surrounded by the oxide, hence the definition of floating gate. By injecting charge in the



(a)

(b)

Figure 1.3: Schematic representation of (a) NOR and (b) NAND storage matrices.

FG the threshold voltage of the cell can be increased, and by removing it, the previous state can be restored: the threshold voltage of the cell can be thus used to store a bit of information. There are two types of Flash architectures, NOR or NAND, schematically depicted in Fig. 1.3. Both are organized in two-dimensional matrix, but in NOR type each cell is connected to its word line, bit line and source line (see Fig. 1.3(a)), whereas in NAND one, cells are connected in a series to form a string and for each string there is only one bit line and one source line, as shown in Fig. 1.3(b). Because NAND architecture uses less contacts per cell, it can be packed more densely, allowing a minimum cell area of $4F^2$ as compared to $10F^2$ of the NOR counterpart.

The programming mechanism consists of a controlled shift of the threshold voltage of the cell: this operation can exploit Channel Hot Electrons injection (CHE) or Fowler-Nordheim tunneling (FN) mechanism [3]. The former approach needs a relatively high drain to source current and is used in NOR Flash, where each cell has its drain contact connected to the bitline. FN programming is instead employed in NAND Flash, where the drain contact is not available, and provides slower single bit operation; however, the much smaller value of the tunneling current allows for parallel programming of several cells in the same array and largely enhances the overall write throughput. As a consequence, NOR memory are mainly used for code storage: applications such as embedded logic that require fast access to data that is modified only occasionally. In contrast, NAND memory is a high density, block-based architecture used for data storage: applications where the random access speed is not a constraint, but where the high data density and the low cost per bit are more important, such as for mass storage applications. The erase operation is achieved in both architectures by FN tunneling and is done by polarizing the substrate of the device: as the substrate is common to all the cells in a block, the whole block is erased at once. The reading mechanism consists in applying a positive bias to the control gate (CG) and subsequently reading of the resulting current that can be high or low as a function of V_T, hence function of the stored charge in the FG.

The Flash success should also be attributed to the fact that the realization process is completely compatible with the CMOS one, only using standard materials and lithography: in fact both the FG and the CG are made of polysilicon, the tunnel oxide is standard silicon oxide, while the Interpoly Dielectric (IPD) is constituted of an ONO stack, i.e., a tri-layers stack with two outer parts made of silicon oxide and an inner one of silicon nitride. The Flash technology has been profitably scaled for almost 30 years, following the CMOS evolution: with the cell dimensions getting smaller and the cells getting closer to each other, there are some problems afflicting the scaled Flash memories [4, 5].

The presence of trap states in tunnel oxide or in the IPD layer contribute to some of the major issues in Flash memories: high-field stress induced by FN tunneling during program and erase (P/E) results in a degradation of the dielectrics and so a generation of traps. Those traps limit the cycling endurance of the Flash memory, for they cause significant V_T shift after about 10^5 P/E cycles [6, 7]. Also the presence of traps in the tunnel oxide may also contribute a trap-assisted-tunneling (TAT) leakage path, thus causing an increased charge-loss from the FG by stress-induced leakage current (SILC) [8, 9]: as the the FG is made of a conductive material, the presence of a leaking path in the tunnel oxide can lead, in principle, to the complete discharge of the FG itself, and so the loss of the stored information. Again, oxide traps may also contribute to V_T shifts induced by charge detrapping from the dielectric layers: electrons trapped during the P/E pulse, are released afterwards, resulting in a thermally-activated charge loss [10]; although the threshold shift is lower in this case with respect to the SILC one, it affects most of the array cells, thus it needs to be carefully predicted for V_T window design, especially in multilevel cell (MLC). This kind of capture/emission processes can also lead to random telegraph noise (RTN) [11, 12]; in this case the drain current I_D fluctuates between two values, as a result of alternated capture/emission of electrons at an oxide trap close to the substrate Fermi level (see Fig. 1.4) a high value (low V_T, empty trap state) and a low value (high V_T, filled trap state). This leads to an

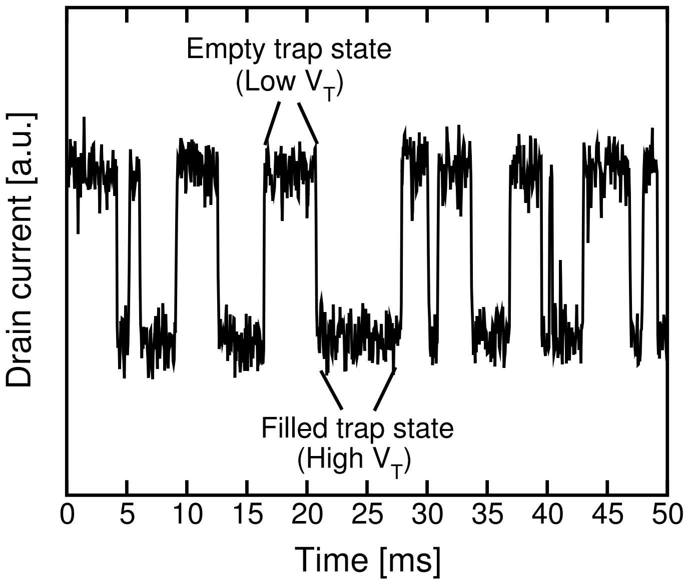


Figure 1.4: RTN two-level fluctuation of the drain current in a Flash memory cell. The low current level corresponds to a filled-trap (high-V_T) state, whereas the high current level is associated with an empty-trap (low-V_T) state (from [11]).

indetermination on the level achieved after the P/E operation, because the V_T randomly fluctuates between two or more levels, in case there is more than one trap giving RTN.

With the dimensions of the cell getting progressively smaller, the RTN impact grows, even more than what could be expected by the simple trapping of a charge at the interface between the silicon channel and the tunnel oxide [12,13] : this is due to the fact that the conduction in the channel is not uniform, but happens through preferred paths. These percolation paths are due to the presence of discrete dopant ions and to the local field enhancements that confine the current at the edges of the cell. Fig. 1.5 shows a TCAD simulation of the current density in a 18 nm cell [14]: the presence of two dopant ions near the surface is highlighted, and the current flows mostly on the left edge of the cell. If there was a RTN trap just over this edge, the current flow would be greatly reduced, and the V_T shift would increase more than expected by simple 1D calculation.

In addition to RTN, there are other sources of V_T spread that should be considered. As the cell scales down, the number of electrons stored in the FG decreases at fixed ΔV_T [15]. Given the stochastic nature of the quantum-mechanical tunneling effect, used for injecting charge in the

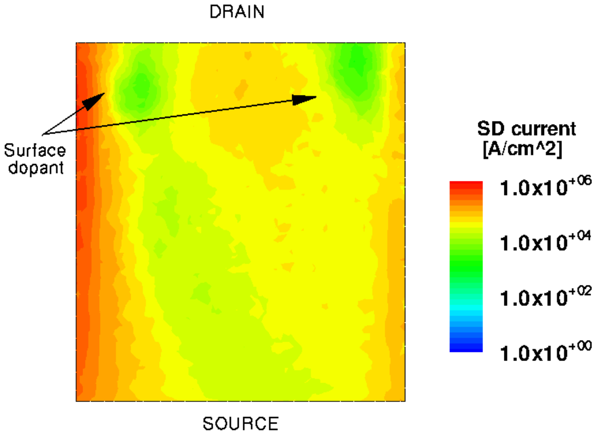


Figure 1.5: TCAD simulation of the current density map in a 18 nm cell in presence of random dopant and edge field enhancement (from [14]).

FG, the actual number of electron injected during a P/E operation may vary considerably, thus affecting the precision of MLC operations [16]. Accurate modeling of spread due to few-electron effects requires Monte Carlo and analytical techniques: advanced programming algorithms can help in reducing the spread, but the problem worsen if MLC are considered, because the V_T difference between two levels decreases with increasing the number of levels to be distinguished in the memory. A further source of V_T instability comes from electrostatic coupling between FGs of adjacent cells, which increases for decreasing distance between cells in the array. With high enough electrostatic coupling, the V_T of a cell not only depends on the charge stored in its floating gate, but also depends on the charge stored in the adjacent cells' FGs: as these cannot be predicted, there is an increase in the spread of a V_T distribution [17, 18]. These effects require careful array level electrostatic 3D modeling, which can provide valuable tools for developing P/E algorithms to minimize cell-cell interferences and improve V_T distributions.

1.3 Beyond Floating-Gate Flash

During the years, many possible solutions have been proposed to the problems exposed in the previous section, and they can be divided into two groups: evolutionary scenario or non charge-based technologies. The former case counts on small changes in the realization of the FG cell,

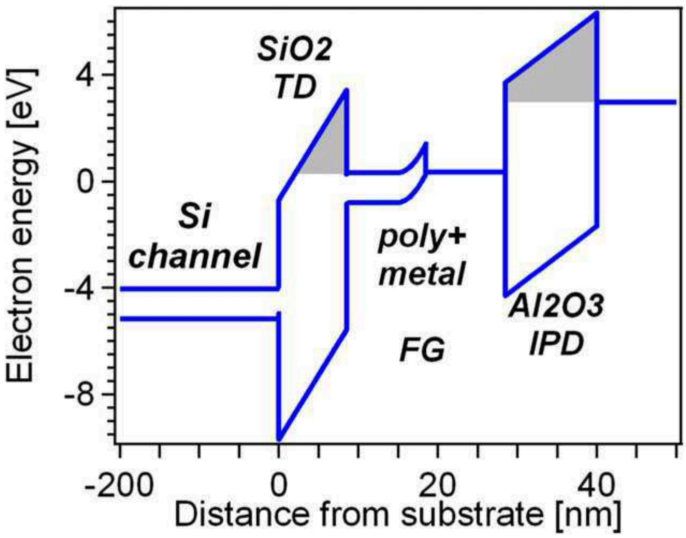


Figure 1.6: Band diagram during erase for the ultrathin FG concept. The FG is made of two layers, n-type polysilicon towards the tunnel oxide and p-type metal toward the IPD (from [19]).

always keeping the information stored in the V_T value of a MOS-like transistor, while the latter is formed by technologies that completely change the way to store the information putting it, for example, in the resistive state of a cell.

1.3.1 Evolutionary scenario

In the evolutionary scenario, the information is still stored in the V_T value of a MOS-like transistor and, ultimately, in the quantity of charge stored in the gate oxide of the cell. Various solutions have been proposed, changing the way the FG cell is realized, and trying to solve one or more of the problems exposed.

The first possibility is to use an ultra-thin floating gate, in order to minimize the electrostatic coupling between adjacent cells [19]. Moreover the FG is made of n-type polysilicon towards the tunnel oxide, to maintain high erase efficiency, and of a p-type metal towards the IPD, as shown in Fig. 1.6 thus reducing the leakage current through the IPD for a given coupling ratio and providing a larger memory window. The use of a p-type metal gate also allows to avoid the low-κ layer that forms when a high-κ dielectric is deposited on top of polysilicon, allowing also to change the IPD material in order to increase the coupling ratio, i.e. how well the FG potential is controlled by the voltage applied to the

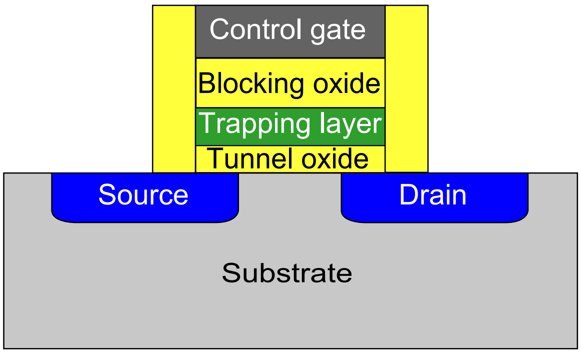


Figure 1.7: Schematic view of a charge trap memory cell: the charge is stored in the trapping layer by applying the proper voltage to the control gate (CG); tunnel and blocking oxides are also shown.

control gate. The main issue of this technology is the feasibility of integrating a thin metal FG in the process flow to realize the gate stack. Simulation show that this solution may allow the 15 nm technology to be realized [5].

Another possibility is to work on the scaling of the physical thickness of the IPD itself; if the FG can be realized with mono-crystalline silicon, the SiO_2 layer grown on top of it can be as thin as 7 nm, just like the tunnel oxide [20]. In this way, even at very small cell dimensions, the architecture can still be realized with the wrapped CG architecture, because there is still space for the two IPD layers on side of the FGs and for the CG between two adjacent cells; this allows to increase the coupling ratio and have higher electron count. Simulations show that this concept could be used to realize the 10 nm technology node [5]. Main issues are the integrity of this oxide on an etched sidewall as well as at the corners of the FG and the cost efficiency of using silicon regrowth techniques in a memory process flow.

A further alternative is to use the charge-trap (CT) memories. From the schematic view shown in Fig. 1.7 it can be seen that the polysilicon FG is replaced by a trapping layer, that is a thin layer of a dielectric material, such as silicon nitride (Si_3N_4), with a high density of trapping sites, in which the charge can be permanently stored. This solution allows to solve, at least in principle, the SILC problem, because the presence of an eventual conductive path in the tunnel oxide only discharges the charge stored in traps above it, and not all the

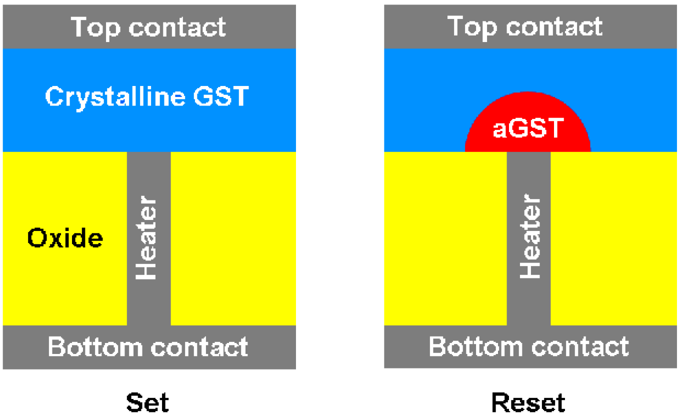


Figure 1.8: Schematic of a PCM cell. The active chalcogenide material is contacted by the heater, which heats it through electrical pulses. The cell is in the low resistive state (set state) when the chalcogenide is all crystalline (left), while the high resistive state (reset) is characterized by the presence of an amorphous active volume (right).

charge in the FG, as it is in standard Flash devices. Also the interference is greatly reduced, because the trapping layer is as thin as a few nm, compared to the around 80 nm of the polysilicon FG [21], thus reducing the electrostatic coupling between adjacent cells. CT technologies, such as Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) and TaN-Alumina-Nitride-Oxide-Silicon (TANOS) memories, will be discussed in more depth in section 1.4.

1.3.2 Non charge-based technologies

A part from the possible evolutions exposed in the previous section, there are also other technologies that are based on different physical phenomena to store the information. Among new proposed storage concepts it is possible to include the magneto-resistive memory (MRAM) [22], ferro-electric memory (FeRAM) [23], phase-change memory (PCM) [24] and resistive-switching memory (RRAM) [25,26]. All this concepts base their working principle on a change of the resistance in the active material, with the exception of the FeRAM, where data are stored in the polarization of a ferro-electric material and the memory element is basically a capacitor. Out of these concepts, the most advanced and most promising is the PCM, that is based on the ability of chalcogenide material to reversibly switch between a crystalline low resistive state and an amorphous high resistive one by simply heating the material with the passage of an electric current, as shown in Fig 1.8. The active chalco-

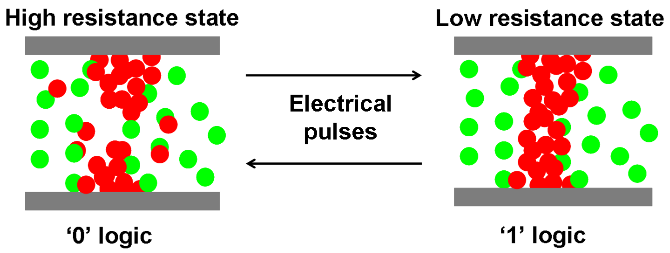


Figure 1.9: Sketch of the operations in a RRAM cell. The active material is normally in a high resistive state (left). By the application of electric pulses, a conductive path can be formed, bringing the cell in the low resistive state (right).

genide material is contacted by a thin metallic plug, the heater, in order to allow the change of phase, from crystalline to amorphous and vice versa, by electric pulses. The information is stored in the resistive state of the cell and can be simply read by applying a fixed voltage to the cell and reading the current flowing through the active material. PCMs are faster than Flash memories, as the P/E and reading operations only require $10 \div 100$ ns, vs the $0.1 \div 1$ ms of the Flash memories. Also PCM are more resistant to degradation, and can be cycled at least 2 orders of magnitude more than the FG counterpart, relaxing thus the need of a wear leveling algorithm in the controller of the memory. Despite all these advantages, the PCM technology is not as widespread as the Flash one, because it is more expensive, for it requires material that are not used in the standard CMOS process flow and it is more power-hungry, as it need a high current pulse for the switching in order to heat the chalcogenide material up to 600°C and allow the switch between the two phases. For these reasons PCMs are not as widespread as the Flash memories, and are generally seen as a replacement for the NOR memories, that require less density of integration, need to be fast and are frequently read but only seldom written. Anyway the scaling of the cell also reduces the volume of the material used and the amount of power needed for the switching, while the mass production can further reduce the production cost: these effects can relieve the actual drawbacks of PCMs with respect to Flash.

Another promising technology is the RRAM: this kind of memory is based on the ability of some dielectric materials to switch between a high resistive state and a low resistive one thanks to the creation of a conductive path that, depending on the dielectric material, the contacts material and the electrical pulses used for the programming of the cell, can be due to defects in the oxide, or to metal migration, etc. (Fig. 1.9).

The conductive path can be afterwards dissolved by the application of a proper electric pulse, making the process reversible, and allowing to take advantage of it to store a bit. Like the PCM cells, the P/E and reading operations are very fast, and also need less power to make the cell switch, but they degrade faster than PCM cells and so the endurance is still quite low. RRAMs are still in the research phase and are forecast to replace the NAND Flash technology below the 10 nm node.

1.4 Charge trap memories

In the evolutionary scenario, one of the best alternatives to the FG Flash memories is to use the Charge Trap (CT) concept. In this technology the polysilicon FG in Flash memories, is replaced by a trapping material, that is an insulator with a high density of defects that can act as traps in which store the charge injected during the programming phase. The main advantage is that the presence of a conductive path in the tunnel oxide, that in FG case is the cause of SILC, can only discharge the traps right above it, because the electrons cannot move freely, but are held in the traps. Another advantage in substituting the FG with a trapping material, is that the layer thickness decreases of about one order of magnitude, from $\sim$ 80 nm [21] to $\sim$ 6 nm [27], also reducing the interference due to electrostatic coupling between adjacent cells. The two main CT technologies are the SONOS and TANOS cells, and one of the interesting advantages of the CT cells is that they can be used to exploit the third dimension and realize 3D structures.

1.4.1 SONOS cells

One of the first CT concept proposed is the Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) cell [28], in which the trapping layer is made of silicon-nitride (Si_3N_4), the blocking oxide is of silicon dioxide, and the control gate is a polysilicon one. The use of a trapping layer in which the charge is stored in traps and, once captured, is basically fixed in that position, requires that the programming operation is as uniform as possible, in order to obtain a threshold that is uniform on all the channel: for this reason SONOS memories are usually programmed using the Fowler-Nordheim (FN) tunneling. Analogously, also the erase should be homogeneous above the channel, in order to avoid charge accumulation after cycling the cell: also for the erase phase, FN tunneling is normally used.

One of the problems of the SONOS cells is the injection of electrons from the gate through the blocking oxide during the erase operation: as

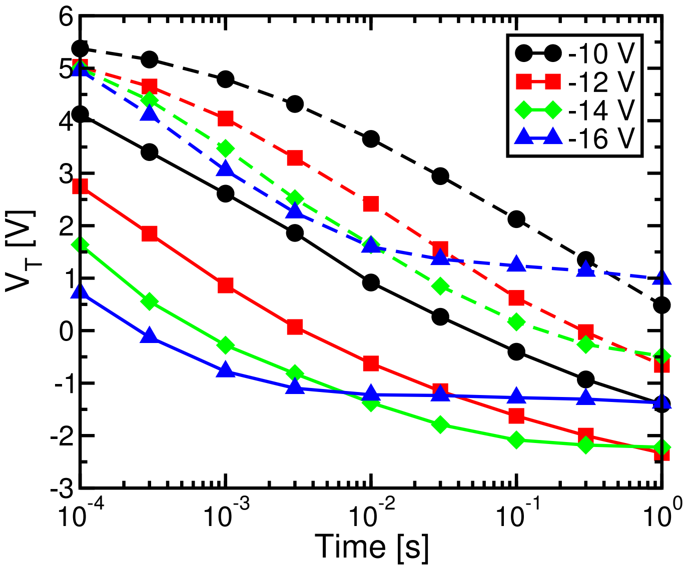
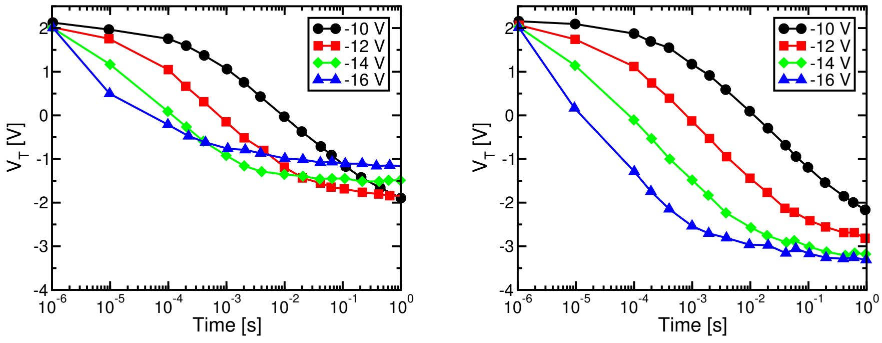


Figure 1.10: Erase transients on SONOS devices with tunnel oxide thickness of 2.0 (solid lines) and 2.5 nm (dashed lines) at different voltages. Nitride layer thickness is 6 nm with 8 nm blocking oxide, and the cell were programmed to $V_T = 6$ V prior the erasing operation. It is evident the saturation of the erase curves, especially for the cell with the thicker tunnel oxide and at higher erasing voltages (data from [29]).

the blocking oxide is made of the same material as the tunnel oxide, also the electric field will be similar. During the erase there will be an outward electron flux from the nitride to the substrate and an inward flux from the gate to the nitride: when the two fluxes balance each other, the threshold voltage saturates, as shown, for example, in Fig. 1.10 by the dashed blue lines [29]. This effect limits the lower threshold voltage achievable during erase to a level that depends on the tunnel oxide thickness and the gate voltage used for the erase operation. To solve the erase saturation problem, one solution is to enhance the hole injection from the substrate: as the hole injection is more difficult due to the higher valence band offset between the substrate and the tunnel oxide, the tunneling barrier must be very thin, in order to enhance considerably the hole current (2 nm or lower [29]). In this way the erase saturation happens at lower V_T values, even if still present (as shown in Fig. 1.10 by the solid lines). The problem in making the tunnel oxide so thin is in the retention phase: both the programmed and the erased states lose charge faster and, after 10 years, the residual threshold window is only hundreds of millivolts [30]. In literature it is possible to find examples of SONOS memories able to guarantee, after 10 years retention at 85°C, a



(a)

(b)

Figure 1.11: Comparison between erase transients in a SANOS cell (a), with polysilicon gate, and a TANOS one (b), with TaN gate. Both cell have the same gate stack with alumina used as a blocking oxide: it is visible as the transient saturation happens at lower voltages for the cells using a metal gate (from [33]).

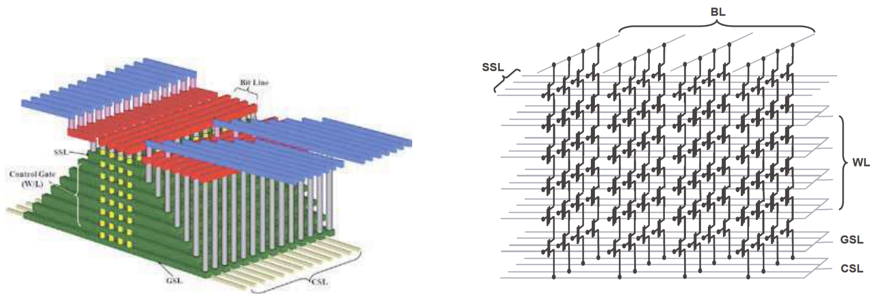
residual threshold window, although quite small [31,32]: these memories can be used in System-on-Chip (SoC) products as embedded memories; thanks to the complete compatibility with the CMOS process, and to the lack of additional masks, as it happens with FG Flash, the SONOS memories are a viable solution for low cost embedded systems.

In order to improve the erase transient without affecting too much the retention, TANOS memories were introduced: by using a higk-κ material as a blocking oxide, the injection of electron from the gate can be greatly reduced.

1.4.2 TANOS cells

By simply changing the blocking oxide material and using a metal gate, the erase saturation problem can be solved. This is obtained in the TaN-Alumina-Nitride-Oxide-Silicon (TANOS) cell [33]. In Fig. 1.11 [33] are shown the experimental erase transients obtained using these devices: it is evident how this solution can help in solving the erase saturation present in SONOS cells.

Using a high-κ dielectric as the blocking oxide, like the aluminum oxide or alumina (Al_2O_3) which has a dielectric constant $\varepsilon \approx 9 \div 10$ [34–36], the same electrostatic coupling between the gate and the nitride can be obtained with a thicker layer, leading to a lower electric field in the blocking oxide, and thus a lower tunneling flux during erase. There are technological problems in using a polysilicon gate over the alumina layer due to fermi level pinning at the interface [37]. For this reason



(a) (b)

Figure 1.12: Principle scheme of a vertial cylindrical memory: (a) Birds-eye view, (b) equivalent circuit of the array.

a high barrier between the Fermi level in the gate and the conduction band of alumina can only be obtained using a metal gate made of a material with a high work function, such as Tantalum-nitride (TaN) that is a metal that can be easily integrated in the CMOS process and has $\phi_{WF} \approx 4.7$ eV [34, 36].

1.4.3 3D structures

Three-dimensional memories are gathering increasing attention as future ultra-high density memory technologies to keep a trend of increasing bit density and reducing bit cost [38]. The simplest approach is to build the cell on a thin polysilicon substrate, and stack more than one level: this solution increases the density with the number of levels, but does not offer any advantage over traditional planar SONOS cells in terms of P/E and retention operations [39, 40].

Another approach is to build a SONOS cell with a cylindrical substrate: the structure can be built by drilling a hole in a alternated polysilicon-oxide stack (the conductive layers will be used for the gates of the cells), and then deposit the ONO stack on the side of the hole and fill the remaining gap with a polysilicon channel [41–43], obtaining a final structure similar to the one shown in Fig. 1.12

These kind of gate-all-around charge-trap (GAA-CT) cell with vertical channel are considered one of the most promising structures for future NAND Flash technologies, showing improved program/erase and retention performance with respect to planar devices [44]. This is due to the curvature effect that relaxes the erase saturation problem: the electric field in the blocking oxide is lower than the one in the tunnel oxide, allowing to increase the out flux from nitride traps toward the substrate

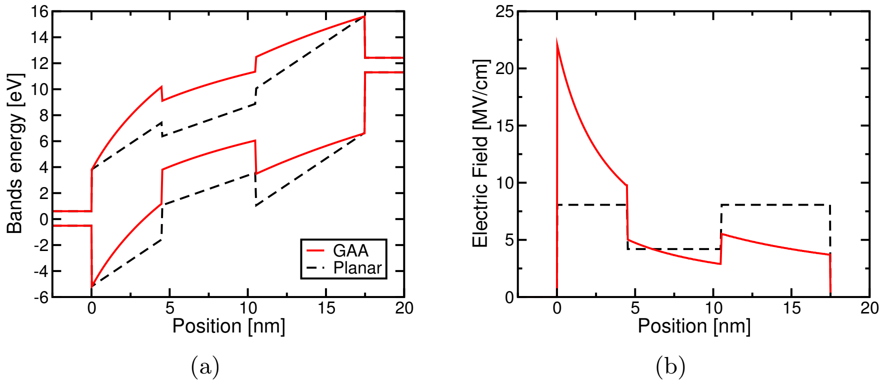


Figure 1.13: Comparison between a cylindrical GAA-CT cell and a planar one with the same gate stack during erase at $V_G = -12$ V and with neutral nitride: (a) energy band profile and (b) electric field.

and reducing the back-tunneling flux coming from the gate. In Fig. 1.13 are reported the energy band diagram and the electric field for a GAA-CT cell with ONO thickness of 4.5/6.0/7.0 nm and an equivalent planar one, under erase conditions with neutral nitride and $V_G = -12$ V: in the planar cell the electric field in the blocking oxide and the one in the tunnel oxide are the same, thus leading to comparable fluxes and to erase saturation. Instead in the GAA case the electric field decreases with increasing radius, so the electric field is higher near the substrate than near the gate, leading to less important charge injection through the blocking oxide.

