Investigation of degradation mechanisms and related performance concerns in 40nm NOR Flash memories

Giulio Torrente

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES Spécialité : **Nanoélectronique et nanotechnologies**

Arrêté ministériel : 7 août 2006

Présentée par

Giulio TORRENTE

Thèse dirigée par **Dr. Gérard GHIBAUDO** et
codirigée par **Dr. David ROY** et **Dr. Jean COIGNUS**

préparée au sein du **STMicroelectronics, CEA-LETI et de l'IMEP-LAHC**
dans **l'École Doctorale d'Électronique, Électrotechnique, Automatique et Traitement du Signal**

Investigation of degradation mechanisms and related performance concerns in 40nm NOR Flash memories

Thèse soutenue publiquement le **11 Juillet 2017**,
devant le jury composé de :

Mireille MOUIS
DR. CNRS Alpes, Présidente

Damien DELERUYELLE
PR. INSA de Lyon, Rapporteur

Raphael CLERC
PR. Université de Saint Etienne, Rapporteur

Gérard GHIBAUDO
DR. CNRS Alpes, Directeur de thèse

David ROY
ING. STMicroelectronics, Co-Encadrant, Invité

Jean COIGNUS
ING. CEA-LETI, Co-Encadrant, Invité

UNIVERSITÉ **GRENOBLE ALPES**

Contents

General introduction

The need for storing large quantity of information and segments of executable code is continuously increasing within all semiconductor industry sectors, involving the development of modern applications and electronic devices into our day life. With the rise of social networks, we need to exchange opinions, pictures, movies and all other types of data with our friends, relatives and with the rest of the world. For this reason, we have to constantly access to our information, which is always stored in large quantities, and we want our devices to be reliable, efficient and fast. This increasing reliance of data storage and device performance made possible the technological progress and development in microelectronics domain so far, especially in semiconductor memory sector.

The semiconductor memory market expanded over the years and evolved, as new technologies and applications became available to industries and mature for the commercialization. First of all, the memory technologies are classified into two main categories: RAMs (Random Access Memories), and ROMs (Read Only Memories). The first one has the feature to be programmed in a short time period, whereas, for the second case, the device is characterized by its non-volatile property, i.e. the capability of retaining the stored data in the absence of power supply. Over the years, the need of developing a technology combining both high access performance of RAMs with the high density and nonvolatile properties of ROMs increased. In addition, it is always necessary to consider additional constraints depending on customer needs and product applications: integration capability and cost, technology scalability, low-power requirements, device performance and reliability and so on.

The main class of Non Volatile Memory (NVM) devices is the Flash memory, whose industrialization has become a \$20 billion/year market [1], [2]. In such a technology, the electrical charge represents the information and the non-volatile characteristic is achieved by storing a certain amount of carriers in an electrically isolated node, called Floating Gate (FG). Flash technology has been invented in 1980 by Fujio Masuoka [3] while working for Toshiba and has been presented at the 1984 IEEE International Electron Devices Meeting (IEDM) held in San Francisco [4]. Then, Intel Corporation saw the massive potential of the invention and introduced the first commercial chip in 1988 [5].

Flash technology is still nowadays the preferred storage memory in many portable consumer and computer applications. In particular, two main Flash technologies can be found: NAND and NOR types. Concerning the first one, the memory can be programmed and read in blocks (or pages), the architecture is characterized by high density and low cost, but without offering the random access capability. Thus, it is suitable mainly for data storage: it can be found primarily in memory cards, USB drives, solid-state drives and similar products for general storage and transfer of data. On the other hand, NOR-type Flash allows a single machine word (byte) to be written or read independently and features high speed and noise immunity. It results to be the technology of choice for embedded applications requiring non-volatile memory properties. In accordance with all these characteristics and specifications, NAND technology is programmed and erased by Fowler-Nordheim (FN) mechanism, whereas NOR technology is usually programmed by Hot Carrier (HC) process and erased via FN [6]. Nowadays, NAND technology is constantly occupying around 21% of the total market of semiconductor memories, while NOR Flash architectures represent a portion equal to 12% [7].

However, the conventional Flash memory is facing technological barriers when scaled below 20nm nodes. Firstly, the technology down-scaling leads the maximum number of electrons stored in the FG to be only few tens making reliability concerns stronger. The excessively high number of Program/Erase (P/E) cycles distorts the 1/0 threshold voltage values, leading to erroneous bit interpretations (endurance concern), and, in addition, induces a loss of cell capability to store electrical charges due to appearance of leakage conduction mechanisms through the oxide (memory retention concern). In particular, additional reductions of gate oxide thickness are made difficult below 6-8 nm by now, because of memory retention constraints. Secondly, further channel length scaling (without gate-oxide thickness scaling) causes poor Ion/Ioff cell ratio, which results not to be high enough to correctly read the cell state. Moreover, no physical space for the contacts is expected and an increasing cell-to-cell interference problem would reduce the operation efficiencies.

For all these reasons, several alternative memory device designs and new materials have been proposed to overcome both scaling limits and reliability issues for Flash memory technology. In particular, programmable resistance devices such as Phase-Change Memory (PCM) [8] and Resistive RAM (RRAM) [9] organized in a 3D cross-point memory architecture are among the most attractive successors.

Waiting for the maturity of such technologies, the Flash cell lifetime has to be optimized pushing its working condition to the intrinsic physical limit. In particular, such an optimization has to be done mainly focusing on reliability concerns, i.e. data retention and endurance, which, as previously underlined, represent the main limiting factors for technology down-scaling. For this reason, several works dealt with data retention concerns,

analyzing, characterizing and modeling the Stress Induced Leakage Current (SILC) [10], [11] with the final aim of limiting or control such an issue. However, on the other hand, there is no work which accurately explored the overall cell evolution during P/E cycling from a microscopic physical standpoint, especially in NOR technology, whose intrinsic 2D degradation nature makes complex the modeling and the analysis of the combined aging mechanisms. An accurate physical comprehension of cell degradation helps not only to find new technological solutions, but also to push the memory cell towards its maximum intrinsic reliability performance. As a matter of fact, the individual degradation mechanisms responsible for the overall cell aging, i.e. Hot Carrier Degradation (HCD) [12] and Fowler-Nordheim Stress (FNS) [13], have been widely studied in the past on transistors. However, the interaction of these processes and their effects on the cell electrostatics and on the loss of P/E operation efficiencies have never been explored.

In this manuscript, an in-depth investigation of P/E degradation mechanisms in 40nm NOR Flash technology issued from STMicroelectronics is conducted. With the help of advanced electrical characterization and proper TCAD simulation, this thesis provides an accurate understanding, evaluation and modeling of the different aging mechanisms involved during P/E cycling. In particular, the respective role of HCD and FNS is pointed out, and their impact on memory cell characteristic drifts and on memory lifetime is assessed. The main challenge is to build a physically-based model which reproduces the Flash cell wear out during P/E cycling. As already pointed out, this enables to push the memory lifetime towards its maximum intrinsic performance, as for example by correctly managing the P/E electrical operations [14], [15]. In addition, such an approach allows to assess the limiting physical mechanism factors for memory cell degradation and consequently to take action for some specific process step optimizations.

The manuscript is divided into 4 chapters:

- **Chapter 1**. NOR Flash cell working principle is firstly introduced. Then, experimental methodologies are proposed in order to properly characterize the device. Such an electrical characterization will be useful in the following, since enabling to properly compare the device structures and making possible the aging analysis. In addition, it is underlined the importance in the extraction accuracy of cell characteristics and the novelty of these approaches. Afterwards, the device model calibration and the computation of tunnel oxide current during both P/E operations are provided. Finally, the advanced experimental setup used in this thesis is presented and the advantage of using ramp-optimized patterns is highlighted. Such an experimental approach, together with the cell model, allows to separately study HCD and FNS on equivalent transistors.
- **Chapter 2**. HCD on equivalent Flash transistor structures is deeply studied. Firstly, the parameter extraction methodologies are presented in order to accurately address

the microscopic wear out of the device. In this part, the limitations of standard approaches are underlined and a new method is provided. In addition, the relationship between macroscopic drifting parameters and microscopic defects is studied, highlighting the channel position dependences. In the second part of the chapter, the aging kinetics of HCD mechanism is addressed, starting from the models present in literature. Then, taking advantage of TCAD tools [16], such a mechanism is accurately modelled using our own approach. In addition, a simple technique for the analysis of trap distribution location, which relies on the relationship between macroscopic electrical parameters, is proposed. The need to calibrate this correlation with experimental data and its strong connection with Si-H bond activation energy are highlighted. Finally, comparing simulations with experiments, physical interpretations on parameter correlation evolution and insights at different stress conditions are presented.

- **Chapter 3**. An accurate analysis of Erase-induced degradation is proposed with the help of equivalent Flash transistors. Concerning the aging of Erase efficiency, the gate current evolution is directly measured on big arrays capturing the impact of trapped charges close to Poly/SiO_2 interface. On the other hand, with the aim of addressing the electrostatic drift of the memory cell, thus capturing the effect of defects close to the bottom oxide interface, AC-stress alternated with Id-Vg measurements is performed on the single device. These two different phenomena are studied separately from a microscopic standpoint, analyzing the impact of each aging mechanism. Finally, accurate models for the related degradation kinetics are proposed with the aim of transferring such a knowledge on the respective Flash cell structures.
- **Chapter 4**. In the last chapter, the overall Flash cell degradation is finally addressed during P/E cycling, taking advantage of the cell model together with the studies on HCD and FNS previously done on equivalent transistor structures. The chapter is divided into three parts. Firstly, the correct separation of macroscopic aging contributions, such as the loss of P/E efficiency and the static drift of the cell, during Flash P/E cycling is provided. Secondly, the interplay between Program- and Erase-induced degradations is analyzed and addressed from an experimental standpoint. Finally, the experimental extraction of the aging contribution drifts and the related physical-based models are presented for cell evolution during Flash endurance experiments. The respective role of Program-efficiency loss, Erase-efficiency loss and static cell aging on the overall cell drift are explored. A proper characterization and understanding of each contribution is performed, and guidelines towards better reliability performance are proposed.

Bibliography

[1] S. Lai, Flash memories: Successes and challenges, IBM Journal of Research and Development, v 52, n 4.5, p 529-35, 2008.

[2] G. Burr, M. Breitwisch, M. Franceschini, D. Garretto et al., Phace change memory technology, Journal of Vaccum Science & Technology B: Microelectronics and Nanometer Structures, v 28, p 223, 2010.

[3] Fulford, Benjamin (24 June 2002). Unsung hero. Forbes. Retrieved 18 March 2008.

[4] F. Masuoka, M. Momodomi, Y. Iwata, R. Shirota, New ultra high density EPROM and flash EEPROM with NAND structure cell, International Electron Devices Meeting, 1987.

[5] Tal, Arie, NAND vs. NOR flash technology: The designer should weigh the options when using flash memory. Retrieved 31 July 2010.

[6] The Chip Collection: 10 - FLASH MEMORY TECHNOLOGY, INTEGRATED CIRCUIT ENGINEERING CORPORATION 10.1-10.16.

[7] B. Barth, Itrs commodity memory roadmap, tech., rep., ITRS, 2003.

[8] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, Phase change memory, in Proc. IEEE Vol. 98, No. 12, pp. 2201-27, 2010.

[9] H. Akinaga and H. Shima, Resistive Random Access Memory (ReRAM) based on metal oxides, in Proc. IEEE Vol. 98, No. 12, pp. 2237-51, 2010.

[10] D. Ielmini, A.S. Spinelli, M.A. Rigamonti, L. Lacaita, Modeling of SILC based on electron and hole tunneling. II. Steady-state, IEEE Transactions on Electron Devices, v 47, n 6, p 1266-72, June 2000.

[11] E. Vianello, F. Driussi, D. Esseni, L. Selmi, et al., Explanation of SILC probability density distributions with nonuniform generation of traps in the tunnel oxide of flash memory arrays, Transactions on Electron Devices, v 54, n 8, p 1953-62, Aug. 2007.

[12] A. Bravaix, V. Huard, F. Cacho, X. Federspiel, D. Roy et al., Hot-carrier degradation in decananometer CMOS nodes: From an energy driven to a unified current degradation modeling by multiple carrier degradation process, in Hot-Carrier Degradation, ed. by T. Grasser (Springer, Wien/New York, 2015).

[13] Y. Nissan-Cohen, J. Shappir, D. Frohman-Bentchkowsky, Measurement of Fowler-Nordheim tunneling currents in MOS structures under charge trapping conditions, Solid-State Electronics, v 28, n 7, p 717-20, July 1985.

[14] V. Della Marca, G. Just, A. Regnier, J.-L. Ogier, R. Simola, S. Niel, J. Postel-Pellerin, F. Lalande, L. Masoero, G. Molas, Push the flash floating gate memories toward the future low energy application, Solid-State Electronics, v 79, p 210-17, January 2013.

[15] J. Coignus, G. Torrente, A. Vernhet, S. Renard, D. Roy, G. Reimbold, Modelling of 1T-NOR Flash Operations for Consumption Optimization and Reliability Investigation, IEEE International Reliability Physics Symposium (IRPS), p PR-1 (4 pp.), 2016.

[16] Synopsys, Zurich, Switzerland, Sentaurus device user guide, J-2014.09.

Chapter 1

Introduction to Flash memory technology characterization and simulation

The aim of this chapter is to firstly present the Flash cell together with the evolution and limitations of the technology. In this context, we present our work and the purpose of this thesis. Then, in order to build solid basis to which we refer, proper Flash cell characterization and modeling is introduced. In particular, new techniques used for the extraction of fundamental cell characteristics and for device model calibration are proposed. Finally, an advanced delay-free experimental setup allowing customizable electrical pattern shapes is presented and its importance for properly characterizing the cell and the related aging evolution is highlighted.

1.1 Flash technology: an overview

In this section, we go through the Flash memory technology: from the working principle and the different architectures, to the technology evolution and limitations.

1.1.1 Flash cell: working principle

The structure of a 1T Flash cell is shown in Fig.1.1: it is composed by a Metal-Oxide Semiconductor Field-Effect Transistor (MOSFET), which has the capability of retaining a previously-stored electrical charge in the Floating Gate (FG) even when the power has been removed. This proficiency is used to set at least two memory states, by modulating the number of charges stored in the FG node, and thus to create a bit. This type of device is a standard MOSFET, which has the peculiarity of experiencing a gate node that is *floating*,

i.e. electrically isolated and inaccessible since completely surrounded by an ensemble of dielectric materials, thus it is called Floating Gate. As shown in the figure, this dielectric layer, which is located between the FG and the Control Gate (CG) nodes, is commonly a trilayer stack formed by a Si_3N_4 insulator (nitride) sandwiched between two layers of SiO_2 and is commonly called Oxide-Nitride-Oxide (ONO) stack. The related FG capacitor is used to couple the FG node to the applied voltage biases in order to access or control the stored information. In particular, by appropriately biasing the four accessible terminals, i.e. Drain (D), Source (S), Control Gate (CG) and substrate Bulk (B), the charge into the FG node can be sensed or modified. The procedure of removing the charge from the FG node is called *Erase* process, whereas the procedure of storing it *Program* operation. On the other hand, the extraction of the information is carried out during the *Read* phase.

The dielectric between the channel and the polysilicon Floating Gate is called tunnel oxide and its thickness Tox results from a trade-off between high Program/Erase (P/E) performance and endurance/retention capabilities. Finally, it is worth noting that triple-well CMOS technologies are usually adopted to independently bias the isolated substrate of the memory device during the different operations.

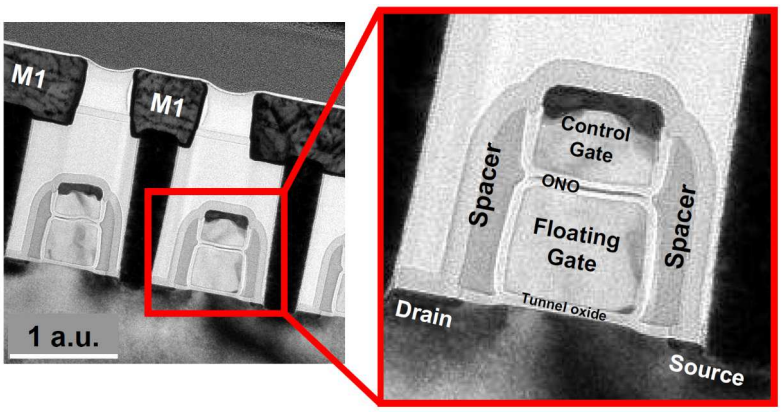


Figure 1.1: Transmission Electron Microscope (TEM) cross-section image along the channel length of a Flash memory cell in NOR configuration. In the zoom, the terminals of the device as well as the oxide layers can be identified [1].

In Fig.1.2 the commonly-used operating modes of Flash devices are schematically illustrated: sensing, storing and removing the electrical charge in FG node. The charge amount at FG can be sensed by associating it with electrical parameters, as for example the drain current Id flowing between S and D terminals. Indeed, in Read operation, the charge stored in FG node modifies the cell threshold voltage Vth, thus two stable reference amounts of charge can be chosen in order to represent the 1 and the 0 memory states, as shown in Fig.1.2(a). Several definitions are present in literature for Vth parameter: in this chapter, the constant-current definition is considered, which is defined as the CG voltage level

that has to be applied in order to obtain an arbitrary reference Id current during the Read operation. In particular, considering a large negative charge amount in the FG, Vth is consequently high (programmed state - 0); as the charge is removed, the cell presents a lower threshold voltage during the following read operation (erase state - 1). Since the FG node is completely surrounded by the dielectric stack, the injected charge is trapped in the FG node and stores the information.

Among the techniques used to remove or inject the charges from/in the FG node, Fowler-Nordheim (FN) tunneling and channel Hot Carrier Injection (HCI) are the most diffused solutions. Cell Program operation can be achieved using both processes. The tunneling mechanism occurring through the oxide layer is applied since mid 70s. For example, Fowler-Nordheim programming regime is achieved when a positive high voltage difference, i.e. 15V - 20V, is applied between the CG, addressed by the word line in a structured matrix array, and the Source or the Bulk of the cell. This is applied in NAND, AND, and DINOR (divided bit line NOR) technologies where high throughput is required sacrificing random-access memory capabilities [1]. Indeed, multiple wordlines can be programmed in parallel using a very small programming current (less then 10nA/cell).

The HCI operation consists in applying a lateral electric field along the channel with the aim of increasing the carrier energy and forcing a transversal electric field between the Floating Gate and the channel in order to enhance the injection [2]-[4]. This operation is performed keeping Bulk and Source grounded and applying a positive high voltage on CG (order of 8-9V) and Drain (order of 3.6-4.2V) terminals, as shown in Fig.1.2(b). Contrary to FN tunneling, this mechanism needs large amount of current (up to $\approx 500\mu$A / cell), thus it cannot be used in high P/E throughput applications, and its efficiency is poor: only few electrons are injected over the total amount of carriers flowing from Source to Drain [5] and the power consumption is high. On the other hand, this mechanism is extremely fast (few microseconds). In NOR technologies, HCI is usually the preferred approach for the Program operation since random-access capability is available and low threshold voltage dispersion is required.

Concerning the Erase operation, the reverse Fowler-Nordheim tunneling is commonly adopted. In this high field regime [6], [7], the Control Gate is biased with high negative voltage values, whereas a positive bias is applied on Source and/or Bulk electrodes. Since the FN process is directly linked with the injected charges, we will perform the Erase operation simply considering a high negative value on CG node and keeping the other terminals to the ground, as shown in Fig.1.2(c), in order to make Erase process suitable with our experimental setup (described in the end of this chapter). However, the Erase operation is relatively slow (order of milliseconds), but the energy consumption can be considered negligible because no current flows in the channel, which gives the possibility to erase in parallel large sections of the memory matrix. The low current consumption also contributes to relax requirements in

embedded applications and thus it allows the integration of more simple circuit blocks for the high voltage bias generations on chip. However, due to the exponential dependence of the gate tunneling current with oxide characteristics (as the thickness for example), accurate process control is required to achieve narrow Vth distributions.

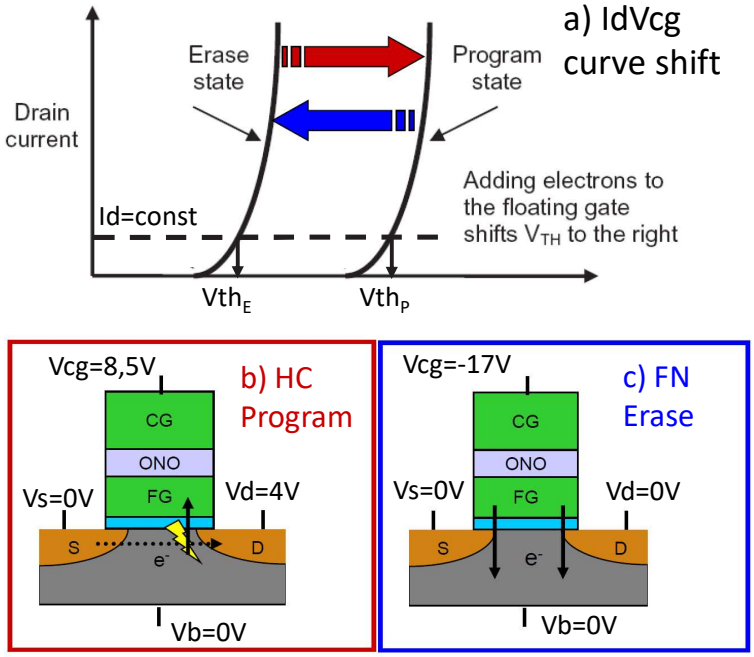


Figure 1.2: Schematic illustrations of Flash memory cell operations [1]. In (c) the current is shown as a function of the CG voltage during Read operation. The device switches between two (or more) states, which are represented by a different amount of carriers stored in FG node. In (b), Program mode operation, performed via HCI mechanism, is shown in Lowly-Doped Drain (LDD) region. Erase mode performed by FN process across the tunnel oxide is represented in (b).

As previously remarked, the FG potential directly depends on applied biases and on the charge stored in the node. A first approach to model this electrostatic system was already proposed by Bhattacharyya [8] and further adopted by Kolodny [9] for EEPROM cells. The charge Qfg stored in FG node in DC conditions can be expressed as:

$$\mathrm{Qfg} = \mathrm{C_{ono}}\,(\mathrm{Vfg} - \mathrm{Vcg}) + \mathrm{C_s}\,(\mathrm{Vfg} - \mathrm{Vs}) + \mathrm{C_d}\,(\mathrm{Vfg} - \mathrm{Vd}) + \mathrm{C_b}\,(\mathrm{Vfg} - \mathrm{Vb}) \tag{1.1}$$

where Vfg is the potential of the FG node, whereas $\mathrm{C_{ono}}$, $\mathrm{C_s}$, $\mathrm{C_d}$, $\mathrm{C_b}$ are the capacitance between the Floating Gate and the Control Gate, the Source, the Drain and the Bulk respectively. From eq.1.1, by dividing the total capacitance $\mathrm{C_T} = \mathrm{C_{ono}} + \mathrm{C_s} + \mathrm{C_d} + \mathrm{C_b}$ seen from the FG node, the FG potential can be directly determined [10]:

$$\mathrm{Vfg} = \alpha_\mathrm{G}\mathrm{Vcg} + \alpha_\mathrm{S}\mathrm{Vs} + \alpha_\mathrm{D}\mathrm{Vd} + \alpha_\mathrm{B}\mathrm{Vb} + \frac{\mathrm{Qfg}}{\mathrm{C_T}} \tag{1.2}$$

where α_j are the coupling coefficients between FG and the nodes j and are defined as the derivative of Vfg respect to the terminal Vj. Thus, it represents the capability of the terminal j to electrically control the FG node. In particular,

$$\alpha_j = \frac{\partial Vfg}{\partial Vj} = \frac{C_j}{C_T} \tag{1.3}$$

$$\sum_j \alpha_j = 1 \tag{1.4}$$

Once the Vfg potential is determined, one can apply any analytical model (currents, charges in the active region etc) for a MOSFET device formed by the terminals FG, D, S, B and replacing the gate voltage with the calculated FG potential.

The main drawback of such an approach is that, due to many extrinsic effects which are dominating at sub-micron scales, the determination of the coefficients can be a challenging task and the techniques usually adopted to experimentally extract them are not always accurate enough. In particular, for the estimation of these parameters, α_G is firstly determined and the others are extracted by measuring the drain leakage currents and the difference between the erase voltages [11], [12]. The most common method to address α_G relies on the determination of the ratios between the threshold voltages, transconductances [9] or subthreshold slopes [13] of FG cells and equivalent transistor devices. It is evident that, since all the coefficients are derived from α_G, a minimal error on the estimation of this parameter affects all the other coupling coefficients. Another problem of such an approach is that the process variation between the compared structures can generate error in the extraction, which can be important then in the calculation of Vfg. In addition, the coupling coefficients are not exactly constant, since channel, drain, source capacitance vary with applied biases [14], [15]. For these reasons, in order to have a more complete physical based model, Charge Based Methology (CBM) is often adopted [16], [17]. It consits in iteratively solving a non-linear charge balance equation at the FG node. The total charge on this node is in fact the sum of the charge stored on the gate of the MOSFET Qg and of the charge stored at the bottom plate of C_{ono} capacitor:

$$Qg(Vfg, Vs, Vd, Vb) + C_{ono}(Vfg - Vcg) = Qfg \tag{1.5}$$

where Qg can be simply determined using standard compact models for MOS transistors [18], [19].

However, it has to be pointed out that such an approach is not always suitable for industry, since exponentially increasing the complexity without bringing a significant improvement in the accuracy. For this reason, it is better to experimentally address the coupling coefficients with very high precision and considering them constant, ensuring that the related extraction error is negligible respect to the intrinsic statistical device dispersion. In the

next section, we will introduce a new methodology that accurately addresses α_G with high precision, with the aim of using the simple expression in eq.1.2 which will be fundamental to link equivalent transistor data with Flash cell results.

1.1.2 Flash memory architectures

Flash memories are organized in arrays of rows (word lines or WL) and columns (bit lines or BL). The array architecture is determined by the type of connections, as shown in Fig.1.3. Two types are usually adopted:

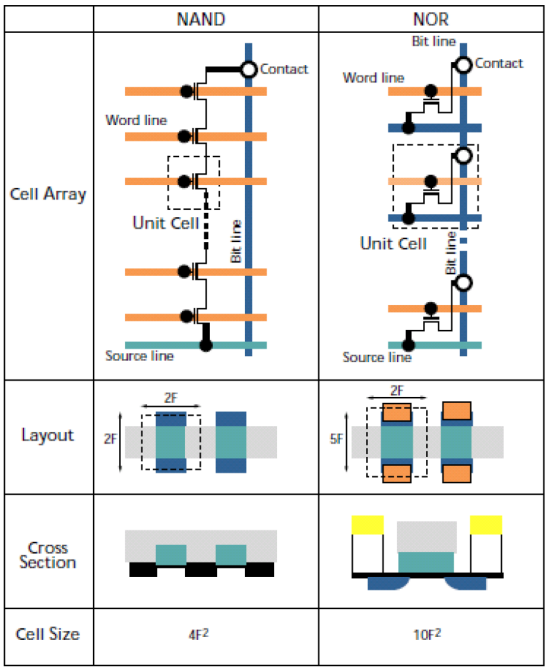


Figure 1.3: Architectures of NAND (left) and NOR (right) memory arrays (source: micron.com).

- **NOR**: The NOR architecture has been introduced by Intel in 1988 for the first time. The memory cells are connected in parallel and the CG electrodes are connected together through the wordline, whereas the Drain is shared along the bitline. The peculiarity of such a structure is that the Drain of each single cell can be selectively addressed. This gives the capability of random accessing any cell in the array. For this reason, the Program operation is performed by HCI and the Erase by FN mechanism. NOR architecture provides fast reading and relatively slow programming mechanisms. The presence of a Drain contact for each cell limits the scaling to $6F^2$, where F is the smallest lithographic feature. Fast read, good reliability and relatively fast Program mechanism make NOR architecture the most suitable technology for the embedded

applications requiring the storage of codes and parameters and more generally for execution-in-place. The memory cells studied in this thesis are integrated in a NOR architecture for embedded applications.

- **NAND**: NAND architecture has been developed in 1987 by Toshiba with the aim of realizing ultra high density EPROM and Flash EEPROM [20]. This architecture has the peculiarity of providing all the cells in series where the CG electrodes are connected together in a wordline, whereas the Drain and the Source nodes are not contacted. This absence of contacts means that the device cannot be selectively addressed and the Program can be uniquely performed by Fowler-Nordheim mechanism. However, it is possible to reach an optimal cell size of $4F^2$, thus a density that is approximately 30% higher respect to NOR architecture. In NAND structure, the Program is relatively fast but the reading process is quite slow, since the reading of a single cell is done by forcing the device in the same BL to the ON state. The high density and the slow Read operation but fast writing speeds make this architecture suitable for USB keys, MP3 audio, GPS, storing digital photos and many other multimedia applications.

1.1.3 Evolution and limits of Flash memories

New applications constantly push the semiconductor market and the research development. Since the invention of Flash memory cell, the progress on device materials and architectures has been significant. Ideally, the memory should have [21]:

- high density solution
- low power consumption
- non-volatility capability
- fast read/write/erase operations
- random read/write access
- endurance along P/E cycles
- low cost scalability
- compatibility with logic circuits and integration

This is the final target of the research in semiconductor industry. However, since the "ideal" device does not exist, different types of memories have been proposed in order to find the best trade-off and thus improve some specific properties. For these reasons, as shown in Fig.1.4, memory technology development did not pursue a single technology solution, but it has been developed in many different directions over time [22], [23]. Anyhow, in this thesis, we will not deal with these architectures, thus their description is skipped.

In the last 30 years, the Flash memory cell size has been shrinked from 1.5μm to 25nm doubling the capacity per surface unit each year approximatively. In Fig.1.5, we show the International Technology Roadmap for Semiconductor (ITRS) 2009 that predicted the future trends of semiconductor technology. We can see that even if the trend is maintained and

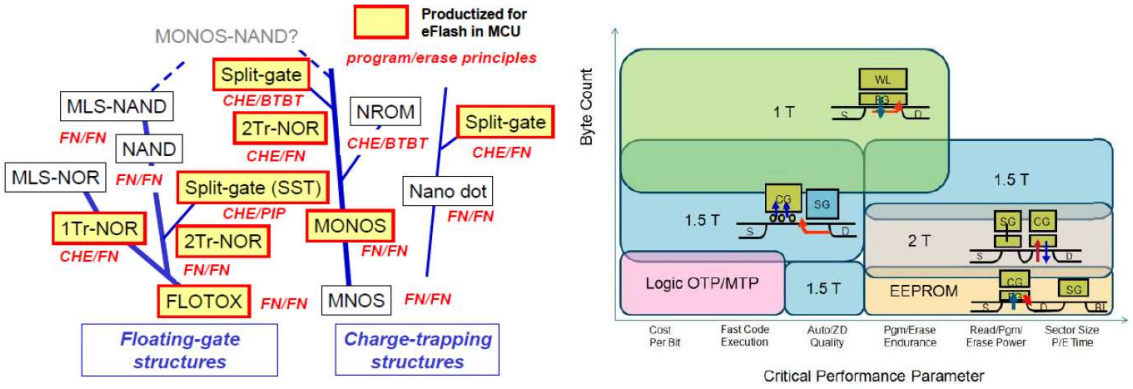


Figure 1.4: Evolution of Flash and embedded-Flash memory technology (left) [22]. Mapping of common eNVM architectures to the NVM byte count and critical characteristics (right) [23].

ITRS 2012 - Process Integration, Devices, and Structures					
	2013	2016	2019	2022	2026
Nor flash technology node – F (nm)	45	38	32	28	22 ?
Cell size – area factor in multipliesof F2	12	12	12-14	14-16	14-16
Physical gate legth (nm)	110	100	90	85	85
Interpoli dielectric thickness (nm)	13-15	13-15	11-13	11-13	11-13

Figure 1.5: Summary of the technological requirements for Flash NOR memories as stated in ITRS 2012 roadmap [24]. White cell color: manufactural solutions exist and are being optimized. Yellow cell color: manufactural solutions are known. Red cell color: unknown manufactural solutions.

the cell scaled down in the years to come, some technological solutions are still not known. Moreover, scaling beyond the 28nm node will be very difficult if no revolutionary technology is adopted. The main issues that limit device miniaturizing are listed in the following and the most important ones are shown in Fig.1.6:

- **Memory retention**: in non-volatile memories, the retention is represented by the amount of time between data storage and the detection of a erroneous readout of the

data [25]. Indeed, during P/E cycling, the stress degrades the tunnel oxide and the cell slowly loses its capacity to store electric charges, as shown in Fig.1.6(a). In particular, in modern Flash cells, FN-induced bulk oxide defects are considered as the main responsible of Stress-Induced Leakage Current (SILC), i.e. an additional dielectric leakage mechanism which decreases the quantity of stored electrons and alter the non-volatile information [26], [27]. The retention depends on the number of cycles and on the tunnel oxide thickness, but the physical scaling of the latter is limited to 6-8 nm. It has to be pointed out that in deep-submicron technologies, i.e. $C_{ono} \approx 1fF$ and $Qfg < 2 \cdot 10^5$, a loss of 5 electrons per day permits to achieve a commonly accepted specification of 10 years in data retention [16]. Such restrictions are even more aggressive in some applications, as in automotive sector, where the cell should operate in a wide range of temperature.

- **Memory endurance**: in general, the memory endurance represents the capability of a memory cell to cope the electrical stress suffered during P/E cycling [25]. Indeed, the device experiences a reduction or a drift of the Programming Window (PW), i.e. the different between Vth at Program and Erase states, after a given amount of P/E operations, as shown in Fig.1.6(b). This represents a serious concern for modern scaled Flash technology, especially for IP designers which develop sense amplifier circuits in order to capture the state of the device. From a microscopic standpoint, these drifts are related to the degradation directly induced by the physical mechanisms responsible for Program and Erase charge flow. For instance, for NOR technology, HCI-induced defects [28],[29] and FN-induced oxide traps [30],[31] are respectively generated by Program and Erase operations and impact the PW.
- **WL/BL disturbs**. They represent another major issue in ultra-scaled technologies. Indeed, in NOR configurations, P/E operations can be achieved only by biasing the entire WL and BL in high voltage conditions. The Program disturb present on the unselected erased devices belonging to the same WL may induce tunnelling of electrons from the channel to the FG. On the other hand, BL disturbs can be generated during both Program or Read conditions. Also in this case, the programmed condition is more sensitive to disturb and Hot Hole Injection (HHI) phenomena occurs from the channel reducing the amount of stored electrons in the FG node [32], [33].
- **Short Channel Effects (SCE).** SCE appear when the gate length is so short that the capability of the gate, i.e. FG in Flash memory cell, to control the channel is reduced, since the influence of S/D potential becomes significant. This parasitic effect produces the well-known Drain Induced Barrier Lowering (DIBL) phenomenona [34], that is translated in a Vth reduction and in a worsening of the subthreshold slope. Accordingly, the OFF current increases, which makes the power consumption reaching

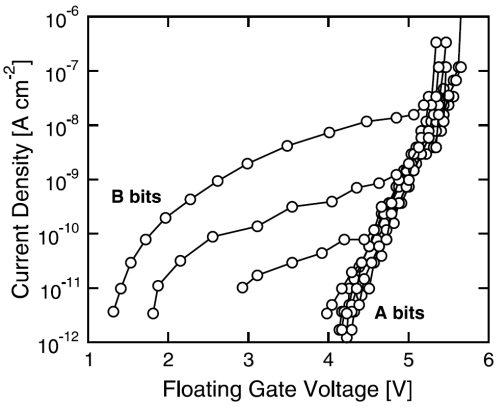
values incompatible with the advanced technology node requirements [35].

- **Vth variability**. An increasing concern with device down-scaling regards the threshold voltage dispersion due to the random variations in quantity and position of doping atoms (Fig.1.6(c)), commonly known as Random Dopant Fluctuation (RDF) phenomena [36], [37]. In particular, considering short gate lengths, the nature of Source-to-Drain conduction is shown to be percolative due to the atomistic configuration of the doping in channel, which is microscopically uncontrollable. This concern increases with technology scaling, also because the doping concentration has to be increased in order to contrast the SCE effects. Moreover, Random Telegraph Noise (RTN), due to border traps generated during P/E cyling, increases this fluctuation [38], [39]. For all these reasons, the Vth statistical dispersion is a serious concern for Flash memory technology. It has been documented [40] that, in 25nm node, the Vth is expected to vary of about 30% due to RDF only.
- **Coupling factor**. As the Flash cell dimensions are reduced, the tunnel oxide and the ONO stack have to be scaled too, even if with a slower rate due to cell retention restrictions. For this reason, in order to maintain a high control of the FG node, i.e. high α_G, most of Flash structures experience a CG electrode surrounding the side-walls of the FG, as shown in Fig.1.6(d). However, the device down-scaling causes a reduction of the distances with other cells and contacts, inducing parasitic capacitances and thus a reduction of α_G.

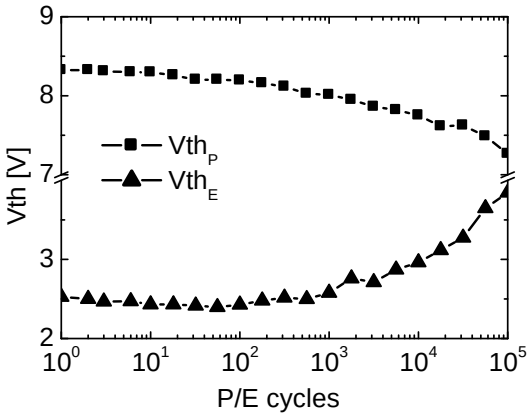
Among these limitations for Flash memory down-scaling, only the first two are exclusively related to the device aging, thus to the reliability. Indeed, all the others concern mainly physical limitations (RDF, SCE) or device architecture issues (Program disturb, parasitic capacitances). Hence, it is fundamental to study and characterize memory retention and endurance, in order to physically understand the aging mechanisms and predict the device working condition after a certain stress time, which simulates the utilization of the cell itself.

Concerning the data retention, several works dealt with this problem in the past. For example, regarding the physical mechanism, DiMaria firstly attribuated the SILC leakage current with the presence of neutral traps in the oxide [43], then Takagi, taking advantage of the Current Separation Technique (CST) [44], linked this phenomena to the Inelastic Trap Assisted Tunneling (ITAT) [45] and finally Ielmini improved this model considering an ITAT which is enhanced by electron-hole recombination in the oxide [46]. In addition, also techniques for correctly extracting position [47], energy level [48] and statistical dispersion [49] of the oxide traps have been proposed. Then hundreds of works followed and improved these approaches.

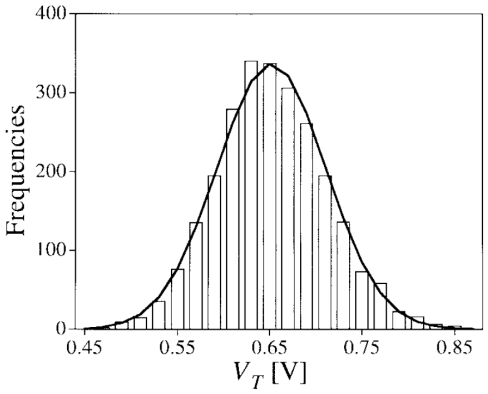
On the other hand, there is a significant lack of knowledge concerning the Flash memory endurance understanding, which is often limited to the electrical characterization and phe-



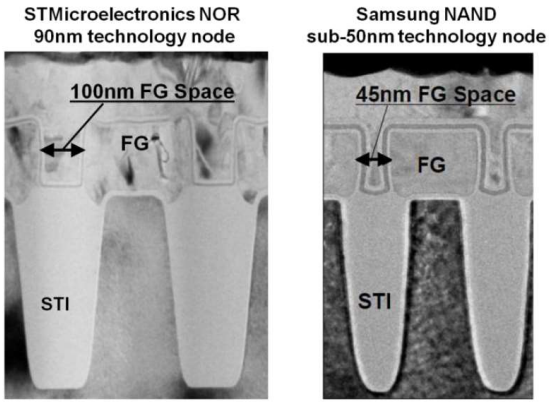
(a) Ifg-Vfg characteristics extracted for different aged cells. "A" cells display pure FN tunneling, whereas "B" cells are dominated by SILC mechanism [41].



(b) Evolution of Vth at Program state (Vth_P) and Vth at Erase state (Vth_E) during P/E cycling. Data refer to 40nm NOR Flash technology considered in this thesis.



(c) Vth distribution computed from the simulations of 2500 MOSFETs (L=W=50nm, Na=$5 \cdot 10^{18}$cm^{-3}, tox=3nm) experiencing different atomistic doping configurations [36].



(d) TEM pictures of STMicroelectronics 90nm NOR Flash (left) and Samsung sub-50nm NAND Flash (right) [42].

Figure 1.6: Well-known Flash memory cell concerns with technology down-scaling. (a) Memory retention, (b) Memory Endurance, (c) RDF, (d) α_G reduction.

nomological evaluation [50], [51]. In particular, endurance models commonly used in industry are often empirical models based on experimental data, thus not relying on the microscopic aging description of the different degradation mechanisms and their interactions. Indeed, the optimization of cell endurance and of P/E patterns are usually based on process trials on Silicon and on the respective experimental observations. Such an approach results to be poorly efficient, slow, costly, especially in the recent past, as new and different technologies are growing up within the non-volatile memory market.

For all these considerations, a physical-based approach is needed for Flash memory endurance optimization: it is fundamental to accurately evaluate the degradation really suffered by the cell, to understand the impact of each defect on cell electrostatics and on P/E operation efficiencies, the physical origin of the wear out and so on. This helps not only to find new technological solutions, but also to push the memory cell towards its maximum intrinsic performance.

Nevertheless, the complexity of such a problem leads the accurate physical modeling to be unusable, since the computational cost would be tremendously high, which, in turn, would cause the loss of the physical validity. For this reason, the real challenge is to a build a simple compact model having the capability to be suitable for industry without losing the physical meaning and thus the accuracy. The simplification is necessary, however all limitations and uncertainties, thus the model validity range, should be perfectly known and controlled.

In this thesis, for the first time, such an approach is adopted. We will show that, starting from physical basis, from the cell model and the respective transistor structure degradation, it possible to predict the Flash cell aging evolution and to optimize the memory technology itself always having a microscopic perspective.

In order to put in place this approach, we will start from the basis, i.e. Hot Carrier Degradation and Fowler-Nordheim Stress on transistors having Flash geometry, and then we will transfer this knowledge on the Flash cell and build a complete physical-based compact model for the memory endurance. For this reason, it is firstly necessary to perfectly know the device, thus a precise characterization of the memory cell has to be performed. In the next section we will introduce some specific techniques in order to acquire the coupling coefficients and the Ifg-Vfg characteristic in both Program and Erase regimes.

1.2 Flash cell electrical characterization

All the experiments presented in this thesis have been performed on Flash cell and nMOSFET structures issued from 40nm embedded 1T-NOR Flash technology developed at STMicroelectronics (Tox=9.4nm, W=60nm, L=140nm). Concerning the "equivalent" transistors, they have Flash technology geometry and are built by shorting the Control and Floating Gates with the objective to reproduce and study the stress induced at the bottom part of the cell, as we will see later. The two devices used are schematically shown in Fig.1.7.