1.5 Motivation of the work

With the increasing problems arising for FG Flash memories as scaling proceeds, it is important to study the achievable performance of the presented alternative solutions. The most promising solution seems to be represented by the CT memories, as they are based on an evolution of the actual mainstream technology. It is important to understand the achievable performance and the limitations of this technology, in order to understand if it can replace the FG Flash and, in case, at which technology node. In order to do that it is important to understand the physics at the basis of the fundamental operations, i.e., program, erase and retention transients. In order to do that various modeling approaches can be adopted, ranging from 1D analytical models, that are useful to understand the main dependences of the transients on the basic parameters of the structure, to full 3D models that can also help in understanding the geometric effects in ultra-scaled cells and optimize, for example, the doping profile in order to avoid problems arising from the fringing field. Also different compositions of the gate stack must be evaluated: the TANOS technology, that seemed able to solve the SONOS problem with erase saturation, cannot be considered a complete solution, as is still has problems due to non-idealities of the alumina layer [36, 45–47]; also nitride engineering, varying the silicon content, gives an erase/retention trade off [48]. BE-SONOS cells and all the modifications on the concept may be interesting, but the trapping in the thin nitride layer should be carefully modeled in order to fully understand the possibilities and the improvement that this technology can give with respect to the standard charge trap memories and how it can help in solving the trade off. Moreover, with the FG Flash scaling projected down to the 10 nm node, other structures are gaining interest, namely 3D concepts, such as the ones using a vertical cylindrical substrate, that can improve the performance thanks to the curvature effect of the device, that can help in improving program and erase despite using relatively thick tunnel and blocking oxides, needed to achieve the retention requirements.

The aim of the present thesis is to analyze some of the previous points and create models able to catch the physics of the various operations, to reproduce experimental data, and predict possible optimizations of the different parameters of the cell. Both planar and cylindrical devices will be analyzed, focusing mainly on the critical points to the realization of the memory cells and trying to give a physical explanation to some aspects in which these memories behave in a different way with respect to standard FG Flash memories.

In particular, chapter 2 will focus on the modeling of planar CT memories: first a simple analytical model for the program operation of these devices will be presented, allowing to explain some of the fundamental differences between the FG and the CT memories. This model is also used to perform a parametric analysis and understand the main dependences of the programming transient on different parameters. Afterwards a more accurate numerical model will be introduced, in order to address some of the peculiarities that cannot be described by the analytical one: the model is able to reproduce program and erase transients, and is tested against experimental data on SONOS devices. Later, in order to understand the differences between the SONOS and TANOS memories, the impact of the introduction alumina layer in TANOS devices will be discussed, starting from experimental evidences and integrating the extracted properties in the previous model, allowing to reproduce experimental program, erase and retention transients on such devices.

Afterwards, in Chapter 3, a study of Incremental Step Pulse Programming (ISPP), that is the most used programming scheme in FG Flash devices, will be presented. This algorithm has some intrinsic advantages in FG Flash, but these are reduced in CT memories. The study is made with a characterization of ISPP on large area SONOS and TANOS capacitors, and using a 3D model on deca-nanometer devices. The analysis made on large area SONOS capacitors allows to better understand the physical differences between these devices and the FG Flash, mainly highlighting and explaining the decrease in the trapping efficiency without considering geometric effects, and then extending the characterization to TANOS devices, pointing out the role of the alumina layer. Finally, an analysis on ultra-scaled devices will be presented, revealing a further decrease of the programming efficiency, that is caused by the fringing field.

Chapter 4 will be dedicated to the modeling of cylindrical CT memories. A physics-based analytical model will be presented, obtaining the electrostatic solution and studying the curvature effect impact on the tunneling current and on the transient dynamics. The model will then be tuned against experimental data on both cylindrical SONOS and TAHOS devices, and finally a parametric analysis of the gate-all-around CT cells will be presented, allowing to point out the achievable performance for the structure depending on the cell parameters.

At the end, in Chapter 5 lateral charge migration in the nitride layer will be studied: in 3D structures the nitride is not cut at the borders of each cell, and this can lead to worse retention transient. In order to study the effect, planar SONOS cells with nitride layer not patterned above the active area are characterized. The obtained experimental data will be presented, explaining how the results can be interpreted in terms of lateral diffusion of the charge out of the active area of the cell. Starting from these results, a 2D model able to simulate retention transients is developed and tested against the experimental data, highlighting the need of a diffusion process to reproduce retention transients on measured cells. The model is then extended to cylindrical geometries, and an analysis of 3D structures is carried on, allowing to understand the impact of the lateral charge migration on these devices. In conclusion, a first order analysis of the scaling limitations, due to retention constraints, will be presented.

Chapter 2

Modeling of planar charge-trap memories

In this chapter the modeling of program, erase and retention operations in charge trap memories will be presented, giving an insight of the physics that rules these basic operations. First a basic 1D analytical model will be introduced to investigate the fundamental differences with the FG cells. Then the model is extended into a numerical one, able to simulate program and erase (P/E) transients in CT cells, and is tested against data obtained on SONOS devices. Finally the alumina non-idealities will be studied and the model will be further extended to reproduce P/E and retention transients on TANOS devices.

2.1 Introduction

Charge trap memories are considered a viable solution for the continuation of the scaling trend of the Flash memories, as they solve some of the problems arising with the scaling of the traditional FG transistor. Replacing the polysilicon floating-gate with a nitride layer for charge storage allows in fact a significant improvement in cell immunity to stress-induced leakage current (SILC) and cell-to-cell parasitic interference, enhancing the scaling perspectives of the technology from the reliability standpoint. The SONOS gate stack has been one of the first architectures implementing the CT memory concept, but showed poor performances due to the compromise between data retention (requiring a relatively thick bottom oxide to avoid the trapped charge to escape from the nitride) and program/erase (P/E) window (requiring a thin tunnel oxide to avoid erase saturation) [49].

The TaN/Al_2O_3/Nitride/Oxide/Silicon (TANOS) structure has been proposed to solve this compromise, featuring a bottom oxide thickness larger than 4 nm for long data retention and disturb immunity and high-κ top dielectric with high work-function metal gate for reduced erase saturation [50]; nevertheless the alumina layer cannot be considered an ideal dielectric, for it has a non negligible density of defects [36, 45, 46], that must be taken into account to correctly reproduce the program, erase and retention transients in TANOS memories.

A modeling of planar large area devices can help in understanding the physical principles behind the basic operations of program, erase and retention, and explain some of the differences with the FG cells. In the recent years several 1D modeling approaches have been developed to describe large area planar charge traps devices [34, 51–56]: most of them treat the trapping/detrapping process in the silicon nitride using the Shockely-Read-Hall (SRH) recombination process [57], usually with the approximation made by Arnett when considering it in insulators at high field [58].

2.2 SONOS modeling

Fig. 2.1 shows the band diagram along a SONOS memory device during programming at a high gate voltage V_G: starting from simple physical equations, a first order analytical model for the programming phase can be derived, allowing the understanding of some basic properties of V_T transients in nitride memories. A more refined numerical model is then presented to achieve a good quantitative agreement between experimental and modeling results. The model implements an accurate description of the trapping/detrapping processes in the nitride layer, including the detailed description of electron and hole transport in the layer. Also the finite number of traps available for charge storage, their energetic and spatial distribution, and the nonzero energy relaxation length of injected carriers [34, 59] are taken into account. With all these features, the model is able to explain the reduced gate control over V_T transients with respect to floating-gate cells during programming, which affects the achievable programming performances, and can also reproduce the erase and retention phases.

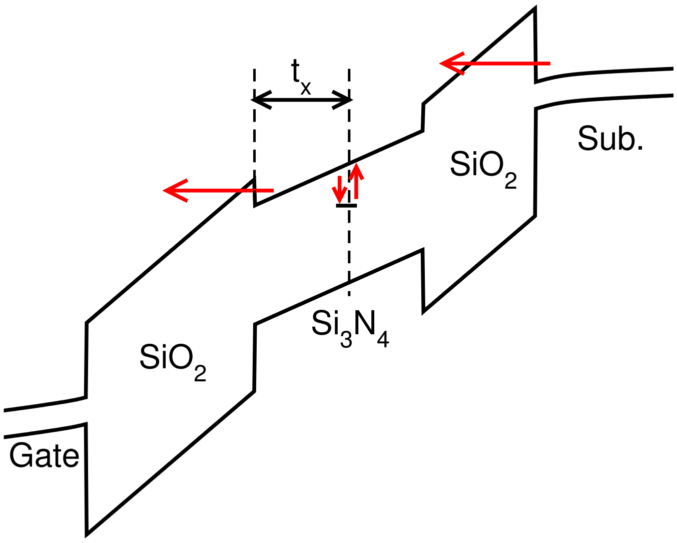


Figure 2.1: Schematics for the band profile in the SONOS structure during a program operation, highlighting the electron fluxes taking place in the device.

2.2.1 First order analytical model for programming

Starting from the band diagram shown in Fig. 2.1, a simple first order analytical model for the programming phase can be derived. Assuming that there is a charge Q trapped in the nitride with centroid at a distance t_x from the blocking oxide interface, this gives rise to a threshold-voltage shift ΔV_T

$$\Delta V_T = -Q\left(\frac{t_{bo}}{\varepsilon_{bo}} + \frac{t_x}{\varepsilon_{sin}}\right) = -\frac{Q}{C_{pp}}. \tag{2.1}$$

In the previous equation, ε_{sin} and ε_{bo} are the nitride and the blocking oxide dielectric constants respectively, t_{bo} is the blocking oxide thickness and $C_{pp} = \left(\frac{t_{bo}}{\varepsilon_{bo}} + \frac{t_x}{\varepsilon_{sin}}\right)^{-1}$ is the capacitance from the trapping point to the gate. The tunneling current entering the silicon nitride layer is strictly related to the electric field F in the bottom oxide, which can be straightforwardly calculated as

$$F = \frac{V_G - \Delta V_T}{EOT}, \tag{2.2}$$

where it was assumed, for the sake of simplicity, that the flat band voltage of the device is $V_{FB} \approx 0$ V, and where EOT is the equivalent oxide thickness of the gate stack, given by

$$EOT = \varepsilon_{ox}\left(\frac{t_{tun}}{\varepsilon_{tun}} + \frac{t_{sin}}{\varepsilon_{sin}} + \frac{t_{bo}}{\varepsilon_{bo}}\right), \tag{2.3}$$

where t_{tun} and t_{sin} are the tunnel oxide and silicon nitride layer thicknesses and where ε_{tun} and ε_{ox} are the the tunnel oxide and the SiO_2 dielectric constants. For the sake of generality, different dielectric constants are considered for the tunnel and blocking oxides and for nitride (respectively ε_{tun}, ε_{bo} and ε_{sin}), though final results for the SONOS devices will consider $\varepsilon_{tun} = \varepsilon_{bo} = \varepsilon_{ox}$. The density of electron-filled traps in the nitride $n'_t = -Q/q$ (units: cm^{-2}) increases as a consequence of the capture of a part of the FN electron flow coming from the substrate, according to the relation [58]

$$\frac{dn'_t}{dt} = \frac{J}{q}\sigma\left(N'_t - n'_t\right) - n'_t\left\langle e\right\rangle, \tag{2.4}$$

where J is the current density of the charge entering the nitride layer (units: $\mathrm{A/cm^2}$), σ is the capture cross section of the nitride traps (units: $\mathrm{cm^2}$), N'_t is the total trap density (units: $\mathrm{cm^{-2}}$), and $\langle e\rangle$ is the electron detrapping rate (units: $\mathrm{s^{-1}}$), e.g., by thermal or tunneling emission. Note that the product $\sigma \times N'_t$ should be lower than one for the previous equation to keep its validity: in fact that term represents the fraction of the injected flux that is trapped per unit of time, when the traps are empty, i.e., when $n'_t = 0$. Taking the time derivative of (2.1) and combining it with (2.4), the following equation for the time evolution of ΔV_T can be obtained:

$$\frac{\Delta V_T}{dt} = \frac{J\sigma\left(N'_t - n'_t\right) - qn'_t\left\langle e\right\rangle}{C_{pp}} = J\sigma\left(\frac{N'_t}{C_{pp}} - \frac{\Delta V_T}{q}\right) - \Delta V_T\left\langle e\right\rangle, \tag{2.5}$$

where C_{pp} is considered as a time-independent constant, i.e. any variation of the charge centroid in the nitride with time is neglected. In order to solve (2.5), the following FN formula for the tunneling current through the bottom oxide can be used:

$$J = AF^2e^{-B/F} \tag{2.6}$$

where A and B depend on the physical parameters of the potential barrier [60, 61]. By means of (2.2) and (2.6), (2.5) can be expressed as a function of the electric field F

$$\frac{dF}{dt} = -AF^2e^{-B/F}\left[\frac{\sigma N'_t}{C_{pp}EOT} - \frac{\sigma}{q}\left(\frac{V_G}{EOT} - F\right)\right] + \left(\frac{V_G}{EOT} - F\right)\left\langle e\right\rangle. \tag{2.7}$$

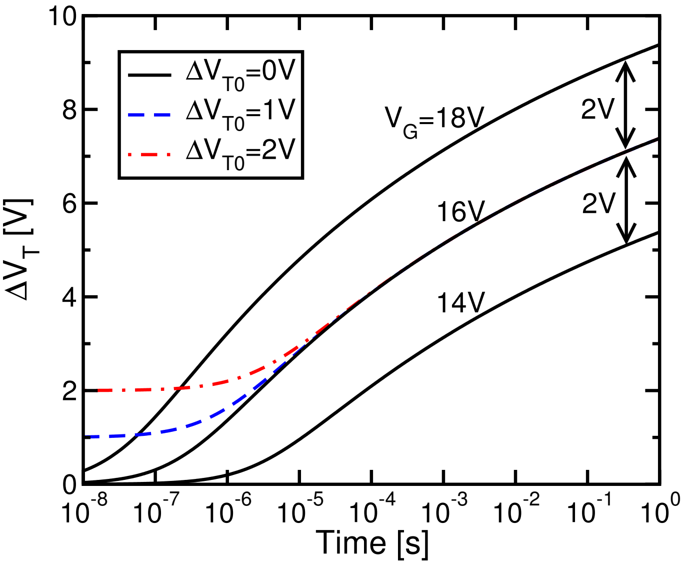


Figure 2.2: Programming ΔV_T transients calculated by (2.10) for different V_G values. Results for $V_G = 16$ V are also shown for $\Delta V_{T0} = 1$ and 2 V.

Large number of trapping sites and no detrapping

Neglecting electron detrapping (i.e., putting $\langle e \rangle = 0$) and assuming that $N_t' \gg n_t'$, only the first term in the square brackets on the RHS of (2.7) can be considered, obtaining

$$\frac{dF}{dt} = -AF^2 e^{-B/F} \frac{\sigma N_t'}{C_{pp} EOT}, \tag{2.8}$$

which can be straightforwardly integrated, with the initial condition $F(t = 0) = F_i$, to obtain the time evolution of the electric field F during programming

$$F = \frac{B}{\ln\left(\frac{\sigma N_t' AB}{C_{pp} EOT} t + e^{B/F_i}\right)}. \tag{2.9}$$

If no charge is initially stored in the nitride, $F_i = V_G/EOT$. Using (2.2), (2.9) gives the ΔV_T evolution with time

$$\Delta V_T = V_G - \frac{EOT \cdot B}{\ln\left(\frac{\sigma N_t' AB}{C_{pp} EOT} t + e^{B/F_i}\right)}. \tag{2.10}$$

Fig. 2.2 shows the calculated ΔV_T transients obtained at V_G equal to 14, 16, and 18 V, assuming the device parameters reported in Table 2.1.

Parameter	Value
t_{tun}	4.0 nm
t_{sin}	6.0 nm
t_{bo}	6.0 nm
ε_{tun}	3.9
ε_{sin}	7.5
ε_{bo}	3.9
A	10^{-5} A/V^2
B	$2.4 \cdot 10^8$ V/cm
N_t'	$3 \cdot 10^{13}$ cm^{-2}
σ	10^{-15} cm^2

Table 2.1: List of the parameters used to calculate the ΔV_T programming transients shown in Fig. 2.2.

For a time t that is sufficiently long to lose the ΔV_T dependence on the different initial electric field F_i, the curves are only vertically shifted by the difference in their programming V_G. This means that, for a fixed time t, the Gate Sensitivity Factor $GSF = \partial \Delta V_T / \partial V_G$ for the transients is equal to one, as normally obtained on floating-gate memory cells. Note that this result does not derive from the trapping of all the electrons injected into the nitride layer, because it was assumed that $\sigma \times N_t' < 1$ (i.e. the fraction of the incoming flux that is trapped per unit time), but from the hypothesis that the trapped charge density is negligible with respect to the number of traps, so the trapped flux remains constant through time. From (2.2), in fact, when the programming voltage is increased from V_{G1} to V_{G2}, the same electric field F is obtained when the difference between the threshold-voltage shift obtained in the two cases (called ΔV_{T1} and ΔV_{T2} respectively), is equal to: $\Delta V_{T2} - \Delta V_{T1} = V_{G2} - V_{G1}$. When this condition is reached, the same tunneling flux enters the nitride layer and the same ΔV_T evolution takes place in both cases, as no limitations in the trapping dynamics come from the charge that has already been stored, owing to the assumptions made. As a result, the ΔV_T transients are only vertically shifted by $V_{G2} - V_{G1}$.

Fig. 2.2 also shows that, for a fixed programming voltage V_G, the ΔV_T transients display a converging behavior when different initial ΔV_T values are assumed (e.g., $\Delta V_{T,i} = \Delta V_T(t = 0) = 1$ and 2 V in the figure for $V_G = 16$ V), as usually obtained in floating-gate memory devices. Moreover, (2.10) states that the shape of these transients is affected only by a change in EOT or B. In fact, any change in σ, C_{pp}, A, or N_t' determines only a horizontal shift of the curves in the logarithmic

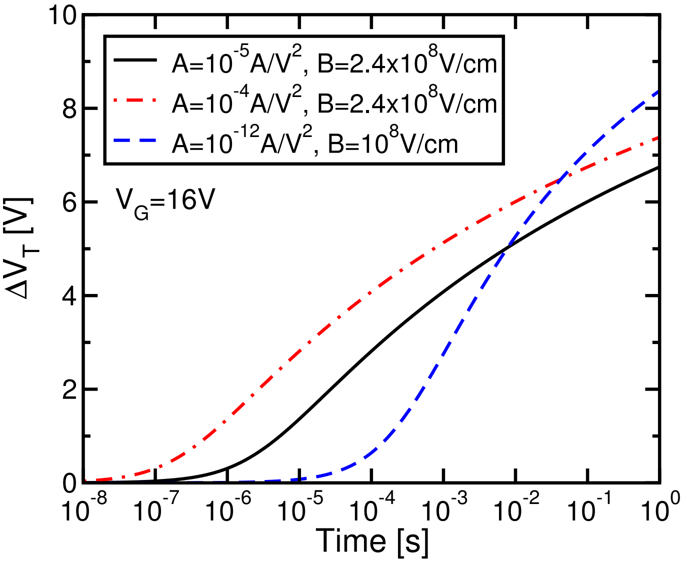


Figure 2.3: Programming ΔV_T transients calculated by (2.10) for different values of the tunneling parameters A and B.

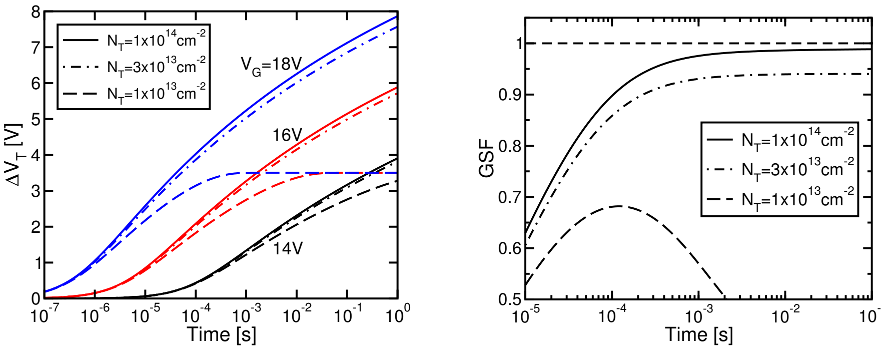
time axis, as shown in Fig. 2.3 when changing A from 10^{-5} to 10^{-4} A/V^2 for a fixed B. Instead, the figure shows that decreasing B from 2.4×10^8 to 10^8 V/cm increases the slope of the transient: this is due to the less field dependent tunneling current characteristics that result from the reduction of B, allowing a smaller decrease of the programming current for a given charge stored in the nitride layer. Finally, note that the change in shape of the ΔV_T curve does not modify the GSF of the transients, which is still equal to one.

Effect of the finite trap density

When the hypothesis of $N_t' \gg n_t'$ is removed, still in the case of no charge detrapping, (2.7) becomes

$$\frac{dF}{dt} = -AF^2 e^{-B/F} \left[\frac{\sigma N_t'}{C_{pp} EOT} - \frac{\sigma}{q} \left(\frac{V_G}{EOT} - F \right) \right] \tag{2.11}$$

Fig. 2.4(a) shows the ΔV_T transients calculated by the numerical integration of (2.11) for different N_t' values, maintaining the $\sigma \times N_t'$ product constant, in order to maintain the same trapping at the beginning of the transient. The reduction of N_t' from 10^{14} to 10^{13} cm^{-2} determines a lowering of the transients as the achieved ΔV_T gets higher;



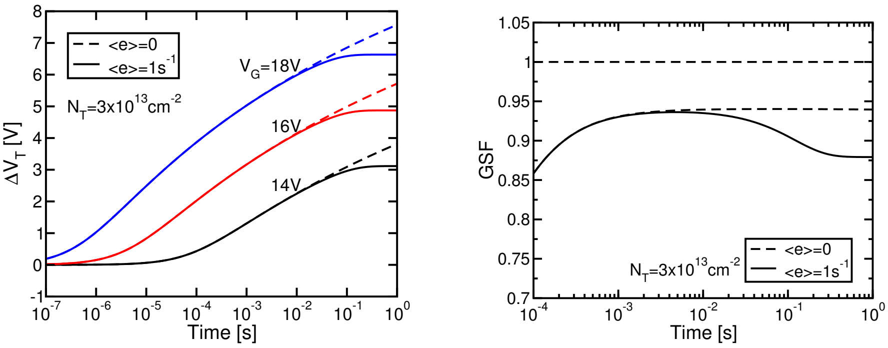
(a) (b)

Figure 2.4: Numerical integration of (2.11) for different N_t' values: (a) ΔV_T and (b) GSF transients.

this comes from the smaller number of empty traps available [62]. The corresponding GSF is shown in Fig. 2.4(b): it is evident that the maximum value for the lower N_t' case is reached for $\sim$ 100 μs and is slightly lower than 0.7. For $N_t' = 3 \times 10^{13}$, the GSF rapidly drops to zero below 1 ms, as the programming time is not long enough to make all the transients lose their dependence on the initial F_i for all the curves, and a $\Delta V_T \approx 0$ V is maintained at $V_G = 14$ V; for longer times the curve converges to about 0.95. Finally, the case with the higher trap density , i.e. with $N_t' = 10^{14}$, can be approximated with the case of infinite traps, presented in the previous section: a convergent behavior of the GSF curve toward one for increasing programming times is obtained; also in this case, for times lower than about 1 ms the GSF drops to zero.

Effect of non-zero detrapping

Fig. 2.5(a) shows the ΔV_T transients calculated by (2.7) when a constant electron detrapping rate $\langle e \rangle = 1$ s^{-1} is assumed, with $N_t' = 10^{14}$ cm^{-2}. Electron detrapping makes the transients saturate at a maximum ΔV_T level which is not determined by the filling of all the available traps but by an equal rate of trapping and detrapping processes. As a consequence, the saturation level depends on the programming voltage, while the time at which saturation occurs is barely affected by V_G. The effect of electron detrapping on the GSF is shown in Fig. 2.5(b): for short programming times, the curves for $\langle e \rangle \neq 0$ coincide with those for $\langle e \rangle = 0$, while they slightly drop to a lower value, determined by the separation of the maximum ΔV_T levels for the different V_G, when saturation occurs.



(a) (b)

Figure 2.5: Numerical integration of (2.7) for different detrapping coefficients: (a) ΔV_T and (b) GSF transients.

2.2.2 Physics based numerical model

The analysis of the program operation presented in the previous section allowed a first-order evaluation of the dependence of the programming ΔV_T transients on the main physical and device parameters of SONOS memories. A reduced gate control on the threshold-voltage shift achieved at a fixed programming time was shown to appear as a result of the finite number of traps in the nitride layer. This reduced gate control was quantified by means of the GSF, representing a fundamental parameter for the determination of the achievable programming performances of the SONOS technology. A value of this parameter lower than one represents a drawback of nitride with respect to floating-gate storage, critically limiting the possibility to scale the programming voltages. This analytical model is then useful for a first order evaluation, but cannot be used as a main tool to predict the real transient in CT memories, as the trapping/detrapping dynamics are approximated, considering all the traps in the middle of the nitride and a fixed emission coefficient. In order to overcome these problems a more accurate numerical model, that self-consistently solves the Poisson, continuity and trapping equations in the nitride layer using a drift-diffusion formalism, has been realized.

Numerical model

The equations implemented for modeling the V_T transients in charge trapping devices are a modification of the Arnett's system [58], where the contribution of holes in presence of amphoteric traps was added [63]. This results in a highly coupled system:

$$\begin{cases} \dfrac{\partial^2\psi(x,t)}{\partial x^2} = \dfrac{q}{\varepsilon_{sin}}\left[n_c(x,t) - p_v(x,t) + n_t(x,t) - p_t(x,t)\right] & \text{(2.12a)} \\ \dfrac{\partial n_c(x,t)}{\partial t} = \dfrac{1}{q}\dfrac{\partial J_n(x,t)}{\partial x} - R_n & \text{(2.12b)} \\ \dfrac{\partial p_v(x,t)}{\partial t} = -\dfrac{1}{q}\dfrac{\partial J_p(x,t)}{\partial x} - R_p & \text{(2.12c)} \\ \dfrac{\partial n_t(x,t)}{\partial t} = \dfrac{J_n(x,t)}{q}\sigma_n\left[N_t - n_t(x,t) - p_t(x,t)\right] + & \\ \qquad - e_n(x,t)n_t(x,t) - \dfrac{J_p(x,t)}{q}\sigma_{rn}n_t(x,t) & \text{(2.12d)} \\ \dfrac{\partial p_t(x,t)}{\partial t} = \dfrac{J_p(x,t)}{q}\sigma_p\left[N_t - p_t(x,t) - n_t(x,t)\right] + & \\ \qquad - e_p(x,t)p_t(x,t) - \dfrac{J_n(x,t)}{q}\sigma_{rp}p_t(x,t) & \text{(2.12e)} \end{cases}$$

In the previous system, N_t, n_t and p_t are the volume densities of traps and of trapped electrons and holes in the nitride, n_c and p_v are the free electron and hole volume densities in the nitride conduction and valence bands, e_n and e_p are the total electron and hole emission rates from the traps as resulting from the different physical mechanisms described later. In order to account for the different trapping rates of amphoteric traps in their neutral, positively- and negatively-charged state, four different trapping cross-sections have been used: σ_n and σ_p are the electron and hole capture cross-sections for neutral traps, while σ_{rn} and σ_{rp} are their counterparts in the case of recombination of a negative- or positive-charged traps, respectively. Finally, R_n and R_p (units: cm^{-3}s^{-1}) describe the net electron and hole recombination rates due to carrier trapping/detrapping in presence of the electron and hole currents J_n and J_p that replace the classical SRH terms in a general semiconductor:

$$\begin{cases} R_n = \dfrac{\partial n_t(x,t)}{\partial t} + \dfrac{J_n(x,t)}{q}\sigma_{rp}p_t(x,t) + \dfrac{J_p(x,t)}{q}\sigma_{rn}n_t(x,t) & \text{(2.13a)} \\ R_p = \dfrac{\partial p_t(x,t)}{\partial t} + \dfrac{J_p(x,t)}{q}\sigma_{rn}n_t(x,t) + \dfrac{J_n(x,t)}{q}\sigma_{rp}p_t(x,t) & \text{(2.13b)} \end{cases}$$

The system (2.12) includes the Poisson equation in the nitride, with electrostatic contributions of both free and trapped charges, and the continuity equations for electrons and holes in the conduction and valence bands. The last two equations of (2.12) describe electron and hole trapping and detrapping in presence of amphoteric traps. Note that, due to the amphoteric assumption, the trap density N_t is unique for both carriers (i.e. both carriers interact with the same physical center). In addition to these equations, the hole and electron currents are calculated with the standard drift-diffusion formalism:

$$\begin{cases} J_n(x,t) = qn_c(x,t)\mu_n F(x,t) + qD_n \dfrac{\partial n_c(x,t)}{\partial x} & \text{(2.14a)} \\ J_p(x,t) = qp_v(x,t)\mu_p F(x,t) - qD_p \dfrac{\partial p_v(x,t)}{\partial x} & \text{(2.14b)} \end{cases}$$

where μ_n and μ_p are the electron and hole mobilities and D_n and D_p are the diffusion coefficients calculated by means of the Einstein relation ($D/\mu = k_B T/q$), being k_B the Boltzmann constant and T the absolute temperature. Due to the high electric fields a modified Einstein relation should be used for the mobility-diffusion ratio (e.g. Arora relation [64] and references therein). However, due to the negligible impact of diffusion process at high fields, this correction can be neglected.