In this section, we propose new methodologies to properly characterize the Flash cell. In particular, with the final aim of using equivalent transistor structures for quantifying the aging induced at the bottom oxide level, cell coupling coefficients and Ifg-Vfg characterstics have to be acquired. The first extraction is performed using our own method, based on the comparison between Flash cell and the respective equivalent transistor, whereas the second one is carried out using the well-known methodology based on the Step Pulse experiment

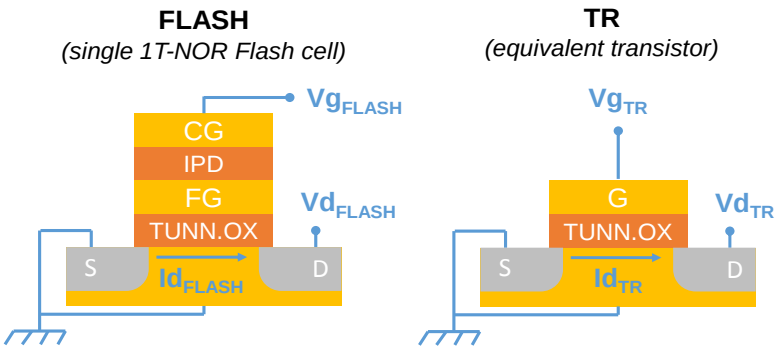


Figure 1.7: Schematic representation of Flash cell and equivalent transistor structures (i.e. TR or MOS_{eq}) considered in this thesis. IPD stands for InterPoly Dielectric stack, i.e. ONO layer in particular.

[52], [53]. Finally, we face the dependence of these characteristics with the geometry, giving important guidelines for scaling strategy.

1.2.1 Coupling coefficient extractions

As previously underlined, it is fundamental to accurately extract the gate coupling coefficient α_G for two main reasons. Firstly, the acquisition of this parameter allows to estimate the other coupling coefficients using eq.1.4. Secondly, it gives a direct calculation of the FG potential within the P/E phase, since linking Vcg (externally controlled) and Vfg (unknown). Thus, a small error on this parameter may lead to a big mistake in the estimation of degradation if the aging itself has a strong dependence on FG level. In the following, we propose our own methodology to extract α_G with high precision.

In Fig.1.8, the extraction methodology of this parameter is schematically shown. The linear Id-Vg trans-characteristic is measured for both Flash cell and for the respective equivalent transistor. Since the gate voltage of the MOSFET structure corresponds to the potential induced at FG node of the cell, it is possible to couple FG-CG voltages for a certain drain current level in linear regime, as shown in Fig.1.8(a). Then, repeating this procedure with a different current level, the relationship Vfg-Vcg of eq.1.2 can be experimentally addressed, as reported in Fig.1.8(b). Now, differentiating this relationship, α_G function of the applied voltage is captured (Fig.1.8(c)). The average value represents our gate coupling coefficient.

This method has the big advantage to be charge insensitive, since a different Qfg would shift the straight line of Fig.1.8(b) without modifying the slope. Similar observations can be done for the other applied voltages (Vd and Vb). However, such a method suffer from two main issues:

- **RTN**. Due to this phenomena, the drain current measurement during the Vg-sweep

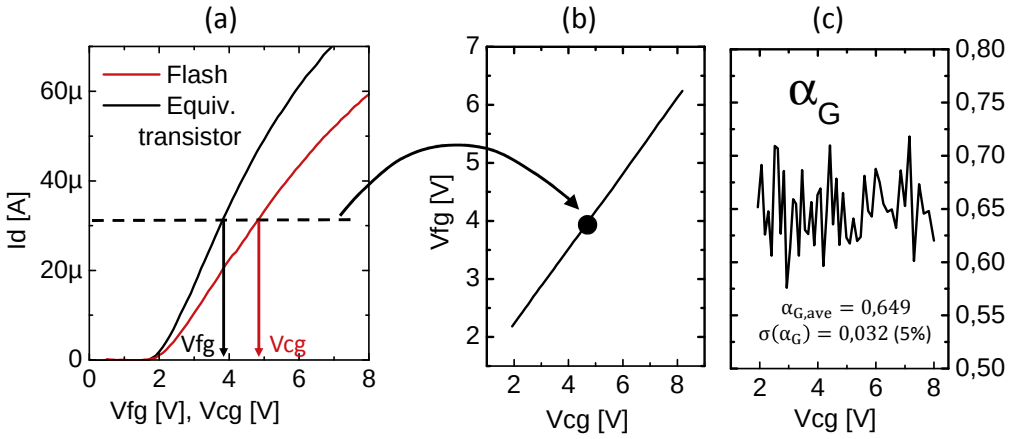


Figure 1.8: α_G extraction methodology. At a given drain current level, FG and CG potentials can be extracted from the linear trans-characteristics (a) acquired respectively on equivalent transistor and Flash cell. This extraction is repeated at different current levels, till forming the Vfg-Vcg linear relationship in (b). Differentiating this relationship, α_G is finally extracted (c).

suffer from abrupt jumps. Indeed, border oxide traps capture and release carriers, which modulates the inversion charge, thus the drain current. This noise is located especially at low frequency ($\leq$ 100Hz), thus a removal is hardly achievable through data filters and treatments. In our case, RTN causes a noise in the acquisition of the gate coupling coefficient. In particular, in Fig.1.8, this parameter has been acquired with a precision of 5%. As we will see in the following, such a precision is not enough to guarantee a correct comparison of the cell with the respective Flash transistor structures.

- **Statistical dispersion**. Although α_G is acquired with a certain precision for one single Flash-MOS_{eq} couple, changing either the transistor either the cell, the resulting extracted coefficient can be very different. This is due to a strong variability between the devices. As already pointed out, the principal reason of such a mismatch comes from RDF. This phenomenon alters not only the Vth, but also the transconductance between the devices. Such a Gm dispersion leads to an erroneous α_G extraction, since the comparison may be performed between two very different devices from a microscopic standpoint.

In order to overcome these limitations, experimental solutions have been found:

- The first problem can be solved thanks to very fast Id-Vg acquisitions (entire Vg-sweep within $\approx 100\mu$s). If such a measurement is performed quickly and is repeated several times, the error is strongly reduced by performing an average. In particular, a fast measurement is almost insensitive to RTN fluctuations but, on the other hand,

a short measurement time causes low precision. For this reason, the measurement is performed several time and an average process is carried out afterwards. The related error induced in the α_G acquisition on one single Flash-MOS$_{eq}$ couple can be considered totally negligible ($\approx 0.1\%$).

- On the other hand, the device characteristic dispersion issue can be overcome by considering a huge statistics, as shown in Fig.1.9. Indeed, the RDF phenomena induces certain channel atomistic doping configurations that, on average, are the same for the two devices, since coming from the same process steps. Thus, extracting the coefficient for each Flash-MOS$_{eq}$ couple, the value at 50% of the population represents the value on which similar channel conductivities have been compared. For this reason, it represents the right α_G, acquired with the precision of the single couple.

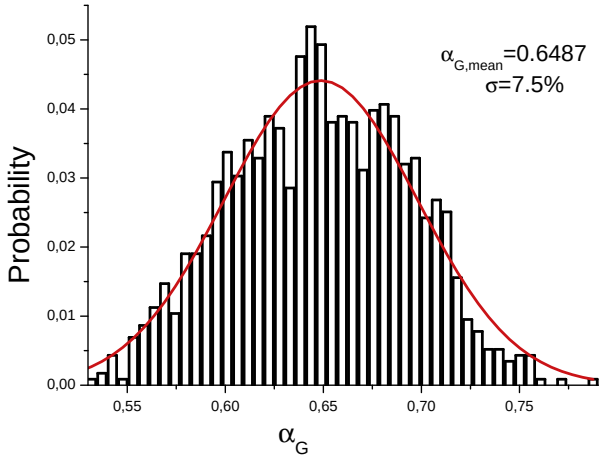


Figure 1.9: Statistical dispersion of α_G extracted following the methodology presented in Fig.1.8 for all the possible combinations Flash-MOS$_{eq}$. 50 nominally equal devices have been considered for both structures giving a total statistics of 2500 Flash/TR combinations.

After having extracted the gate coupling coefficient, α_D can be roughly estimated as $\alpha_D = \alpha_S \approx (1 - \alpha_G)/2$, assuming that the parasitic effects are negligible. In any case, a high precision acquisition of such a parameter is not fundamental in our case. Indeed, during the Erase phase, the drain potential is set to zero, whereas Vd $\approx$ Vfg within the Program pulse, thus the total error is surely close to zero. However, in order to extract the coupling coefficient separately, we tried to directly address this parameter.

The difficulty to acquire α_D coefficient relies on the fact that the Flash cell and the respective equivalent transistor should be compared at fixed Vfg varying the drain potential. However, during the Vd-sweep on the memory cell, the injection current becomes important and Qfg varies. This makes the Flash Id-Vd curve stretched since the cell is being programmed during the read operation. For this reason, an alternative method is applied

to our devices and is shown in Fig.1.10. The Id-Vg curve is measured at two different Vd levels for both devices. For the equivalent transistor, a fixed Gate potential, i.e. representing the FG of the cell, is chosen, which hence corresponds with two current levels, as shown in Fig.1.10(a). Assuming again that a certain Drain current is induced by the same Vfg for the two devices, these two current levels are used to extract the effect of Vd on the CG potential. In particular, the related Vcg shift can be easily demonstrated to be equal to

$$\Delta \mathrm{Vcg} = \frac{\alpha_\mathrm{D}}{\alpha_\mathrm{G}} \Delta \mathrm{Vd} \tag{1.6}$$

Since α_G has been previously acquired, α_D can now be extracted as shown in Fig.1.10(b). However, it is worth noting that such an approach does not lead to a high precision in the extraction. In addition, changing either Vd couple, either devices, the extracted coefficient may be different. For this reason, we considered a certain statistics for both applied Vd and device couple and we finally got an average $\alpha_\mathrm{D} \approx 0.18$. It is worth mentioning that α_D is surely not constant with applied potentials, since coming from the parasitic overlap drain capacitance. For instance, a high drain potential induces a higher drain coupling coefficient. However, as previously remarked, the related error on our endurance results is surely negligible, thus we will consider this parameter constant with applied voltages.

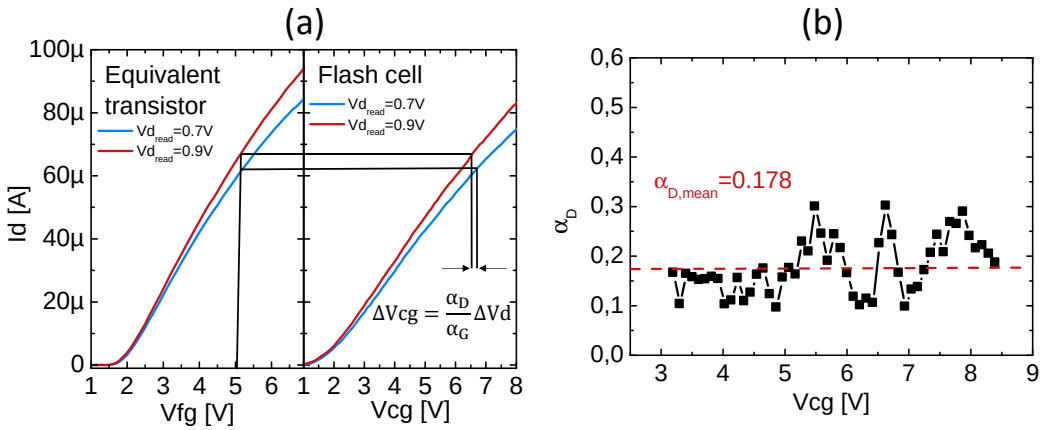


Figure 1.10: α_D extraction methodology. At a given Vfg potential, a couple of drain currents is extracted from the linear trans-characteristics measured at two different Vd levels for the equivalent transistor. These current values are used to measure the effect of the drain voltage increase on the CG potential taking advantage of the respective Flash linear Id-Vcg curves (a). This procedure is repeated at different Vfg (i.e. Vcg) levels and α_D is finally extracted using eq.1.6 (b).

1.2.2 Step Pulse experiment

Another fundamental device characteristic to be acquired is the injection Ifg-Vfg curve in both Program and Erase phases, i.e. Hot Carrier and Fowler-Nordheim characteristics. Unfortunately, those currents are too small to be directly measured on nominally-scaled equivalent transistors. This is also valid for Hot Carrier current: indeed, in that case, the integration time has to be increased so much that the related measurements would be already affected by a strong degradation. For this reason, two possible strategies can be adopted to address the gate current: a direct measurement with "ad-hoc" transistor structures, i.e. matrices of devices connected in parallel, or indirect measurement on Flash cells looking at the Vth variation induced by P/E pulses.

This last possibility makes use of the standard electrostatic relationships between Vfg and Vcg and between the FG charge Qfg and the cell threshold voltage Vth. In particular, it consists in applying a train of short Program or Erase pulses and between them measuring the cell Vth. Thus, the evolution of such a parameter can be tracked as a function of the cumulated P/E duration, as shown in Fig.1.11(a) for the Program case. Now, it is obvious that the Vth evolution is linked with Qfg dynamics, in particular, using eq.1.2 at threshold condition:

$$\mathrm{Vth_{mos}} = \alpha_{\mathrm{G}}\mathrm{Vth(t)} + \alpha_{\mathrm{D}}\mathrm{Vd_{lin}} + \frac{\mathrm{Qfg(t)}}{\mathrm{C_T}} \tag{1.7}$$

where the time "t" is the cumulative Program or Erase time and $\mathrm{Vth_{mos}}$ is the threshold voltage at Vfg level, i.e. FG potential which induces 8μA during the Read operation[1]. It is worth noting that the acquisition of $\mathrm{Vth_{mos}}$, performed on equivalent transistors, is necessary. On the other hand, within P/E pulses, the FG potential follows the law:

$$\mathrm{Vfg(t)} = \alpha_{\mathrm{G}}\mathrm{Vcg_{P/E}} + \alpha_{\mathrm{D}}\mathrm{Vd_{P/E}} + \frac{\mathrm{Qfg(t)}}{\mathrm{C_T}} \tag{1.8}$$

where CG and Drain potentials are the applied voltages during the P/E phase. Now, computing the difference between these two equations in order to be charge insensitive, we get:

$$\mathrm{Vfg(t)} = \mathrm{Vth_{mos}} - \alpha_{\mathrm{G}}\mathrm{Vth(t)} + \alpha_{\mathrm{G}}\mathrm{Vcg_{P/E}} + \alpha_{\mathrm{D}}(\mathrm{Vd_{P/E}} - \mathrm{Vd_{lin}}) \tag{1.9}$$

which computes the FG potential dynamics. On the other hand, differentiating the eq.1.7, the dynamics of the gate current can be easily addressed:

$$\mathrm{Ifg(t)} = \mathrm{C_{ono}}\frac{\mathrm{dVth(t)}}{\mathrm{dt}} \tag{1.10}$$

Having computed Vfg(t) and Ifg(t) with eqs.1.9, 1.10 respectively, the Ifg-Vfg characteristic is finally extracted, as shown in Fig.1.11(b) for the Program case. This method is commonly

[1] such a parameter can be measured directly on equivalent transistor structures. This is true in general, even for a degraded device, as long as Flash cell and the respective equivalent transistor experience exactly the same degradation.

called Step Pulse Program (SPP) experiment (or SPE for the Erase case) and is used since more than 30 years [52]. A common example of application is the detection of very low current levels such as SILC [53].

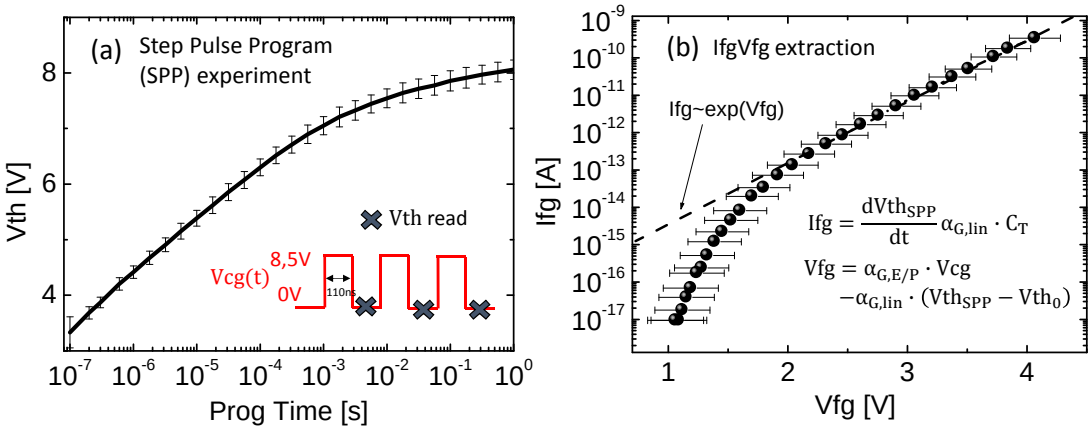


Figure 1.11: SPP experimental results. The Vth evolution is measured during a train of short HC Program pulses (a). From these experimental data, the Ifg-Vfg curve can be computed (b). Error bars refer to the statistical variation of Vfg acquisition, which represents the limiting factor in the Ifg-Vfg extraction accuracy.

Till now we implicitly assumed that the train of short pulses is exactly equivalent to a single large pulse, whose time corresponds to the sum of the short pulse durations. This has been experimentally verified in our case (not shown). However, this means that such an experiment allows to address the cell dynamics within a constant CG pulse. Indeed, in general, the cell dynamics is computed solving the differential equation:

$$\frac{dVfg}{dt} = \alpha_G \frac{dVcg}{dt} + \frac{Ifg(Vfg)}{C_T} \tag{1.11}$$

which directly comes from eq.1.8. In particular, Ifg(Vfg) represents the FN/HC characteristic. For a constant CG pulse, the equation reduces to:

$$\frac{dVfg}{dt} = \frac{Ifg(Vfg)}{C_T} \tag{1.12}$$

Looking at Fig.1.11(b), we can assume that Ifg exponentially depends on the FG potential, i.e.

$$Ifg(Vfg) = -A \cdot \exp(B \cdot Vfg) \tag{1.13}$$

It is now easy to demonstrate that, after a certain time, the Vfg dynamics follows the law:

$$Vfg(t) = -\frac{1}{B} \cdot \ln\left(t \cdot \frac{A \cdot B}{C_T}\right) \tag{1.14}$$

This expression means that, in Steady-State (SS) condition, the cell evolution within the pulse is completely independent from the initial state, i.e. Qfg stored before the pulse, and on the constant CG level. This is intuitive, since after a certain time, the cell experiences a given Qfg amount, which induces a given Vfg, which means a certain Ifg, thus a ΔQfg and a new $\mathrm{Vfg} = \mathrm{Vfg_{old}} - \Delta\mathrm{Vfg}$, which leads to a lower FG current and so on. Thus, in SS condition, the Vfg dynamics must follow the universal law in eq.1.14. This is confirmed by experimental data. As shown in Fig.1.12, if the Program pulse has been fixed, the final Vth at the Program state will always be the same independently from the initial Erase state and vice versa, unless the initial state is very close to the final one. In other words, the final P/E state of the cell depends uniquely on P/E phase itself respectively. This is an important peculiarity that we will use in the following. In particular, whenever we will consider the degradation of the device, the final Vth at Program state, for example, will depend uniquely on the degradation of HC current and of cell electrostatics.

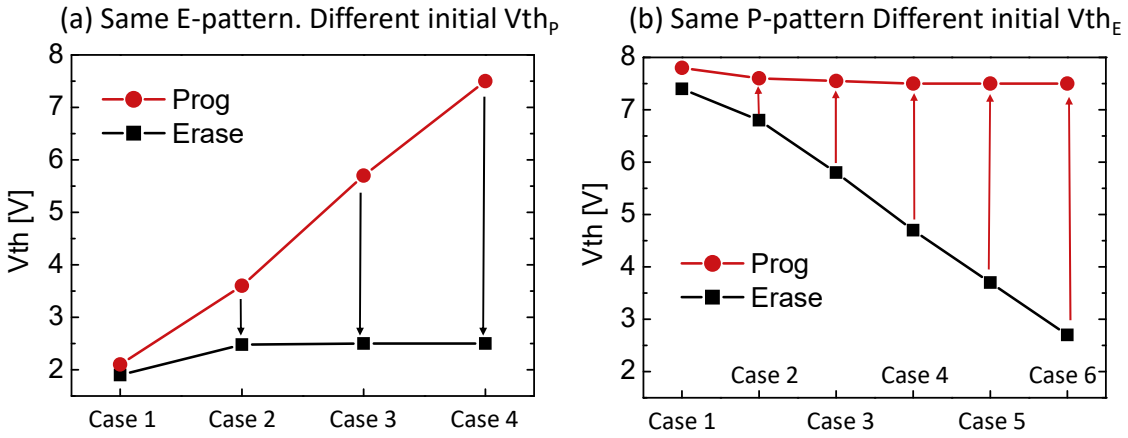


Figure 1.12: In (a) $\mathrm{Vth_E}$ state is shown for different initial $\mathrm{Vth_P}$ levels, i.e. fixed Erase pattern and different Program pulse durations. In (b) the opposite situation is reported.

For the sake of completeness, in Fig.1.13 the FN Ifg-Vfg characteristic is shown. This curve has been extracted using the SPE method, similarly as done before for the Program case, with $\delta\alpha_G \approx 5\%$. In the figure, the high values of uncertainty for the Vfg acquisition have been highlighted. In particular, in accordance with eq.1.9, the error induced in the Vfg extraction for a fixed Ifg level is:

$$\delta\mathrm{Vfg} = \sqrt{\delta\alpha_G^2\mathrm{Vcg}_{P/E}^2 + \alpha_G^2\delta\mathrm{Vth}^2 + \delta\alpha_G^2\mathrm{Vth}^2 + \delta\mathrm{Vth}_{mos}^2} \tag{1.15}$$

This represents the worst case, since assuming that the errors are decorrelated. In any case, the uncertainty resulting in Fig.1.13 turns out ot be huge. This comes from the first term of eq.1.15: assuming an α_G extracted with a precision of 5%, the Vfg dispersion increases up to 0.56V, since Vcg is set at -17V during SPE experiment. The error in the estimation of

Vfg is then translated directly in an exponential error of the gate current, which is directly responsible of the FN degradation. For this reason, the related FN aging dependence on FG potential is significant and needs very accurate extraction of the curve. Fortunately, the extraction of such a parameter has been previously refined, which brings the FN Ifg-Vfg extraction to be precise enough for our purpose.

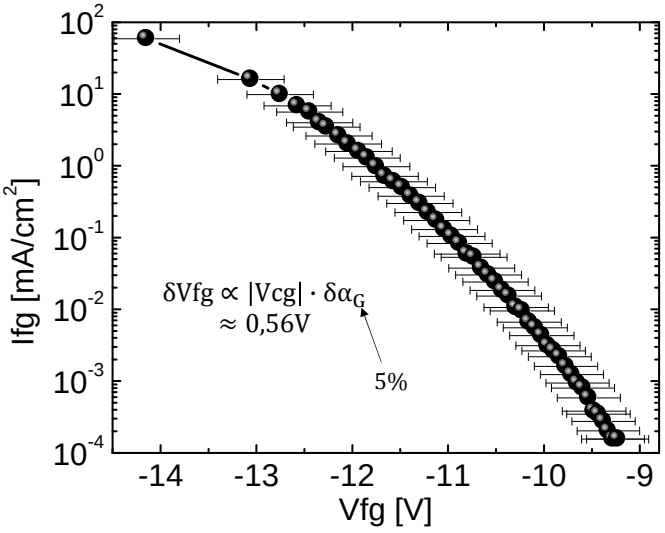


Figure 1.13: FN Ifg-Vfg curve extracted during SPE experiment. The error bars in the Vfg acquisition correspond to an α_G precision of only 5%.

After having properly defined the Step Pulse experiment and seen the related features, the bias dependences, i.e. Vth(t) variation with applied potentials, have to be addressed. In particular, the Vcg dependence results to be very interesting. In Fig.1.14(a), the variation of Vth dynamics with the CG potential is shown during the Program phase. To understand this dependence, we take advantage of eqs.1.12, 1.13, 1.14. Indeed, as previously highlighted, the Vfg evolution is independent from the CG potential after a certain time, as TCAD simulations confirm in Fig.1.14(b). This means that, ΔVth(t) must coincide with the applied ΔVcg, as experimentally verified looking at the three curve of Fig.1.14(a).

It is worth noting that these observations are true only if the gate coupling coefficient is the same during P/E regime and in read phase. If for example $\alpha_{G,P/E}$ is lower than the respective parameter in read mode, the resulting ΔVth(t) < ΔVcg. In particular:

$$\Delta \mathrm{Vth(t)} = \frac{\alpha_{G,P/E}}{\alpha_{G,lin}} \Delta \mathrm{Vcg} \tag{1.16}$$

This should be the case during the Program operation, since the drain capacitance is higher in this regime than in read operation. However, as experimentally verified, this is completely negligible. On the other hand, $\alpha_{G,E} \equiv \alpha_{G,lin}$, because in both strong inversion and accumula-

tion regions the total capacitance seen at FG node must be close to Cox, since V-dependent depletion capacitance in Si-substrate results to be negligible in both cases.

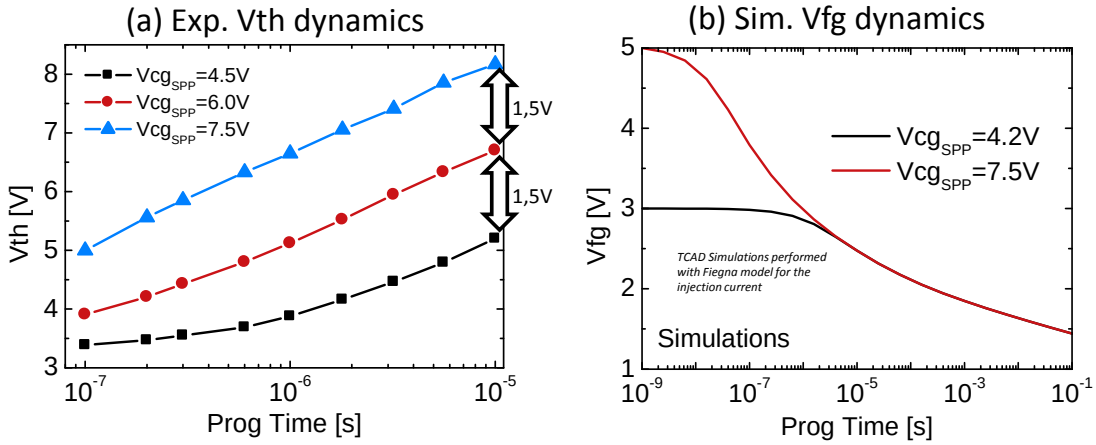


Figure 1.14: SPP carried out at different Vcg levels. In (a) experimental results, in (b) TCAD simulations have been performed with Fiegna model (see next section).

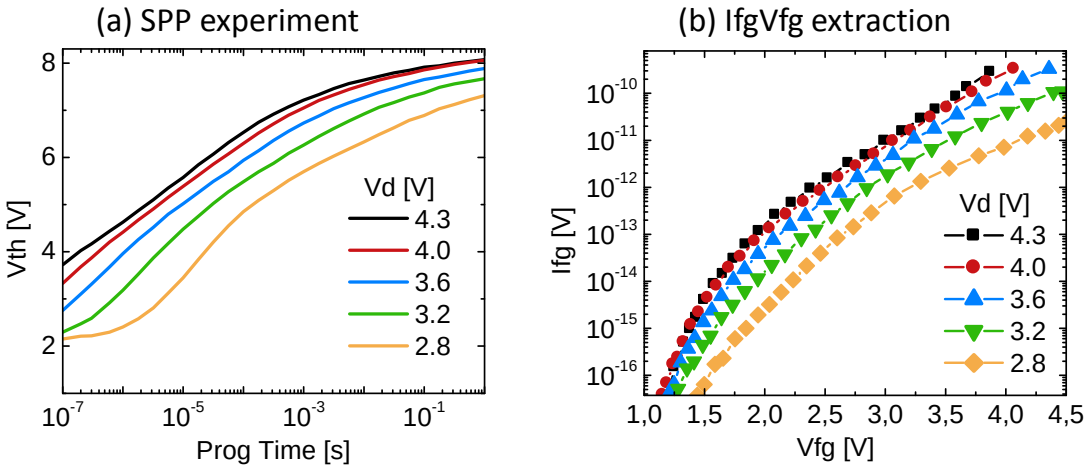


Figure 1.15: SPP performed at different Vd levels. In (a) experimental Vth evolution and in (b) the resulting extracted Ifg-Vfg curves are shown.

For the sake of completeness, the Vd dependence of Vth dynamics during SPP experiment is shown in Fig.1.15(a). It has to be remarked that these curves come from different Ifg-Vfg characteristics, as reported in Fig.1.15(b). The Vd dependence during the Program phase is addressed with very high precision since the resulting Ifg-Vfg curves are extracted on the same device reducing the uncertainty to zero. It has been verified that performing such an experiment several times on the same cell does not degrade significantly the device itself and thus alter the extraction (not shown here).

To conclude the subsection, the Step Pulse approach has been used to finely extract cell characteristics during P/E operations. This helps to define the pulse shape and duration for a given PW. Moreover, we will show in the following that such a method results to be fundamental for Flash cell aging understanding and modeling. Indeed, thanks to this approach, we will be able to directly extract the Flash Vth degradation induced by the P/E efficiency loss.

1.2.3 Impact of device geometry

After having shown the Step Pulse experiment and the gate coupling coefficient extraction, we now address the related W/L dependences. In Fig.1.16 the extracted α_G is shown function of geometry dimensions. The gate coupling coefficient results to linearly decrease with the device width and to be constant with the length. These dependences are easily understandable considering the Flash cell geometry along the device length (Fig.1.1) and width (Fig.1.6(d)).

Considering firstly the W dependence, we would not expect any significant dependence if the transistor can be approximated as a perfect 2D structure, since $\alpha_G \approx C'_{ono} \cdot W/(C'_T \cdot W)$. However, most of Flash structures experiences a CG electrode surrounding the side-walls of the FG along W, as shown in Fig.1.6(d). In particular, the oxide thickness is not constant along this dimension: at the edges Tox is higher. Thus, increasing the width, the average Tox $\searrow$, which induces an increase of the overall oxide capacitance Cox. Hence, the gate coupling coefficient $\alpha_G \approx C'_{ono}/(\cdots + C'_{ono} + Cox')$ decreases accordingly, as experimentally verified in Fig.1.16(a) .

Concerning instead the cell length dependence, the situation is more complex. Similarly to before, the ideal case would be $\alpha_G \approx C'_{ono} \cdot L/(C'_T \cdot L)$. However, with device down-scaling, parasitic capacitances Cpar do not scale with L and the Short Channel Effects reduce the substrate capacitance Csub. Since these two phenomena weakly impact α_G and have an opposite effect with L scaling, i.e. $\alpha_G \approx C'_{ono} \cdot L/(\cdots + C'_{ono} \cdot L + Cpar(=) + Csub'(\searrow) \cdot L)$, the gate coupling coefficient does not significantly vary with this dimension, as experimentally verified in Fig.1.16(b).

Concerning now the Step Pulse experiment with device dimensions, the situation is summarized in Fig.1.17. In this case, the dependences are several: Vth_{mos}, Ifg-Vfg and α_G. Considering the device length variation, just the first two dependences have to be considered, since the gate coupling coefficient has been observed almost constant with L variation. Thus, if Vth_{mos} does not vary with L neither, we would expect an increase in the injection efficiency when reducing this dimension, as similarly observed with Vd increase, since the lateral electric field, and consequently the gate current, increases. However, as previously remarked, after a certain time $Vth(t) \propto Vth_{mos}$ and the threshold voltage of the transistor

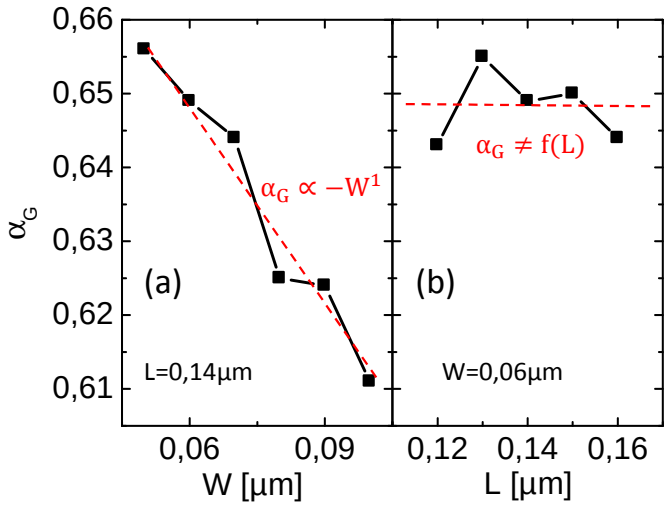


Figure 1.16: Extracted gate coupling coefficient as a function of device dimensions: width W (a) and length L (b).

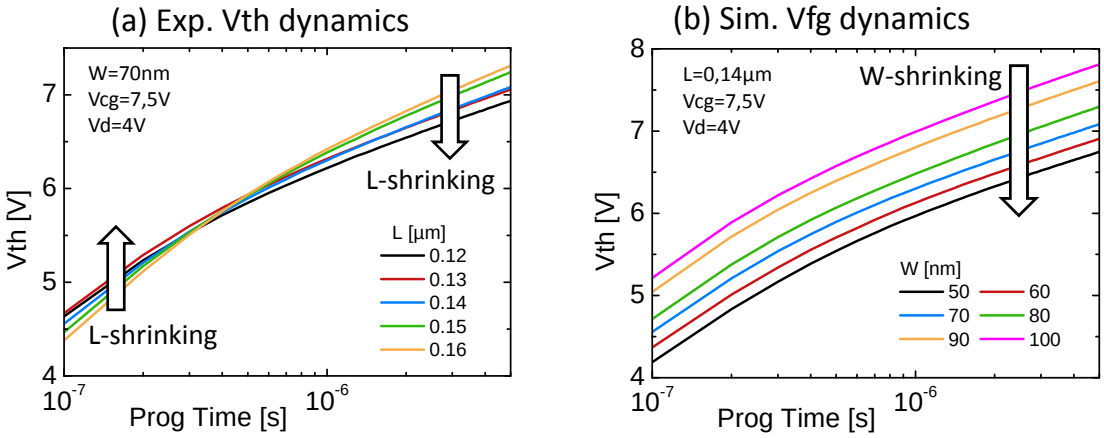


Figure 1.17: SPP experiments performed with different device lengths L (a) and widths W (b).

decreases with L-scaling due to the well-known SCE. Looking at the Fig.1.17(a), the two effects are clear. An improvement is observed in the beginning of the dynamics with L-scaling, since the injection current is (slightly) higher. On the other hand, after a certain time, the injection efficiency is reduced since the Vth evolution is proportional to the MOS threshold voltage. This is a strong limitation for device down-scaling. Indeed, a loss in the injection efficiency is directly translated in a decrease of cell lifetime, since the cell Program pulse duration has to be increased, in order to have the same PW, and the induced HC degradation per cycle is consequently accelerated.

Considering the W dependence of SPP, the situation is more simple. The main contri-

bution affecting the SPP variation is $\alpha_G(W)$. Coming back to eq.1.7, we can write:

$$Vth(t) = \frac{1}{\alpha_G} Vth_{mos} - \frac{\alpha_D}{\alpha_G} Vd_{lin} - \frac{Qfg(t)}{C_{ono}} \tag{1.17}$$

In a first approximation, the last two terms are almost constant with W variation, whereas the first one varies in accordance with $\alpha_G(W)$ and $Vth_{mos}(W)$. In particular, decreasing W, both contributions make the Vth(t) lower (especially $\alpha_G \nearrow$), which explains the experimental results in Fig.1.17(b).

Similar observations can be done for the Vth dependence on geometrical dimensions in SPE case. However, we skip this representation since redundant.

To conclude the subsection, the Step Pulse experiment has been shown to be an important tool for addressing geometrical device dependences. In particular, SCE have been demonstrated to be the limiting factor for Program efficiency, and thus for technology scaling itself. Such problems can be easily handled taking advantage of TCAD simulations, which represent a powerful tool especially for device electrostatic computation. In the next section, we make use of these tools in order to properly simulate the Flash cell and have a device model reference for the thesis.

1.3 Flash cell modeling

In this section, we briefly go through Flash cell modeling, from device calibration to gate current calculation. These models and the physical observations done in this section represent a solid basis on which next chapters are based.

1.3.1 TCAD device model

Taking advantage of Synopsys tools [54], we calibrated our device model. Firstly, we matched the real device dimensions (measured during the process) with the process simulation. The resulting structure is shown in Fig.1.18. It is worth noting that a 2D geometry has been considered for simplification purpose.

Afterwards, the cell characteristics have to be calibrated, i.e. Vth, Gm and so on. To address this point, the sequent approach has been adopted:

- Linear Id-Vg equivalent transistor calibration. For better comprehension and for avoiding misunderstandings, such a characteristic is called Id-Vfg from now on.

  Vth_{mos} calibration considering a Si/SiO_2 interface fixed charge.

  Gm_{mos} calibration tuning mobility parameters for Canali [55], [56] and Lombardi [57] models and considering ad-hoc series resistances Rsd at Drain and Source contacts.

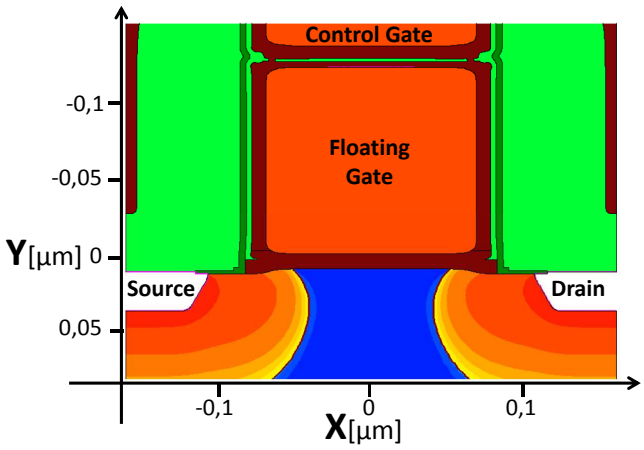


Figure 1.18: TCAD model of the structure considered in this thesis. The doping concentration scale is skipped since confidential.

- Linear Id-Vcg Flash cell calibration.

  Vth calibration varying the FG charge.

  α_G calibration tuning an additional capacitance placed at FG node, called Cfg_{3D}.

It is worth noting that in this approach the Sub-threshold Slope calibration has been skipped: there is no need for this, since the simulations already reproduce this parameter with the resulting doping configuration. On the other hand, the piezoresistance effect [58], [59], [60] has not been accounted for in the mobility calculation. Indeed, the available models are too approximated and add a lot of complexity increasing exponentially the computational time and the number of fitting parameters.

Considering firstly the equivalent transistor calibration, the Coulomb Scattering model for the electron mobility in channel has been taken from [61], [62], [63], where a fine calibration has been performed on FDSOI technology. On the other hand, for high field dependence, the Canali model has been considered:

$$\mu(\mathrm{E}) = \frac{\mu_0}{\left[1 + \left(\frac{\mu_0 \mathrm{E}}{\mathrm{v_{sat}}}\right)^{\beta}\right]^{\frac{1}{\beta}}} \tag{1.18}$$

where E is the electric field, μ_0 is the low field mobility, $\mathrm{v_{sat}}$ is the electron saturation velocity in Silicon and β is an empiric parameter. This last parameter has been set to 1.25 (the nominal value is equal to 1.213). Then, in order to perfectly match the experimental Gm-Vfg curve, Rsd has been considered equal to 25Ω.

Concerning the $\mathrm{Vth_{mos}}$, a positive Si/SiO_2 interface charge has been tuned in order to match this electrical parameter without significantly modifying the channel mobility. In

particular, a value of $Q_{surface} = 7 \cdot 10^{10} cm^{-2}$ has been chosen. The Id-Vfg curves at different device lengths are shown in Fig.1.19(a): simulations well reproduce the experimental results.

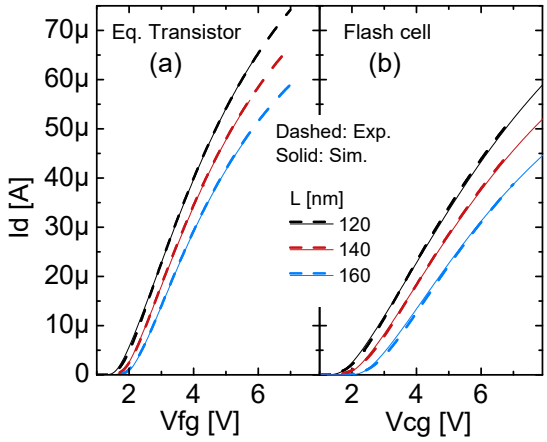


Figure 1.19: Linear Id-Vg simulations vs experiments (Vd=0.5V) at Erase state. Equivalent transistor and Flash cell are considered in (a) and (b) respectively. It is worth noting that, for each device length, each couple TR(a)/Flash(b) represents one simulation.

Once equivalent transistor calibration is achieved, it is simple to reproduce the cell Id-Vcg curve too. Knowing the relationship:

$$\alpha_G = \frac{C_{ono,2D} + Cfg_{3D}}{C_{T,2D} + Cfg_{3D}} \tag{1.19}$$

the gate coupling coefficient can be easily matched. Indeed, the two parameters $C_{ono,2D}$, $C_{T,2D}$ can be extracted varying twice Cfg_{3D}. Then, the right value of Cfg_{3D} can be addressed by matching the experimental gate coupling coefficient using eq.1.19. The resulting Id-Vcg simulations are compared with the experimental data in Fig.1.19 and a good fitting quality is achieved.

A calibrated device model helps not only in the physical comprehension at microscopic level, but also it allows to easily predict the effects of process variations on currents and on the electrostatics. This is fundamental for industry, since the application on wafer of process variations takes time and financial efforts.

1.3.2 HC current modeling

The modeling of the injection current during Hot Carrier regime is an important and difficult challenge. Fiegna was the first to propose an approximated model of this phenomena in 1991 [64]. The total gate current can be written as:

$$Ig = q \int P_{ins} \left(\int_{Eb}^{\infty} v_{\perp}(\epsilon) f(\epsilon) g(\epsilon) d\epsilon \right) ds \tag{1.20}$$

where ϵ is the electron energy, Eb is the height of the semiconductor-insulator barrier, $v_\perp$ is the velocity normal to the interface, $f(\epsilon)$ is the electron energy distribution function, $g(\epsilon)$ is the density-of-state of the electrons, P_{ins} is the probability of scattering in the image force potential well [54] and s is the coordinate along the channel interface. The difficulty of modeling a device in HC regime is to accurately calculate the electron distribution in energy. Fiegna simplified this task assuming that the phonon scattering is the limiting mechanism factor, the lattice and the carriers are in equilibrium and the band structure is parabolic and isotropic. Making these simplifications, he ended up with an analytical expression for the energy distribution function:

$$f(\epsilon) = A \cdot \exp\left(\chi \frac{\epsilon^3}{F_{lat}^{1.5}}\right) \tag{1.21}$$

where χ and A are semi-empirical parameters nominally equal to $4.87 \cdot 10^4 cm/(s \cdot eV^{2.5})$ and $1.3 \cdot 10^8 (V/(cm \cdot eV))^{1.5}$ respectively. Considering this expression for the energy distribution function in eq.1.20, the gate current expression can be rewritten as:

$$Ig = q\frac{A}{3\chi}\int P_{ins} n \frac{F_{lat}^{1.5}}{\sqrt{Eb}} \exp\left(\frac{\chi Eb^3}{F_{lat}^{1.5}}\right) ds \tag{1.22}$$

This model has been calibrated with our experimental data, in particular by matching the SPP results. With this aim, two possible strategy can be adopted. One solution is to simulate the Vfg (i.e. Qfg) dynamics on the cell structure with a constant CG pulse, then sample some Qfg_i at certain times t_i and afterwards simulate the Id-Vcg curves considering Qfg_i at FG node. This allows to calculate Vth_i at t_i. A different method relies on the assumption of constant coupling coefficients. In particular, from the simulated Vfg(t) and Vth_{mos}, it is directly possible to calculate the Vth dynamics from eq.1.9. These two simulation approaches are shown in Fig.1.20. It is worth noting that the results are very similar, thus the second method has been preferred, since demanding less simulations.