The carrier emissivity phenomena must be carefully modelled due to different ways how a carrier can be emitted from a trap and the unknown relative weight of the different phenomena. The emission coefficient e_n can be written in the following way (and similarly for e_p):

$$e_n = \nu_{th} P_{th} + \nu_{tb} P_{tb}, \tag{2.15}$$

where P_{th} is the thermal emission probability, P_{tb} is the trap-to-band tunneling probability (calculated by means of the Wentzel-Kramers-Brillouin (WKB) equation [65]) and ν_{th} and ν_{tb} are the corresponding attempt-to-escape frequencies. The thermal emission coefficient has been calculated according to [66]:

$$P_{th} = e^{-(E_T - \beta_{pf}\sqrt{F})/k_B T}, \tag{2.16}$$

where E_T is the trap energy position with respect to the nitride conduction band and β_{pf} is a constant value dependent on the material properties (for the silicon nitride, $\beta_{pf} \approx 3.8 \cdot 10^{-4}\ \sqrt{\text{Vcm}}$). As final details over the numerical model, the traps in the nitride have a gaussian energy distribution [67] with average trap depth $E_T = 1.5$ eV and standard deviation $\sigma_E = 0.2$ eV, and the injected electron flux has a finite relaxation rate $\lambda = 3.1$ eV/nm [34], so that only electrons on the edge of the conduction band can be trapped.

In order to solve numerically system (2.12) and determine the V_T transient of the memory device under program, erase or retention conditions, the free and trapped electron and hole concentrations must be calculated as a function of x and t. To this aim, equations have been discretized in the spatial and in the time domains using the standard central-difference scheme for the Poisson equation and the Scharfetter-Gummel approximation for the current density equations [68].

In order to test the performance of the model presented in this section, it has been tested with a test case coming from published data. The parameters used for the fitting are the ones of the traps: the trap density N_t, the four capture cross sections σ and the emission escape frequencies ν, as they depend on the process used to deposit the silicon nitride layer.

Numerical treatment

In order to numerically solve system (2.12) and determine the V_T transient of the memory device under program, erase or retention conditions, the free and trapped electron and hole concentrations must be calculated as a function of x and t. To this aim, equations have been discretized in the spatial (index i, space step h_i) and in the time (index k, time step τ_k) domains using the standard central-difference scheme for the Poisson equation and the Scharfetter-Gummel approximation for the current density equations [68]. Space discretization allows the use of a non-uniform mesh to refine, for example, the current injection boundaries during program and erase simulations. Referring to electrons (holes are treated in a similar way), the following system results, where $B(y) = y/(e^y - 1)$ is the Bernoulli function and $B_p = B(\mu/D(\psi_{i+1} - \psi_i))$ and $B_n = B(\mu/D(\psi_i - \psi_{i+1}))$:

$$\begin{cases} \dfrac{\psi^k_{i-1} - 2\psi^k_i + \psi^k_{i+1}}{h_i^2} = q\dfrac{n^k_{c,i} + n^k_{t,i} - p^k_{v,i} - p^k_{t,i}}{\varepsilon_{sin}} \\ \dfrac{n^k_{c,i} - n^{k-1}_{c,i}}{\tau} = \dfrac{D^k_{n,i}}{h_i^2}\left[n^k_{c,i-1}B_{n,i-1} + n^k_{c,i+1}B_{p,i+1} - n^k_{c,i}(B_{p,i} + B_{n,i})\right] - R^k_i \\ \dfrac{n^k_{t,i} - n^{k-1}_{t,i}}{\tau} = \dfrac{J^k_{n,i}}{q}\sigma_n\left[N_{t,i} - n^k_{t,i} - p^k_{t,i}\right] - e^k_{n,i}n^k_{t,i} - \dfrac{J^k_{p,i}}{q}\sigma_{rn}n^k_{t,i} \end{cases}$$

The trapping dynamics is hidden in the term R^k_i allowing the use of tridiagonal solver for both the Poisson and continuity equations. The time solutions stability is obtained using implicit numerical methods: backward-Euler scheme for the continuity equations and two-step backward-differential-formula for the trapping equations.

Fig. 2.6 shows the adopted Gummel-map. The initial guess is used to solve the Poisson and continuity equations, then solving the coupled electron/hole trapping equations until convergence. The convergence of the full system including Poisson, continuity and trapping equations is then checked: in the case convergence is not reached, the calculated functions are used as the new guess for the solution of the Poisson and continuity equations. The residuals of the main cycle for each equation are shown in Fig. 2.7: Poisson equation rapidly converges, while trapping equations require a higher number of iterations, though eventually converging even for strongly-coupled systems.

Program/Erase performance

In order to simulate the programming transient, appropriate boundary conditions should be given for solving the system (2.12). The origin of the x axis is set at the nitride/blocking oxide interface, so that the tunnel oxide/nitride interface is at $x = t_{sin}$; the electron tunneling currents J_e^{tun} flowing from the substrate to the nitride, due to the finite relaxation rate, forces a boundary condition at $x = x_r$, rather than $x = t_{sin}$, where x_r is the coordinate at which the incoming flux relaxes on the bottom of the conduction band. The boundary conditions for the programming transient are:

$$\begin{cases} J_n(x_r,t) = J_e^{tun} + qn_c(x_r,t)\mu_n F(x_r,t) + qD_n \dfrac{\partial n_c(x_r,t)}{\partial x} \\ J_p(t_{sin},t) = 0 \\ J_n(0,t) = qn_c(0,t)\mu_n F(0,t) TC_e^{bo} \\ J_p(0,t) = 0 \\ n_c(x,0) = p_v(x,0) = 0 \\ n_t(x,0) = p_t(x,0) = 0, \end{cases}$$

where TC_e^{bo} is the tunneling transmission coefficient of conduction-band electrons through the top dielectric, calculated with the WKB approximation. The boundary condition applied to the hole current is strictly valid when a metal gate is used, but represents also a good approximation for SONOS-type devices, having a rather large top oxide thickness (e.g. $5-8$ nm) and therefore negligible hole injection from the gate to the nitride.

For the erase transient, the bottom interface is characterised by both hole injection from the substrate (J_h^{tun}) and electron emission from traps directly to the substrate or to the conduction band; electron injection (J_e^{bo}) and possible hole emission must be accounted for the gate interface.

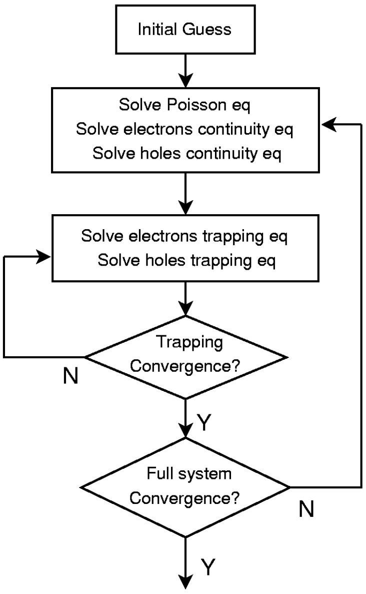


Figure 2.6: Modified Gummel map used: the inital guess is used to solve Poisson, electron and hole continuity equation; after that, the coupled trapping equations are brought to convergence and the full systems convergence is checked.

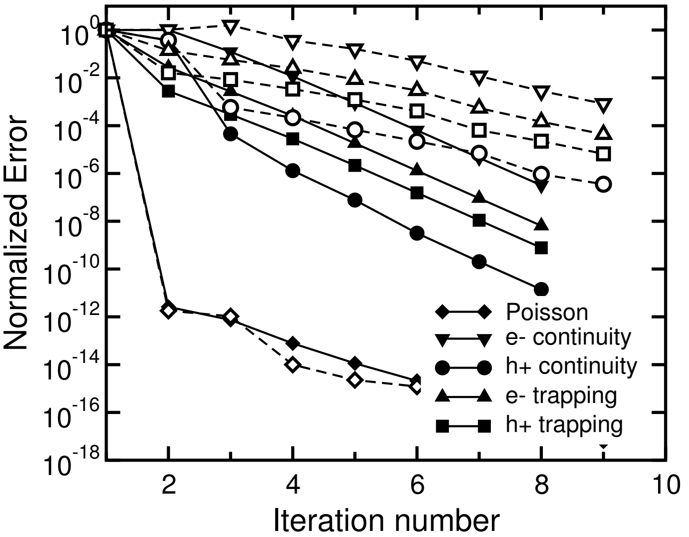


Figure 2.7: Residual error of the different equations as a function of the iteration number for a strongly coupled system (empty symbols) and for weakly coupled one (full symbols).

The relaxation position for electrons injected from the gate is at $x = x_r$, while zero-length thermalization was assumed for holes. The boundary conditions become:

$$\begin{cases} J_n(x_r,t) = J_e^{bo} + qn_c(x_r,t)\mu_n F(x_r,t) + qD_n \dfrac{\partial n_c(x_l,t)}{\partial x} \\ J_p(0,t) = qp_v(0,t)\mu_p F(0,t) TC_h^{bo} \\ J_n(t_{sin},t) = qn_c(t_{sin},t)\mu_n F(t_{sin},t) TC_e^{tun} \\ J_p(t_{sin},t) = J_h^{tun} + qp_v(t_{sin},t)\mu_p F(t_{sin},t) + qD_p \dfrac{\partial p_v(t_{sin},t)}{\partial x} \\ n_c(x,0) = p_v(x,0) = 0 \\ n_t(x,0) = n_0(x,0) \\ p_t(x,0) = p_0(x,0). \end{cases}$$

Here, TC_h^{bo} and TC_e^{tun} are the tunneling transmission coefficients through the top and bottom dielectrics for valence band holes and conduction band electrons, respectively. Moreover, $n_0(x,0)$ and $p_0(x,0)$ represent the trapped electron and hole concentrations in the nitride at the beginning of the erase as obtained at the end of programming.

Fig. 2.8 shows the comparison between our numerical results for P/E transients and the experimental data extracted from [69], referring to a SONOS device with a 3/5/4.5 nm ONO stack. The trap spatial profile is assumed to pile-up toward the top oxide as proposed in [69] for the silicon-rich nitride. Calculated results were obtained with a bulk trap density $N_t = 4 \times 10^{19}$ cm^{-3}. Other parameters are $\sigma_n = 10^{-15}$ cm^2, $\sigma_{rn} = 5 \times 10^{-15}$ cm^{-2} and $\sigma_p = \sigma_{rp} = 10^{-19}$ cm^{-2}. The obtained values, quite high for electron trapping and small for hole trapping, are consistent with the range obtained for coulombic centers resulting in donor traps [70, 71]. Note that the simulator is able to reproduce the saturating behavior during program (due to the balance between trapping and emission of electrons in the traps) and in erase, due to the leakage current through the top oxide: this is a typical behavior of the SONOS known as erase saturation. Fig. 2.9(a) shows the free and trapped electron concentration in the nitride of program at a fixed time $t = 1$ ms. The free electron concentration slightly decreases for deeper positions in the nitride due to the capture mechanism, while the trapped electron profile is correlated with the U-shape spatial distribution of the trap density which comes from the fact that, at the interfaces with the tunnel and the blocking oxides, the change in material gives rise to a higher defect density and, in turn, to higher trap density. We have

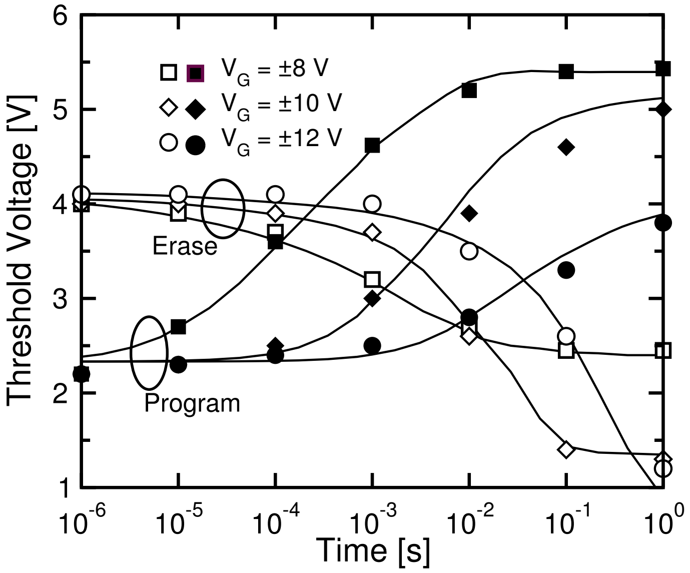


Figure 2.8: Fitting of SONOS (3/5/4.5 nm of ONO stack) program/erase experimental transient (data taken from [69]) with the presented model: line is the simulation while symbols are the experimental data.

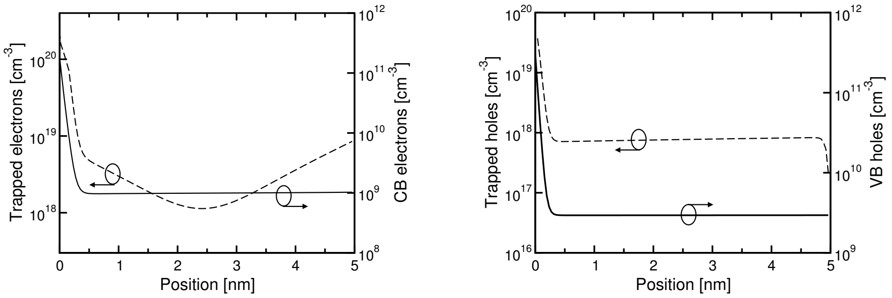


Figure 2.9: Simulated concentration of carriers in the nitride traps (left axis) and in the nitride band (right axis) of a SONOS device (3/5/4.5 nm of ONO stack) after 1 ms: (a) electrons during program at $V_G = +10$ V and (b) holes during erase at $V_G = -10$ V.

assumed that the defects are not only at the interface, but extend also in the bulk nitride with a tail, from which the U shape [34, 51]. At the interface with the blocking oxide, there is also a larger trapping rate, due to the larger free electron concentration in this region determined by the output barrier [72]. The higher trap density together with the large free electron concentration in the region, gives rise to the peak in the trapped charge distribution near the blocking oxide interface. A similar result for holes is shown in Fig. 2.9(b) for the erase condition. The slight decrease of the hole concentration in the valence band is due to the low values of the trapping and recombination cross-section σ_h and σ_{rp}. The effect of the barrier for hole tunneling toward the gate, creating an accumulation layer, is also evident. The V_T variation during erase is caused by hole trapping and by electron loss from the nitride conduction band and traps to the substrate.

2.3 Impact of alumina introduction in TANOS cells

TANOS memories are considered a viable solution to overcome the erase–data retention compromise of SONOS stack [33]. One of the key concepts is the use of a high-k dielectric (usually Al_2O_3) on top of the nitride trapping layer, to both increase the coupling efficiency between the trapping layer and the metal gate and reduce the electron injection from the gate during erase. Notwithstanding the clear benefits in terms of program/erase (P/E) speed and threshold voltage window brought by the introduction of the alumina layer, it has usually been considered as an ideal dielectric, neglecting trapping properties and all the other reliability constraints, as most of the research activity has been focused on the nitride layer. In particular, the alumina trapping properties have been studied in only a few works [36, 45–47].

2.3.1 Experimental evidences

Alumina has been characterized on large-area capacitors with TaN metal gate and Al_2O_3 directly grown by ALD on a 1 nm chemical SiO2 layer on a p-type substrate. An n-type ring provides the minority carriers for inversion. A 1100°C inert anneal was used for alumina crystallization. Fig. 2.10 shows the V_{FB} evolution during constant-voltage stress, applying to all samples the same initial electric field $F_{Al_2O_3} = 4.6$ and 4.9 MV/cm for positive and negative stress, respectively. A large shift in V_{FB} (more than 1 V) can be observed in Fig. 2.10 for the 15 nm

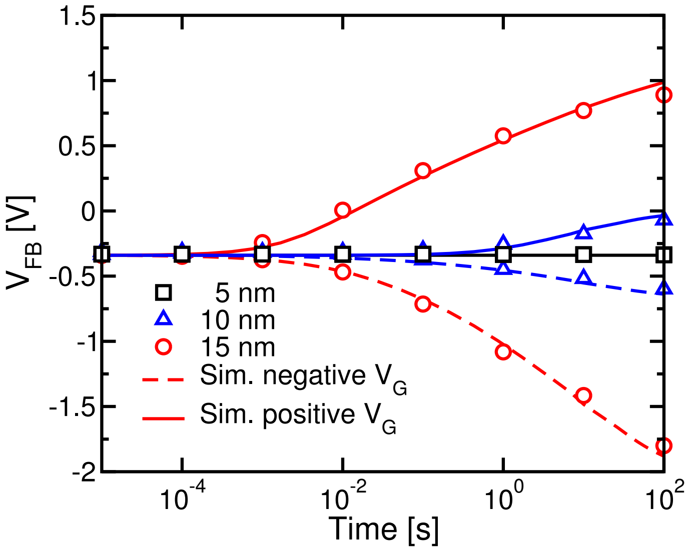
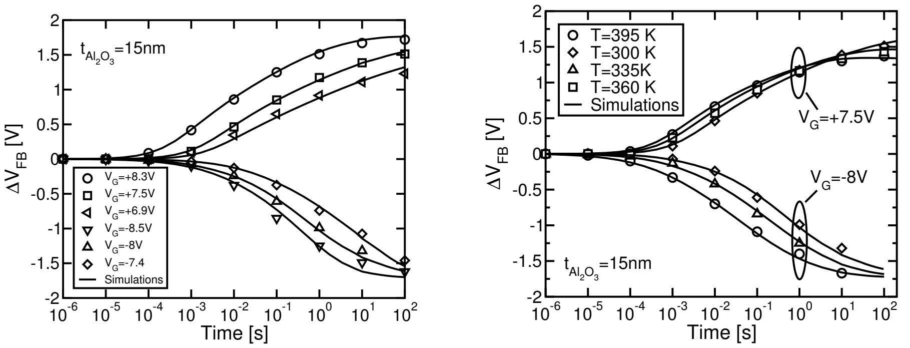


Figure 2.10: V_{FB} transients for different TAOS capacitors during positive and negative V_G stress at the same field and $T = 25°\text{C}$.

stack, as reported also in [46], while the V_{FB} shift decreases when reducing $t_{Al_2O_3}$ until no visible variation appears in the 5 nm case up to 100 s stress time. This reveals an increase of the bulk trap density as the Al_2O_3 thickness increases beyond 5 nm. The behavior of the 15 nm sample is further investigated in Fig. 2.11(a), showing the room temperature V_{FB} variation (ΔV_{FB}) during positive and negative stress of 100 s at different gate voltages, starting from the same $V_{FB} \approx -0.1$ V. Note that different behaviors can be observed for the two polarities: for positive stress, a saturation level is neatly reached only at the highest electric field, whereas for lower values a saturation is not reached within the observation time of 100 s. In the case of negative stress, instead, the ΔV_{FB} curves seem to converge toward the same level, independently of the electric field strength. A different behavior for program and erase can also be seen in Fig. 2.11(b), where the ΔV_{FB} transients are shown as a function of temperature. A negligible variation with temperature can be seen for positive V_G, where only the saturation value changes slightly, while results for the negative stress polarity feature a much stronger temperature dependence. Figs. 2.10 and 2.11 make clear that both positive and negative charges can be stored in the Al_2O_3 layer. While negative charges are usually interpreted as electrons being trapped into acceptor-like traps, positive charges could arise from both hole injection from the substrate and trapping into acceptor-like



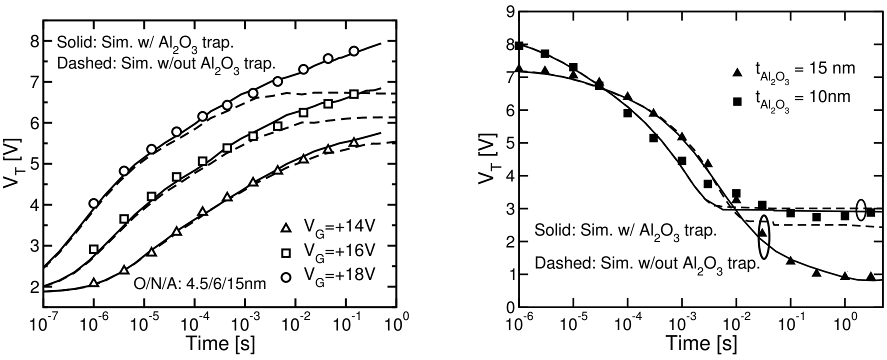
(a)

(b)

Figure 2.11: ΔV_{FB} transients on a TAOS capacitor with $t_{Al_2O_3} = 15$ nm: (a) stresses at different V_G at $T = 25°$C and (b) stresses at different temperatures.

traps or from electron detrapping from donor-like trap sites. To discriminate between these phenomena, the model previously introduced has been used, slightly modified in order to simulate the new structure. Numerical simulations confirmed that hole injection plays a negligible role during erase due to the high injection barrier and long injection distance, and can thus be neglected. To reproduce both program and erase transients, acceptor-like and donor-like traps must be considered in the alumina, having densities N_{TA} and N_{TD} respectively. From the transients fitting $N_{TA} = 1.5 \times 10^{19}$ and 5×10^{18} cm^{-3} have been obtained for the 15 and 10 nm samples respectively, with a negligible density in the 5 nm case [47]. The donor-like traps were assumed to be interface traps located in correspondence of the alumina/SiO$_2$ interface, with density $N_{TD} = 6$ and 4×10^{12} cm^{-2} for the 15 and 10 nm respectively. Trap energy was extracted to be $E_T = 1.8$ eV below the alumina conduction band and the electron capture cross-section is $\sigma = 1.5 \times 10^{-14}$ cm^2.

Simulation results are shown in Figs. 2.10 and 2.11 (solid lines), featuring a good agreement with data. In particular, the program and erase variations with temperature shown in Fig. 2.11(b) are correctly accounted for. The difference in the behavior is due to the strong dependence on temperature of the electron emission from traps, as opposed to a program transient which is ruled by electron capture. The latter is not dependent on temperature except for the final part of the transient, where a balance between capture and emission is reached. This explains the slightly-different program saturation levels found with increasing temperature as well as the stronger dependence of erase, and provides further support to the existence of donor-like traps, as opposed



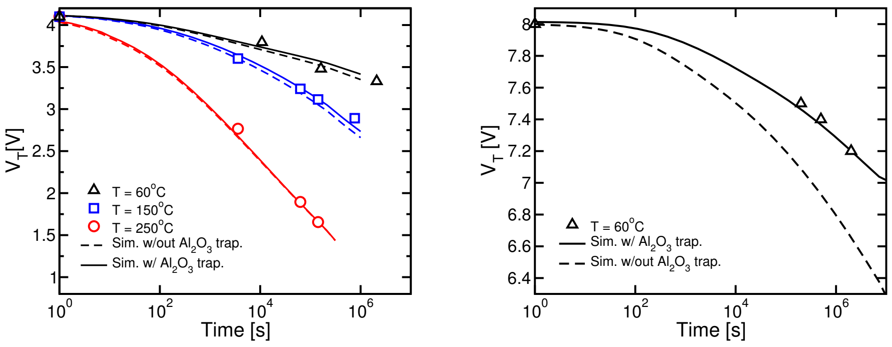
(a) (b)

Figure 2.12: Effect of alumina trapping on (a) programming transients and (b) erasing transients for TANOS devices. Simulation results with and without alumina trapping are reported.

to hole injection and capture.

2.3.2 Impact on TANOS reliability

To directly evaluate the impact of the previous effects on TANOS device performances, we have characterized TANOS cells with 4.5 nm bottom oxide, 6 nm nitride, and different Al_2O_3 layer thicknesses ($t_{Al_2O_3}$). To account for the trapping phenomena in alumina, the numerical model for CT simulations presented in Sect. 2.2.2 was extended to include the trapping/detrapping dynamics for this layer, too. Fig. 2.12(a) shows a comparison between experimental data and simulations for the programming transients of samples with $t_{Al_2O_3} = 15$ nm. Simulation results are shown with and without alumina trapping. Note that trapping in the Al_2O_3 allows to explain the non-saturating behavior of the programming transients, which is typically observed in TANOS stacks (as opposed to SONOS devices): when alumina trapping is included, V_T continues to increase for large programming times, while simulations predict a saturation when accounting only for the trapping in the nitride layer, in particular when high V_G is applied. The previously-extracted parameters were used for the alumina, with the exception of the donor-like traps, which have been here neglected as a consequence of the different interface (alumina on nitride rather than over oxide). For the nitride, the following parameters were used: $N_T = 6 \times 10^{19}$ cm^{-3} and $\sigma_n = 5.5 \times 10^{-15}$ cm^2. Fig. 2.12(b) reveals that Al_2O_3 trapping should also be invoked in the erase operation, in particular for explaining the



(a) (b)

Figure 2.13: Retention transients for TANOS devices at (a) different temperatures and (b) 60°C at higher program level. Simulation results with or without alumina trapping are also included.

position of the saturation level: given the trapping characteristics shown in Fig. 2.10, this effect is stronger for thicker Al_2O_3 layers.

Finally the impact of the alumina trapping properties on retention were investigated. Fig. 2.13(a) shows that the impact of alumina trapping on data retention are quite small, at least for small programmed levels: this is in accordance with the fact that the trapping in alumina becomes important only for high programmed levels (cf. Fig. 2.12(a)). The effect becomes important for large initial ΔV_T, as shown in Fig. 2.13(b), with more charge trapped in the blocking oxide; misleading results can be obtained for data retention if Al_2O_3 trapping is neglected: deep traps in the alumina lead to a slower charge-loss transient, in better agreement with experimental data, which cannot be reproduced by the model if alumina is considered an ideal dielectric and all the V_T shift is ascribed only to charge trapped in the nitride.

2.4 Conclusions

This chapter dealt with the modeling of program, erase and retentions operations in charge trap memories. A first order analytical model allowed to investigate the fundamental results obtained for this kind of cells and to explain the main differences with the FG. The model was then extended to a more complete numerical one that was tested against data obtained on SONOS devices. Then, in order to explain some peculiarities of TANOS devices, trapping in the alumina layer was investigated on specific samples and the results were integrated into the numerical model, allowing to reproduce experimental data on TANOS devices, not explainable by only using trapping in the nitride layer. The developed model can be used to simulate large area devices, and predict the intrinsic performance and, potentially, limitations of charge trap devices.

Chapter 3

Incremental Step Pulse Programming behavior

A detailed investigation of the Incremental Step Pulse Programming (ISPP) dynamics of charge-trap memory capacitors is presented. Differently from the floating-gate case, results on nitride-based memories show that the flat-band increase per step does not equal the step amplitude of the gate staircase, decreasing, moreover, as programming proceeds. As a consequence, the electric field and tunneling current through the bottom oxide are shown to largely increase. Then, using a 3D simulator, ultra scaled devices are also analyzed, revealing a further decrease of the programming efficiency.

3.1 Introduction

The incremental step pulse programming (ISPP) algorithm represents a mandatory programming scheme for multi-level deca-nanometer NAND Flash memories, allowing to obtain very narrow threshold voltage (V_T) distributions [16, 73]. With this algorithm, accurate placement of cell V_T by Fowler-Nordheim tunneling is achieved by applying short pulses of equal duration τ_s and increasing amplitude to the gate, keeping the channel grounded. In the case of floating-gate devices, a constant increase V_s of the pulse amplitude gives the possibility to reach a stationary condition where the V_T variation per step ($\Delta V_{T,s}$) equals V_s [73,74], which can be exploited to tighten the V_T distribution by means of verify operations after each programming pulse [16,75]. Different results are, instead, reported for nitride-based memories, displaying a V_T variation per step that is usually lower than V_s [76,77]. A comprehensive understanding of this behavior and of its impact on device performance represents a mandatory issue for the development of multi-level SONOS and TANOS memories.

3.2 ISPP analysis on large area capacitors

In order to understand the main differences between the CT cells and the FG ones, a detailed analysis of the ISPP dynamics of charge-trap memory capacitors is carried out, investigating the tunnel oxide electric field (F_{tun}) and the tunneling current (J_{tun}) evolution during programming. Fig. 3.1(a) shows the V_{FB} transient as a function of the gate bias V_G during ISPP with $V_s = 500$ mV and $\tau_s = 10$ μs, for a SONOS capacitor featuring p-type substrate and poly-gate, n^+-ring, thickness of the bottom oxide $t_{bot} = 4$ nm, of the stoichiometric nitride $t_N = 6$ nm and of the top oxide $t_{top} = 5$ nm. When V_G rises above 12 V, a significant increase of V_{FB} starts taking place as a consequence of electron storage in the nitride, with $\Delta V_{FB,s}$ remaining nearly equal to the 80% of V_s up to $V_G = 17$ V. However, for higher V_G a strong reduction of $\Delta V_{FB,s}$ appears, resulting in a clear sub-linear V_{FB} growth.