The calibration has been performed for three channel lengths. The results are shown in Fig.1.21 and a good fitting quality can be noticed. Such simulations allow to address the cell dynamics within the first 100ns. Indeed, lower dynamic times are impossible to capture experimentally with high accuracy, since the pulse potential on CG is applied with a certain delay due to parasitic capacitance effects, which limit the minimum pulse duration.

Such an approach represents a good compromise between complexity and accuracy for the gate current calculation during the Program operation. However, this model may be not suitable for the degradation computation. Indeed, for the HC-aging simulation, we will see that the accurate knowledge of the energy distribution function is necessary in order to get the physical nature of the degradation process. For this reason, we will use the Spherical Harmonic Expansion (SHE) method [65] to simulate the device under HC-aging regime. Such a model computes the energy distribution function by solving the lowest-order SHE of the Boltzmann Transport Equation (BTE). This method has the advantage of performing

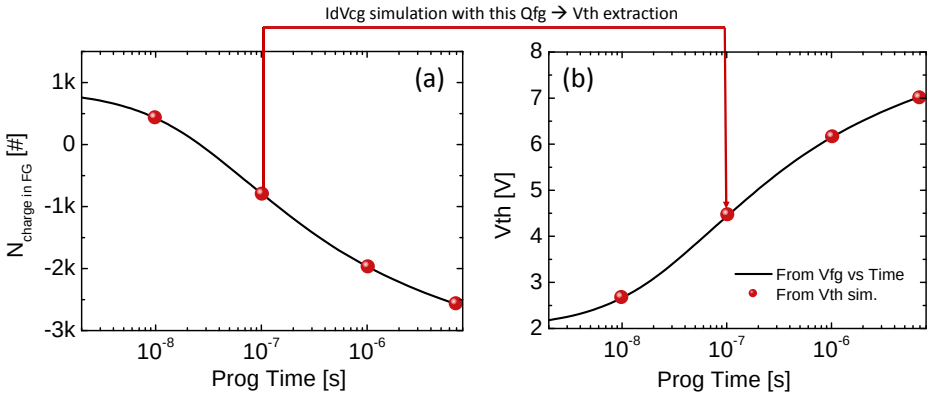


Figure 1.20: Vth dynamic simulation using Fiegna model [64]. From the Qfg dynamics (a), Vth evolution can be computed (b). This can be done simulating the Id-Vcg with the obtained Qfg or directly from eq.1.9.

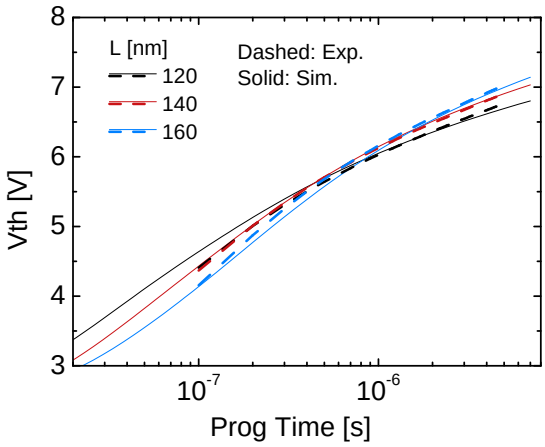


Figure 1.21: SPP simulations [64] vs experiments at different channel lengths.

the simulation in non-equilibrium regime, with the drawback of exponentially increasing the computational time [66].

1.3.3 FN current modeling

Concerning the Fowler Nordheim erase current, the simulation is much more simple. Indeed, it is necessary just to simulate the device structure using the classical Drift-Diffution model and then the current is computed:

$$\mathrm{Ig} = \mathrm{q} \int_0^\infty \mathrm{T}(\epsilon)\mathrm{n}(\epsilon)\nu(\epsilon)\mathrm{d}\epsilon \tag{1.23}$$

where $T(\epsilon)$ is the barrier transparency, $n(\epsilon)$ is the electron concentration at the Si/SiO_2 interface and $\nu(\epsilon)$ is the impact frequency. The 1D approximation has been implicitly done, since Vfg is quite high in absolute value making the channel uniform from an electrostatic standpoint, as we will see later. The electron density is simply equal to $f(\epsilon) \cdot g(\epsilon)$, where f is the Fermi-Dirac statistics, g the electron Density of States (DOS) and standard classic approximations have been considered. On the other hand,

$$T(\epsilon) = \exp\left[\frac{-4\sqrt{2m_{SiO_2}}Tox}{3q\hbar Vox}(Eb-\epsilon)^{\frac{3}{2}}\right] \tag{1.24}$$

$$\nu(\epsilon) = \frac{Q_{Si}}{2\epsilon_{Si}\sqrt{2m_{Si}\epsilon}} \tag{1.25}$$

where m_{SiO_2}, m_{Si} are the effective masses in insulator and in Silicon respectively, Vox is the potential drop across the oxide and Q_{Si} is the total charge in Silicon substrate (inversion and depletion). The calculation of the impact frequency has been performed with the assumption of a triangular barrier at the interface [67], [68], whereas the computation of $T(\epsilon)$ comes from the well-known Wentzel-Kramers-Brillouin (WKB) approximation [69], [70], [71]. This approach, in our case, gives results very close to the exact solution based on Airy functions [72], [73] (not shown here), since the oxide thickness Tox is significant. Such a simulation has been performed on our devices: an example is shown in Fig.1.22.

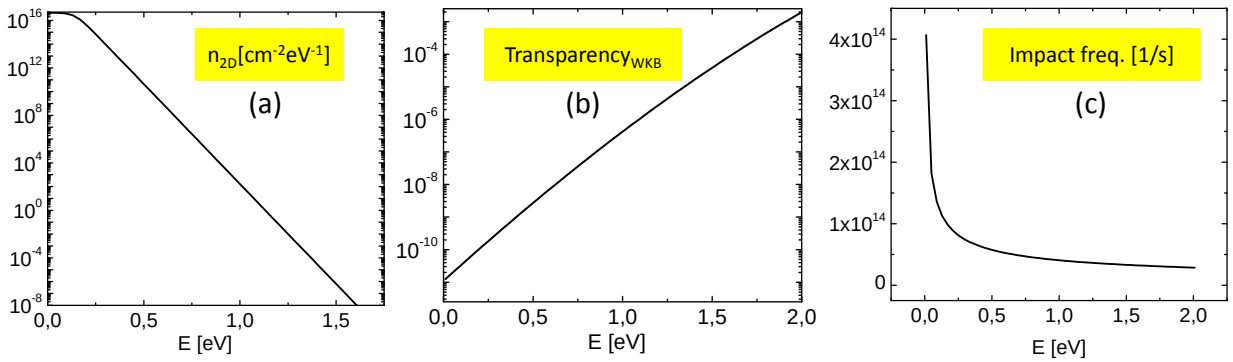


Figure 1.22: WKB "microscopic" approach. The integral in energy of the electron concentration (a) multiplied by the tunnel transparency (b) and by the impact frequency (c) gives the gate current, in accordance with eq.1.23. Vfg=-12V has been considered.

The task can be facilitated considering the well-known simplified WKB analytical expression [74], i.e. denoted "macroscopic" WKB model in the following, which differs from the previous "microscopic" one. In particular, the current for thick oxides reduces to:

$$Ig = A \cdot F^2 \exp\left(-\frac{B}{F}\right) \tag{1.26}$$

where F is the oxide electric field and A and B are equal to:

$$A = \frac{q^3 m_{Si}}{8 \cdot \pi \cdot m_{SiO_2} \cdot h \cdot Eb} \tag{1.27}$$

$$B = \frac{4\sqrt{2 \cdot m_{SiO_2} \cdot Eb^3}}{3 \cdot q \cdot \hbar} \tag{1.28}$$

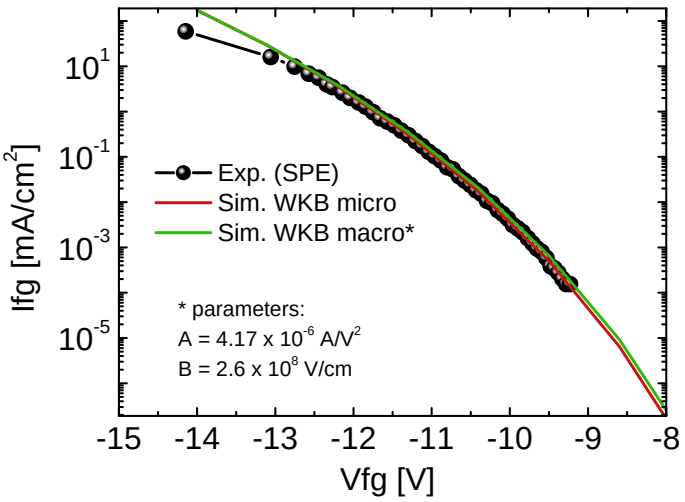


Figure 1.23: Experimental Ifg-Vfg obtained with SPE method is compared with WKB approaches, i.e. standard (micro) and simplified (macro) WKB models.

The comparison between these results and the respective experimental data, obtained through the SPE approach, are shown in Fig.1.23. It is worth noting the accuracy of the simulations and the perfect match between the two approaches. This means that we are allowed, from now on, to consider the "macroscopic" approach, which will be useful in the following for the degradation computation. The small mismatch for high current comes from a high incertitude in the experimental Ifg calculation after the first pulse of SPE experiment, since the gate current cannot be considered constant within it.

Concerning the 1D approximation, an additional simulation has to be performed. Taking advantage of the TCAD structure model, we simulated the FN regime and looked at the electrostatic results, which are shown in Fig.1.24. It is worth noting that the conduction band Ec is constant along the interface with a certain precision, which makes the 1D approximation of the system reasonable. For the sake of the precision, the LDD regions experience higher electric field, thus higher electron flow. However, these regions do not significantly influence the device currents and characteristics, thus this feature will not be taken into account in the following. On the other hand, it can be noticed that the electron quasi-Fermi level is not constant in these regions. This justifies the S/D currents which can be experimentally

observed during FN regime on large devices, since a portion of the electron flow injected in the substrate is "captured" by the junctions. Anyhow, the current experimentally addressable on S/D contacts does not correspond to the flow in the overlap region, since a large part of the electron flux close to the channel center is then deviated by the electrostatics of the system and captured by S/D contacts.

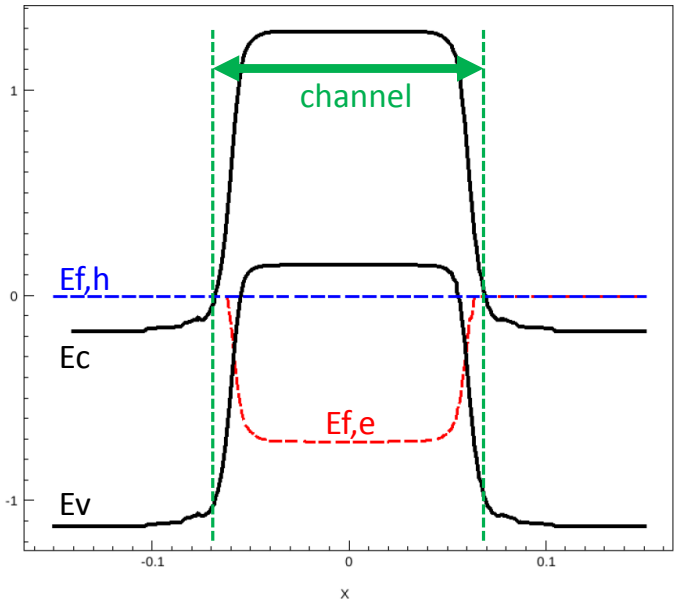


Figure 1.24: TCAD simulation of band structure in FN regime with Vfg=-12V. In the figure, the bands and the quasi-Fermi levels are shown along the channel interface. Vfg=-12V has been considered.

1.4 Experimental approach

In this section we present the experimental setup, which has been developed during the thesis, and the modeling of the CG pattern of the cell. Details of these characterization approaches are reported in our publications [75], [76].

1.4.1 The experimental setup

In conventional Flash endurance characterization setup, an electromechanical switch is needed in order to combine current measurement (Read operation) and pulse generation capability (Program and Erase operations) [77]. However, this leads to a poorly controlled delay, i.e. in the 1-100ms range, before Read operation. Such an uncontrolled delay may lead to an important uncertainty in the measurement [75], because of relaxation effects on degraded cells.

The new Flash NOR characterization setup used in this thesis is shown in Fig.1.25: it allows endurance measurements with controlled delays between Flash operations (Program, Erase and Read) down to 50ns. It features 4 synchronized Keysight B1530 Fast-Measurement Units (FMUs) enabling pulse generation up to 10V together with aggressive current sampling capability (time resolution = 5ns). In particular, FMUs are dedicated to Program and Read operations, whereas a high-voltage pulse generator (PG) is specific for Erase operation. A unique CMOS switch connected to cell CG is needed for accurately synchronizing FMUs and PG: such a switch has been developed offering ns-range switching durations and no parasitic voltage glitches. CMOS off-leakage is not a concern for this experimental setup, since CG shows high impedance.

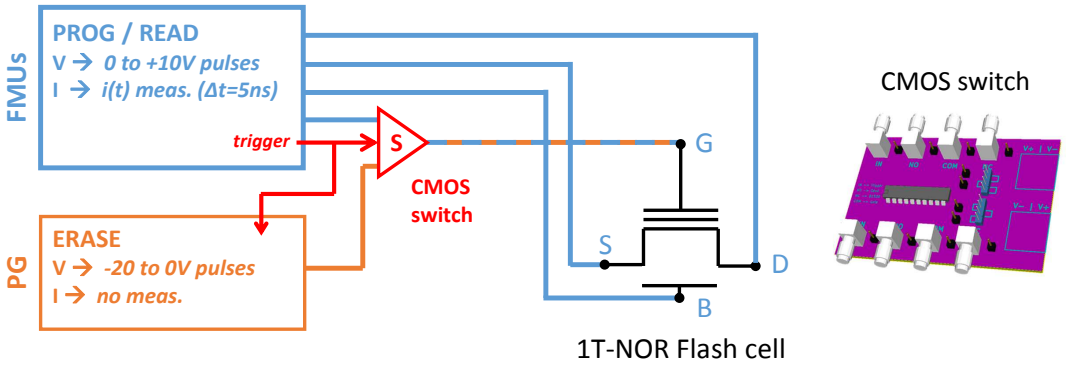


Figure 1.25: Schematic representation of the advanced Flash endurance characterization setup used in this thesis: FMUs (Fast Measurement Units), PG (Pulse Generator), which are paired with a CMOS switch (represented on the right).

During P/E cycling, this setup provides fast Id-Vcg trans-characteristics (measurement duration in the 10-100μs range) during Read operations in order to avoid parasitic stretches due to device trapping [78]. In addition, during Program operations, dynamic Drain, Bulk and Source currents (Id(t), Ib(t) and Is(t)) are systematically measured. This provides cell dynamics during Program operation thanks to a specific extraction methodology [75].

Reference fast Id-Vcg characteristics have been measured and averaged on a set of equivalent Flash transistor structures at $Vd_{TR} = Vd_{Program}$, allowing to link cell drain current $Id_{Flash}(t)$ measured during Program phase with its FG potential Vfg(t). It is worth noting that fast measurement technique is considered in order to get rid of any measurement-induced HC degradation. Fig.1.26 shows the Vfg(t) extraction principle, applied on a fresh Flash cell before endurance experiment, i.e. P/E cycling stress.

From Vfg(t) dynamics extracted during Program phase, conventional Flash equations are applied [10] in order to extract Qfg(t), Vth(t) and Ifg(t). Thus, such an approach allows to estimate the Ifg-Vfg characteristic, as similarly done with the SPP case. It has to be

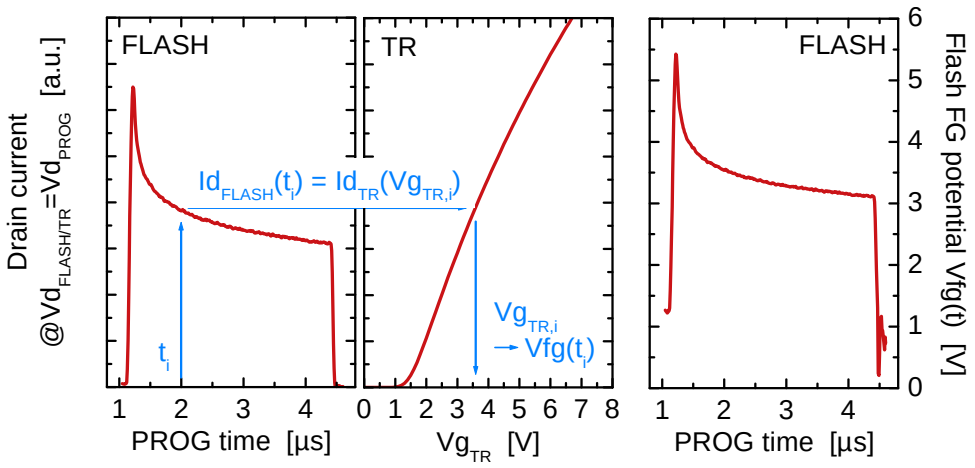


Figure 1.26: (left) $Id_{Flash}(t)$ measured during standard Flash Program operation, (middle) Id-Vfg characteristic of the equivalent transistor (i.e. TR in the figure) measured at $Vd_{TR} = Vd_{Program}$ and (right) the resulting extracted Vfg evolution during Flash Program phase.

pointed out that such a method represents the complementary approach of SPP experiment: it does not bring accurate results for low gate currents (where SPP is very precise) and it calculates high current levels with high precision (where SPP is weak in the accuracy). This occurs since in the beginning of the pulse the drain current is high and the related decrease is fast, whereas, for long Program time, the current is low and almost constant, thus the noise can strongly disturb the Vfg(t) extraction. Combining this approach with SPP method, the Ifg-Vfg curve can be captured in a wide Vfg range.

However, it has to be pointed out that this procedure can not be accurately applied along cell cycling, since we should have the corresponding equivalent transistor stressed exactly in the same way. We will see how to overcome this problem by taking advantage of the optimized ramp patterns, which are described in the next subsection.

1.4.2 CG pattern modeling

While important integration efforts are in progress for reducing biases and currents, a proper management of Flash operation schemes seems to be a promising path for consumption reduction [79]. In this part, we propose a physically-based model providing optimized CG pulse patterns for control and managing operation efficiency and Program current, which is of prime interest since determining charge pump scalability. In the following, different optimized patterns are simulated and their impact on Flash cell characteristics is electrically assessed. This is performed using the Flash endurance experimental setup previously presented, which guarantees intrinsic device characterization deprived from any relaxation effect

and which provides Id(t) measurement during the Program phase. From this last feature, it is possible to extract Id_{MAX} and the cell consumption E_p during the Program phase. The last one is equal to

$$Ep = \int_{t_P} Id(t) Vd_P dt \tag{1.29}$$

These cell characteristics will be used in the following.

A standard, i.e. square-shaped, CG Program pattern induces a Vfg dynamics as shown in Fig.1.26(right). The decrease of Vfg during Program time is due to FG filling of electrons. Thus, for this standard pulse, the overall cell programming occurs in the first μsecs, whereas a clear Vfg saturation appears at the end of the operation. This indicates efficiency loss when reaching low FG values, and, in addition, power waste and pointless HC stress. Moreover, such a Vfg evolution leads to non-constant Id(t) characteristic and to the presence of a sharp Id_{MAX} peak (Fig.1.26(left)). This means that the standard Program scheme also appears to poorly benefit from charge pumps capabilities.

The CG pulse modeling aims to compensate Vfg decrease during the Program phase thanks to a continuous increase of the Control Gate potential itslef. Contrary to the Incremental Step Pulse Programming (ISPP), in which narrow pulses of increased amplitudes are considered [80], a single ramp-shaped CG pulse guarantees constant Vfg, Id and thus operation efficiency during the entire Program phase, making this approach suitable with fast HC-programmed NOR Flash cells. In Fig.1.27, the optimized Program CG pattern is represented. The pulse shape is defined by 3 parameters:

- Vcg_{start} is set in order to reach $Vfg = Vfg_{target}$ after the pulse rise:

$$Vcg_{start} = \frac{Vfg_{target} - \alpha_D Vd_P - \alpha_G (Vth_0 - Vth_E)}{\alpha_G} \tag{1.30}$$

- The pulse slope S_p is set in order to maintain the gate current target and is given by:

$$S_p = \frac{Ifg(Vfg_{target})}{C_{ono}} \tag{1.31}$$

- The pulse duration t_P is set in order to guarantee the desired Programming Window (PW_{target}), i.e. $Vth_P - Vth_E$, and is given by:

$$t_p = \frac{PW_{target}}{S_p} \tag{1.32}$$

The model input parameters are Vfg_{target} (or Id_{target}) and PW_{target}, whereas Vth_E directly comes from the Erase pulse and is independent from the Program phase, as previously demonstrated.

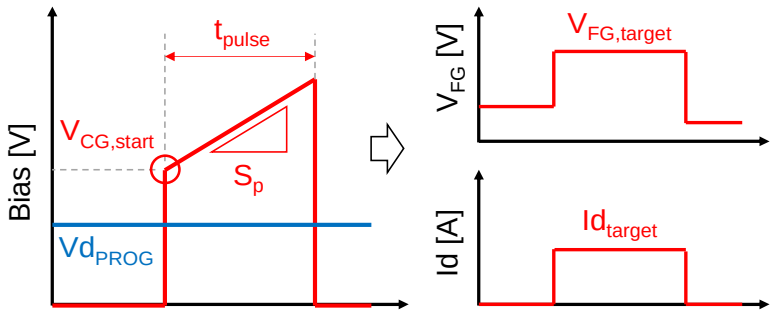


Figure 1.27: Schematic representation of optimized CG pattern, aiming to provide constant $Vfg = Vfg_{target}$ and $Id = Id_{target}$ during the whole Program phase.

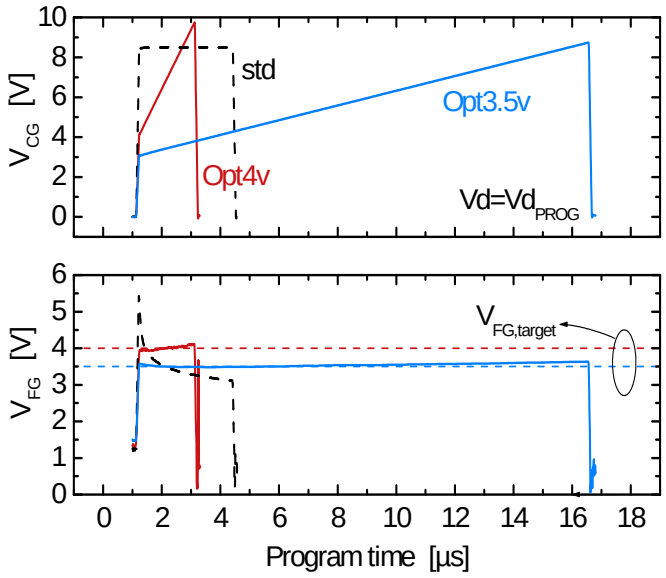


Figure 1.28: (top) Control Gate pulse model outputs, for 2 different Vfg_{target} (3.5V and 4V), and (bottom) associated experimental Vfg extraction from Id(t) measurements.

Two optimized Program patterns have been considered with $PW_{target} = 5.75V$ (for comparison with standard program scheme): Vfg_{target}=3.5V and Vfg_{target}=4V (denoted Opt4v and Opt3.5v in the following). Fig.1.28 provides a time-representation of the optimized CG patterns together with the corresponding Vfg(t) extracted from Id(t) measurements. It is worth noting that $t_{pulse,3.5V} > t_{pulse,4V}$ because the first pattern provides less efficiency and same PW amplitude. Experimental Vfg(t) extractions show the expected constant behavior and are close to targeted values, which highlights the accuracy of the methodology used.

Furthermore, CG pulses have been modelled in the simulation-only domain, leading to the representation of the best Id_{MAX}/E_p trade-off for different PW (Fig.1.29). This is of prime interest for technology design since addressing application-dependent scaling/consumption

trade-off. It has to be pointed out that, looking at the Fig.1.29, a large range of optimized Program patterns even improves both characteristics respect to the standard scheme.

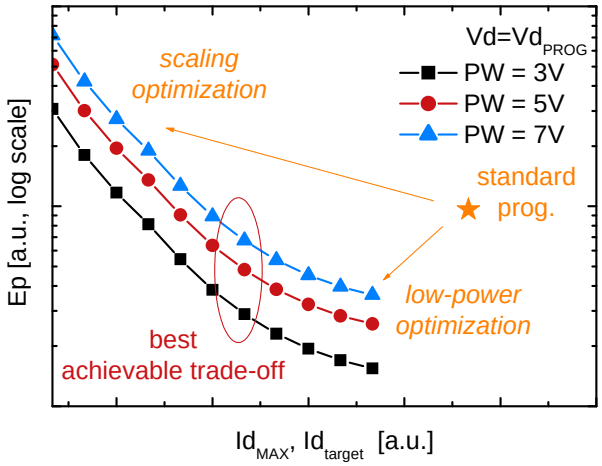


Figure 1.29: Simulated Id_{MAX} / E_p relationship for different PW amplitudes, showing the best achievable scalability / power consumption trade-off for a given application.

This ramp-pattern methodology will be uniquely considered from now on, also for the Erase phase, whose representation has been skipped since redundant. It is worth noting that, for the Erase case, the SPE approach has been used as a reference to model the CG ramp pattern. It can be easily demonstrated that the resulting pulse modeling is independent from the total capacitance C_T, which makes such an approach strongly robust.

We will take advantage of such "constant Vfg" approach to study, on the equivalent transistors only, the degradation mechanisms suffered by the cell separately. Indeed, studying the Program-only or Erase-only induced aging directly on the cell is not possible. However, the optimized ramp-pattern approach developed in this section allows to compare Flash results with the ones measured on the equivalent transistors, which are stressed with the respective square patterns. The advantage of using such devices is that we can consider P/E operations together or separately, since the FG potential is fixed, directly applied. This allows to study the impact of each aging mechanism, i.e. Hot Carrier Degradation and Fowler-Nordheim Stress, and the related interactions.

In particular, in next chapter, the Hot Carrier Degradation will be considered, whereas the aging induced by the Fowler-Nordheim process will be studied in the third one. Finally, we will come back to the Flash cell in the last chapter and we will evaluate and address the different aging contributions induced during the endurance experiment, which will be performed with the optimized ramp-patterns.

1.5 Conclusions

In this chapter, a general overview of Flash cell technology, together with the device electrical characterization and modeling has been proposed.

After having shown the memory working principle and the technology architectures and limitations, we focused our attention on the Flash cell characteristic extractions. In particular, we showed how to extract the gate coupling coefficient taking advantage of equivalent transistor structures. This has been performed with high accuracy overcoming the well-known extraction noise coming from RDF and RTN concerns. On the other hand, Step Pulse experiment has been presented. Such a methodology has been used to estimate the Ifg-Vfg characteristics during P/E operations. In particular, a theoretical study has been done on the evolution of Vth during such an experiment, highlighting that the device dynamics does not depend on initial state and applied CG potential whenever reaching the Steady-State condition. In addition, geometrical dependences are considered, giving important insights on scaling strategy.

In the third part of the chapter, the modeling of the Flash cell and of the related tunnel oxide current during P/E phases is addressed. An accurate methodology for calibrating the device model has been proposed. On the other hand, the Fiegna model has been used and calibrated on SPP results for Hot Carrier current; whereas a 1D WKB simplified model has been considered and validated for FN current.

In last part of the chapter, we showed our advanced experimental setup used in this thesis, which allows customizable shape waveforms for CG patterns during P/E phases, together with delay-free system for keeping P/E cycling tests within reasonable durations and for getting read of SILC concern. Taking advantage of the cell model and of such an experimental setup, we optimized P/E operations in order to have constant equivalent FG patterns, which makes possible a perfect control of cell power consumption and maximum drain current.

Such an approach gives the possibility to study separately FNS and HCD mechanisms on the respective equivalent transistors by applying constant gate potentials. For this reason, before dealing with Flash cell endurance, such mechanisms are studied individually in next two chapters. We will go through those aging processes starting from the basis of the physical degradation principles, and then we will simplify the picture in order to build a model as simple as possible without losing the physical and microscopic representation. As we will see in the last chapter of the thesis, this will allow to have a complete picture of Flash cell endurance and to build a physical-based model for the PW evolution along cycling which is suitable for industry and, thus, for technology optimization.

Bibliography

[1] Davide Garretto, Numerical and compact modeling og embedded flash memory devices oriented to IC design, PhD thesis 2011, Ecole polytechnique Federale de Lausanne.

[2] T. H. Ning, Hot-electron emission from silicon into silicon dioxide, Solid-State Electronics, vol. 21, 1978, pp. 273-282.

[3] E. Takeda, Y. Ohji, and H. Kume, High field effects in MOSFETS, in Electron Devices Meeting, 1985 International, 1985, pp. 60-63.

[4] H. Chenming, C. T. Simon, H. Fu-Chieh, K. Ping-Keung, C. Tung-Yi, and K. W. Terrill, Hot-electron-induced MOSFET degradation - Model, monitor, and improvement, Electron Devices, IEEE Transactions on, vol. 32, 1985, pp. 375-385.

[5] T. Simon, K. Ping-Keung, and H. Chenming, Lucky-electron model of channel hot-electron injection in MOSFET'S, Electron Devices, IEEE Transactions on, vol. 31, 1984, pp. 1116-1125.

[6] R. H. Fowler and L. Nordheim, Electron Emission in Intense Electric Fields, Proceedings of the Royal Society of London. Series A, vol. 119, May 1, 1928, pp. 173-181.

[7] C. Chang, M. Liang, x, S, C. Hu, and R. W. Brodersen, Carrier tunneling related phenomena in thin oxide MOSFET's, in Electron Devices Meeting, 1983 International, 1983, pp. 194-197.

[8] A. Bhattacharyya, Modelling of write/erase and charge retetion characteristics of floating gate eeprom devices, Solid-state electronics, v 27, n 10, p 899-906, 1984.

[9] A. Kolodny, S. Nieh, B. Eitan and J. Shappir, Analysis and modeling of FG eeprom cells, IEEE Transaction on Electron devices, v 33, n 6, p 835-44, 1986.

[10] P. Pavan, R. Bez, P. Olivo, E. Zanoni, Flash memory cells-an overview, Proceedings of IEEE, v 85, n 8, p 1248-71, August 1997.

[11] K. San, C. Kaya, D. Liu, T. Ma, and P. Shah, A new technique for determining the capacitive coupling coefficiencts in flash operations, Electron device letters, IEEE, v 13 , n 6, p 328-31, 1992.

[12] W. Choi and D. Kim, A new technique for measuring coupling coefficients and 3-d capacitance characterization of floating-gate devices, IEEE Transaction on Electron devices, v 41, n 12, p 2337-42, 1994.

[13] M. Wong, D. Liu, S. Huang, T. Inc and T. Dallas, Analysis of the subthreshold slope and the linear transconductance techniques for the extraction of the capacitance coupling coefficients of FG devices, IEEE Electron Device Letters, v 13, n 11, p 566-68, 1992.

[14] R. Duane, A. Concannon, P. O' Sallivan and A. Mathewson, Advanced numerical modelling of non-volatile memory cells, IEEE Solid-State Device Research Conference, 1998, p 304-7, 1998.

[15] L. Larcher, P. Pavan, L. Albani and T. Ghilardi, Bias and w/l dependence of capacitive coupling coefficients in FG memory cells, IEEE Transaction on Electron Devices, v 48, n 9, p 2081-9, 2001.

[16] P. Pavan, L. Larcher and A. Marmiroli, Floating gate devices: operation and compact modeling. Kluwer Academic Publisher 2004.

[17] L. Larcher, P. Pavan, S. Pietri, L. Albani and A. Marmiroli, A new compact DC model of FG memory cells without capacitive coupling coefficiencts, IEEE Transaction on Electron devices, v 49, n 2, p 301-7, 2002.

[18] B. Sheu, D. Sharfetter, P. Ko and M. Jeng, Bsim: Berkeley short-channel igfet model for mos transistors, IEEE Journal of Solid-State Circuits, v 22, n 4, p 558-66, 1987.

[19] C. Enz, F. Krummenacher and E. Vittoz, An analystical mos transistor model valid in all regions of operating and dedicated to low-voltage and low-current applications, Analog integrated circuits and signal processing, v 8, n 1, p 83-114, 1995.

[20] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, New ultra high density EPROM and flash EEPROM with NAND structure cell, in Electron Devices Meeting, 1987 International, 1987, pp. 552-555.

[21] Vincenzo Della Marca, CHARACTERIZATION AND MODELING OF ADVANCED CHARGE TRAPPING NON VOLATILE MEMORIES, PhD thesis, 2013.

[22] H. Hidaka, Evolution of embedded flash memory technology for MCU, in IC Design & Technology (ICICDT), 2011 IEEE International Conference on, 2-4 May 2011, pp. 1-4.

[23] K. Baker, Embedded Nonvolatile Memories: A Key Enabler for Distributed Intelligence, in Memory Workshop (IMW), 2012 4th IEEE International, 20-23 May 2012, pp. 1-4.

[24] ITRS, INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS, Process integration, devices, and structures. Technical report 2012.

[25] J.E. Brewer and M. Gill, Nonvolatile Memory Technologies with Emphasis on Flash: A Comprehensive Guide to Understanding and Using Flaash Memory Devices. Wiley, 2008.

[26] H. P. Belgal, N. Righos, I. Kalastirsky, J. J. Peterson, R. Shiner, and N. Mielke, A new reliability model for post-cycling charge retention of flash memories, in Reliability Physics Symposium Proceedings, 2002. 40th Annual, 2002, pp. 7-20,

[27] A. Chimenton, P. Pellati, and P. Olivo, Analysis of erratic bits in flash memories, Device and Materials Reliability, IEEE Transactions on, vol. 1, 2001, pp. 179-184

[28] A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, Hot carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature, International Reliability Physics Symposium (IRPS), 2009, pp. 531-546.

[29] S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, et al., Interface traps density-of-states as a vital component for hot-carrier degradation modeling, Microelectron. Reliab. 50, 1267-1272 (2010).

[30] C. Papadas, P. Morfouli, G. Ghibaudo, G. Pananakakis, Analysis of the trapping characteristics of silicon dioxide after Fowler-Nordheim degradation, Solid-State Electronics, v 34, n 12, p 1375-9, December 1991.

[31] Young-Bog Park, D.K. Schroder, Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM, IEEE Transactions on Electron Devices, v 45, n 6, p 1361-8, June 1998.

[32] C. Yih, Z. Ho, M. Liang and S. Chung, Characterization of hot-hole injection induced silc and related disturbs in flash memories, IEEE Transaction on Electron Devices, v 48, n 2, pp 300-6, 2001.

[33] D. Ielmini, A. Ghetti, A. Spinelli and A. Visconti, A study of hot-hole injection during programming drain disturb in flash memories, IEEE Transaction on Electron Devices, v 53, n 4, pp 668-76, 2006.

[34] L. D. Yau, Simple I/V model for short-channel i.g.f.e.t.s in the triode region, Electronics Letters, vol. 11, 1975, pp. 44-45.

[35] J. R. Brews, W. Fichtner, E. H. Nicollian, and S. M. Sze, Generalized guide for MOSFET miniaturization, Electron Device Letters, IEEE, vol. 1, 1980, pp. 2-4.

[36] A. Asenov, Random Dopant Induced Threshold Voltage Lowering and Fluctuations in Sub-0.1 m MOSFET's: A 3-D Atomistic Simulation Study, IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 45, NO. 12, DECEMBER 1998 2505

[37] Jun Geun Kang, Boram Han, Kyoung-Rok Han, Sung Jae Chung, Gyu-Seog Cho, Sung-Kye Park and Woo Young Choi, Dependency of NAND flash memory cells on random dopant fluctuation (RDF) effects, Proceedings of the 2012 12th Annual Non-Volatile Memory Technology Symposium (NVMTS), p 37-40, 2012.

[38] G. Torrente, N. Castellani, A. Ghetti, C.M. Compagnoni, A.L. Lacaita, A.S. Spinelli, A. Benvenuti, Assessment of the statistical impedance field method for the analysis of the RTN amplitude in nanoscale MOS devices, International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), p 21-4, 2013

[39] G. Torrente, N. Castellani, A. Ghetti, C.M. Compagnoni, A.L. Lacaita, A.S. Spinelli, A. Benvenuti, Investigation of the RTN amplitude statistics of nanoscale MOS devices by the statistical impedance field method, Journal of Computational Electronics, v 12, n 4, p 585-91, Dec. 2013

[40] D. J. Frank, T. Yuan, I. Meikei, and H. S. P. Wong, Monte Carlo modeling of threshold variation due to dopant fluctuations, in VLSI Technology, 1999. Digest of Technical Papers. 1999 Symposium on, 1999, pp. 169-170.

[41] D. Ielmini, A.S. Spinelli, A.L. Lacaita, A. Modelli, Modeling of anomalous SILC in flash memories based on tunneling at multiple defects, Solid-State Electronics 46 (2002) 1749-1756.

[42] K. Kim, Technology for sub-50nm DRAM and NAND flash manufacturing, in Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International, 5-5 Dec. 2005, pp. 323-326

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

[44] C. Cheng, C. Hu, and R. W. Brodersen, Quantum yield of electron impact ionization in silicon, J. Appl. Phys., vol. 57, pp. 302-309, 1985

[45] S. Takagi, N. Yasuda, and A. Toriumi, A new I-V model for stress-induced leakage current including inelastic tunneling, IEEE Trans. Electron Devices, vol. 46, pp. 348-354, Feb. 1999

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

[47] D. Ielmini, A.S. Spinelli, A.L. Lacaita, R. Leone, A. Visconti, Localization of SILC in flash memories after program/erase cycling, IEEE International Reliability Physics Symposium. Proceedings. 40th Annual (Cat. No.02CH37320), p 1-6, 2002

[48] D. Ielmini, Member, A.S. Spinelli, A. Visconti, Characterization of Oxide Trap Energy by Analysis of the SILC Roll-Off Regime in Flash Memories, IEEE Trans. Electron Devices, vol. 53, no.1, pp. 126-134, 2006

[49] D. Ielmini, A.S. Spinelli, A.L. Lacaita, L. Confalonieri, A. Visconti, New technique for fast characterization of SILC distribution in flash arrays, IEEE International Reliability Physics Symposium Proceedings. 39th Annual (Cat. No.00CH37167), p 73-80, 2001

[50] S. Aritome, R. Shirota, G. Hemink, T. Endoh, F. Masuoka, Reliability issues of Flash memory cells, Proceedings of the IEEE, v 81, n 5, p 776-88, May 1993.

[51] Hong Yang, Hyunjae Kim, Sung-il Park, Jongseob Kim, et al. Reliability issues and models of sub-90nm NAND Flash memory cells, 2006 8th International Conference on Solid-State and Integrated Circuit Technology (IEEE Cat. No. 06EX1294), 3 pp., 2006.

[52] B. Eitan and D. Frohman-Bentchkowsky, Hot-electron injection into the oxide in n-channel MOS devices, IEEE Transaction on. Electron Devices, vol. ED-28, pp. 328-340, March 1981.

[53] A. Modelli, F. Gilardoni, D. Ielmini, and A. S. Spinelli, A new conduction mechanism for the anomalous cells in thin oxide flash EEPROMs, IEEE International Reliability Physics Symposium (IRPS), p 61-6, April-May 2001.

[54] Synopsys, Zurich, Switzerland, Sentaurus device user guide, J-2014.09

[55] D. M. Caughey and R. E. Thomas, Carrier Mobilities in Silicon Empirically Related to Doping and Field, Proceedings of the IEEE, vol. 55, no. 12, pp. 2192-3, 1967.

[56] C. Canali et al., Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature, IEEE Transactions on Electron Devices, vol. ED-22, no. 11, pp. 1045-7, 1975.

[57] C. Lombardi, S. Manzini, A. Saporito, M. Vanzi, A physically based mobility model for numerical simulation of nonplanar devices, Proc. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 7 (11) (November 1988) 1164-1171

[58] M. Lades et al., Analysis of Piezoresistive Effects in Silicon Structures Using Multidimensional Process and Device Simulation, in Simulation of Semiconductor Devices and Processes (SISDEP), vol. 6, Erlangen, Germany, pp. 22-5, September 1995.

[59] K. Matsuda et al., Nonlinear piezoresistance effects in silicon, Journal of Applied Physics, vol. 73, no. 4, pp. 1838-47, 1993.

[60] Z. Wang, Modelisation de la piezoresistivite du Silicium: Application a la simulation de dispositifs M.O.S., Ph.D. thesis, Universite des Sciences et Technologies de Lille, France, 1994.

[61] D. Rideau et al., Mobility in high-K metal gate UTBB-FDSOI devices: From NEGF to TCAD perspectives. IEEE IEDM, p.328, 2013.

[62] O. Nier et al., Multi-scale strategy for high-k metal gate UTBB-FDSOI Devices modeling with emphasis on Back Bias impact on mobility, Journal of Computational Electronics 12.4 (2013), 675-684.

[63] TCAD modeling challenges for 14nm FullyDepleted SOI technology performance assessment C. Tavernier, F. G. Pereira, O. Nier, D. Rideau, F. Monsieur, G. Torrente, M.Haond, H.Jaouen, O.Noblanc, Y.M.Niquet, SISPAD, 2015.

[64] C. Fiegna et al., Simple and Efficient Modeling of EPROM Writing, IEEE Transactions on Electron Devices, v 38, n 3, p 603-10, 1991.

[65] Gnudi et al., Two dimensional MOSFET Simulation by Means of Multidimensional Spherical Harmonics Expansion of the Voltzmann Transport Equation, Solid-State Electronics, v 36, n 4, p 575-81, 1993.

[66] S. Jin et al., Gate Current Calculations using Spherical Harmonic Expansion of Boltzmann Equation, International Conference on Simulator of Semiconductor Processes and Devices (SISPAD), p 202-5, September 2009.

[67] B. Govoreanu, P. Blomme, K. Henson, J. V. Houdt, and K. D. Meyer, An Effective Model for Analysing Tunneling Gate Leakage Currents Through Ultrathin Oxides and

High-k Gate Stacks from Si Inversion Layers, Solid-State Electronics, vol. 48, no. 4, pp. 617-25, 2004

[68] F. Li, S. P. Mudanai, Y.-Y. Fan, L. F. Register, and S. K. Banerjee, Physically Based Quantum-Mechanical Compact Model of MOS Devices Substrate Injected Tunneling Current Through Ultrathin (EOT 1 nm) SiO2 and High-k Gate Stacks, IEEE Transactions on Electron Devices, vol. 53, no. 5, pp. 1096-1106, 2006.