In order to study in more detail the programming dynamics, F_{tun} and the average trapped current density (J_{trap}) were calculated during each programming step of the ISPP staircase from the V_{FB} transient according to:

$$F_{tun} = \frac{V_G - \phi_i - (V_{FB} - V_{FB,0})}{EOT} \tag{3.1}$$

$$J_{trap} = \frac{C_{pp} \cdot \Delta V_{FB,s}}{\tau_s} \tag{3.2}$$

where $\phi_i \simeq 1.3$ V takes into account the voltage drop on the substrate and on the gate during programming (assumed constant for simplicity), $V_{FB,0}$ is the neutral device flat-band voltage and $EOT = t_{tun} + t_{sin}\varepsilon_{ox}/\varepsilon_{sin} + t_{bo}$ is the equivalent oxide thickness of the gate stack (ε_{ox} and ε_{sin} are the oxide and nitride dielectric constants, respectively). C_{pp} represents the capacitance (per unit area) from the position of the stored-charge centroid in the nitride (assumed in the middle of the nitride) to the gate. Results are reported in Fig. 3.1(b) and in Fig. 3.2,

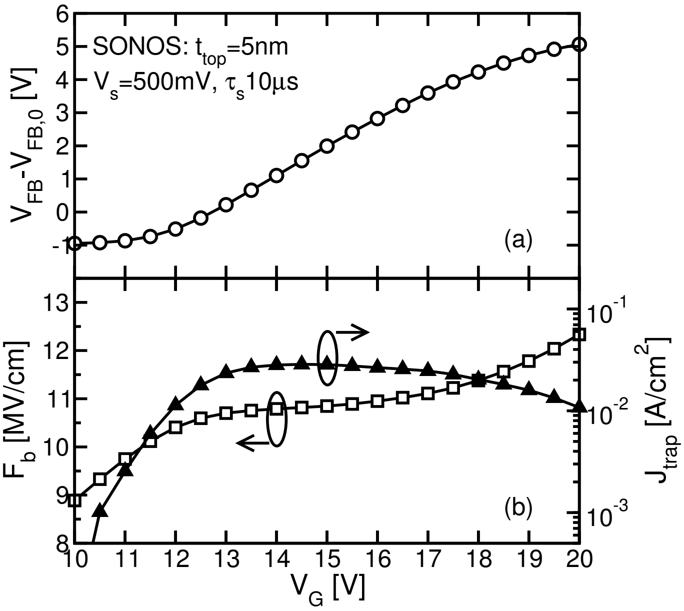


Figure 3.1: V_{FB} transient during ISPP (V_s = 500 mV, τ_s = 10 μs) of a SONOS capacitor (a) and calculated F_{tun} and J_{trap} (b). $V_{FB,0}$ is the neutral (no charge in the nitride) flat-band voltage of the device.

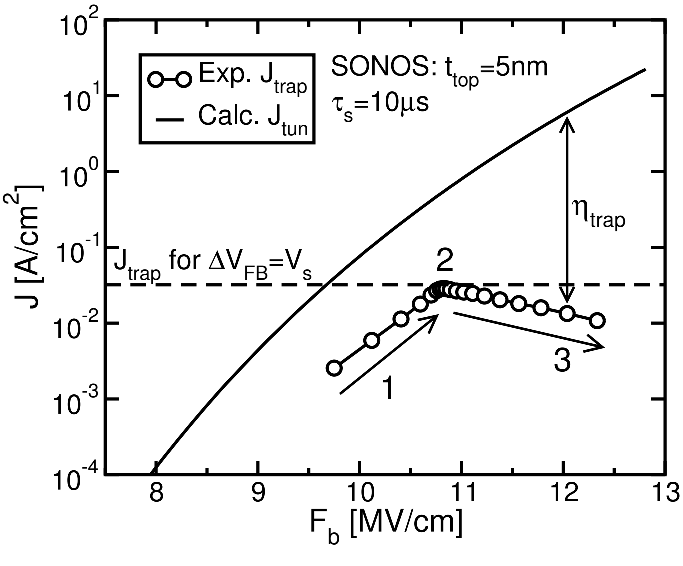


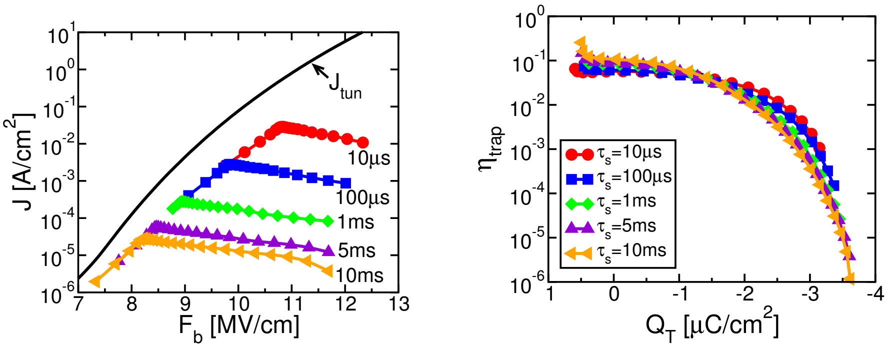
Figure 3.2: Experimentally extracted J_{trap} as a function of F_{tun} (symbols) from the V_{FB} transient of Fig. 3.1. The calculated J_{tun} through the bottom oxide is also shown (solid line).

where J_{trap} is directly shown as a function of F_{tun}. Three programming phases clearly appear during the ISPP transient, as highlighted in Fig. 3.2. In phase 1, corresponding to the V_G increase from 10 to 12 V in Fig. 3.1, F_{tun} and J_{trap} rapidly increase, allowing J_{trap} to become nearly equal to $\overline{J_{trap}} = C_{pp}\Delta V_s/\tau_s = 32$ mA/cm^2, representing the stationary trapping current allowing the condition $\Delta V_{FB,s} = V_s$ to hold during ISPP. When this condition is reached, F_{tun} from (3.1) remains nearly constant and a nearly stationary working point for ISPP is reached in the F_{tun}–J_{trap} diagram (phase 2, V_G from 12 to 17 V), similarly to what is usually reported for floating-gate devices [16, 74]. However, as programming proceeds J_{trap} starts decreasing (phase 3, V_G higher than 17 V), resulting into a growth of F_{tun} due to the V_G increase from one step to the next. Note that the electric field increase given by the V_G growth in this phase is not enough to increase J_{trap} to maintain a stationary condition during ISPP. An additional strong difference between the ISPP dynamics of FG and SONOS devices can be derived from the comparison of the J_{trap} characteristics in Fig. 3.2 with the calculated [78] J_{tun} through the bottom oxide (solid line). In all the ISPP phases of Fig. 3.2, J_{trap} is always lower than J_{tun}, revealing a trapping efficiency $\eta_{trap} = J_{trap}/J_{tun}$ lower than 1, as commonly reported for charge-trap devices [79].

3.2.1 Investigation of the trapping efficiency

The decrease of J_{trap} and the consequent increase of F_{tun} during phase 3 of Fig. 3.2 represents an important feature of the ISPP dynamics of charge-trap memories. In fact, besides being a drawback for the programming speed and efficiency, the increase of F_{tun} and of the tunneling current J_{tun} flowing through the bottom oxide is a potential drawback also for device reliability. In order to understand the physical reasons behind this effect, we investigated η_{trap} for different τ_s on the previously considered SONOS capacitors. Fig. 3.3(a) shows the experimentally extracted J_{trap} vs. F_{tun} characteristics for τ_s ranging from 10 μs to 10 ms.

The overlap of the growing part of the curves (phase 1) associated to different τ_s confirms the carefulness of our analysis, revealing that J_{trap} depends in this phase only on F_{tun} and so on J_{tun}. The curves reach, then, a different $\overline{J_{trap}}$ during phase 2 as a result of the different τ_s and, finally, decrease during phase 3 displaying, at the same F_{tun}, a different distance from the J_{tun} curve. This reveals that during phase 3, η_{trap} is not driven by F_{tun}, as instead observed during phase 1. To better understand the previous results, Fig. 3.3(b) shows η_{trap} extracted from Fig. 3.3(a) as a function of the calculated charge stored in the silicon



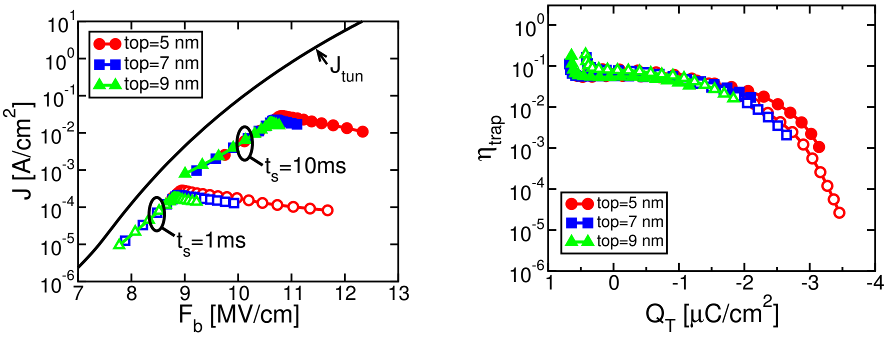
(a) (b)

Figure 3.3: Evaluation of the ISPP transient dependence on τ_s for SONOS devices: (a) experimentally extracted J_{trap} as a function of F_{tun} (computed with (3.2) and (3.1), respectively), and (b) trapping efficiency as a function of the charge stored in the nitride.

nitride layer $Q_T = -(V_{FB} - V_{FB,0}) \cdot C_{pp}$.

At the beginning of the programming transient ($|Q_T| < 2$ μC/cm^2), η_{trap} remains nearly constant and equal to 10%, in agreement with typical trapping efficiencies reported for nitride memories [79]. This corresponds to phase 1 and 2 of the ISPP transients, where J_{trap} is ruled only by F_{tun} and J_{tun}. However, when $|Q_T|$ increases above 2 μC/cm^2, a rapid reduction of η_{trap} on the logarithmic axis clearly appears, corresponding to phase 3 of the ISPP dynamics of Figs. 3.2 and 3.3(a). The good matching of the η_{trap} results for different τ_s even in this regime reveals that the J_{trap} reduction is driven by the amount of charge stored in the nitride and, in turn, by the number of free electron traps available for trapping in the silicon nitride layer.

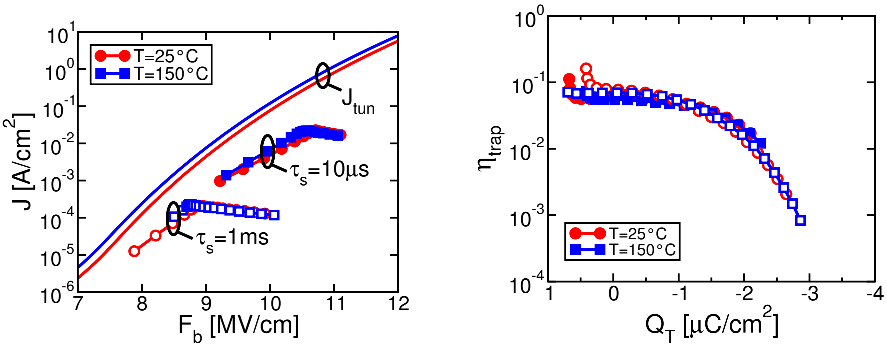
To further check the previous results, Fig. 3.4 shows the J_{trap} vs. F_{tun} characteristics (a) and η_{trap} vs. Q_T (b) for SONOS capacitors with different t_{top} and the same t_{bot} and t_{sin} of the previous devices. Results confirm that during phase 1 and 2, a constant efficiency is obtained, whose value is independent from the stack composition. During this phases, therefore, J_{trap} depends only on F_{tun} and J_{tun}, as highlighted by the overlap of the J_{trap} curves associated to different t_{top}. Moreover, Fig. 3.4(b) shows that the η_{trap} curves overlap also in the high Q_T regime (phase 3 of the ISPP transients), revealing that the efficiency drop does not change when modifying the top oxide layer. This confirms that the decrease of J_{trap} in this regime is mainly a result of the finite number of traps in the nitride and not, for instance, of a significant electron emission from the filled traps when large fields in the top oxide layer are



(a)

(b)

Figure 3.4: J_{trap} vs. F_{tun} characteristics for SONOS capacitors of different t_{top} (a) and corresponding η_{trap} (b).



(a)

(b)

Figure 3.5: Same as in Fig. 3.4, but on the same SONOS stack having $t_{top} = 7$ nm and at different temperatures T.

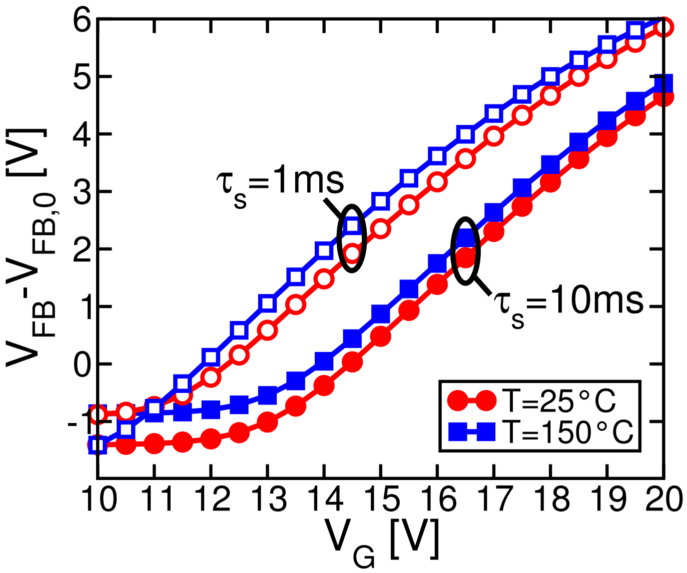
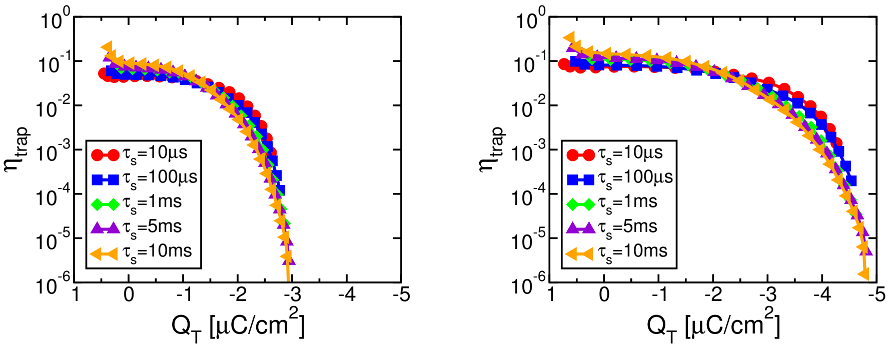


Figure 3.6: V_{FB} transients during ISPP on SONOS capacitors with t_{top} = 7 nm. Results at different temperatures are reported.

reached.

This is further confirmed by the results of Fig. 3.5, referring to ISPP transients at different temperatures T on SONOS capacitors with t_{top} = 7 nm. In this case, the J_{trap} curves do not perfectly overlap in their rising regime, revealing a larger electron storage as T is increased. Note, however, that this is due to the increase of J_{tun} at higher T, as shown in Fig. 3.5(a), and that the resulting η_{trap} is almost constant with T (see Fig. 3.5(b)) , confirming that the trapping dynamics are ruled by a temperature-independent capture cross-section [54]. As a consequence, the V_{FB} transients during ISPP are a little bit faster at higher temperatures, as shown in Fig. 3.6. Finally, note that similar results are found for η_{trap} in Fig. 3.5(b) even in the high Q_T regime, confirming that electron emission from the traps is not responsible for the J_{trap} reduction during phase 3 of the ISPP transient.

The assumption that the charge centroid is in the middle of the nitride layer does not impair the results. In fact Fig. 3.7 shows the results displayed in Fig. 3.3(b) by changing this assumption: Fig. 3.7(a) shows the results if the charge centroid is at the interface with the tunnel oxide, while Fig. 3.7(b) shows the results if the charge centroid is at the interface with the blocking oxide. It is evident that the curves corresponding to different programming times still in good agreement, just slightly



(a) (b)

Figure 3.7: Same as in Fig. 3.3(b), but changing the charge centroid position: at the nitride/tunnel oxide interface (a), and at the nitride/blocking oxide interface (b).

shifted vertically and saturating at different trapped charge levels.

3.2.2 Investigation of TANOS devices

ISPP dynamics on TANOS capacitors have been investigated at last, in order to highlight any possible impact of alumina non-idealities (mainly, trapping and leakage) on the programming transients. The characterized devices have tunnel oxide thickness $t_{tun} = 4.5$ nm, nitride thickness $t_{sin} = 6$ nm and top alumina thickness ranging from 10 to 15 nm. The analysis has been carried on using the η_{trap} results obtained from SONOS capacitors to describe the nitride trapping dynamics in the TANOS devices, solving the following equation [54] for the subsequent steps of the ISPP algorithm:

$$\frac{dV_{FB}}{dt} = \frac{J_{tun}}{C_{pp}} \eta_{trap} \tag{3.3}$$

where η_{trap} is adjusted at the beginning of each step to follow the dependence on Q_T, and in turn V_{FB}, obtained from Fig. 3.3(b). Calculated results are shown in Fig. 3.8, where they are compared with experimental results. A good agreement between data and calculations appears for $t_{top} = 10$ and 12.5 nm in the initial part of the transient, with differences appearing only when very large V_{FB} shifts nearly equal to 6 V are reached, where electron emission through the alumina may become significant. This reveals that thin alumina layers behave as ideal dielectrics on a large part of the ISPP transient, which is still ruled by the electron trapping dynamics in the nitride. However, when a thick alumina of

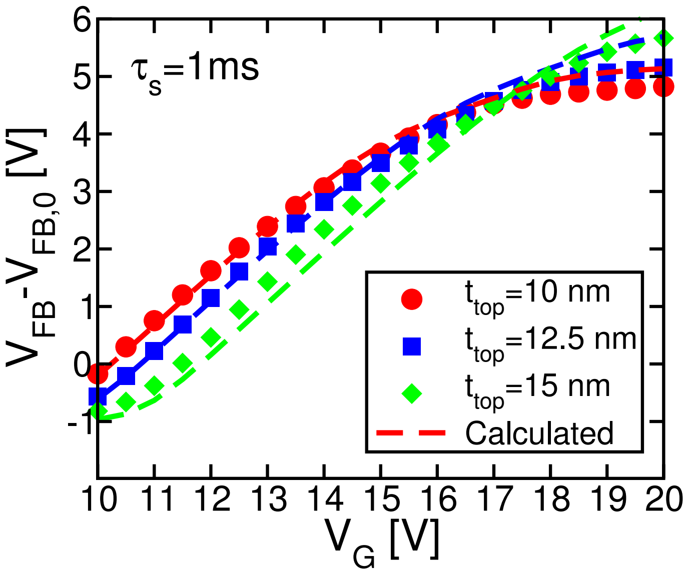


Figure 3.8: Experimental V_{FB} transients during ISPP on TANOS capacitors with different alumina thickness. Calculated results obtained assuming the same trapping efficiency of SONOS capacitors are also shown.

15 nm is used, differences between calculations and experimental results appear since the beginning of the ISPP transient, revealing that alumina trapping is not negligible and impacts the ISPP dynamics. This is in agreement with what was shown in Fig. 2.12(a), where a sample with $t_{top} = 15$ nm was measured and compared with the simulations obtained by only considering the trapping in the nitride layer or by also including the trapping in the alumina layer: it is visible that for times longer than 1 ms the two simulations depart from each other.

3.3 ISPP analysis on ultra-scaled devices

The programming efficiency of charge-trap memories has been shown to largely decrease when device dimensions are reduced to the decananometer scale [76, 77, 80–82]. Referring to the incremental step pulse programming (ISPP) algorithm [73, 83], in fact, significant differences have been shown to appear in the ratio between the threshold-voltage increase per step ($\Delta V_{T,s}$) and the step amplitude (V_s) for large area capacitors and nanoscale cells. While the ratio is only slightly below 1 for the former, at least far from the saturation of the available traps in the storage layer [84], quite lower values in the 0.5–0.65 range are typi-

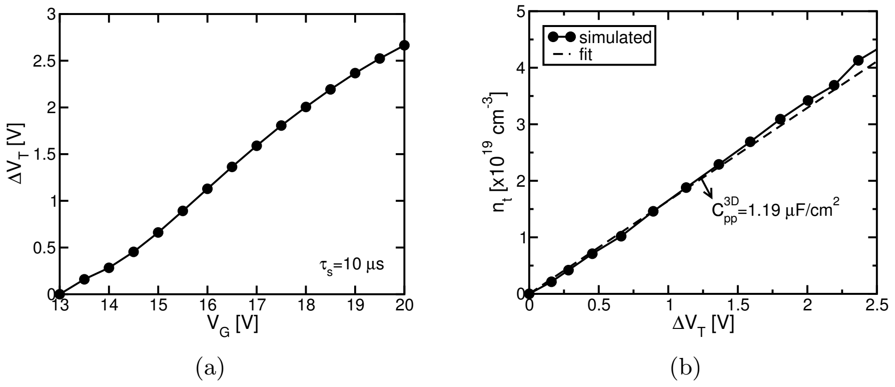


Figure 3.9: Simulated ISPP transient for a 20 nm cell with ONO thicknesses 4/4.5/5 nm: (a) ΔV_T transient and (b) trapped charge density as a function of the V_T shift.

cally reported for the latter [76,80,81]. The decrease of the ISPP slope for small area cells has been clearly correlated not only to cell dimensions but also to cell geometry [77]: the low programming efficiency of nanoscale charge-trap cells mainly results from the low impact of locally stored electrons on V_T due to fringing fields at the cell edges [85].

In deca-nanometer devices the statistical effects, due to the discrete nature of traps in the storage layer and of the electron flow charging it [16,74,85–89], and due to atomistic doping, creating percolation paths in the substrate [90–96]. Nevertheless, understanding the average behavior is still important, and can be done with a continuous 3D model, including only the localized nature of charge storage but neglecting the discreteness of traps, dopants and electron flow. The simulations were obtained by a numerical tool extending to 3D geometries the model for charge-trap memory programming that has been presented in chapter 2, implementing electron trapping as [58,97]

$$\frac{dn_t\left(x,y,z\right)}{dt}=\frac{J\left(x,y,z\right)}{q}\,\sigma\left[N_t-n_t\left(x,y,z\right)\right] \tag{3.4}$$

where $n_t(x,y,z)$ is the trapped electron density in the silicon nitride. The emissivity has been supposed negligible, as was obtained for large area SONOS capacitors in the previous section.

In Fig. 3.9(a) are shown the simulated ISPP transients for a 20 *nm* 3D cell with ONO thicknesses 4/4.5/5 nm; the transient was simulated for a gate voltage increase per step V_s = 500 mV with time step τ_s = 10 μs: after the first steps, the ΔV_T transient becomes linear, but the programming efficiency is quite low, $\Delta V_{T,s}/V_s \simeq 0.5$. In Fig. 3.9(b) is shown the average trapped charge density in the storage layer as a function of the ΔV_T during the programming transient: an almost linear curve is obtained and the average capacitance per unit area can be calculated. The slope of the linear fitting is indeed proportional to the 3D capacitance:

$$C_{pp}^{3D} = q\frac{dn_t}{dV_T}t_{sin}. \tag{3.5}$$

A capacitance per unit area of 1.19 μF/cm^2 is obtained. This value should be compared with the 1D expected capacitance per unit area:

$$C_{pp}^{1D} = \frac{\varepsilon_{ox}}{t_{bo} + t_{sin}\varepsilon_{ox}/2\varepsilon_{sin}}, \tag{3.6}$$

in which case a value of 0.56 μF/cm^2 is obtained. The difference is the result of the fringing fields in the 3D electrostatics, increasing the gate coupling both with the nitride stored charge and the channel [87]. The non uniform filling of the traps in the nitride, as the electric field is higher at the STI cell corners, gives a deviation from the linear behavior for large densities of stored charge, and reflects a change in the impact of the stored charge on V_T as the programming proceeds.

The low $\Delta V_{T,s}/V_s$ is mainly the result of the low impact exerted by stored electrons on V_T and, in turn, of the large effective C_{pp}^{3D}. Note that assuming for the electrons an electrostatic control as in the 1D case, the resulting ISPP slope would be increased by the ratio $C_{pp}^{3D}/C_{pp}^{1D} \simeq 2.12$, i.e. it would be even a little bit larger than 1. This means that the non-uniform tunneling profile over the active area not only do not degrade the electron injection process, but enhance indeed the process with respect to the 1D case.

Fig. 3.10(a) shows a scaling analysis of the ISPP ΔV_T transient, obtained using the continuous 3D model for cell programming and assuming a modification of the cell area keeping the same gate stack of the previous: cells of 30 nm and 40 nm length are simulated and compared to the 20 nm cell previously analyzed. A reduction of the ISPP efficiency with cell scaling clearly appears, in terms both of slope and of horizontal delay of the curves. This is due to a decreasing impact of electrons stored in the nitride on V_T as cell dimensions are reduced, for the increasing of the fringing effects. Fig. 3.10(b) shows, in fact, that the average n_t curves display a faster electron injection and storage in the nitride as scaling proceeds, due to a larger field enhancement at the corners of the cell area. This is also confirmed from the 3D capacitance per unit area extraction: with a procedure similar to the one used for the 20 nm cell, for the 30 nm cell, $C_{pp}^{3D} = 0.84$ μF/cm^2, while for the

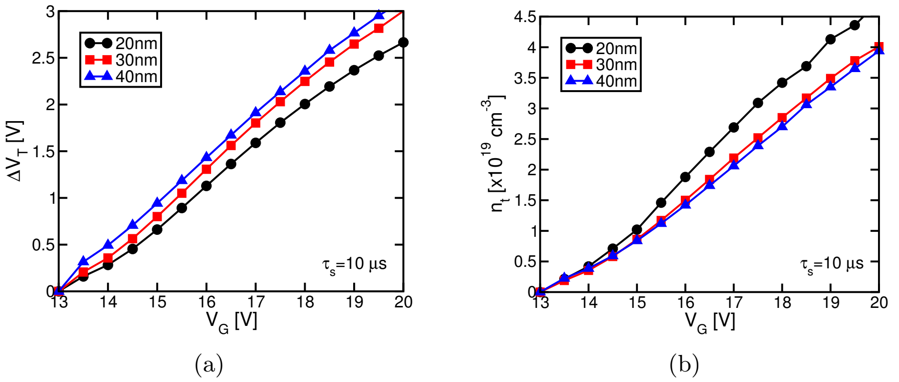


Figure 3.10: Simulated ISPP transient for different cells size: (a) ΔV_T and (b) trapped charge density.

40 nm case it is 0.73 μF/cm^2. As expected, the fringing effect is higher for smaller cells, leading also to higher capacitance.

3.4 Conclusions

A detailed investigation of the Incremental Step Pulse Programming (ISPP) dynamics of charge-trap memory capacitors was presented, investigating the tunnel oxide electric field (F_{tun}) and the tunneling current (J_{tun}) evolution during programming. Differently from the floating-gate case, results on nitride-based memories show that, after the stationary working point of the ISPP algorithm is almost reached on the J_{trap}-F_{tun} diagram, the V_{FB} increase per step does not equal the step amplitude of the gate staircase, decreasing, moreover, as programming proceeds. The reduction of J_{trap} determines an increase of F_{tun} and J_{tun}, potentially compromising device reliability. Results from ISPP experiments with different τ_s, at different temperatures and on samples with different stack compositions show that this is due to a drop of η_{trap} as more and more charge is stored in the nitride layer, as a result of the reduction of the number of free traps available in the nitride. Afterwards, using a 3D simulator, ultra scaled devices were also analyzed with simulations, revealing a further decrease of the programming efficiency, caused by the increasing impact of the fringing field. For this reason ISPP algorithm, used in nitride based memories, does not have the same behavior as it has in FG Flash. Moreover, a careful cell engineering should be done in ultra-scaled nodes, as the V_T increase per step worsen with decreasing dimensions and, in turn, increasing impact of the fringing field.

Chapter 4

Modeling of cylindrical CT memory

This chapter deals with modeling of cylindrical charge trap devices: a detailed analytical investigation of the transient dynamics of gate-all-around charge-trap memories is presented. To this aim, the Poisson equation is solved in cylindrical coordinates and a modification of the well-known Fowler-Nordheim formula is proposed for tunneling through cylindrical dielectric layers. Analytical results are then validated by experimental data on devices with different gate stack compositions, considering a quite extended range of gate biases and times. Finally, the model is used for a parametric analysis of the gate-all-around cell, highlighting the effect of device curvature on both program/erase and retention.