[69] G. Wentzel, Eine Verallgemeinerung der Quantenbedingungen fur die Zwecke der Wellenmechanik, Zeitschrift fur Physik A Hadrons and Nuclei, vol. 38, no. 6, pp. 518-29, 1926.

[70] B. Majkusiak, Gate Tunnel Current in an MOS Transistor, IEEE Trans.Electron Devices, vol. 37, no. 4, pp. 1087-92, 1990.

[71] A. Hadjadj, G. Salace, and C. Petit, Fowler-Nordheim Conduction in Polysilicon (n+)-Oxide-Silicon(p) Structures: Limit of the Classical Treatment in the Barrier Height Determination, J.Appl.Phys., vol. 89, no. 12, pp. 7994-8001, 2001.

[72] K. F. Brennan and C. J. Summers, Theory of resonant tunneling in a variably spaced multiquantum well structure : An Airy function approach, Journal of Applied Physics, vol. 61, no. 2, pp. 614-23, 1987.

[73] S. S. Allen and S. L. Richardson, Improved airy function formalism for study of resonant tunneling in multibarrier semiconductor heterostructures, Journal of Applied Physics, vol. 79, no. 2, pp. 886-94, 1996.

[74] R. H. Fowler and L. Nordheim, Electron Emission in Intense Electric Field, Proceedings of the Royal Society of London, v 119, n 781, p 173-81, 1928.

[75] J. Coignus, A. Vernhet, G. Torrente, S. Renard, D. Roy, G. Reimbold, Relaxation-free Characterization of Flash Programming Dynamics along P-E Cycling, IEEE International Integrated Reliability Workshop Final Report (IIRW), p 119-21, October 2015.

[76] J. Coignus, G. Torrente, A. Vernhet, S. Renard, D. Roy, G. Reimbold, Modelling of 1T-NOR Flash Operations for Consumption Optimization and Reliability Investigation, IEEE International Reliability Physics Symposium (IRPS), p PR-1 (4 pp.), 2016.

[77] Keysight Technologies, Improving Flash Memory Cell Characterization Using the Keysight B1500A, Application note B1500-9, 2014.

[78] A. Subirats, X. Garros, J. Mazurier, J. El Husseini, O. Rozeau, G. Reimbold, O. Faynot and G. Ghibaudo, Impact of dynamic variability on SRAM functionality and performance in nano-scaled CMOS technologies, IEEE International Reliability Physics Symposium (IRPS) Proceedings, pp. 4A.6.1-4A.6.5, 2013.

[79] V. Della Marca, G. Just, A. Regnier, J.-L. Ogier, R. Simola et al., Push the flash floating gate memories toward the future low energy application, Solid-State Electronics, vol. 79, pp. 210-217, 2013.

[80] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome and R. Shirota, Fast and accurate programming method for multi-level NAND EEPROMs, in 1995 Symposium on VLSI Technology Tech. Digest, pp. 129-130.

Chapter 2

Hot Carrier Degradation

Hot Carrier (HC) injection mechanism is a physical process used to program the Flash cell in NOR technology. Whenever a voltage between Source and Drain contacts of a MOSFET device is applied, the carriers are accelerated by lateral electric field and can gain substantially high energies, depending on applied biases. Since such carriers overcome the oxide barrier at the Si/SiO_2 interface, they have a certain probability to be injected in the FG node programming the cell. However, such a process produces damage, which is known as "Hot-Carrier Degradation" (HCD). The induced defects impact both the Program Flash operation efficiency, i.e. the HC-injection current, and the cell electrostatics, causing its failure. Thus, the study of HCD for NOR Flash cells and of the related impact on Flash characteristics is fundamental for such a technology.

2.1 Introduction

During HC regime, energetic carriers interact with the silicon-insulator interface of the transistor. In particular, they exchange energy causing damage at or near this interface, as schematically shown in Fig.2.1.

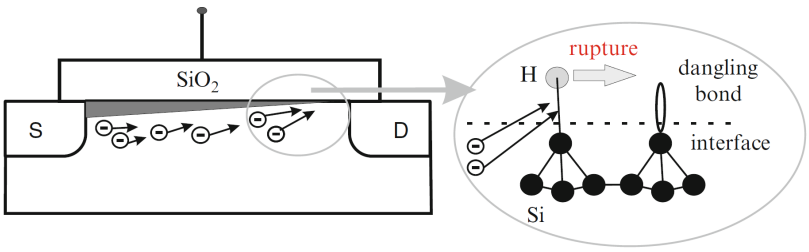


Figure 2.1: A schematic representation of hot-carrier degradation [89]. Energetic carriers collide with the dielectric-silicon interface breaking the weak Si-H bond. As such a bond is broken, a dangling bond remains.

Such a wear out process was initially reported in 1970s [1],[2],[3] and is still one of the major reliability concerns for CMOS technology [4],[5] nowadays. The term "hot" suggests that the carriers which activate this process are severely in non-equilibrium and are characterized by high energies. It is well-known that the hot-carrier damage is due to the dissociation of Si-H bonds at the interface [3],[6], which is triggered by channel carriers with a certain rate that strongly depends on the carrier energies [7],[8].

The structural disorder at the interface and the lattice mismatch between materials gives rise to dangling silicon bonds. These bonds are electrically active and can capture charge carriers. In order to passivate them, hydrogen species are intentionally incorporated during device fabrication forming passive Si-H bonds. This interface bond is not strong, indeed its energy ϵ_a varies between 1.2-2.5eV [9],[10]. Under hot carrier regime, the electrons may gain enough energy to overcome ϵ_a and thus dissociate the bond. Hence, due to these interactions with "hot" channel carriers, Si-H bonds are depassivated during HCD process, turning back into dangling Si- bonds. These dangling bonds act as trap centers and are called "amphoteric" traps [11]: they can capture charge carriers, thus distorting the electrical properties of the device. Two different types of defects can be found: donor and acceptor. The first is uncharged when unoccupied and it carries one hole when fully occupied, while the second one is uncharged when unoccupied and carries one electron when fully occupied. Their occupation depends on the energy level of the interface trap. Considering for example donor traps: in a first approximation, if the trap energy level is above the fermi level E_F, the trap is empty, whereas it hosts one electron in the opposite situation.

In order to detect and track the degradation induced by hot carrier mechanism, the hot carrier electrical stress is interrupted n times per stress time decade and device characteristics in linear regime are measured, which are very sensitive to HC-induced aging. Indeed, the amphoteric traps strongly affect the electrostatics, (since lowering the inversion charge), as well as the mobility (because inducing coulomb scattering). For such low electron energy levels, the latter is the limiting scattering mechanism [12].

It has to be pointed out that HCD not only produces interface amphoteric traps, but also trapping in the oxide [13]. Considering for example a n-MOSFET, an energetic electron accelerated by high Vds has a high probability to be injected through the oxide, thus to be trapped in the oxide bulk if its thickness is important. Usually, this mechanism is neglected since trapped electrons in the oxide bulk are mainly above the n^+ drain pocket and have no electrical impact on the drain current in linear regime. However, they may have a significant effect on Flash memory performance.

Considering NOR Flash technology, HCD remains one of the major reliability concerns [14],[15]. Indeed, the programming current degradation is mainly due to hot carrier damage, which means loss of programming efficiency, thus a lowering of Vth at Programming state.

In addition, HCD has a strong effect on the cell electrostatics in linear regime, significantly increasing the Vth at Erase state. These two distinct effects contribute to the closure of the Programming Window (PW), thus to the failure of the cell itself.

This chapter faces the HC reliability issues for NOR Flash technology induced by the Program operation. In order to uniquely study this degradation mechanism independently from Erase-induced aging, we take advantage of equivalent Flash transistors on which only Hot Carrier Stress is applied. Since we will uniquely deal with equivalent transistors, from now on all electrical parameters present in this chapter will just refer to MOSFET structures[1] (unless specified) for simplification purpose in terminology.

The chapter is divided into two parts: the extraction of physical device parameters during the stress, which aims to capture HCD physics, and the modeling of the aging kinetics, which targets the reproduction of the degradation evolution as experimentally observed. In the first section, after showing conventional extraction procedures, a new method will be proposed [13], which gives important insights on the physical aging mechanisms occurring during HCD. In the last part, an accurate physical-based model of the degradation kinetics for HCD is proposed [16],[17] which perfectly reproduces the device aging.

2.2 Parameter extraction and physical dependences

The precise extraction of MOSFET parameters is a major concern for both characterization and circuit design purposes in microelectronics. With the continuous device downscaling, an accurate extraction of physical parameters recorded as a function of stress time is fundamental to get the device behavior for any kind of degradation. An accurate extraction helps not only the physical comprehension of the degradation, but also gives important guidelines for a physical-based calibration of any kind of aging model.

In literature, very complex numerical/TCAD HCD models are calibrated on Vth, whose extraction method is rarely specified. On the other hand, compact models are often calibrated on Gm_{max} drift [7],[8], assuming that the dynamics of this parameter represents the HCD kinetics. In any case, before starting any kind of model, it is always mandatory to be sure about the physics of the wear out mechanism and to accurately choose the most relevant parameters which represent it.

Normally, Capacitance measurement (CV) [18] or Charge Pumping techniques (CP) [19]-[22] are used to directly extract physical parameters. However, for our nominally-scaled devices, these techniques would not lead to relevant results because of geometrical issues. Indeed, the signal would be too low not only for CV measurement, but also for CP technique since $W < L$ [23]. A solution to overcome such a problem would be to consider very

[1] as for example Id, Vg, but also the extracted Vth, Id_{ON} and so on

large equivalent structures, i.e. large W, in order to increase the signal during CV or CP experiment. However, we want to keep nominal devices in order to provide a relevant study performed on devices having real geometrical and 3D issues. For this reason, efforts have to be done in order to extract physical information from the IdVg linear trans-characteristics measured during the stress.

In this section, we focus the attention on HCD considering equivalent Flash transistors. The stress has been performed up to 1s (or till 10s in some cases), which is the equivalent of 10^6 NOR Flash Program only cycles, and the linear trans-characteristic IdVg has been measured 5 times per decade. Measurements have been performed in AC (alternate stress and delay periods of 1μs, duty ratio of 0.5), and the degradation displayed is exactly the same as observed with the corresponding DC pattern (not shown), since the device does not suffer from strong electron detrapping and self-heating effect.

HC stress has been applied on equivalent Flash transistors with different stress conditions and IdVg is measured in linear regime (Vd=50mV). In Fig.2.2, the evolution of this curve is shown for the stress condition Vg/Vd = 6V/4V and it exhibits an important drift towards higher Vg values. Starting from these experimental data, extraction procedures are performed in order to extract physical parameters, which give us precise information on the HC aging mechanism and evolution. In particular, extraction methodologies, described in this section, are applied on IdVg curves in order to probe and distinguish the electrostatic from transport aging.

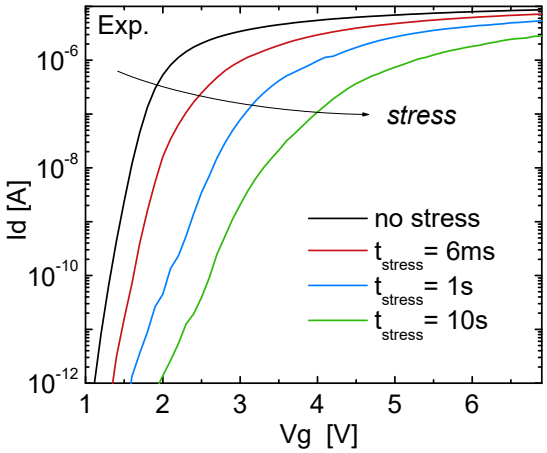


Figure 2.2: Evolution of IdVg characteristic measured in linear regime (Vd=50mV) during HCS condition (Vg=6V and Vd=4V).

2.2.1 Conventional extraction methodologies

2.2.1.1 Description and comparison of standard methods

In literature, several methods for the extraction of physical parameters from IdVg in linear regime can be found [24]-[29]. All of them are based on analytical models for drain current, and through an extraction procedure or fitting the experimental data, physical parameters are extracted. The most famous one is called *Y-function* method [24], and is simply based on an extraction procedure in the strong inversion region of the transistor in linear regime. It assumes that the inversion charge can be approximated as $C_{ox}(Vg - Vth)$, where C_{ox} is the oxide capacitance and Vg and Vth are the applied gate voltage level and the threshold voltage respectively. On the other hand, the mobility is approximated as $\frac{\mu_0}{1+\theta(Vg-Vth)}$, where θ is an empirical parameter representing the overall scattering dependence with the inversion charge together with the lowering of the current due to series resistances, whereas μ_0 represents the mobility at low field. Thus, the simple model of the drain current becomes:

$$Id = \frac{W}{L} Q_{inv}(Vg)\mu(Vg)Vd = \frac{W}{L} C_{ox}(Vg - Vth)\frac{\mu_0}{1 + \theta(Vg - Vth)}Vd \tag{2.1}$$

Assuming these simple approximations, it can be easily demonstrated that the ratio between the drain current and the square root of the transconductance Gm has a linear dependence with Vg-Vth. This function is called *Y-function* and is equal to:

$$Y = \frac{Id}{\sqrt{Gm}} = \left(\frac{W}{L} C_{ox}\mu_0 Vd\right)^{1/2} (Vg - Vth) \tag{2.2}$$

Whenever this function is computed from experimental data, the parameters Vth and μ_0 (and consequently θ) can be easily extracted. However, with technology down scaling, roughness scattering has become important, thus the mobility model has to be modified as $\frac{\mu_0}{1+\theta_1(Vg-Vth)+\theta_2(Vg-Vth)^2}$ [26]. In this case, the *Y-function* becomes:

$$Y = \sqrt{\frac{\beta Vd}{1 - \theta_2 Vgth^2}} Vgth \tag{2.3}$$

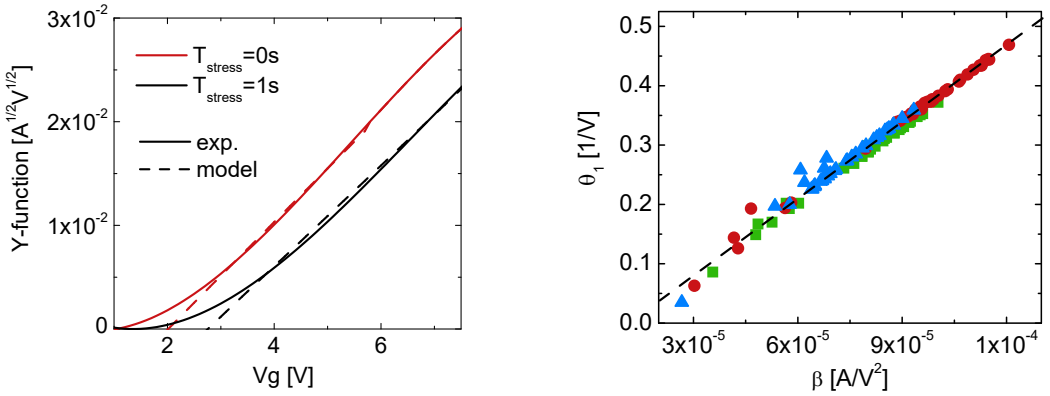
where $\beta = \frac{W}{L} C_{ox}\mu_0$ and Vgth=Vg-Vth. Making the assumption that θ_2 is not significant, $Y_0 \approx \sqrt{\beta Vd} \cdot Vgth$ can be assumed and (μ_0,Vth) are initially extracted. Then, considering that [26]:

$$\theta = \frac{\beta Vd}{Id} - \frac{1}{Vgth} = \theta_2 Vgth + \theta_1 \tag{2.4}$$

$$Y_1 = Y_0\sqrt{1 - \theta_2 Vgth^2} \tag{2.5}$$

(θ_1,θ_2) can be extracted (eq.2.4) and the new couple (μ_0,Vth) is updated (eq.2.5). Iterating such extractions using eqs.2.4-2.5, in few cycles the procedure converges.

This procedure is repeated during the degradation of the transistor under hot carrier regime, in order to physically address HCD kinetics looking at physical-based parameters. In Fig.2.3(a) the comparison between experiments and *Y-function* model is shown for fresh and aged device. It is worth noting that the *Y-function* is almost a straight line, which means that θ_2 is almost zero, and this non linearity does not change with stress time. This is easily explained considering that the surface roughness should not degrade since linked to the geometrical non uniformity of the Si/SiO_2 interface.



(a) Comparison of extracted and experimental Y-function.

(b) Evolution of the correlation between extracted θ_1 and β during stress time.

Figure 2.3: *Y-function* method applied during Hot Carrier Stress. $Vg_{stress} = 6V$, $Vd_{stress} = 4V$ and $Vd_{read} = 0.5V$.

Another important insight concerns the drift of θ_1. Indeed, it can be demonstrated that $\theta_1 = \theta_{1,0} + \beta R_{sd}$ [24], where R_{sd} are the series resistances at source and drain contacts and $\theta_{1,0}$ is an empirical parameter addressing the phonon scattering. It is commonly believed that the series resistances degrade whenever the interface amphoteric traps are created inside the drain/source regions [30]-[32]. However, looking at the results in Fig.2.3(b), the evolution of β-θ_1 correlation during the stress time is observed to be a perfect straight line indicating a negligible degradation of series resistances. Indeed, simulating the trans-characteristics in linear region with Synopsys TCAD tools [12], no electrical signature is observed considering defects in drain/source regions and a negligible shift of the extracted R_{sd} is detected whenever traps are located within the active channel (not shown here).

The general problem of such a method is the extraction noise. Indeed, it is necessary to consider the square root of the transconductance in the denominator of eq.2.2, which adds a lot of numerical noise. To simplify the extraction, a simple fitting of IdVg can be done in strong inversion region [25] using the same model for the drain current (eq.2.1). This method is called *Hamer* and gives similar outputs respect to *Y-function*. Thus, we skip the

representation of the related results since redundant.

The last method we consider is based on capturing the incremental drift between consecutive IdVg curves during stress time. It can be easily demonstrated that for small variations:

$$\Delta Vg_{i+1,i} = \frac{\Delta Id_{i+1,i}}{Gm_i} = -\Delta Vth_{i+1,i} + \frac{Id_i}{Gm_i}\frac{\Delta\mu_{i+1,i}}{\mu_i} \tag{2.6}$$

where i and i+1 are two consecutive IdVg measurement points, which correspond to two time stress durations. Considering different Vg levels, $\Delta Vg_{i+1,i}$ can be computed function of Id_i/Gm_i and, from this linear relationship, $\Delta Vth_{i+1,i}$ and $\Delta\%\mu_{i+1,i}$ can be extracted. Then, in order to compute the total degradation from fresh state, the sequent computation has to be performed:

$$\Delta Vth_{i,0} = \sum_{k=0}^{i-1} \Delta Vth_{k+1,i} \tag{2.7}$$

$$\Delta\%\mu_{i,0} = 1 - \prod_{k=0}^{i-1} (1 - \Delta\%\mu_{k+1,k}) \tag{2.8}$$

As evident, this method is based on a cumulative extraction of the drift, thus, from now on, we call it *Cumul*. It is worth noting that the aging of the mobility in 2.6,2.8 is not exactly the aging of low field mobility, but is the overall aging of the transport in strong inversion. For this reason, in order to properly compare this extraction procedure with the previous ones, an average of μ(Vg) has to be performed in a large Vg range for the conventional methods previously seen:

$$\mu_{TOT} = \frac{1}{b-a}\int_a^b \mu(Vg)dVg = \frac{\mu_0}{(b-a)\theta_1}\ln\left(\frac{(b-Vth)\theta_1+1}{(a-Vth)\theta_1+1}\right) \tag{2.9}$$

where a and b are two Vg > Vth values. The comparison between the different methods is shown in Fig.2.4. Looking at the results, the effect of HC-induced degradation is evident. At fixed Vg, the amount of inversion charge decreases, which induces an increase of Vth. On the other hand, the coulomb scattering lowers the overall transport in strong inversion. In addition to the methods described in this section, the standard extractions commonly used in industry have been plotted too. Concerning the mobility aging, the drift of the maximum of the trans-conductance Gm_{max}, which has been demonstrated to represent the overall transport degradation [33], has been considered. On the other hand, ΔVg at Constant Current level (ΔVth_{CC}), calculated at Id = 8μA, i.e. close to the "real" threshold condition at fresh state, has been considered for the electrostatic shift. It is worth noting that the aging of the total mobility for all the methods is very similar. This is not surprising, since the model behind all of them is exactly the same. Concerning the electrostatic drift, the situation is slightly different. Indeed, the aforementioned methods are applied in different regimes, either mobility-limited or not. In any case, Vth is strongly suspected to be only an

effective drifting parameter, at least for CC and *Cumul* extractions, since sensitive to both electrostatic and lateral transport degradations.

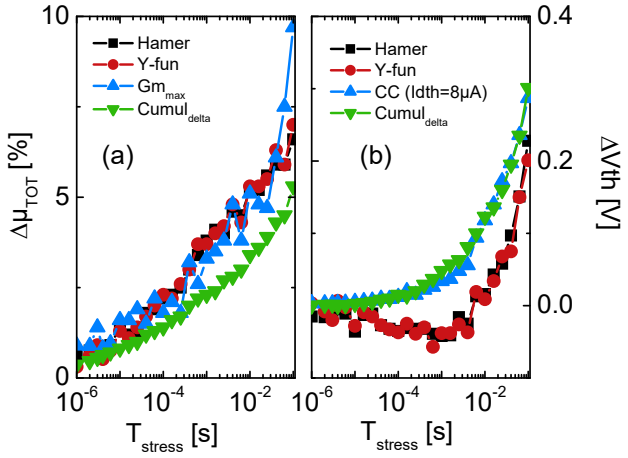


Figure 2.4: Comparison between the different extraction methods, where *Cumul* being the cumulative extraction performed as in eqs.2.7,2.8.