4.1 Introduction

Three-dimensional architectures appear today as viable solutions for the integration of non-volatile memory cells in Tera-bit arrays [39–43, 98]. In particular, the gate-all-around charge-trap (GAA-CT) cell with vertical channel is considered one of the most promising structures for future NAND Flash technologies, showing improved program/erase and retention performance with respect to planar devices [99–102]. Moreover, thanks to the reduction of corner and fringing field effects during both program/erase and read, GAA-CT cells allow more uniform trapped charge distributions in the storage layer and provide, in turn, steeper incremental step pulse programming (ISPP) transients than planar cells [76,77]. GAA-CT performance have been investigated by many 1-D numerical models, exploiting the cylindrical symmetry of the device [103,104]. However, despite implementing physics comprehensively, all the proposed numerical models lack computational efficiency, mainly due to the computational load needed to correctly evaluate the tunneling current to/from the storage layer in presence of non-constant electric fields in cylindrical dielectrics.

In this chapter, an accurate physics-based analytical model for both the program/erase and retention dynamics of GAA-CT memory cells is presented. The model is developed by solving the Poisson equation in cylindrical coordinates and modifying the well-known Fowler-Nordheim formula to deal with tunneling through cylindrical dielectrics. Analytical results are validated against experimental data for different gate stack compositions. Then, a detailed analysis of the GAA-CT cell performance is presented, investigating the program/erase and retention transients as a function of the parameters of the cylindrical structure and highlighting the effect of device curvature.

4.2 Physics-based analytical model

Fig. 4.1 shows the template MONOS device used, without any lack of generality, to develop the analytical model for the transient dynamics of GAA-CT cells. Throughout this section, the following parameters are used, unless otherwise specified: substrate radius $R_0 = 3$ nm, tunnel oxide thickness $t_{tun} = R_{tun} - R_0 = 4.5$ nm, nitride thickness $t_{sin} = R_{sin} - R_{tun} = 6$ nm, blocking oxide thickness $t_{bo} = R_{bo} - R_{sin} = 7$ nm. Aluminum was assumed for the gate. For the sake of generality, different dielectric constants will be considered for the tunnel and blocking oxides and for nitride (respectively ε_{tun}, ε_{bo} and ε_{sin}), though final results will consider $\varepsilon_{tun} = \varepsilon_{bo} = \varepsilon_{ox}$ (the SiO2 dielectric constant). Axial symmetry is assumed, and the model is developed only as a function of the radial coordinate r.

4.2.1 Electrostatic solution

The electrostatics of the GAA cell can be straightforwardly calculated by solving the Poisson equation in cylindrical coordinates:

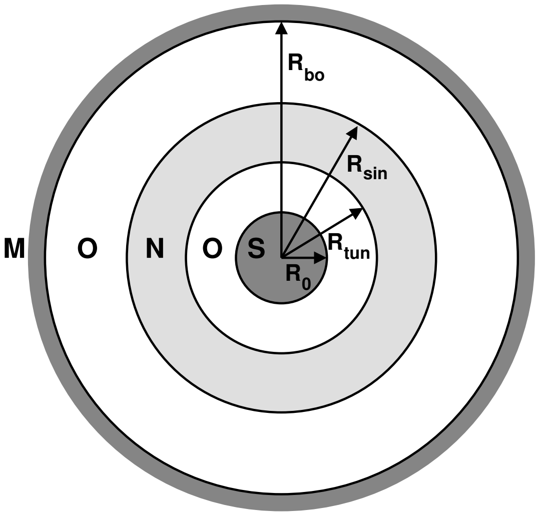


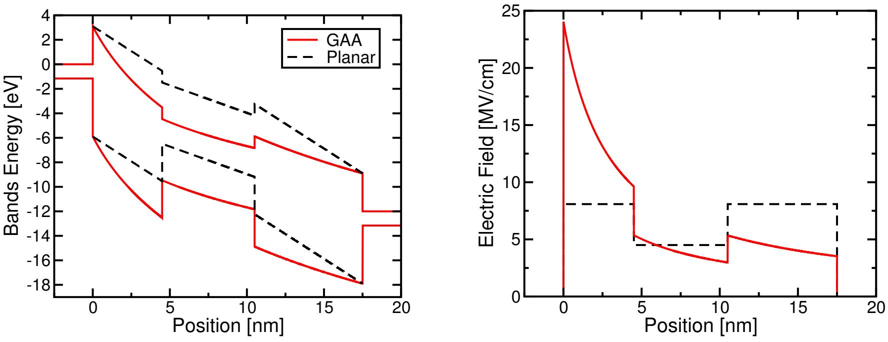
Figure 4.1: Schematic cross-section of the template GAA-CT MONOS cell: R_0, R_{tun}, R_{sin} and R_{bo} are the radii corresponding to the interfaces between the materials in the stack, as highlighted in the figure.

$$\frac{\partial^2 V(r)}{\partial r^2} + \frac{1}{r}\frac{\partial V(r)}{\partial r} = -\frac{q n_t}{\varepsilon_{sin}} \left[H\left(r - R_{tun}\right) - H\left(r - R_{sin}\right) \right] \tag{4.1}$$

where q is the electron charge, $V(r)$ is the electrostatic potential along the radial coordinate and n_t is the volumetric trapped electron density in the nitride (assumed constant, units: [cm^{-3}]). The Heavyside functions H in (4.1) are used to include electron trapping only in the nitride volume and not in the cell oxides layers. Initially neglecting any potential drop in the silicon substrate, whose impact will be addressed in Section 4.2.4, (4.1) can be analytically integrated to obtain $V(r)$ along the tunnel oxide, the nitride and the blocking oxide (namely, $V_{tun}(r)$, $V_{sin}(r)$ and $V_{bo}(r)$, respectively), when a gate bias V_G is applied:

$$\begin{cases} V_{tun}(r) & = C_1 \ln \dfrac{r}{R_0} \\ V_{sin}(r) & = C_1 \ln \dfrac{R_{tun}}{R_0} + C_2 \ln \dfrac{r}{R_{tun}} + \dfrac{q n_t}{4 \varepsilon_n} \left(r^2 - R_{tun}^2 \right) \\ V_{bo}(r) & = V_G - C_3 \ln \dfrac{R_{bo}}{r} \end{cases} \tag{4.2}$$

where the explicit expression for the constants C_i (i = 1–3) is:



(a) (b)

Figure 4.2: Comparison between the templare GAA MONOS cell and the planar CT cell having the same thickness of the gate dielectrics in terms of energy-band profile (a) and electric field (b), for $V_G = 12$ V and neutral nitride.

$$\begin{cases} C_1 & = \frac{V_G}{\alpha} + \frac{qn_t}{2\varepsilon_{sin}\alpha}\left[R_{tun}^2 \ln \frac{R_{sin}}{R_{tun}} + \right. \\ & \left. - \frac{1}{2}\left(R_{sin}^2 - R_{tun}^2\right)\left(1 + 2\frac{\varepsilon_{sin}}{\varepsilon_{bo}} \ln \frac{R_{bo}}{R_{sin}}\right)\right] \\ C_2 & = \frac{\varepsilon_{bot}}{\varepsilon_{sin}} C_1 - \frac{qn_t}{2\varepsilon_{sin}} R_{tun}^2 \\ C_3 & = \frac{\varepsilon_{tun}}{\varepsilon_{bo}} C_1 - \frac{qn_t}{2\varepsilon_{bo}}\left(R_{tun}^2 - R_{sin}^2\right) \end{cases}$$

where α is given by:

$$\alpha = \frac{\varepsilon_{tun}}{\varepsilon_{bo}} \ln \frac{R_{bo}}{R_{sin}} + \ln \frac{R_{tun}}{R_0} + \frac{\varepsilon_{tun}}{\varepsilon_{sin}} \ln \frac{R_{sin}}{R_{tun}}$$

From (4.2), the electric field in the tunnel oxide ($F_{tun}(r)$), blocking oxide ($F_{bo}(r)$) and nitride ($F_n(r)$) region results:

$$\begin{cases} F_{tun}(r) & = -C_1 \frac{1}{r} \\ F_n(r) & = -C_2 \frac{1}{r} - \frac{qn_t}{2\varepsilon_n} r \\ F_{bo}(r) & = -C_3 \frac{1}{r} \end{cases} \tag{4.3}$$

Fig. 4.2 shows the comparison between the electrostatics of the template GAA MONOS cell (solid) and of the planar CT cell having the same thickness of the gate dielectrics (dashed) in the case of $V_G = 12$ V and neutral nitride (i.e., $n_t = 0$). The energy band profile of Fig. 4.2(a), clearly shows a reduction of the thickness of the energy barrier preventing electron tunneling from the substrate to the nitride. As highlighted in (4.3) and differently from the planar case, in fact, the electric field is

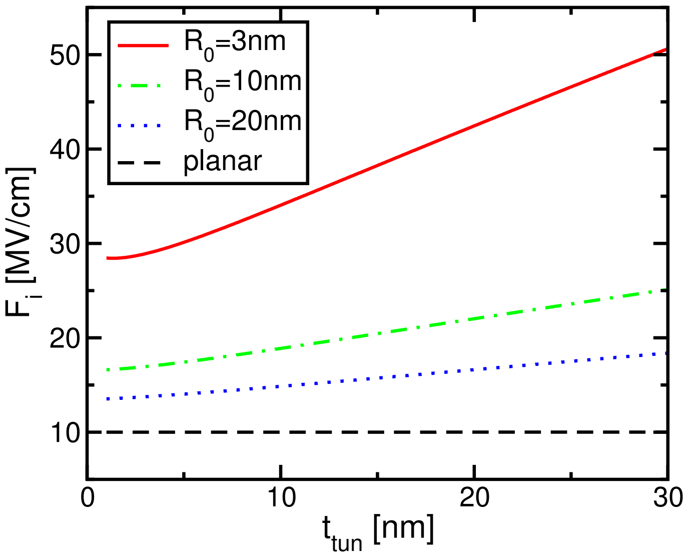


Figure 4.3: Maximum electric field at the substrate/tunnel oxide interface of the template GAA cell when modifying t_{tun} (with fixed t_{sin} and t_{bo}). A gate voltage V_G allowing a constant electric field $V_G/EOT = 10$ MV/cm in the tunnel oxide of the planar device is assumed.

not constant in the dielectrics of the GAA device, with a maximum value F_i appearing at the substrate/tunnel oxide interface (see Fig. 4.2(b)). The maximum value is about three times larger with $R_0 = 3$ nm than the electric field in the tunnel oxide of the planar device, resulting in the possibility for a strong improvement of the programming dynamics [44], as will be discussed in Section 4.2.3. In addition to that, the electric field in the top oxide is lower in the GAA than in the planar case, resulting in the possibility for a lower electron leakage from the nitride to the gate during programming.

The maximum electric field F_i of the template GAA cell is reported in Fig. 4.3 as a function of t_{tun} (with fixed t_{sin} and t_{bo}), assuming a constant electric field in the tunnel oxide of the planar case $V_G/EOT =$ 10 MV/cm, where $EOT = \varepsilon_{ox}\left(t_{tun}/\varepsilon_{tun} + t_{sin}/\varepsilon_{sin} + t_{bo}/\varepsilon_{bo}\right)$ is the equivalent oxide thickness of the planar dielectric stack (ε_{ox} is the oxide dielectric constant). An increase of F_i with t_{tun} clearly appears, representing a main feature of the cylindrical system. Moreover, the field increase is enhanced as the substrate radius is reduced, therefore allowing the possibility for large improvements in both the programming and the retention dynamics.

From (4.3), the threshold voltage shift ΔV_T resulting from electron

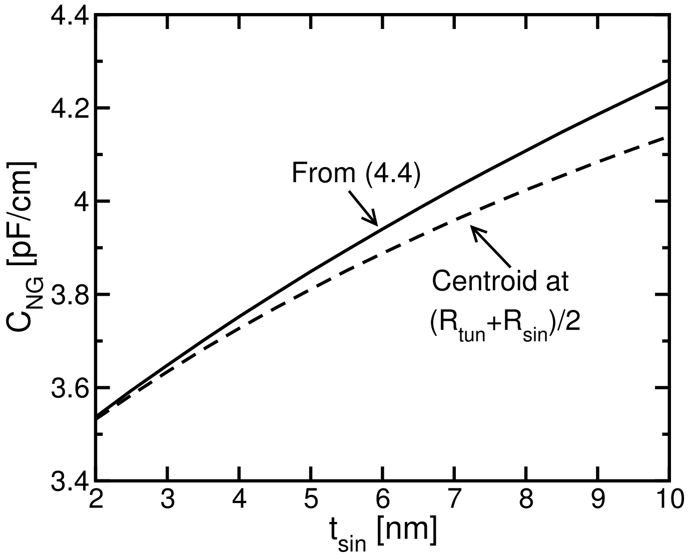


Figure 4.4: C_{NG} calculated from (4.4) and assuming the trapped electron centroid at the middle of the nitride layer, for increasing t_{sin} and fixed R_0, t_{tun} and t_{bo}.

storage in the nitride volume can be easily calculated as the increase of V_G required to maintain the same F_i present in the cell when $n_t = 0$. From a straightforward analysis of (4.3) and of the coefficient C_1, the threshold voltage shift can be written as:

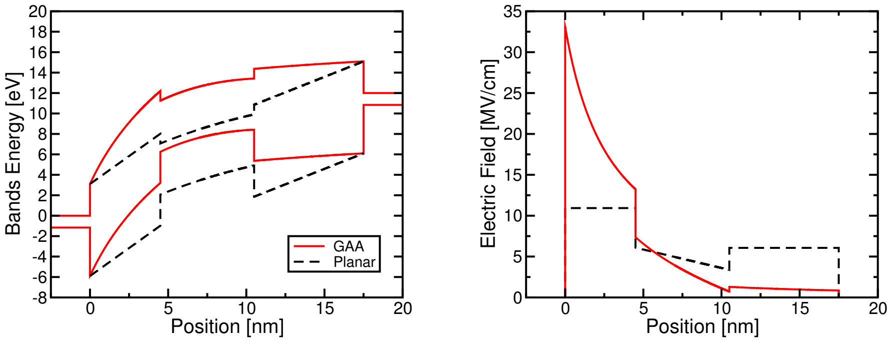
$$\Delta V_T = -\frac{qn_t}{2\varepsilon_{sin}}\left[R_{tun}^2 \ln\frac{R_{sin}}{R_{tun}} - \frac{1}{2}\left(R_{sin}^2 - R_{tun}^2\right)\left(1 + \frac{2\varepsilon_{sin}}{\varepsilon_{bo}}\ln\frac{R_{bo}}{R_{sin}}\right)\right]$$

This equation directly gives ΔV_T in the case of uniformly distributed electrons in the nitride volume and allows the extraction of the capacitance per unit length in the wire direction (C_{NG}, units: [F/cm]) between the centroid of stored electrons and the gate:

$$C_{NG} = -\frac{Q}{\Delta V_T} = \frac{qn_t\pi\left(R_{sin}^2 - R_{tun}^2\right)}{\Delta V_T} \tag{4.4}$$

where $Q = -q\int_{R_{tun}}^{R_{sin}} 2\pi r n_t dr$ is the stored charge per unit length in the nitride (units: [C/cm]). Fig. 4.4 shows that a rather negligible error is committed (less than 2.5% in the considered t_{sin} range), in the evaluation of C_{NG} if this is calculated assuming the trapped electron centroid in the middle of the nitride layer.

Finally, Fig. 4.5 shows a comparison between the GAA and the planar cell electrostatics during erase for $V_G = -12$ V and $\Delta V_T = 3$ V.



(a) (b)

Figure 4.5: Same as in Fig. 4.2, but for $V_G = -12$ V and $\Delta V_T = 3$ V.

Similarly to the case of positive V_G, in the case of the GAA cell the electric field reaches a maximum at the substrate/tunnel oxide interface, which is quite larger than the electric field present in the planar device. This enhances the hole tunneling current from the substrate to the nitride during erase, as will be discussed in Section 4.2.3. In addition, the quite lower electric field at the gate/blocking oxide interface should prevent electron injection from the gate, therefore relieving the erase saturation issues [44, 49].

4.2.2 Tunneling current calculation

When quantization effects are accounted for in a cylindrical geometry, the energy eigenvalues are given by [105]:

$$E_{l,i} = \frac{\hbar^2 \lambda_{l,i}^2}{2m^* R_0^2} \tag{4.5}$$

where $\lambda_{l,i}$ is the i-th zero of the l-th order Bessel function. The electron concentration per unit length on each level, $n_{l,i}$, is given by

$$n_{l,i} = \sqrt{\frac{8m_D^* k_B T}{\pi \hbar^2}} \mathcal{F}_{-1/2}\left(\frac{E_F - E_{l,i}}{k_B T}\right), \tag{4.6}$$

where $\mathcal{F}_{-1/2}$ is the Fermi-Dirac integral of order $-1/2$ and m_D^* is an “effective” density-of-states mass in the axial direction. Its value was computed requiring that the quantum charge concentration approaches the classical value for large quantization radii: $m^* = m_l$ was chosen (the longitudinal mass of silicon), obtaining $m_D^* = 36m_t^2/m_l$ (m_t is the silicon transverse mass). The choice of m^* affects the results by less than an order of magnitude, which can be considered quite good for the analysis. The tunneling current density (per unit length) can now be calculated as

$$J'_{tun} = q \sum \frac{n_{l,i}}{\tau_{l,i}} T_{l,i} \tag{4.7}$$

where $T_{l,i}$ is the tunneling probability and $\tau_{l,i}$ the inverse of the attempt to escape frequency. The tunneling probability is computed with the transfer matrix method, following the work of [106], while for $\tau_{l,i}$ the radial round-trip time was taken, adopting the same approach as in a planar geometry (which is somewhat justified by the similarity between the cylindrical and planar tunneling times reported in [107]).

Such a numerical approach may become unsuitable when fast evaluation of the device performance is needed; for this reason, a simplified WKB approximation mimicking a planar behavior has been proposed [103]. Another approach is instead to analytically express J'_{tun} via a Fowler-Nordheim (FN) equation [60, 61], like in the planar case:

$$J'_{tun} = A' F_{eq}^2 exp\left[-\frac{B}{F_{eq}}\right] \tag{4.8}$$

where A' and B are constants including the physical parameters of the potential barrier, and F_{eq} is an effective electric field. It is worth noting, first of all, that for $R_0 \geq 3$ nm the tunnel oxide tunneling transparency can be calculated under the WKB approximation as [103]:

$$T_{cyl}(E_r) = \frac{R_0}{R_{tun}} exp\left[-\frac{2\sqrt{2m_{ox}}}{\hbar} \int_0^{r1} \sqrt{E_c(r) - E_r} dr\right] \tag{4.9}$$

where E_r is the electron energy in the radial direction, $E_c(r)$ is the conduction-band energy profile and m_{ox} is the electron tunneling mass in the oxide. The integral in (4.9) is performed from the substrate surface to the radius r_1 where E_c becomes equal to E_r. Assuming $E_r = E_c(R_0) = 0$, the tunneling transparency calculated by (4.9) for the hyperbolic $E_c(r)$ profile of the GAA cell, computed by (4.2), can be equaled to that of a planar structure by defining an equivalent electric field (F_{eq}) so that:

$$\int_0^{r_2} \sqrt{E_{FN}(r)} dr = \int_0^{r_1} \sqrt{E_c(r)} dr \tag{4.10}$$

where $E_{FN}(r)$ is the conduction band profile in the case of constant electric field in the tunnel oxide, as schematically shown in Fig. 4.6, and r_2 is the radius at which $E_{FN}(r)$ becomes equal to 0. From (4.10), F_{eq}

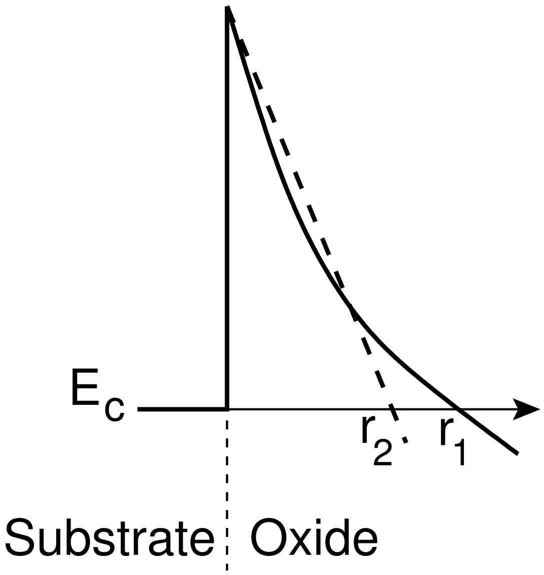


Figure 4.6: Schematics for the conduction band profile in the GAA cell (continuous line) and linear band profile (dashed line) giving the same tunneling transparency for electron energy level in the radial direction $E_r = E_c(R_0) = 0$.

can be straightforwardly evaluated as:

$$F_{eq} = \frac{2\sqrt{\Phi_B^3}}{3q\int_0^{r_1}\sqrt{E_c(r)}dr} \tag{4.11}$$

where $\Phi_B = 3.1$ eV is the electron tunneling barrier height (only tunneling from the bottom of the band is assumed for simplicity).

Fig. 4.7 shows F_{eq} as a function of F_i for the template GAA MONOS cell when changing t_{tun} (a) or R_0 (b). A good linear relation of unit slope between the two fields clearly appears for sufficiently high F_i, with negligible dependence on t_{tun}. Note, moreover, that the displacement of the F_{eq} curves from the straight line at very low F_i in Fig. 4.7(a) is due to the change of the tunneling regime from Fowler-Nordheim to direct tunneling. This change takes place at smaller F_i for thicker t_{tun} and is not addressed in this analysis, as the resulting tunneling currents are too low to meet the programming specifications of non-volatile devices. Fig. 4.7(b) shows, in addition, that the straight line describing the F_{eq} vs. F_i relation shifts towards higher F_i values when R_0 is increased. As a result, F_{eq} can be calculated as

$$F_{eq} = F_i - \frac{V_0}{R_0} \tag{4.12}$$

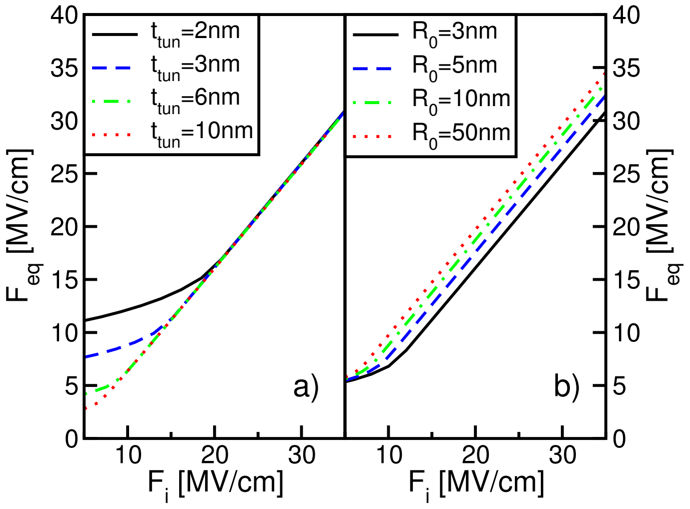


Figure 4.7: Relationship between the F_{eq} computed with (4.11) and F_i for different t_{tun} (a) and R_0 (b).

where $V_0 = 1.2$ V is directly obtained from fitting the R_0 dependence of Fig. 4.7(b).

Fig. 4.8 shows a comparison between results obtained via (4.7) and via (4.8). To achieve a better fit, however, the FN equation was not applied to J'_{tun}, but rather to the areal current density at the oxide/nitride boundary:

$$J_{tun} = \frac{J'_{tun}}{2\pi R_{tun}} = A F_{eq}^2 \exp\left(-\frac{B}{F_{eq}}\right) \qquad (4.13)$$

Note that a very good fit is achieved in the investigated range of $t_{tun} = 3$ to 6 nm and R_0 ranging from 3 to 10 nm (beyond this value the structure can be treated as almost planar). Moreover, the parameters values $A \approx 10^{-7}$ A V^{-2}, $B \approx 215$ MV/cm are basically the same that are extracted from planar structures. This is a consequence of the adoption of the effective field (4.11), which captures the main effect of the curvature on the tunneling barrier. The previous analysis was then extended to hole tunneling under negative V_G, yielding the following fitting parameters for (4.12) and (4.13): $V_0 = 1.5$ V, $B \approx 275$ MV/cm and A about a factor of two smaller than the electron value. However, it is worth pointing out that the values of A depends on the adopted approximation and parameters values, and should not be regarded as definitive.

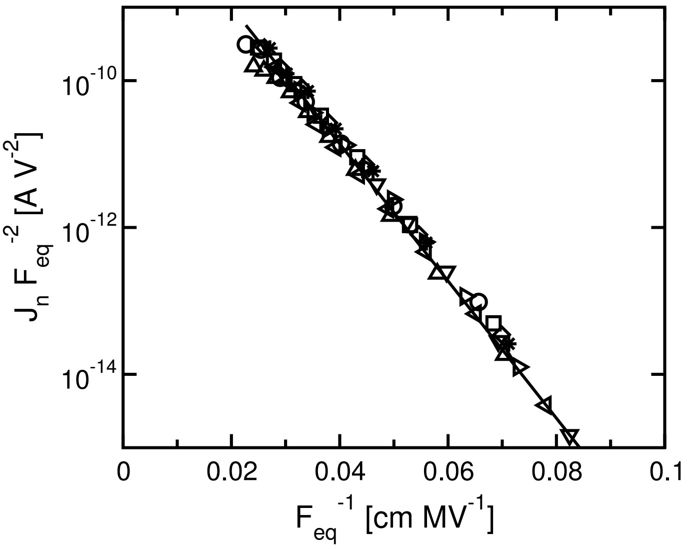


Figure 4.8: Comparison between (4.7) (symbols) and (4.13) (line) for different R_0 ranging from 3 to 10 nm and *tbot* = 3 and 6 nm.

4.2.3 Transient dynamics

Once analytical solutions for the electrostatic and tunneling problems are found, the program transients of the GAA-CT cell can be calculated with the following equation [54]:

$$\frac{dn_t}{dt} = \frac{J_{tun}}{q}\left(\frac{R_{tun}}{R_{tun} + t_{sin}/2}\right)\sigma\left(N_t - n_t\right) - en_t, \tag{4.14}$$

where N_t is the trap density in the nitride (units: [cm^{-3}]), σ is the electron trapping cross-section (units: [cm^2]), and e is the Poole-Frenkel emissivity (units: [s^{-1}]) from filled nitride traps:

$$e_n = \nu_0 \; exp\left[-\frac{E_T - \beta\sqrt{F_{sin}}}{kT}\right] \tag{4.15}$$

Here, ν_0 is the attempt-to-escape frequency from nitride filled traps [units: s^{-1}], E_T is the trap depth from the nitride conduction band, $\overline{F_{sin}}$ is the average electric field in the nitride and β is the Poole-Frenkel coefficient [66,108]. Note that (4.14) assumes that all the empty nitride traps see the same tunneling current, i.e. it neglects both distributed trapping along the nitride thickness and electron transport in the nitride conduction band, which were shown to barely impact the program operation of planar cells [54]. As a consequence, traps are concentrated in the

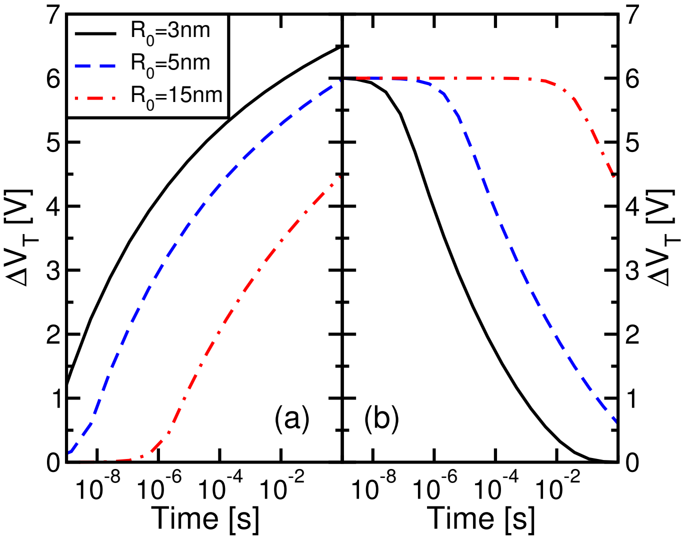


Figure 4.9: Calculated program (a) and erase (b) transients at $V_G = \pm 12$ V, for different R_0, on the template GAA-CT MONOS cell considered in Section 4.2, assuming $N_t = 6 \times 10^{19}$ cm^{-3}, $\sigma = 5 \times 10^{-13}$ cm^2, $\sigma_r = 5 \times 10^{-13}$ cm^2, $E_T = 1.5$ eV, $\nu_0 = 5 \times 10^8$ s^{-1}.

middle of the nitride, and conservation of J_{tun} leads to a correcting factor $R_{tun}/(R_{tun} + t_{sin}/2)$ in (4.14). Fig. 4.9(a) shows the calculated programming transient on the template GAA MONOS cell at $V_G = 12$ V, assuming $N_t = 6 \times 10^{19}$ cm^{-3}, $\sigma = 5 \times 10^{-13}$ cm^2, $\nu_0 = 5 \times 10^8$ s^{-1}. The programming dynamics are faster for smaller R_0, thanks to a larger F_i (hence J_{tun}), as shown by (4.12) and Fig. 4.3. This effect overrides the decrease of ΔV_T that is predicted by (4.4) for smaller R_0 and fixed thickness of the dielectrics.

In order to describe the erase operation of GAA-CT cells, (4.14) was modified according to:

$$\frac{dn_t}{dt} = -\frac{J_p}{q}\left(\frac{R_{bot}}{R_{bot} + t_n/2}\right)\sigma_r n_t - e_n n_t, \tag{4.16}$$

where σ_r is the electron/hole recombination cross-section (units: [cm^2]). Note that (4.14) and (4.16) do not have an analytical solution. However, they can be directly solved discretizing the time variable, and updating the electric field (and in turn J_{tun}, e and $J_{tun,p}$) at each time step.

Fig. 4.9(b) shows the calculated erase transients at $V_G = -12$ V for different R_0 when assuming $E_T = 1.5$ eV and $\sigma_r = 5 \times 10^{-13}$ cm^2. Also in this case, faster ΔV_T dynamics appear when reducing R_0, thanks to

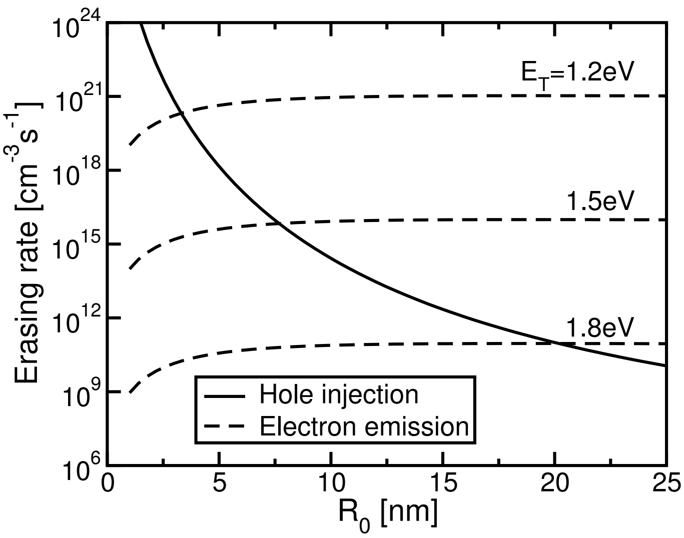


Figure 4.10: Dependence on the internal radius R_0 of the trap emptying rates due to hole injection and electron emission for different trap energies at the beginning of the erase transient ($\Delta V_T = 6$ V, $V_G = -12$ V) on the template GAA-CT MONOS cell.

the larger F_{eq}. The roles of holes and electrons on the erase transients are highlighted in Fig. 4.10, where the trap emptying rate given by hole recombination ($J_{tun,p}\sigma_r n_t/q$) and electron emission (en_t) at the beginning of the erase transient ($\Delta V_T = 6$ V, $V_G = -12$ V) are shown as a function of R_0. Considering typical values of $E_T = 1.2 \div 1.8$ eV [44, 54,104], electron emission appears as the dominant erase mechanism for large R_0, while hole injection gains importance for small radii, due to the increase of F_{eq} and $J_{tun,p}$. The value of R_0 marking the transition between the two mechanisms decreases as lower E_T are considered, due to the larger emission rate given by (4.15).

Finally, note that (4.16) can be also used to investigate the retention transients on the GAA-CT devices, at least at high temperature, where the Poole-Frenkel emission is the dominant mechanism. Typical results at $T = 85°$C are shown in Fig. 4.11, where the ΔV_T loss from the programmed state appears strongly dependent on E_T and more weakly on R_0. This is due to the dominant role played by electron emission from the nitride over hole injection from the substrate [104], as evident in Fig. 4.12, where the trap emptying rate given by hole recombination ($J_{tun,p}\sigma_r n_t/q$) and electron emission (en_t) at the beginning of the retention transient ($\Delta V_T = 6$ V, $V_G = 0$ V) are shown as a function of R_0 for the template GAA-CT MONOS cell.

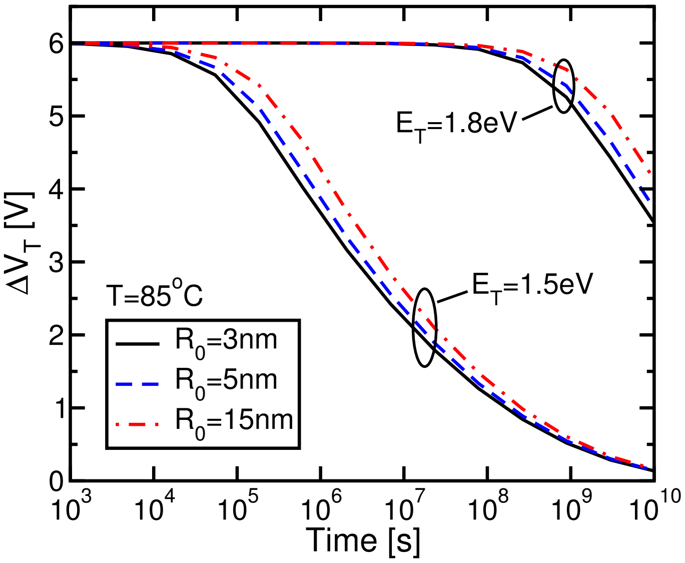


Figure 4.11: Calculated retention transients at 85°C on the template GAA-CT MONOS cell for different E_T and R_0.

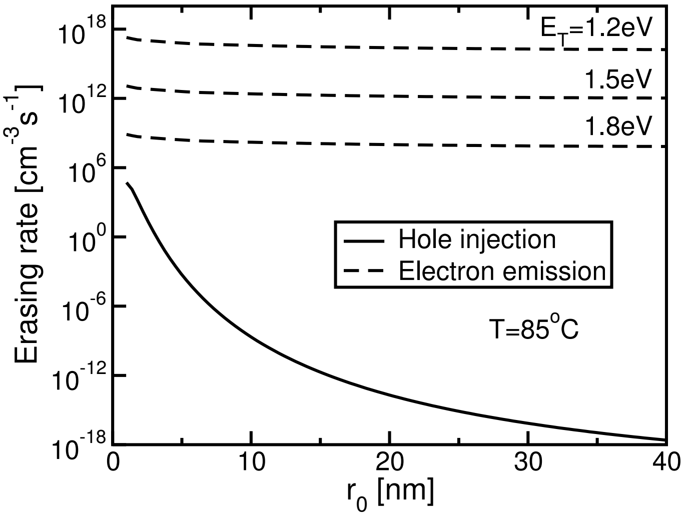


Figure 4.12: Dependence on the internal radius R_0 of the trap emptying rates due to hole injection and electron emission for different trap energies at the beginning of the retention transient ($\Delta V_T = 6$ V, $V_G = 0$ V) on the template GAA-CT MONOS cell, at $T = 85$°C.

4.2.4 Substrate effects

A detailed analysis of the electrostatic of GAA MOSFETs [109, 110] showed that the surface potential saturates at around 0.6 V for increasing V_G, with rather negligible dependence on R_0. As a consequence, a first-order account of both the potential drop in the substrate and the built-in potential deriving from the work-function difference between the metal gate and the silicon can be obtained by adding a correcting factor to the gate bias V_G. Such a term should be constant for programming and long-term retention, while a dependence on V_G should be included for cell erase. In fact, the potential profile in the substrate strongly depends on the hole supply mechanism during the short erase pulses, which is related to the carrier generation process. This may depend on physical cell details such as source/drain junction doping, and its accurate inclusion is not straightforward even in advanced models [104]. As a result, the approach of [44] was followed and a linear increase of the surface potential with V_G was adopted for the erase transients.

4.3 Modeling results

The model will now be compared with experimental data on GAA-SONOS cells taken from literature, to assess its validity. A parametric analysis of the program, erase and retention performance of GAA-CT memory devices will be presented afterwards, focusing on the dependence on R_0.

4.3.1 Comparison with experimental data

Fig. 4.13 shows a comparison of the modeling results with experimental data for the program/erase transients of a GAA SONOS cell with $R_0 =$ 3 nm and oxide/nitride/oxide layers of 6/5/8 nm [44]. A reasonably good agreement appears, using the following parameters: $N_t = 5.3 \times 10^{19}$ cm^{-3}, $\sigma = 2 \times 10^{-12}$ cm^2, $\sigma_r = 10^{-12}$ cm^2, $\nu_0 = 5 \times 10^8$ s^{-1}, $E_T = 1.5$ eV. Note that the mismatch between data and modeling results during programming at $V_G = 18$ V can be attributed to a reduction of the trapping efficiency at large bias, as previously reported in [54], which is not included in the current model. Moreover, at the shortest times experimentally investigated in the figure, the possibility for spurious delays to compromise the pulse shape reaching device gate should be accounted for.

Recently, the employment of high-k dielectrics in conjunction with a metal gate has been shown feasible for the GAA technology [102].

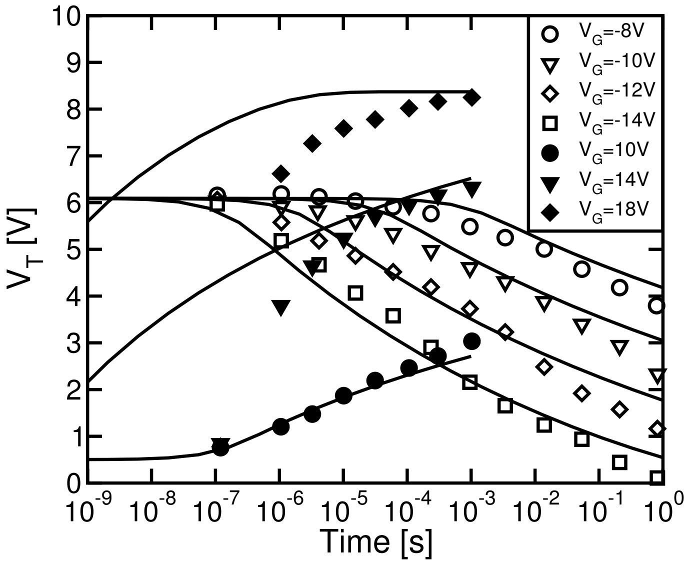


Figure 4.13: Experimental data [44] (symbols) and calculated results (lines) for the program/erase transients on a GAA-CT SONOS cell with $R_0 = 3$ nm. Note that the starting time of all the simulations is 10^{-12} s.

Fig. 4.14 shows that a good agreement between modeling and experimental data is also achieved for a GAA TAHOS (TaN/Al_2O_3/HfO_2/SiO_2/Si) cell having $R_0 = 12$ nm and a A/H/O gate stack of 10/7/5 nm (data taken from [102], see this paper for more details on the device structure). Parameter values used for modeling are $N_t = 5 \times 10^{19}$ cm^{-3}, $\sigma = 5 \times 10^{-13}$ cm^2, $\sigma_r = 5 \times 10^{-13}$ cm^2, $\nu_0 = 5 \times 10^8$ s^{-1}, $E_T = 1.5$ eV, $\varepsilon_{bo} = 10$, $\varepsilon_{Hf} = 18$, where ε_{Hf} is the HfO_2 relative dielectric constant. Note that alumina was considered as an ideal dielectric, neglecting any possibility of charge trapping in this layer. In Chapter 2 it was shown that alumina trapping may play a role for long pulse durations, when large ΔV_T are reached [36, 45, 111, 112]. This can slightly impact the evaluation of N_t, without changing the overall conclusions. These results confirm that the simple model can reproduce the main features of the program/erase dynamics on different GAA-CT cells and can be used as an easy-to-implement tool for device optimization.

4.3.2 Parametric analysis of GAA-CT cells

The impact of R_0 and t_{tun} on ΔV_T after a 1 ms program or erase pulse on the template GAA-CT cell, keeping $t_{sin} = 6$ nm and $t_{bo} = 7$ nm and retaining the same parameters used in Section 4.2.3, which are similar

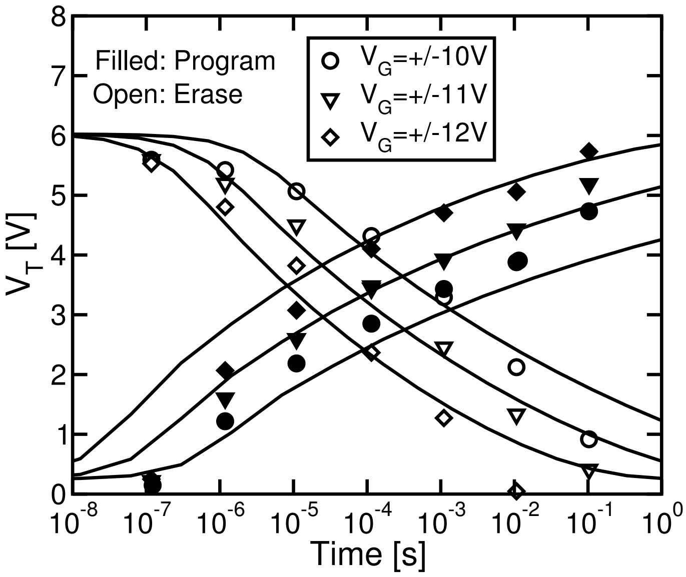


Figure 4.14: Experimental data [102] (symbols) and calculated results (lines) for the program/erase transients of a GAA-CT TAHOS cell with $R_0 = 12$ nm.

to those extracted from the experimental data and in good agreement with [54, 104]. Fig. 4.15 shows the iso-ΔV_T curves in the $t_{tun} - R_0$ plot, revealing that a reduction of these parameters accelerates the program/erase dynamics, with a stronger sensitivity to t_{tun}. For a careful cell design, however, these results should be coupled with those shown in Fig. 4.16, displaying the iso-ΔV_T curves after 10^6 s of data retention at 85°C from a 6 V programmed state. A non-negligible increase of the retention loss appears when R_0 is reduced below 5 nm for the whole explored range of t_{tun}, making clear that a trade-off must be considered. This is due to the fact that, at low R_0 values, there is a strong increase on the capacitance, requiring thus more trapped charge for a given ΔV_T, but this also leads to an increased emission during the retention transient. The considerations here presented are based only on the charge emission from the nitride to the substrate or the gate: a more detailed analysis of the retention transient will be presented in the following Chapter.

4.4 Conclusions

This chapter dealt with modeling of cylindrical charge trap devices: a detailed analytical investigation of the transient dynamics of gate-all-

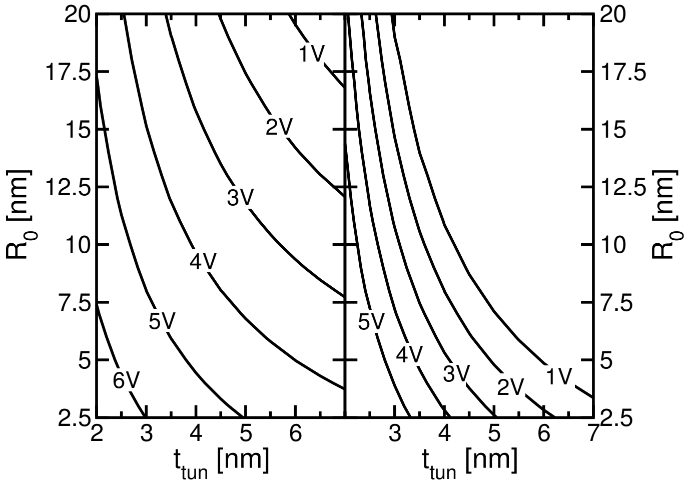


Figure 4.15: Iso-ΔV_T curves corresponding to the V_T shift after 1 ms program (left) or erase (right) pulse at $V_G = \pm 12$ V for the template GAA-CT cell.

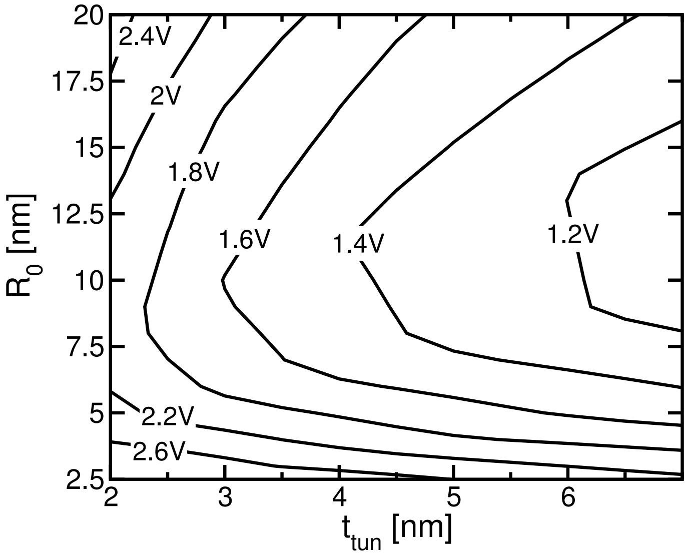


Figure 4.16: Same as in Fig. 4.15, but for the iso-ΔV_T curves after 10^6 s at 85°C from a 6 V programmed level.

around charge-trap memories was presented. After solving the Poisson equation in cylindrical coordinates, a modification of the well-known Fowler-Nordheim formula was proposed for tunneling through cylindrical dielectric layers, in order to obtain a simple formula that can be used to simulate program, erase and retention transients. The model was then validated and tuned against experimental data on devices with different gate stack compositions, considering a quite extended range of gate biases and times. Afterwards a parametric analysis of the gate-all-around cell was performed, highlighting the effect of device curvature on both program/erase and retention. The model represents a computationally efficient tool for the electrical investigation of the GAA-CT memory technology.

Chapter 5

Lateral charge migration in charge trap devices

This chapter investigates the impact of lateral charge migration on the retention performance of charge-trap memories whose storage layer is not patterned self-aligned with the channel area of each cell. Experimental results on planar SONOS devices, revealing an important contribution of lateral charge migration at 150° C, are used to calibrate a numerical model accounting for both the vertical and the lateral charge loss from the silicon nitride. Modeling results aims at estimating the trade-off between the minimum channel length and the maximum temperature required for 3D SONOS cells to fulfill the retention specifications. Disturbs to neighboring cells and impact of the lateral migrated charge on the string resistance are also evaluated in detail.

5.1 Introduction

In chapter 4, a simple analytical model for the basic operation on vertical Gate-All-Around Charge-Trap (GAA-CT) cells [38, 42, 43, 98] has been presented. That model is a 1D model that only considers the radial direction (from the substrate to the gate), and takes advantage of the cylindric symmetry. However, a constraint for the 3D vertical arrays, not present in planar devices, is that the charge-trap layer cannot be easily interrupted between different planes in a BiCS-like (Bit Cost Scalable) approach [42, 113]. This creates additional leakage pathways for the charge, migrating out of the cell active area toward other devices sharing the same string, as schematically illustrated in Fig. 5.1(a). Lateral charge leakage represents an extra source of retention loss and disturbs for cells in 3D vertical arrays, which should be carefully considered for the reliability assessment of the technology.

In this chapter an in-depth investigation of the lateral charge migration in SONOS memories is presented. Starting from experimental results on planar test-vehicle, clearly revealing the lateral charge migration during retention at 150°C, a new numerical model accounting for both the vertical and the lateral charge loss from the silicon nitride is presented, allowing a detailed analysis of the retention transients of both planar and 3D SONOS arrays. This analysis aims at estimating the trade-off between the minimum channel length and the maximum temperature required for 3D SONOS cells to fulfill the retention specifications. Disturbs to neighboring cells and impact of the lateral migrated charge on the cell's string resistance are also evaluated in detail.

5.2 Experimental results on planar devices

In order to investigate the impact of lateral charge migration on data retention, measurements were carried out on planar SONOS devices, consisting in single memory cells where the gate stack is not etched self-aligned with the device active area along the cell width (W), in the x direction, as schematically depicted in Fig. 5.1(a). Therefore, also the silicon nitride storage layer extends beyond the STI edges, being asymmetrically cut after 230 nm on the left side and after more than 10 μm on the right side. The gate stack is instead self aligned to the cell active area along the length (L), in the z direction. Despite being planar, this test-vehicle allows a lateral pathway for electrons stored above the channel region after Fowler-Nordheim (FN) programming, enabling the possibility to investigate lateral charge migration along the nitride [114]. To this aim, long term retention transients from the programmed state were measured and compared on cells with different shape factors, namely with W and L of either 1 μm or 10 μm. The use of large area cells allows to minimize short-channel effects and fringing fields typical of a 3D electrostatics [85].

Retention was also measured on a second type of device (Fig. 5.1(b)), consisting in a large area (10^{-3} cm^2) planar capacitor surrounded by a

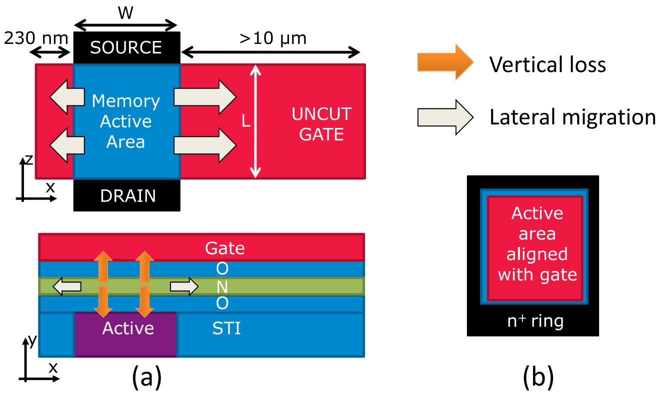


Figure 5.1: Schematic representation of the used planar test structures: (a) cells (shown in top-down and cross section view) with different W and L and overrunning gate stack over the STI and (b) large area capacitor (top-down view) with surrounding n$^+$-doped diffusion and gate stack etched aligned with the active area.

n$^+$-doped ring diffusion. In this vehicle the gate stack, and hence the charge trapping layer is completely etched aligned to the device active area, preventing any possible lateral pathway for charge loss, thus representing a reference structure for evaluating the vertical charge loss through the tunnel and blocking oxides. Moreover, the large area of these capacitors makes border effects negligible. All the samples have the same ONO gate stack, consisting of 4 nm thermal SiO$_2$ as tunnel oxide, 6 nm LP-CVD Si$_3$N$_4$ as charge-trap layer and 6 nm HTO as blocking oxide. The gate is p$^+$-doped polysilicon. In cells the threshold voltage (V_T) is obtained from the I_D-V_G curve extracting it at a fixed drain current, while in capacitors the flat-band voltage (V_{FB}) is obtained from the C-V characteristic extracting it at a fixed capacitance.

Fig. 5.2 shows the retention transients measured at room temperature (RT) on both cells and capacitors: given the adopted programming mechanism (FN tunneling) and the large area of all devices, a nearly uniform trapped electron profile over the active area is expected. Still very low V_T losses can be detected after $\sim$ 8 weeks in all devices, revealing a small impact of both the vertical and the lateral (for cells) charge loss at low temperature. In order to enhance the vertical loss and validate the assumption of equal vertical loss for cells and capacitors, retention transient at RT was field-accelerated by means of a low negative bias of -6 V applied to the gate terminal [115]. Fig. 5.3 shows that very similar transients were obtained for both the capacitor and the 10 μm cell, confirming that the vertical charge loss through the tun-

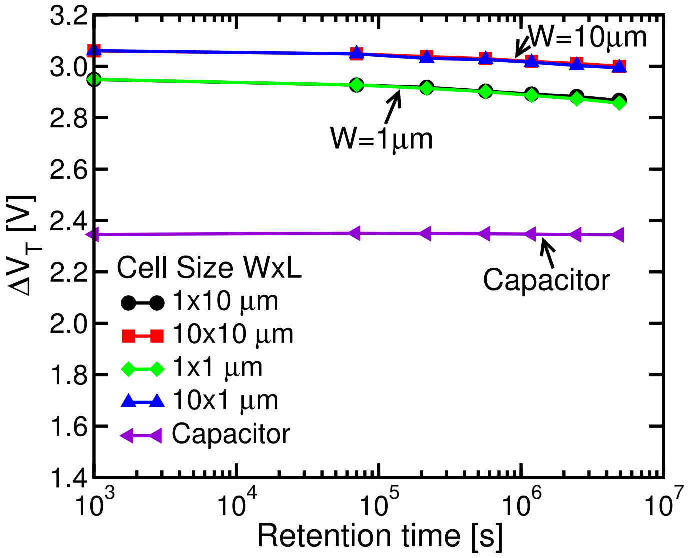


Figure 5.2: Retention transients measured in all the test devices at room temperature.

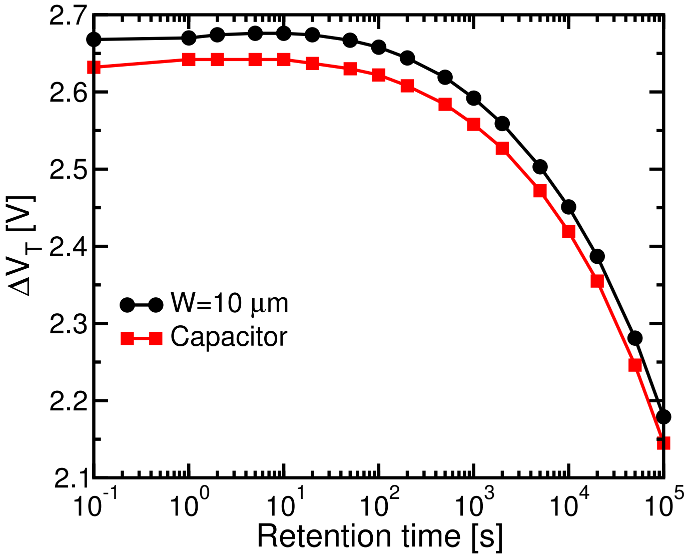


Figure 5.3: Comparison of retention transients measured in the capacitor and the larger area cell at room temperature while enhancing the vertical loss by applying −6 V as gate bias.

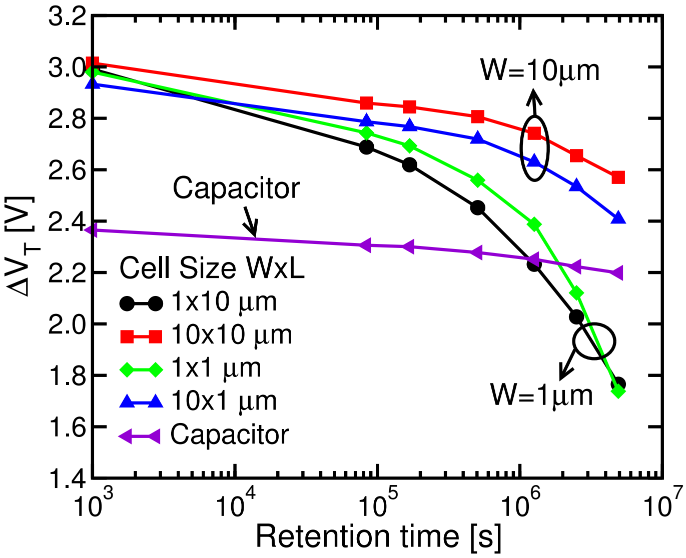


Figure 5.4: Retention transients measured in all the test devices at a temperature of 150°C.

nel oxide is the same for these devices, as expected given the same gate stack composition. As no lateral charge migration takes place in the capacitors, this result allows the quantitative investigation of the vertical charge loss for both cells and capacitors directly from the V_T transients on the latter devices. Fig. 5.4 shows that the retention transients of all the samples are strongly temperature accelerated, with larger V_T losses obtained at 150 °C with respect to those shown in Fig. 5.2. This is due to the increase of the Poole-Frenkel emission of trapped electrons from the traps to the nitride conduction band when temperature is increased, representing the bottleneck for both the lateral and the vertical charge loss from the silicon nitride at high temperatures. Note, in fact, that direct tunneling of charge from the silicon nitride traps to the substrate is small because of relatively thick oxide barriers and is almost temperature independent. Therefore, this component is overwhelmed by the thermal activation of electrons in the silicon nitride conduction band at high temperatures [116]. The larger V_T-loss of cells with respect to capacitors is attributed to the additional contribution of lateral charge migration in the former devices. This is further confirmed by the dependence of the V_T-loss on cell geometry: a higher loss is clearly evident for cells with smaller W, while L dependence is negligible. Indeed, the increase of W reduces the impact on V_T of the charge trapped near the STI corners: assuming the same lateral charge loss from this region, a lower

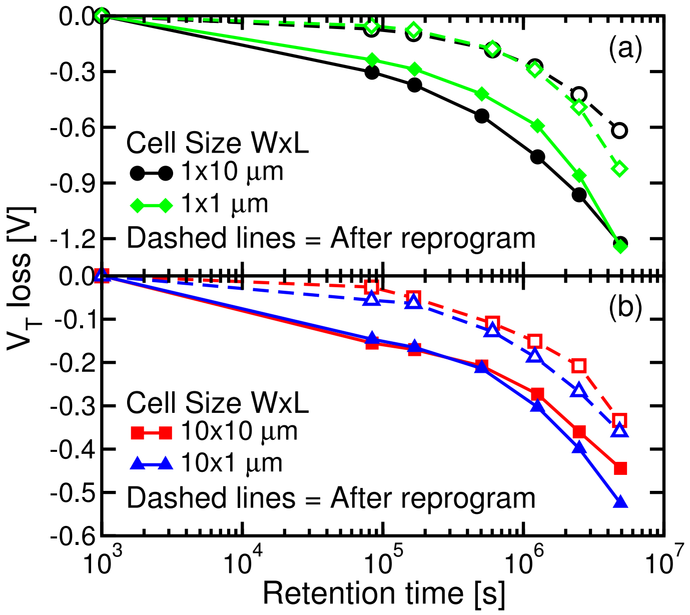


Figure 5.5: Comparison of retention loss measured on cells at 150 °C of devices programmed once (solid line) and devices reprogrammed after a previously performed retention (dashed line) in (a) $W = 1$ μm cells and (b) $W = 10$ μm cells.

V_T loss during data retention is expected As a final evidence of the presence of lateral charge migration along the cell nitride, a second retention experiment was performed on the same cells of Fig. 5.4. After the first high temperature retention transient, all cells were reprogrammed to the initial level and a second bake test was carried out at 150 °C: Fig. 5.5 compares the two retention transients, showing a V_T loss for the second one about 50% lower w.r.t. the first one. In case of pure vertical charge loss, this cannot be justified, hence being a consequence of the lateral charge migration during the first retention experiment, yielding some electrons to diffuse and to be stored in the silicon nitride layer outside the active area. This charge affects the second retention experiment, slowing down the lateral migration of charge from the active area.

5.3 Modeling and simulation of planar devices

Based on the previous experimental observations, a simulation tool that computes both the vertical and the lateral charge loss from the silicon nitride of the devices was developed. Synopsys Sentaurus Device is used to solve the Poisson equation in a 2D SONOS structure, in the x-y

plane shown in Fig. 5.1, and the new tool accesses the electric field and potential values of the solved mesh through the simulator's Physical Model Interface (PMI). In the tool, the trapped charge was assumed to be stored in the silicon nitride as a result of a previous (not simulated) program operation and its evolution during retention is computed taking into account carrier emission and recapture, according to the equation:

$$\frac{dn_t}{dt} = n_c v_{th} \sigma (N_T - n_t) - n_t e_n \tag{5.1}$$

where, n_t represents the trapped charge density, N_T the total trap density, n_c the concentration of carriers in the conduction band, v_{th} the thermal velocity, σ the neutral trap capture cross section and e_n the emission coefficient. The latter is computed considering the loss by trap-to-band tunneling toward the gate or the substrate [116] and the Poole-Frenkel emission to the silicon nitride conduction band [116]:

$$e_n = \nu_{PF} \exp\left(-q\frac{E_T - \sqrt{\frac{qF}{\pi\epsilon}}}{KT}\right) + \nu_T (T_{sub} + T_{gate} - 2T_{sub}T_{gate}) \tag{5.2}$$

with F the modulus of the electric field, E_T the trap energy, K the Boltzmann constant, T the absolute temperature, ϵ the SiN optical permittivity, T_{sub} the transmission probability from the traps toward the substrate, and T_{gate} the one toward the gate; ν_{PF} and ν_T are the emission frequencies for Poole-Frenkel emission and tunneling emission, respectively. A gaussian energy profile is used for traps in the silicon nitride, with the same parameters extracted in [67] for the same type of stoichiometric silicon nitride.

Between capturing events, carriers in the nitride band can be vertically emitted by tunneling through the tunnel or the blocking oxide. Moreover, carriers can migrate laterally along the nitride by a charge concentration driven diffusion mechanism, as already discussed in previous works in which electrostatic force microscopy (EFM) analyses were carried out [117, 118]:

$$\frac{\partial n}{\partial t} = -D_n \frac{\partial^2 n}{\partial x^2} \tag{5.3}$$

To achieve better numerical stability, lateral diffusion is applied both to trapped and emitted carrier, by implementing a temperature dependent diffusion coefficient [117]:

$$D_n = D_{n0} e^{-q\frac{E_A}{KT}} \tag{5.4}$$

where the activation energy $E_A = E_T$ for the trapped carriers, while $E_A = 0$ for charge in the silicon nitride conduction band. Eq. 5.3 is

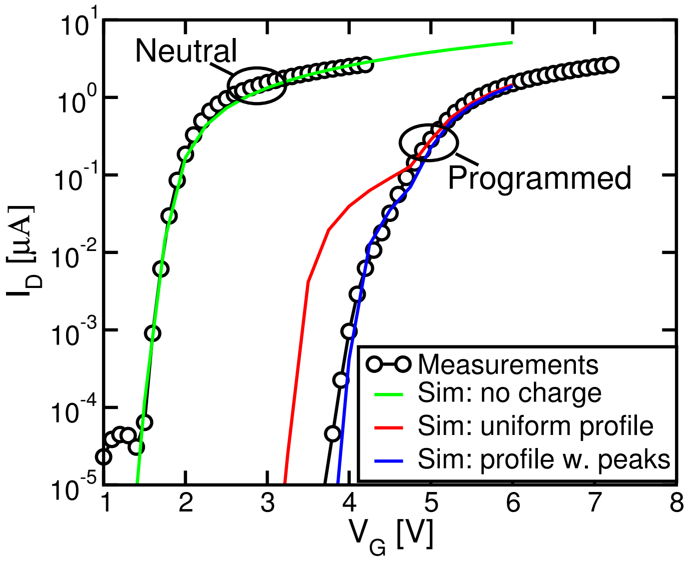


Figure 5.6: Validation of the proportionality of measured I_D-V_G curves (circles) with simulated Q_{inv}-V_G curves (solid lines) in the neutral and in the programmed state.

solved in the x direction only, by adopting Neumann boundary conditions for the charge, i.e., by forcing zero flux at the border of the nitride (230 nm on the left side and 10 μm on the right). The diffusivity D_{n0} is instead directly extracted through fitting of measured data.

The simulated structure is bi-dimensional, considering only the vertical direction, along the y axis, and the lateral direction, along the x axis (measurements in Fig. 5.2 and 5.4 showed negligible dependence of retention on L). To compare experimental and modeling results, Q_{inv}-V_G curves have been computed, by integrating the inversion charge calculated by Sentaurus Device in the active area of the device. Indeed, given the low drain voltage of the measurements (100 mV), Q_{inv}-V_G are expected to be proportional to I_D-V_G representing the measured data from which device V_T is extracted. Fig. 5.6 shows good proportionality of simulated Q_{inv}-V_G and measured I_D-V_G when the cell is in the neutral state, while a strong deformation of the simulated Q_{inv}-V_G (Fig. 5.6, red curve) is observed at the programmed state by using a constant trapped charge profile along x and resulting in a clear hump close to the device V_T. A similar hump is also observed in the measured I_D-V_G curves in the programmed state, but of a much smaller magnitude. Simulations of the electric field allowed to identify the origin of such hump: at the STI edges of the device the electric field is enhanced by corner effects, hence

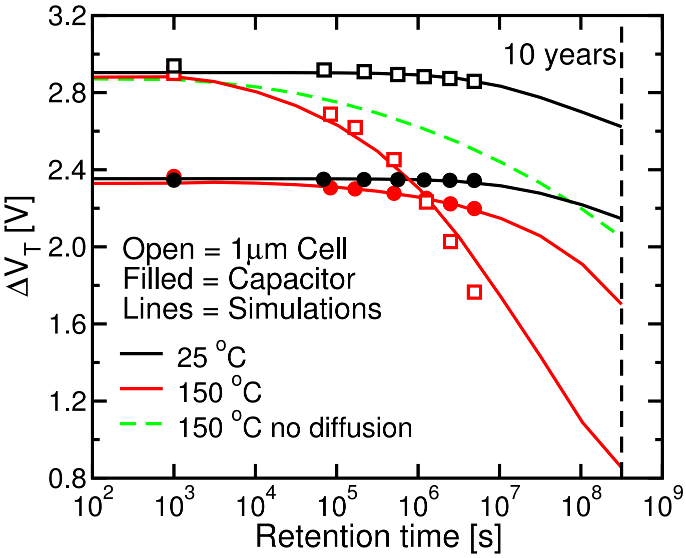


Figure 5.7: Achieved fitting of measurements (symbols) through the developed model (lines) of RT and 150°C retention for capacitors (filled circles) and 1 μm cells (open squares).

the channel in that region is created at lower V_G, ultimately resulting in a structure with two side transistors with lower V_T in parallel with the main one [77]. At the same time the enhanced electric field is also present during program operation, resulting in a higher trapped charge concentration above the side transistors. The electric field with a program gate bias of 14 V, the same used in the experiments, was simulated and implemented an initial charge profile including two charge peaks at the side of the device, extending along x in agreement with the simulated electric field enhancement. These peaks increase the V_T of the two side transistors, yielding a much better agreement with the measured I_D-V_G curves (Fig. 5.6, blue curve). In both measurement and simulation a large enough I_D was taken, so that the V_T takes into account both the main and the side transistors.

Fig. 5.7 shows the fitting of the experimental results for capacitors (circles) and 1 μm cells (squares), as obtained from the developed model at RT and 150 °C. As capacitors do not suffer from lateral charge migration, they were used to tune the model parameters related to the vertical loss, *i.e.*, trap-to-band tunneling at RT and Poole-Frenkel emission at 150 °C. Good fitting was obtained with parameters similar to those already used for the study performed in [116] on TANOS devices. Table 5.1 lists the numerical values for the different parameters used to

N_T	6×10^{19} cm^{-3}	ν_{PF}	2×10^8 Hz
ν_T	1×10^8 Hz	E_T	1.65 eV
σ	1×10^{-14} cm^2	D_{n0}	4.4×10^3 cm^{2}s^{-1}

Table 5.1: Parameters used in the model

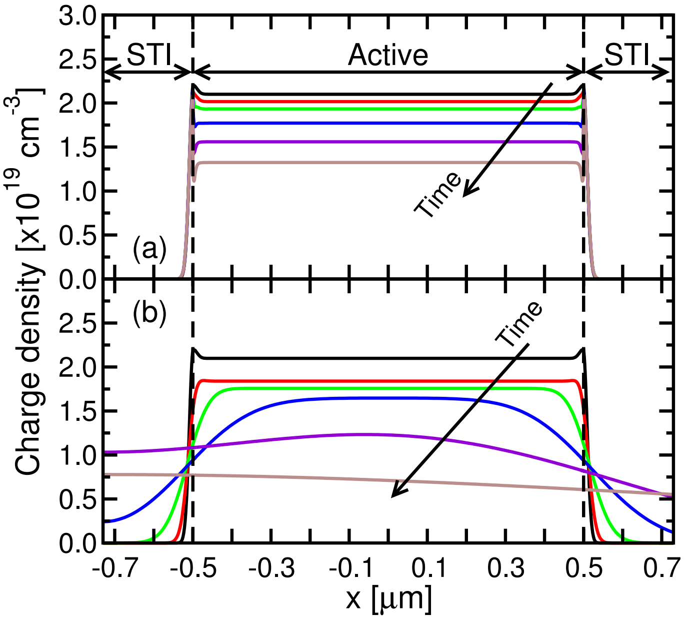


Figure 5.8: Simulated lateral charge profile evolution during the retention transient at 150 °C while (a) not considering and (b) considering the lateral charge migration.

fit all the measurements. Modeling results for the cells were obtained by maintaining the same parameters for the vertical loss of the capacitors and by tuning D_{n0}, that regulates the lateral charge redistribution. It is interesting to notice that cell fitting at RT (black curve in Fig. 5.7) can be obtained without the contribution of the lateral charge migration, which is predicted to be negligible at RT by the developed model. At 150 °C, instead, the lateral charge migration component is much more important and, if neglected, the simulation would result in the green dashed curve of Fig. 5.7, which is clearly not in agreement with the experiments.

Fig. 5.8 shows the lateral evolution of the trapped charge profile during the retention transients at 150 °C on the 1 μm cells, either considering or not the contribution of the lateral charge migration. As

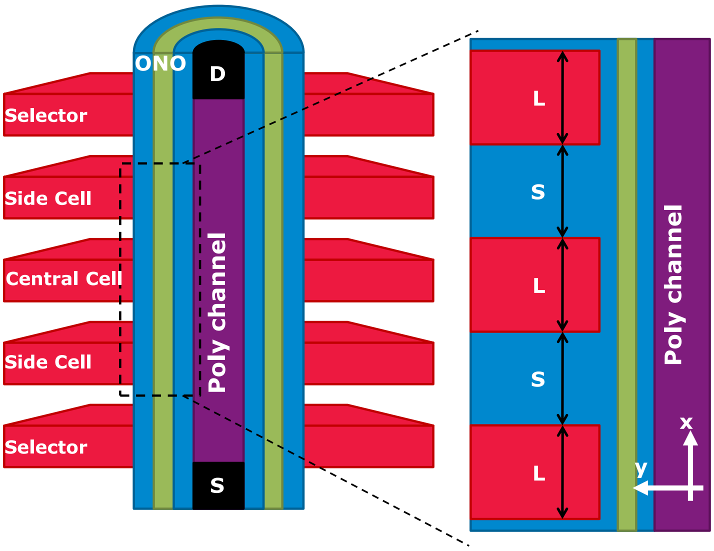


Figure 5.9: Schematic representation of the simulated 3D cylindrical structure; L is the cells length, while S is spacing between the gates.

explained above, the two small charge peaks at the corner of the STI in the initial distributions result from the enhanced electric field at the edge of the active area. When lateral diffusion of charge is neglected (Fig. 5.8(a)), the charge profile is preserved along x, being just scaled in amplitude. Lateral charge diffusion (Fig. 5.8(b)), instead, severely affects the trapped charge profile up to the center of the device and results in an almost complete charging of the left corner of the silicon nitride (which is only 230 nm wide) and in the spread of the charge up to 1 μm from the right STI corner (where the nitride is longer than 10 μm, not shown).

5.4 Analysis of 3D SONOS arrays

The model developed in the previous section was extended for investigating lateral charge migration in 3D SONOS devices. A portion of a vertical memory string is implemented in Sentaurus Device, with the cross-section schematically illustrated in Fig. 5.9. The cylindrical geometry is simulated by computing the solution of the Poisson equation in cylindrical coordinates, through rotation of the structure around the

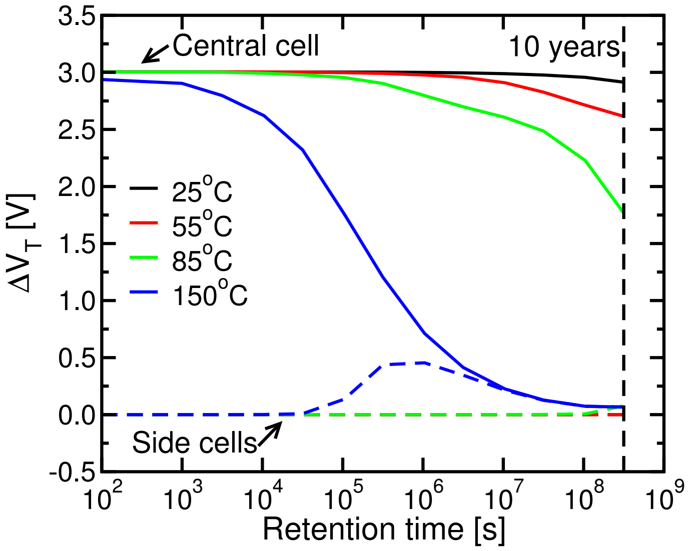


Figure 5.10: Simulated retention on the $L = 100$ nm devices at different temperatures, assuming the central cell programmed and the side cells unprogrammed.

symmetry axis passing through the center of the channel. The outcome is a portion of a BiCS-like structure [113], consisting of three adjacent cells. Different channel lengths were considered, from $L = 100$ nm to $L = 25$ nm, while maintaining a gate spacing $S = 50$ nm. The gate stack consists of 4 nm tunnel oxide, 5 nm trapping layer and 5 nm blocking oxide. The diameter of the inner undoped polysilicon substrate is 40 nm and the gate is heavily p$^+$-doped polysilicon. Such a device would have an equivalent feature size $F_{EQ} \simeq 17$ nm [41, 113] assuming 16 cells in a string and single level cell operation. As the lateral charge loss in such a structure is oriented along the x axis, Source and Drain junctions were implemented at the end of the polysilicon substrate with two select transistor partially overlapped with them. I_D-V_G curves are simulated applying 500 mV at the drain, and the V_T is extracted at a fixed I_D value. Unlike the planar case, in the vertical structure the substrate is continuous and the gate is cut: for this reason the tunneling probability during program is constant in the central region of the cell but decreases near its corners. To take this effect into account, a lower initial charge was used in the simulations at the cell corners, matching the field profile during programming.

Fig. 5.10 reports simulation results for data retention assuming $L = 100$ nm when the central cell is programmed to 3 V and the side cells are in the neutral (uncharged) state. The simulation has been performed at

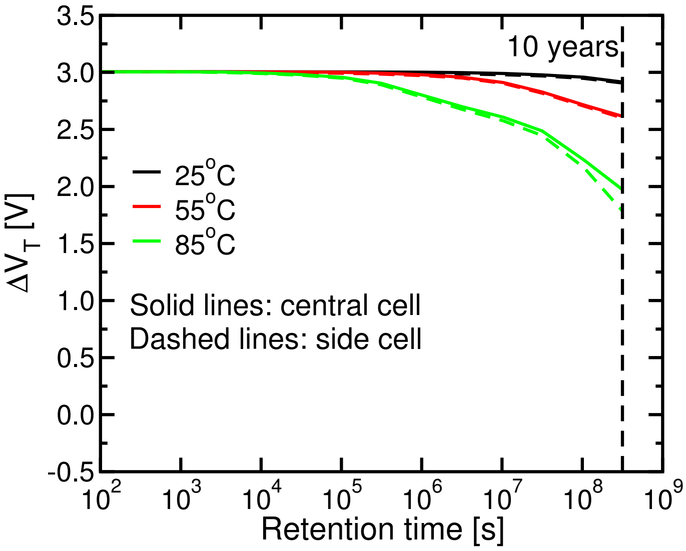


Figure 5.11: Simulated retention on the $L = 100$ nm devices at different temperatures, assuming both the central and the side cells in the programmed state.

four different temperatures: room temperature (RT), 55, 85 and 150 °C, where the last temperature represents the typical specification for automotive applications, and 55 or 85 °C are instead typical for consumer applications. It is immediately clear from Fig. 5.10 that the presence of lateral charge migration compromises the functionality of 3D SONOS at 150 °C, leading to an excessive charge loss from the programmed cell and also to an important disturb for adjacent unprogrammed cells. Results at 85 °C reveal a much smaller loss for the central cell, which may be still considered acceptable as V_T lowers by 40% of its initial value in 10 years. With the proposed spacing (50 nm), the impact on the neighboring cells is almost negligible, indicating that laterally migrated charge cannot reach the side cells. Simulations at even lower temperatures show that a charge loss below 10% affects the central cell, and obviously no relevant impact on the side cells is observed.

Fig. 5.11 shows the simulation results obtained still with $L = 100$ nm, but with all three cells programmed to 3 V. Dashed curves represent the retention transient for the side cells, while solid lines are for the central cell. The loss is almost identical in all cases, being slightly higher only at 85 °C for the side cells. The reason can be found in Fig. 5.12, where the charge profile evolution with retention time along the x direction is compared between the case of side cells neutral (Fig. 5.12(a)) and the case of side cells programmed (Fig. 5.12(b)): in the latter, each

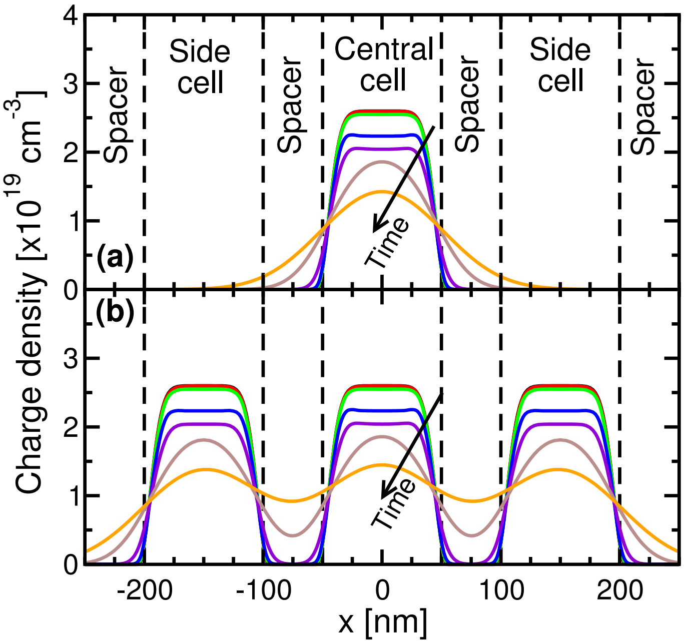


Figure 5.12: Evolution of the lateral charge profile during retention transient at 85 °C for the case of (a) neutral side cells and (b) programmed side cells.

spacing is completely filled by the charge coming from the surrounding programmed cells, resulting in an overall slowdown of the lateral diffusion for longer retention times. This is true mainly for the central cell that has a programmed cell on both sides, while the side ones can still discharge more toward the string edge. Therefore, the most critical case for retention is a programmed cell surrounded by two neutral ones.

Fig. 5.13 reports retention simulations on the programmed central cell at 55 °C assuming neutral side cells and for channel length ranging from 100 to 25 nm, keeping the cell spacing fixed at 50 nm. Smaller L are beneficial for scaling, allowing to stack more layers for the same string height, which is limited by the etching capability of the used technology. The performed test confirms that even for L = 25 nm the charge loss is quite limited and less than 700 mV V_T loss is observed. Therefore, 55 °C appears to be the optimal maximum operational temperature to provide good scaling of 3D SONOS devices.

As shown in Fig. 5.12, operations at 85 °C results in severe lateral migration, with subsequent charging of the silicon nitride in the spacing between the gates. This charge cannot be easily removed during erase operations, because of the reduced gate control in the spacers, hence it

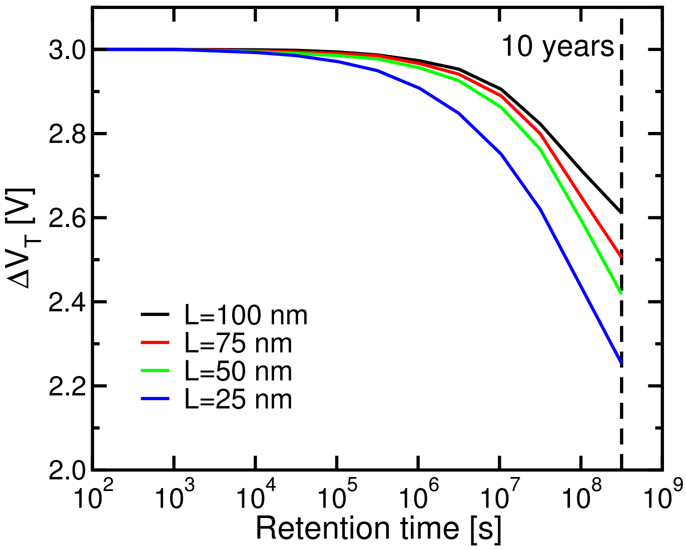


Figure 5.13: Simulated retention at 55 °C for devices featuring different channel length, assuming the central cell in the programmed state and the side cells neutral.

is prone to remain trapped and to keep on accumulating with time. The effect of this charge is detrimental for channel resistivity, as it masks the fringing field created by side gate during read operation, required to invert the portion of the substrate in correspondence of the spacers.

The impact of this charge on the simulated I_D-V_G curves for the devices of Fig. 5.12 was also investigated. Five lateral charge profiles (labeled 1 to 5), summarized in Fig. 5.14, were analyzed, representing the initial programmed state (1), the state after 3 years retention (2) and 10 years retention (3) assuming neutral side cells; the case of 3 years (4) and 10 years retention (5) in the case of programmed side cells were also included. The inset of Fig. 5.15, shows that the analyzed cases corresponds (in order) to increasing total charge trapped in the cell spacing. Fig. 5.15 reports the corresponding linear I_D-V_G curves (translated in order to have the same V_T), showing a clear degradation of the series resistance of the channel with increasing charge in the spacing between the gates. The degradation is small in subthreshold regime (not shown), which was used to extract the V_T shifts, with the only exception of case (5), where the charge difference between spacing and active areas is small and the I_D-V_G distortion is very pronounced. It is therefore clear that the charge stored between cells has a major impact on the total resistance of the string, and can even hamper proper extraction of the V_T of the cells in the string.

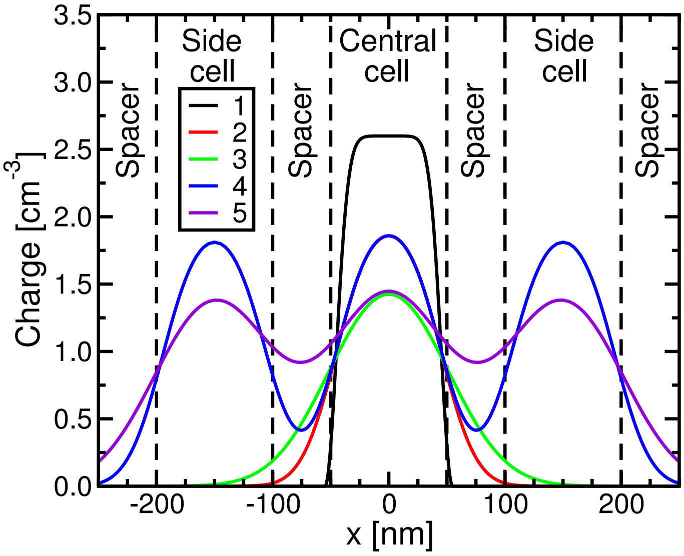


Figure 5.14: Lateral charge profile after 85 °C retention. Legend as follows: (1) Initial programmed state for the case of neutral side cells, (2)/(4) after 3 years and (3)/(5) after 10 years assuming side cells neutral/programmed.

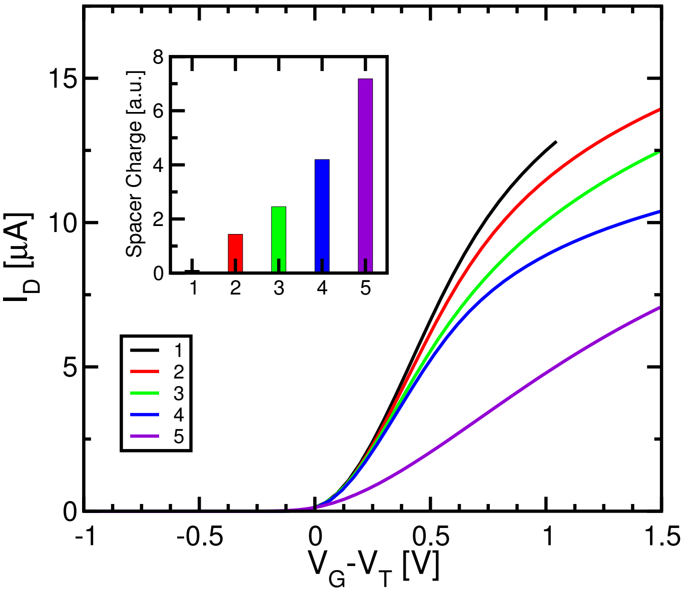


Figure 5.15: I_D-V_G curves distorsion due to the presence of charge trapped in the spacing between the gates. Inset shows the integral of the total charge stored in the nitride region between the gates. Legend as in Fig. 5.14.

5.5 Conclusions

This chapter investigated the impact of lateral charge migration on the retention transients of charge-trap memories. Experimental results on planar test structures revealed a major impact of lateral charge loss when the silicon nitride layer is not cut at the active area edges. These results were used to calibrate a new simulation tool including the vertical charge loss toward the gate and the substrate and the lateral diffusion of charge along the nitride. Simulation results were then used to predict the impact of lateral charge migration on 3D SONOS devices, finding a minimum L of 100 nm required for cells to assure sufficiently long data retention at 85 °C, whereas at 55 °C, L can be scaled at least down to 25 nm without impacting too much the retention. 85 °C are anyway critical as they may result into large charge stored in the nitride in the spacing between the cells after long retention times, increasing the string resistance and eventually hampering the proper detection of the device V_T.

Summary of results

The present work contribute to get a deeper insight on the performance and limitations of charge-trap (CT) devices for non-volatile memory technologies: the experimental characterization and the modeling efforts were focused on understanding the peculiarities, the challenges and the opportunities offered by CT technologies, on both planar and 3D cylindrical structures. Different models were developed, with different complexity and also aiming to understand, reproduce and predict different aspects of the cells, from large area planar devices, to deca-nanometer cells, to cylindrical structures.

A simple first order 1D analytical model was first developed that, despite the various approximation made, allowed to understand the fundamental parameters ruling the programming transient of SONOS cells. The approximations made not only limit the precision of the simulator itself, but also limit its application to only the Fowler-Nordheim regime of the programming phase; in order to overcome these limitations, a more complete numerical model was developed, to include more in detail the various physical processes affecting the program, erase and retention transients of the SONOS devices. The introduction of the alumina layer in the TANOS devices to avoid the erase saturation problem of the SONOS, also brings with it the non-idealities of this layer: it was shown that thick alumina layers are not trap-free, and this impacts the program, erase and retention transients, that cannot be fully described by only changing the dielectric constant of the blocking oxide layer, and considering it as a ideal insulator, like it is done for the SONOS cells. A careful experimental characterization allowed to understand the impact of this layer, and to include it in the model previously developed.

The Incremental Step Pulse Programming (ISPP) was then analyzed on CT devices: this programming scheme is the most widely used in FG Flash together with a verify operation after each programming step, for it allows to obtain tight V_T distributions, mandatory in case of Multi-Level Cells (MLC). The ISPP was studied on SONOS devices, in order to explain the widely reported non-ideal programming efficiency, with respect to the FG devices. An analysis on large area SONOS devices, allowed to understand that the decrease in the programming efficiency comes from the progressive filling of the traps in the nitride layer: this leads to an increase of the electric field in the tunnel oxide, and also of the injected current as the programming proceeds that can lead to reliability problems, not present in FG devices, in which the electric field and the tunneling current remains constant throughout the transient. Using the results obtained on the SONOS devices, the analysis was then extended to TANOS cells, attributing the differences between the simulations and the experimental results to the non-idealities of the alumina layer. In literature it is often reported that the programming efficiency in scaled CT devices can be as low as $0.5 \div 0.6$ V/V, and only part of it can be explained with the results previously obtained: to better understand the programming efficiency on deca-nanometer devices, a 3D model was used and allowed to attribute the further reduction in the efficiency to the fringing field effect that happens in very small devices.

The CT technology is one of the most promising technologies, also because it can be easily extended to vertical cylindrical architectures, allowing to develop a 3D structure that, taking advantage of the curvature effect, allows a further increase in the storage density. A cylindrical model was developed in order to understand and analyze the program, erase and retention transients: also in this case the model was tuned with experimental data, and a parametric analysis has been carried on, in order to predict the substrate diameter and the ONO thickness that can fulfill the requirements for program erase and retention of a non-volatile memory.

One of the characteristics of vertical cylindrical technologies is to have a continuous nitride: the trapping layer is not cut at the cell's edges, but extends also in the spacer between the word planes. This creates a path for the charge to diffuse, and worsens the retention transient. In order to understand the impact of the lateral charge migration, a study on planar devices with nitride layer uncut was performed by measuring cells with different form factors and by comparing the results obtained on large area capacitors, in which the nitride is cut at the edges of the active area. The results showed an important contribution of the lateral charge diffusion at high temperature; this was modeled with a 2D simulator, that included the possibility for the charge to diffuse in the region above the STI. The model was then extended to cylindrical structures to study the impact of lateral charge migration in this architecture: a study of the operating temperature limitation due to retention constraints was presented, also analyzing the impact of the migrating charge on the neighboring cells V_T and on the string resistance.

Bibliography

[1] S. Lai, “Non-volatile memory technologies: The quest for ever lower cost,” in *IEDM*, 2008, pp. 1–6.

[2] D. C. Guterman, I. Rimawi, R. Halvorson, and D. McElroy, “An electrically alterable nonvolatile memory cell using a floating-gate structure,” *IEEE Journal of Solid-State Circuits*, vol. 14, no. 2, pp. 498 – 508, 1979.

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

[4] D. Ielmini, “Overview of modeling approaches for scaled non volatile memories,” in *Proc. SISPAD*, 2009, pp. 9–15.

[5] J. V. Houdt, “Charge-based nonvolatile memory: Near the end of the roadmap?” *Current Applied Physics*, vol. 11, no. 2, Supplement 1, pp. e21 – e24, 2011.

[6] N. Mielke, “NVM reliability,” in *IRPS Tutorial*, 2008.

[7] 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.

[8] 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.

[9] ——, “A statistical model for SILC in Flash memories,” *IEEE Trans. Electron Devices*, vol. 49, pp. 1955–1961, 2002.

[10] 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.

[11] 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.

[12] 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*, 2008, pp. 610–615.

[13] D. Ielmini, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita, and C. Gerardi, “A new channel percolation model for V_T shift in discrete-trap memories,” in *Proc. IRPS*, 2004, pp. 515–521.

[14] S. M. Amoroso, A. Maconi, A. Mauri, C. Monzio Compagnoni, E. Greco, E. Camozzi, S. Viganò, P. Tessariol, A. Ghetti, A. S. Spinelli, and A. L. Lacaita, “3D Monte Carlo simulation of the programming dynamics and their statistical variability in nanoscale charge-trap memories,” in *IEDM Tech. Dig.*, 2010, pp. 540–543.

[15] 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.

[16] 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.

[17] 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.

[18] K. Prall, “Scaling non-volatile memory below 30 nm,” in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 5–10.

[19] P. Blomme, M. Rosmeulen, A. Cacciato, M. Kostermans, C. Vrancken, 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. on VLSI Technology*, 2010, pp. 129–130.

[20] R. Degraeve, F. Schuler, B. Kaczer, M. Lorenzini, D. Wellekens, P. Hendrickx, M. van Duuren, G. Dormans, J. V. Houdt, L. Haspeslagh, G. Groeseneken, and G. Tempel, "Analytical percolation model for predicting anomalous charge loss in flash memories," *IEEE Transactions on Electron Devices*, vol. 51, no. 9, pp. 1392 – 1400, 2004.

[21] 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.

[22] 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*, 2003, pp. 703–714.

[23] A. Sheikholeslami and P. Gulak, "A survey of circuit innovations in ferroelectric random-access memories," *Proceedings of the IEEE*, vol. 88, no. 5, pp. 667–689, may 2000.

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

[25] I. Baek, M. Lee, S. Seo, M. Lee, D. Seo, D.-S. Suh, J. Park, S. Park, H. Kim, I. Yoo, U.-I. Chung, , and J. Moon, "Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses," 2004, pp. 587 – 590.

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

[27] H. Lue, S. Lai, T. Hsu, Y.-H. Hsiao, P.-Y. Du, S.-Y. Wang, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "A critical review of charge-trapping NAND flash devices," in *9th International Conference on Solid-State and Integrated-Circuit Technology*, 2008, pp. 807 –810.

[28] M. White, D. Adams, and J. Bu, "On the go with SONOS," *IEEE Circuits and Devices Magazine*, vol. 16, no. 4, pp. 22 –31, 2000.

[29] 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.

[30] J. Bu and M. H. White, "Design considerations in scaled SONOS nonvolatile memory devices," *Solid-State Electron.*, vol. 45, pp. 113–120, 2001.

[31] ——, "Retention reliability enhanced SONOS NVSM with scaled programming voltage," in *Aerospace Conference Proceedings*, 2002, pp. 5–2383 – 5–2390.

[32] R. van Schaijk, M. van Duuren, P. Goarin, W. Mei, and K. van der Jeugd, "Reliability of embedded SONOS memories," in *Proc. of ESSDERC*, 2004, pp. 277 – 280.

[33] 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.*, 2003, pp. 613–616.

[34] 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.*, 2008, pp. 555–558.

[35] V. Gritsenko, "Design of SONOS memory transistor for terabit scale EEPROM," in *IEEE Conference on Electron Devices and Solid-State Circuits*, 2003, pp. 345 – 348.

[36] A. Padovani, L. Larcher, V. D. Marca, P. Pavan, H. Park, and G. Bersuker, "Charge trapping in alumina and its impact on the operation of metal-alumina-nitride-oxide-silicon memories: Experiments and simulations," *Journal of Appl. Physics*, vol. 110, no. 1, p. 014505, 2011.

[37] C. Hobbs, L. Fonseca, V. Dhandapani, S. Samavedam, B. Taylor, J. Grant, L. Dip, D. Triyoso, R. Hegde, D. Gilmer, R. Garcia, D. Roan, L. Lovejoy, R. Rai, L. Hebert, H. Tseng, B. White, and P. Tobin, "Fermi level pinning at the polySi/metal oxide interface," in *Digest of Tech. Papers. Symp on VLSI Technology*, 2003, pp. 9 – 10.

[38] Y. Hsiao, H. Lue, T. Hsu, K. Hsieh, and C. Lu, “A critical examination of 3D stackable NAND Flash memory architectures by simulation study of the scaling capability,” in *IMW*, 2010, pp. 1 – 4.

[39] J. Kim, A. J. Hong, S. M. Kim, E. B. Song, J. H. Park, J. Han, S. Choi, D. Jang, J.-T. Moon, and K. L. .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.*, 2009, pp. 186–187.

[40] 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.*, 2009, pp. 188–189.

[41] Y. Fukuzumi, R. Katsumata, M. Kito, M. Kido, M. Sato, H. Tanaka, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, “Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable Flash memory,” in *IEDM Tech. Dig.*, 2007, pp. 449–452.

[42] 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.*, 2009, pp. 136–137.

[43] 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.*, 2009, pp. 192–193.

[44] E. Nowak, A. Hubert, L. Perniola, T. Ernst, G. Ghibaudo, G. Reimbold, B. D. Salvo, and F. Boulanger, “In-depth analysis of 3D silicon nanowire SONOS memory characteristics by TCAD simulations,” in *IMW Tech. Dig.*, 2010, pp. 116–119.

[45] 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*, 2010, pp. 966–969.

[46] M. Specht, H. Reisinger, F. Hofmann, T. Schulz, E. Landgraf, R. J. Luyken, W. Rosner, M. Grieb, and L. Risch, “Charge trapping memory structures with Al_2O_3 trapping dielectric for high-temperature applications,” *Solid-State Electron.*, vol. 49, pp. 716–720, 2005.

[47] A. Kerber, E. Cartier, R. Degraeve, P. Roussel, L. Pantisano, T. Kauerauf, G. Groeseneken, H. Maes, and U. Schwalke, “Charge trapping and dielectric reliability of SiO2-Al2O3 gate stacks with TiN electrodes,” *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1261 – 1269, 2003.

[48] G. V. den bosch, A. Furnemont, M. Zahid, R. Degraeve, L. Breuil, A. Cacciato, A. Rothschild, C. Olsen, U. Ganguly, and J. V. Houdt, “Nitride engineering for improved erase performance and retention of TANOS NAND Flash memory,” in *Non-Volatile Semiconductor Memory Workshop, 2008 and 2008 International Conference on Memory Technology and Design. NVSMW/ICMTD 2008. Joint*, 2008, pp. 128–129.

[49] S.-Y. Wang, H.-T. Lue, P.-Y. Du, C.-W. Liao, E.-K. Lai, S.-C. Lai, L.-W. Yang, T. Yang, K.-C. Chen, J. Gong, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, “Reliability and processing effects of bandgap-engineered SONOS (BE-SONOS) Flash memory and study of the gate-stack scaling capability,” *IEEE Trans. Electron Devices*, vol. 8, pp. 416–425, June 2008.

[50] Y. Park, J. Choi, C. Kang, C. Lee, Y. Shin, B. Choi, J. Kim, S. Jeon, J. Sel, J. Park, K. Choi, T. Yoo, J. Sim, and K. Kim, “Highly manufacturable 32Gb Multi-Level NAND Flash memory with 0.0098 μm^2 cell size using TANOS (Si-Oxide-Al_2O_3-TaN) cell technology,” in *IEDM Tech. Dig.*, 2006, pp. 29–32.

[51] A. Paul, C. Sridhar, and S. Mahapatra, “Comprehensive simulation of program, erase and retention in charge trapping Flash memories,” in *IEDM Tech. Dig.*, 2006, pp. 393–396.

[52] A. Furnemont, M. Rosmeulen, A. Cacciato, L. Breuil, K. De Meyer, H. Maes, and J. Van Houdt, “Physical understanding of

SANOS disturbs and VARIOT engineered barrier as a solution," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 94–95.

[53] Y. C. Song, X. Y. Liu, Z. Y. Wang, K. Zhao, G. Du, Z. L. Xia, D. Kim, and K.-H. Lee, "Evaluating the effects of physical mechanism on program, erase and retention in charge trapping memory," in *Proc. SISPAD*, 2008, pp. 41–44.

[54] 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.

[55] E. Vianello, M. Bocquet, F. Driussi, L. Perniola, G. Molas, and L. Selmi, "Program efficiency and high temperature retention of SiN/high-K based memories," *Microelectronic Engineering*, vol. 86, no. 7–9, pp. 1830 – 1833, 2009.

[56] A. Mauri, S. M. Amoroso, C. Monzio Compagnoni, A. Maconi, and A. S. Spinelli, "Comprehensive numerical simulation of threshold-voltage transients in nitride memories," *Solid-State Electron.*, vol. 56, pp. 23–30, 2011.

[57] W. Shockley and W. T. Read, "Statistics of recombination of holes and electrons," *Phys. Rev.*, vol. 87, pp. 835–842, 1952.

[58] P. C. Arnett, "Transient conduction in insulators at high fields," *J. Appl. Phys.*, vol. 46, pp. 5236–5243, 1975.

[59] T. Tomita, Y. Kamakura, and K. Taniguchi, "Energy relaxation length for ballistic electron transport in SiO_2," *Phys. Stat. Sol.*, vol. 204, pp. 129–132, 1997.

[60] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, pp. 278–283, 1969.

[61] Y. L. Chiou, J. F. Gambino, and M. Mohammad, "Determination of the Fowler-Nordheim tunneling parameters from the Fowler-Nordheim plot," *Solid-State Electron.*, vol. 45, pp. 1787–1791, 2001.

[62] A. Arreghini, F. Driussi, D. Esseni, L. Selmi, M. J. van Duuren, and R. van Schaijk, "Experimental extraction of the charge centroid and of the charge type in the P/E operation of SONOS memory cells," in *IEDM Tech. Dig.*, 2006, pp. 499–502.

[63] M.Peterson and Y.Roizin, "Density functional therory study of deep traps in silicon nitride memories," *Applied Physics Letters*, vol. 89, p. 053511, 2006.

[64] V. K. Arora, "Drift diffusion and einstein relation for electrons in silicon subjected to a high electric field," *Applied Physics Letters*, vol. 80, pp. 3763–3765, 2002.

[65] L. Landau and E. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory. Vol. 3.* Pergamon Press., 1977.

[66] J. Frenkel, "On pre-breakdown phenomena in insulators and electronic semi-conductors," *Phys. Rev.*, 1938.

[67] A. Suhane, A. Arreghini, R. Degraeve, G. V. den bosch, L. Breuil, M. Zahid, M. Jurczak, K. D. Meyer, and J. V. Houdt, "Validation of retention modeling as a trap-profiling technique for SiN-based charge-trapping memories," *Electron Device Letters, IEEE*, vol. 31, no. 1, pp. 77 –79, 2010.

[68] D.Scharfetter and H.Gummel, "Large-signal analysis of a silicon read diode oscillator," *Transaction on Electron Devices*, vol. 16, pp. 64–77, 1969.

[69] K. Wu, H. Chien, T. Tsai, J. Chang, C. Chan, T. Chen, C. Kao, and C. Chien, "Phenomenal SONOS performance for next-generation Flash memories," in *Proc. of Symposium on nano device technology*, 2004, pp. 35–40.

[70] P. Arnett and B. Yun, "Silicon nitride trap properties as revealed by charge-centroid measurements on mnos devices," *Applied Physics Letters*, vol. 26, no. 3, pp. 94–96, feb 1975.

[71] A. Arreghini, F. Driussi, D. Esseni, L. Selmi, M. van Duuren, and R. van Schaijk, "New charge pumping model for the analysis of the spatial trap distribution in the nitride layer of sonos devices," *Microelectronic Engineering*, vol. 80, no. 0, pp. 333–336, 2005.

[72] E. Vianello, M. Bocquet, F. Driussi, L. Perniola, G. Molas, and L. Selmi, "Program efficiency and high temperature retention of sin/high-k based memories," *Microelectronic Engineering*, vol. 86, no. 7–9, pp. 1830–1833, 2009.

[73] 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.*, 1995, pp. 129–130.

[74] 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.

[75] 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.*, 2008, pp. 831–834.

[76] 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*, 2008, pp. 693–694.

[77] H.-T. Lue, T.-H. Hsu, Y.-H. Hsiao, S.-C. Lai, E.-K. Lai, S.-P. Hong, M.-T. Wu, F. H. Hsu, N. Z. Lien, C.-P. Lu, S.-Y. Wang, J.-Y. Hsieh, L.-W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "Understanding STI edge fringing field effect on the scaling of charge-trapping (CT) NAND Flash and modeling of incremental step pulse programming (ISPP)," in *IEDM Tech. Dig.*, 2009, pp. 839–842.

[78] 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.

[79] A. Suhane, A. Arreghini, G. Van den bosch, L. Breuil, A. Cacciato, A. Rothschild, M. Jurczak, J. Van Houdt, and K. De Meyer, "Experimental evaluation of trapping efficiency in silicon nitride based charge trapping memories," in *Proc. ESSDERC*, 2009, pp. 276–279.

[80] H.-T. Lue, T.-H. Hsu, S.-Y. Wang, Y.-H. Hsiao, E.-K. Lai, L.-W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "Study of local trapping and STI edge effects on charge-trapping NAND Flash," in *IEDM Tech. Dig.*, 2007, pp. 161–164.

[81] M. F. Beug, T. Melde, M. Czernohorsky, R. Hoffmann, J. Paul, R. Knoefler, and A. T. Tilke, "Analysis of TANOS memory cells

with sealing oxide containing blocking dielectric," *IEEE Trans. Electron Devices*, vol. 57, pp. 1590–1596, July 2010.

[82] S.-H. Ku, K.-F. Chen, L.-H. Chong, Y.-J. Chen, T.-H. Yeh, S.-W. Lin, T.-T. Han, N.-K. Zous, I. Huang, M.-S. Chen, W.-P. Lu, K.-C. Chen, and C.-Y. Lu, "Investigation of geometric effect impact on SONOS memory in a NAND array structure," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 315–324, 2011.

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

[84] A. Maconi, C. Monzio Compagnoni, S. M. Amoroso, E. Mascellino, M. Ghidotti, G. Padovini, A. S. Spinelli, A. L. Lacaita, A. Mauri, G. Ghidini, N. Galbiati, A. Sebastiani, C. Scozzari, E. Greco, E. Camozzi, and P. Tessariol, "Investigation of the ISPP dynamics and of the programming efficiency of charge-trap memories," in *Proc. ESSDERC*, 2010, pp. 444–447.

[85] 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.

[86] 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.

[87] 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.

[88] 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.*, 2007, pp. 165–168.

[89] 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.

[90] 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.*, 1993, pp. 705–708.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] E.-S. Choi, H.-S. Yoo, K.-H. Park, S.-J. Kim, J.-R. Ahn, M.-S. Lee, Y.-O. Hong, S.-G. Kim, J.-C. Om, M.-S. Joo, S.-H. Pyi, S.-S. Lee, S.-K. Lee, and G.-H. Bae, "Modeling and characterization of program/erasure speed and retention of TiN-gate MANOS (Si-Oxide-SiN$_x$-Al$_2$O$_3$-Metal gate) cells for NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 83–84.

[98] A. Hubert, E. Nowak, K. Tachi, V. Maffini-Alvaro, C. Vizioz, C. Arvet, J.-P. Colonna, J.-M. Hartmann, V. Loup, L. Baud, S. Pauliac, V. Delaye, C. Carabasse, G. Molas, G. Ghibaudo, B. D. Salvo, O. Faynot, and T. Ernst, "A stacked SONOS technology, up to 4 levels and 6 nm crystalline nanowires, with gate-all-around or independent gates (ϕ-Flash), suitable for full 3D integration," in *IEDM Tech. Dig.*, 2009, pp. 637–640.

[99] J. Fu, K. D. Buddharaju, S. H. G. Teo, C. Zhu, M. B. Yu, N. Singh, G. Q. Lo, N. Balasubramanian, and D. L. Kwong, "Trap layer engineered gate-all-around vertically stacked twin Si-nanowire nonvolatile memory," in *IEDM Tech. Dig.*, 2007, pp. 79–82.

[100] K. H. Yeo, K. H. Cho, M. Li, S. D. Suk, Y.-Y. Yeoh, M.-S. Kim, H. Bae, J.-M. Lee, S.-K. Sung, J. Seo, B. Park, D.-W. Kim, D. Park, and W.-S. Lee, "Gate-all-around single silicon nanowire MOSFET with 7 nm width for SONOS NAND Flash memory," in *Symp. VLSI Tech. Dig.*, 2008, pp. 138–139.

[101] M. Chen, H. Y. Yu, N. Singh, Y. Sun, N. S. Shen, X. Yuan, G.-Q. Lo, and D.-L. Kwong, "Vertical-Si-nanowire SONOS memory for ultrahigh-density application," *IEEE Electron Device Lett.*, vol. 30, pp. 879–881, Aug. 2009.

[102] J. Fu, N. Singh, C. Zhu, G.-Q. Lo, and D.-L. Kwong, "Integration of high-k dielectrics and metal gate on gate-all-around Si-nanowire-based architecture for high-speed nonvolatile charge-trapping memory," *IEEE Electron Device Lett.*, vol. 30, pp. 662–664, Jun. 2009.

[103] E. Nowak, M. Bocquet, L. Perniola, G. Ghibaudo, G. Molas, C. Jahan, R. Kies, G. Reimbold, B. D. Salvo, and F. Boulanger, "New physical model for ultra-scaled 3D nitride-trapping non-volatile memories," in *IEDM Tech. Dig.*, 2008, pp. 559–562.

[104] E. Gnani, S. Reggiani, A. Gnudi, G. Baccarani, J. Fub, N. Singh, G. Lo, and D. Kwong, "Modeling of gate-all-around charge trapping SONOS memory cells," *Solid-State Electron.*, vol. 54, pp. 997–1002, 2010.

[105] L. Wang, D. Wang, and P. Asbeck, "A numerical Schrödinger-Poisson solver for radially symmetric nanowire core-shell structures," *Solid State Electron.*, vol. 50, pp. 1732–1739, 2006.

[106] E.-X. Ping, “I-V characteristics by radial tunneling in double-barrier tunneling diodes with cylindrical barriers,” *IEEE J. Quantum Elec.*, vol. 31, pp. 1210–1215, July 1995.

[107] ——, “Büttiker-Landauer traversal times in the radial direction of cylindrical single and double barriers,” *J. Appl. Phys.*, vol. 76, pp. 1929–1931, 1994.

[108] S. Manzini and F. Volonté, “Charge transport and trapping in silicon nitride-silicon dioxide dielectric double layers,” *J. Appl. Phys.*, vol. 58, pp. 4300–4306, 1985.

[109] J. He, Y. Tao, F. Liu, J. Feng, and S. Yang, “Analytic channel potential solution to the undoped surrounding-gate MOSFETs,” *Solid-State Electron.*, vol. 51, pp. 802–805, 2007.

[110] D. Jiménez, B. Iñíguez, J. Suñé, L. F. Marsal, J. Pallarès, J. Roig, and D. Flores, “Continuous analytic I-V model for surrounding-gate MOSFETs,” *IEEE Electron Device Lett.*, vol. 25, pp. 571–573, Aug. 2004.

[111] L. Larcher, A. Padovani, V. della Marca, P. Pavan, and A. Bertacchini, “Investigation of trapping/detrapping mechanisms in Al_2O_3 electron/hole traps and their influence on TANOS memory operation,” in *Proc. VLSI-TSA*, 2010, pp. 52–53.

[112] G. Molas, L. Masoero, P. Blaise, A. Padovani, J. P. Colonna, E. Vianello, M. Bocquet, E. Nowak, M. Gasulla, O. Cueto, H. Grampeix, F. Martin, R. Kies, P. Brianceau, M. Gely, A. M. Papon, D. Lafond, J. P. Barnes, C. Licitra, G. Ghibaudo, L. L. S. Deleonibus, and B. De Salvo, “Investigation of the role of H-related defects in Al_2O_3 blocking layer on charge-trap memory retention by atomistic simulations and device physical modelling,” in *IEDM Tech. Dig.*, 2010, pp. 536–539.

[113] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, and A. Nitayama, “Bit cost scalable technology with punch and plug process for ultra high density Flash memory,” in *Symp. VLSI Tech. Dig.*, 2007, pp. 14–15.

[114] J. S. Sim, J. Park, C. Kang, W. Jung, Y. Shin, J. Kim, J. Sel, C. Lee, S. Jeon, Y. Jeong, Y. Park, J. Choi, and W.-S. Lee, “Self aligned trap-shallow trench isolation scheme for the reliability of TANOS (TaN/AlO/SiN/Oxide/Si) NAND Flash memory,”

in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 110–111.

[115] C. Monzio Compagnoni, A. S. Spinelli, and A. L. Lacaita, "Experimental study of data retention in nitride memories by temperature and field acceleration," *IEEE Electron Device Lett.*, vol. 28, pp. 628–630, July 2007.

[116] A. Arreghini, N. Akil, F. Driussi, D. Esseni, L. Selmi, and M. van Duuren, "Long term charge retention dynamics of SONOS cells," *Solid-State Electronics*, vol. 52, no. 9, pp. 1460 – 1466, 2008.

[117] S. J. Baik, K. S. Lim, W. Choi, H. Yoo, and H. Shin, "Charge diffusion in silicon nitrides: Scalability assessment of nitride based flash memory," in *Proc. IRPS*, 2011, pp. 6B.4.1 – 6B.4.6.

[118] E. Vianello, E. Nowak, D. Mariolle, N. Chevalier, L. Perniola, G. Molas, J. Colonna, F. Driussi, and L. Selmi, "Direct probing of trapped charge dynamics in SiN by Kelvin Force Microscopy," in *IEEE International Conference on Microelectronic Test Structures (ICMTS)*, 2010, pp. 94 –97.

List of publications

A. Mauri, C. Monzio Compagnoni, S. M. Amoroso, A. Maconi, F. Cattaneo, A. Benvenuti, A. S. Spinelli, A. L. Lacaita. "A new physics-based model for TANOS memories program/erase". IEDM 2008, pp. 555–558

C. Monzio Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi, A. S. Spinelli. "Physical modeling for programming of TANOS memories in the Fowler-Nordheim regime". IEEE Transactions on electron devices, vol. 56, no. 9, pp. 2008–2015

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, G. Ghidini. "Reliability constraints for TANOS memories due to alumina trapping and leakage". IRPS 2010, pp. 966–969

A. Mauri, C. Monzio Compagnoni, S. M. Amoroso, A. Maconi, A. Ghetti, A. S. Spinelli, A. L. Lacaita. "Comprehensive investigation of statistical effects in nitride memories - Part I: Physics-based modeling". IEEE Transactions on electron devices, vol. 57, no. 9, pp. 2116–2123

C. Monzio Compagnoni, A. Mauri, S. M. Amoroso, A. Maconi, E. Greco, A. S. Spinelli, A. L. Lacaita. "Comprehensive investigation of statistical effects in nitride memories - Part II: Scaling analysis and impact on device performance". IEEE Transactions on electron devices, vol. 57, no. 9, pp. 2124–2131

A. Maconi, C. Monzio Compagnoni, S. M. Amoroso, E. Mascellino, M. Ghidotti, G. Padovini, A. S. Spinelli, A. L. Lacaita, A. Mauri, G. Ghidini, N. Galbiati, A. Sebastiani, C. Scozzari, E. Greco, E. Camozzi, P. Tessariol. "Investigation of the ISPP dynamics and of the programming efficiency of charge-trap memories". ESSDERC 2010, pp. 444–447

S. M. Amoroso, A. Maconi, A. Mauri, C. Monzio Compagnoni, E. Greco, E. Camozzi, S. Vigano', P. Tessariol, A. Ghetti, A. S. Spinelli, A. L. Lacaita. "3D Monte Carlo simulation of the programming dynamics and their statistical variability in nanoscale charge-trap memories". IEDM 2010, pp. 540–543

G. Ghidini, N. Galbiati, E. Mascellino, C. Scozzari, A. Sebastiani, S. M. Amoroso, C. Monzio Compagnoni, A. S. Spinelli, A. Maconi, R. Piagge, A. Del Vitto, M. Alessandri, I. Baldi, E. Moltrasio, G. Albini, A. Grossi, P. Tessariol, E. Camerlenghi, A. Mauri. "Charge retention phenomena in charge transfer silicon nitride: impact of technology and operating conditions". Journal of vacuum science & technology. B, vol.29, no.1, pp.01AE01–01AE01–4

A. Mauri, S. M. Amoroso, C. Monzio Compagnoni, A. Maconi, A. S. Spinelli. "Comprehensive numerical simulation of threshold-voltage transients in nitride memories". Solid-state electronics, vol. 56, no. 1, pp. 23–30

S.M. Amoroso, A. Maconi, A. Mauri, C. Monzio Compagnoni, A. S. Spinelli, A.L. Lacaita. "Three-dimensional simulation of charge-trap memory programming - Part I: Average behavior". IEEE Transactions on electron devices, vol. 58, no. 7, pp. 1864–1871

A. Maconi, S.M. Amoroso, C. Monzio Compagnoni, A. Mauri, A. S. Spinelli, A.L. Lacaita. "Three-dimensional simulation of charge-trap memory programming - Part II: Variability". IEEE Transactions on electron devices, vol. 58, no. 7, pp. 1872–1878

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

G. Van den bosch, A. Arreghini, G. S. Kar, A. Maconi, J. Van Houdt. "Scalability investigation of 3D SONOS NAND Flash by experiment and simulation". Accepted for SISC 2011

POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci, 32 I 20133 – Milano