Looking at these results, the conventional extraction procedures applied during hot carrier regime do not fully satisfy. Indeed, using *Y-function* and *Hamer* procedures, no additional information is really added respect to Vth_{CC}-Gm_{max} drift extractions. However, these methods give us important physical insights on device characteristics and degradation. Indeed, the roughness scattering results to be negligible respect to the phonon's one, and no drift of the series resistance has been observed. For this reason, the series resistance can be extracted at $T_{stress} = 0$, comparing for example test structures having different lengths [34], also accounting for the increase of low field mobility when decreasing the length as in [35], and then keeping it constant during stress time.

In the next paragraph we will study the validity of these procedures, since the comprehension of HCD evolution relies on an accurate extraction and separation of electrostatic and mobility drifts.

2.2.1.2 TCAD evaluation of conventional extractions

In order to investigate the validity of the extracted parameters, a TCAD model has been used as reference. Indeed, the parameters extracted from the simulated IdVg curves (Vd=50mV) using conventional techniques and TCAD input parameters have been compared. The TCAD model of the structure has been developed considering the real process steps and is electrically calibrated. The mobility model being considered [12],[36] takes into account the Coulomb scattering occurring with interface charges. In order to further

understand the impact of mobility lowering on the extracted Vth, a theoretical case has been considered with the help of TCAD. The MOS device has been simulated with different degradation levels, i.e. different negative interface charges Nit uniformly distributed at the interface, and the IdVg curve has been acquired turning on or off the Nit-induced mobility degradation modeling. Fig.2.5 shows the drift of Vth extracted from simulated IdVg using different methodologies.

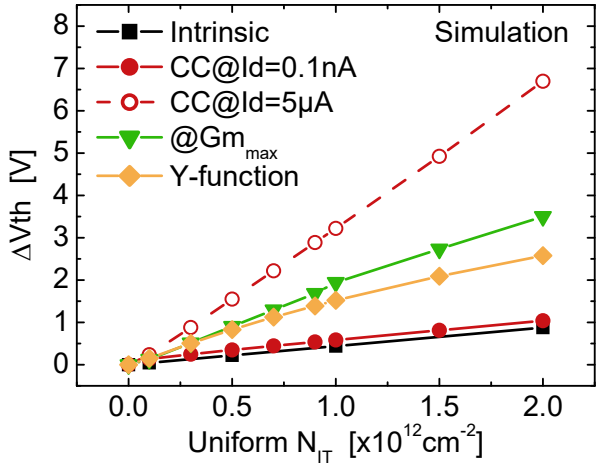
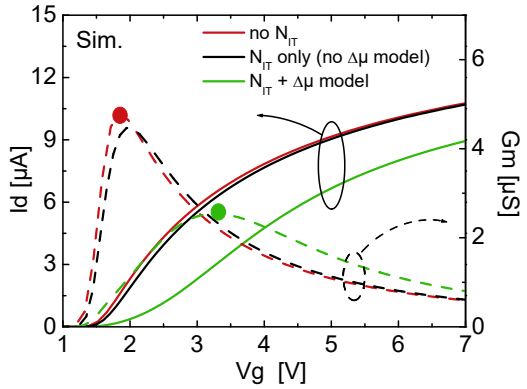


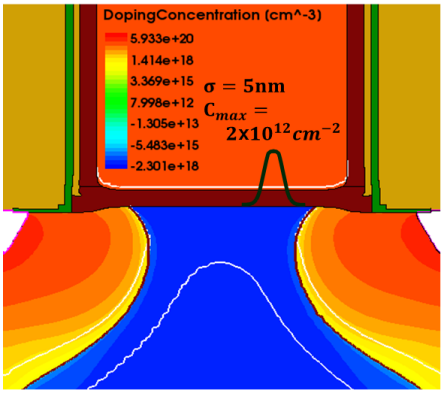
Figure 2.5: ΔVth extractions on simulated IdVg (Vd=50mV) curves: this has been done for different densities of uniform interface negative charge.

Looking at the figure, ΔVth_{CC} appears to be close to the intrinsic threshold voltage drift $\Delta Vth_{intrinsic}$ (i.e. the Vth shift simulated by turning off the model for the mobility reduction with Nit, i.e. representing the pure electrostatic shift) whenever the Idth considered is quite low, as for Idth=0.1nA case. This is because Vth extraction is less sensitive to mobility degradation in this region, otherwise it mixes electrostatic and transport shifts (it obviously fails at Idth = 5μA). It is also shown that the Gm_{max} position (Vg_{max}) drifts much more than the $Vth_{intrinsic}$, meaning that Gm_{max} is not a good sensor of the physical device aging: its position drifts more than the pure electrostatic shift, leading to a not fair mobility comparison (different Vg-Vth overdrive). Analogously, it is observed that $\Delta Vth_{Y-function}$ overestimates $\Delta Vth_{intrinsic}$, which similarly leads to an underestimation of low field mobility drift.

Considering now the case of a localized fixed negative charge close to the drain region (as for hot carrier damage) in Fig.2.6, the situation is even more clear, as transport degradation is mainly observed. Indeed, TCAD simulation gives a negligible electrostatic shift component ($\Delta Vth_{intrinsic} = 107mV$), while Vth values extracted by conventional methods are strongly impacted by transport degradation ($\Delta Vth_{CC,5\mu A} = 1.4V$, $\Delta Vth_{Y-function} = 0.98V$ and $\Delta Vg_{max} = 1.16V$).



(a) simulated IdVg (Vd=50mV) with Nit=0 and with Nit < 0 distributed close to the drain region



(b) Nit < 0 considered in the simulation is distributed close to the drain region with a Gaussian shape as depicted

Figure 2.6: IdVg simulations for localized interface negative charge at the drain edge.

To conclude this part, conventional extraction methodologies do not satisfy. As previously remarked, they do not add significant physical information respect to simple parameter tracking, as Gm_{max} case. In addition, as demonstrated with the help of TCAD tools, they result not to be appropriate whenever considering localized interface charges. Thus, they do not give relevant results if applied during HCD process. For this reason, in the next subsection, we introduce our own extraction methods to accurate address HC-induced damage.

2.2.2 A new methodology for parameter extraction

2.2.2.1 Method 1

As evidenced in Fig.2.5, drain current below threshold is a good indicator of pure electrostatic drifts. However, interface defects have an amphoteric nature [11] and so one cannot easily get the ΔVth_{CC} below threshold because of the subthreshold slope (SS) evolution with Dit, as experimentally observed in Fig.2.2. The first proposed approach is to quantify the physical degradation in a wide Vg range keeping a phenomenological model for μ(Vg) and improving Q_{inv} modeling from the conventionally used strong inversion approximation $C_{ox}(Vg - Vth)$ (i.e. the assumption lying behind both *Y-function* and Gm_{max} extractions). The procedure is based on a good fitting of the whole IdVg curve in order to extract both electrostatic and transport degradations, mainly in weak and in strong inversion respectively. The model used is the classical 1D model for long-channel transistors [18]:

$$Q_s = \sqrt{2q\epsilon_{Si}\left(N_A\phi_S + \frac{n_i^2}{bN_A}\left(\exp\left(b\phi_S\right) - 1\right)\right)} \tag{2.10}$$

$$Q_{inv} = Q_S - \sqrt{2q\epsilon_{Si}N_A\phi_S} \tag{2.11}$$

$$Vg = Vfb_0 + \phi_S + \frac{\phi_S}{Cox} + \frac{q\Delta Dit\,(\phi_S - \phi_B) + \Delta Qox}{Cox} \tag{2.12}$$

$$Id(Vg) = \frac{W}{L}Vd\frac{\mu_0}{1 + \theta_1\frac{Q_{inv}(Vg)}{Cox}}Q_{inv}(Vg) \tag{2.13}$$

where $b = kT/q$ [V], $\phi_B = b \cdot \ln(N_A/n_i)$ [V], ϕ_S [V] is the surface potential, Vfb_0 [V] is the flat band potential for the fresh case, Q_S [$C \cdot cm^{-2}$] is the total charge in the silicon bulk, Q_{inv} [$C \cdot cm^{-2}$] is the inversion charge, ΔDit [$cm^{-2}eV^{-1}$] and ΔQox [$C \cdot cm^{-2}$] are respectively the amount of interface traps (considered uniformly distributed with energy between the intrinsic Fermi level and conduction band Ec) and equivalent bulk oxide charges reported at the interface.

The methodology proposed here [13] is based on an iterative fitting procedure of the experimental curves Log(Id)-Vg and IdVg with the solution of the system eqs.2.10-2.13. For a fresh device (i.e. before stress) $\Delta Dit = \Delta Qox = 0$ is assumed and the fitting procedure is done via the following steps:

- fit of Log(Id)-Vg varying N_A&Vfb_0, because the SS is sensitive to C_D, keeping μ_0 and θ_1 fixed.
- fit of linear IdVg varying μ_0 and θ_1, because the current in strong inversion is more sensitive to the transport, keeping N_A&Vfb_0 fixed.

The convergence occurs very fast, after a couple of loops. For a stressed device the scheme is the same, with the difference that N_A-Vfb_0 are fixed and ΔDit-ΔQox vary for the Log(Id)-Vg fitting. In this way the four degradation parameters ΔDit, ΔQox, $\Delta\%\mu_0$ and $\Delta\theta_1$ are extracted. Examples of comparison between fitting and experimental IdVg curves are shown in Fig.2.7.

The resulting fitting is quite good for a fresh device and in general for low degradation level, while it might be lost around threshold for highly stressed device. Possible explanations for such a mismatch are the following:

- Constant Dit vs E considered in eq.2.12 is not appropriate. In this case a good fitting can only be achieved with a trap concentration increase of 2 order of magnitude close to the conduction band, which is unlikely.
- A distorted experimental curve due to parasitic trapping during the Vg-sweep. This effect is known to be particularly present in strong inversion [37], which is not the case here.

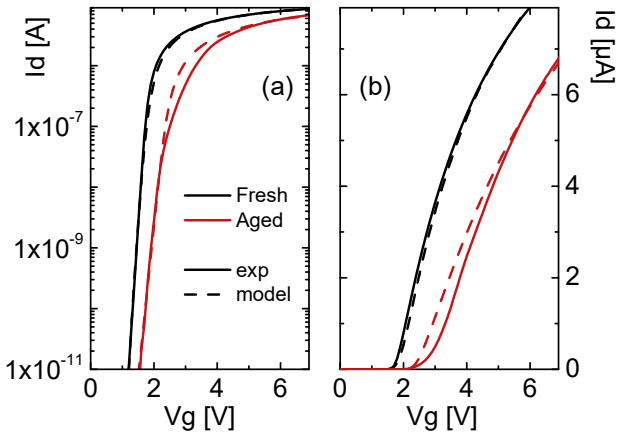


Figure 2.7: Confront between the experimental and fitting curves for Method 1: (a) in Log (b) in Lin, both for fresh and stressed cases.

- Mobility model in eq.2.13 is not appropriate whenever a strong degradation occurs. It is known that the Coulomb scattering reduces the mobility more at low field than in strong inversion: indeed, for highly stressed device, the field-dependence of the mobility is not monotonous anymore, $\mu(\mathrm{Vg})$ looking like a bell-shape. This effect is well known looking at any result from split CV technique [38]. Since this behavior is not expected by the mobility model in eq.2.13, a good fitting could not be achieved. Coming back to the conventional extraction methods, it is now clear why they cannot provide an accurate extraction: in both cases the transport model behind is the same as in eq.2.13.

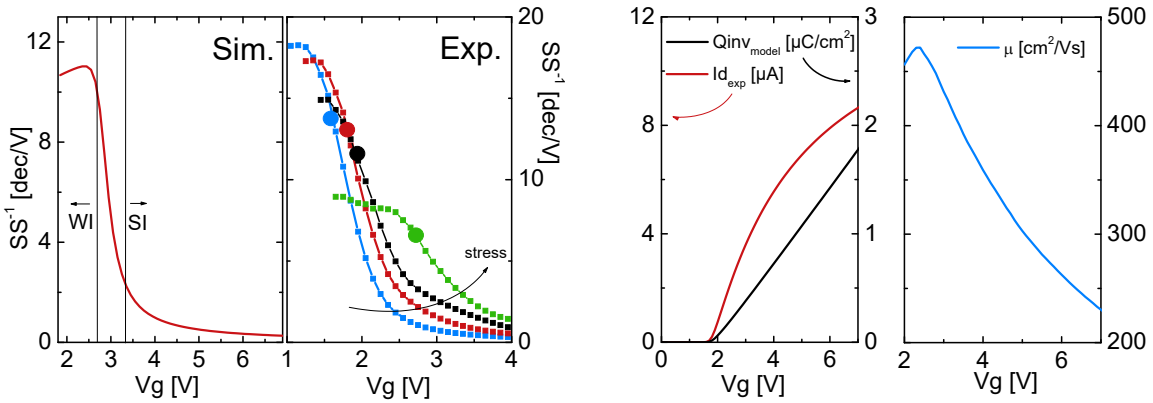
2.2.2.2 Method 2

Since the divergence in the fitting occurs because the mobility model does not take into account correctly the real scattering events, the objective is to get the transport directly from experimental data without assuming any mobility model. A complete separation of electrostatic and transport extractions has been proposed during this PhD [13] by getting the real, i.e. electrostatic-induced, threshold voltage shift first and then obtaining the field-dependent mobility in strong inversion. For the fresh case nothing changes with respect to *Method 1*, but whenever one considers a stressed device, the electrostatic $\Delta\mathrm{Vth}$ has to be estimated and the mobility is simply given by:

$$\mu(\mathrm{Vg}) = \mathrm{Id}(\mathrm{Vg})_{\mathrm{exp}} \left(\frac{\mathrm{W}}{\mathrm{L}} \cdot \mathrm{Vd} \cdot \mathrm{Q}_{\mathrm{inv}}(\mathrm{Vg}) \right)^{-1} = \mathrm{Id}(\mathrm{Vg})_{\mathrm{exp}} \left(\frac{\mathrm{W}}{\mathrm{L}} \cdot \mathrm{Vd} \cdot \mathrm{Q}_{\mathrm{inv,0}}(\mathrm{Vg} - \Delta\mathrm{Vth}) \right)^{-1} \tag{2.14}$$

where $Q_{inv,0}(Vg)$ is the inversion charge extracted for the fresh case and ΔVth is the real threshold voltage shift. The difficulty of this methodology is to accurately determine ΔVth. To achieve this, a given Vg bias at constant electrostatic state (i.e. constant charge in the channel) has to be considered in the IdVg curve and possibly close to the real Vth. Since ΔVth_{CC} and $\Delta Vth_{Y-function}$ overestimate the threshold voltage shift, leading to a non-physical increase of $\mu(Vg)$ during stress, an alternative way is proposed.

This point is addressed by considering the curve dLog(Id)/dVg vs Vg, that in practice is $SS^{-1}(Vg)$ [dec/V]. The observation of the curve shape in Fig.2.8(a)(left) allows to distinguish the two transistor regions: Weak inversion (WI) and Strong Inversion (SI). Since an exponential increase of the mobility close to the SI region can affect the validity of the extraction, Vg has been tracked at $SS^{-1} = 0.8 \cdot SS^{-1}_{max}$. This 0.8 factor comes from the electrostatic situation corresponding to $\phi_S = 2\phi_B$ for these devices. Whenever considering an aged device, this point (i.e. Vg at $0.8 \cdot SS^{-1}_{max}$) remains in the same electrostatic state with a very good approximation: simulating the case with Dit $= 5 \cdot 10^{12} cm^{-2} eV^{-1}$ the error on ϕ_S is only 17mV. This occurs because the decay in Fig.2.8(a) between the two plateaus is mainly driven by the fact that ϕ_S loses the linear dependence with Vg.



(a) Left: simulation of $SS^{-1}(Vg)$ (case Dit $= 5 \cdot 10^{12} cm^{-2} eV^{-1}$). Right: acquisition of ΔVth from experimental data, the circles indicate the estimated Vth (Vg at $0.8 \cdot SS^{-1}_{max}$). Different colors indicate different stress time.

(b) from the extracted $Q_{inv}(Vg)$ and the experimental IdVg (left), the $\mu(Vg)$ has been calculated (right).

Figure 2.8: Extraction of electrostatic shift (a) and mobility degradation (b) using *Method 2*

As shown in Fig.2.8(a)(right), the drift of this parameter is thus approximated as the real ΔVth, and then used in eq.2.14 in order to get $\mu(Vg)$, as shown in Fig.2.8(b). Thereafter, the extraction method aims to separate the respective contributions of ΔQox and ΔDit on the overall ΔVth. In order to isolate the effect of the first one, ΔVth_{ox} is extracted

by extrapolating the IdVg curve towards the "midgap current" calculated for fresh case. This technique [39] is based on the assumption that whenever $\phi_S = \phi_B$ the current is not sensitive to any amphoteric interface trap, so any drift that is observed at this current level is attributed to trapped electrons in the bulk oxide. This is true when considering a uniform distribution of traps between the intrinsic Fermi level and the conduction band.

In order to validate this procedure, $Q_{inv}(Vg)$ extracted by fitting the experimental IdVg (Fig.2.9(top-left)) and $Q_{inv}(Vg)$ coming from the integral of the $C_{inv}(Vg)$ (Fig.2.9(bottom-left)) for a large equivalent device having W=L=10μm have been compared (Fig.2.9(right)). The two curves are in good agreement, highlighting the validity of the procedure for a fresh device.

Similarly, $\mu(Vg)$ characteristics calculated by TCAD and estimated by the proposed extraction procedure, applied on the simulated linear IdVg, are compared. In the first case, the inversion charge has been found by integrating the electron density in the center of the channel from the $Si/Si0_2$ interface to the silicon bulk: $Q_{inv}(Vg) \approx q \int_{Si/Si0_2}^{\infty} n(0, y; Vg)dy$. In Fig.2.10 the results are observed to be very close not only for the case without any defects, but also considering both interface negative charges and amphoteric traps, that clearly highlights that the electrostatic drift is well captured.

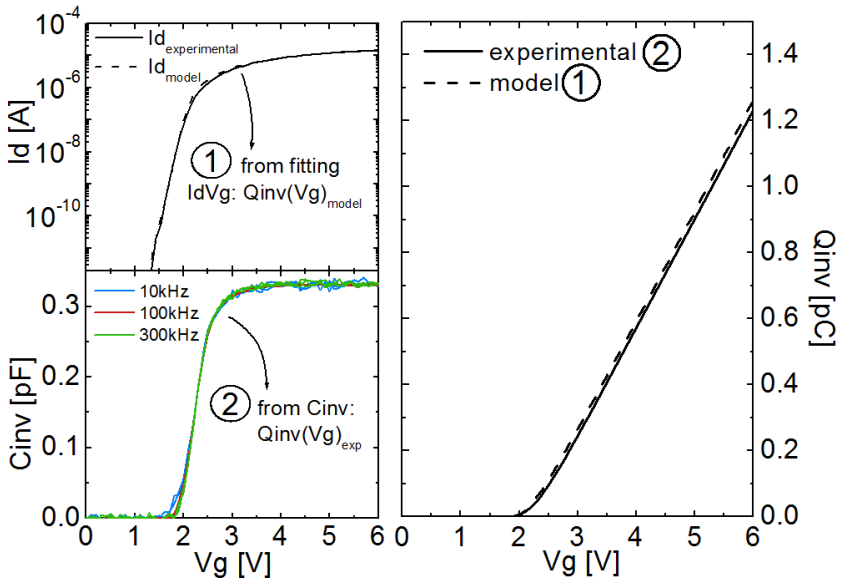


Figure 2.9: Comparison between split-CV and Method 2 extractions on a large equivalent device having W=L=10μm: (right) confront between $Q_{inv}(Vg)$ extracted for a fresh device after fitting the IdVg, in (top-left), and the $Q_{inv}(Vg)$ obtained from $C_{inv}(Vg)$, in (bottom-left).

The weaknesses of the extraction procedure proposed here mainly lie on the ΔVth as-

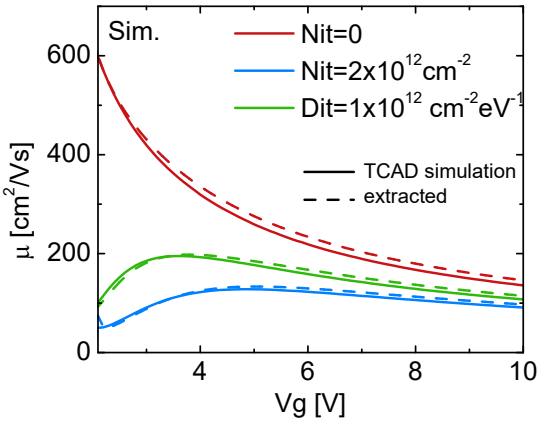


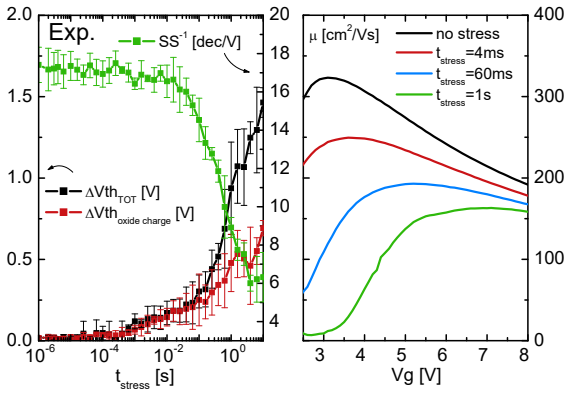
Figure 2.10: Comparison between μ(Vg) extracted with TCAD and Method 2 on the same simulated device for different uniform negative interface charges/traps (Nit/Dit). In both cases the mobility has been extracted from the inversion charge. The Q_{inv}(Vg) considered in TCAD case has been extracted integrating the electron density at the channel center

sessment. Indeed, the sub-threshold current has to be measured: since its magnitude might be very low, an increase of the measurement time could be necessary. Contrary, if trapped electrons in the oxide quickly relax in the silicon bulk, leading to a Vth recovery, the measurement time has to be as short as possible. However, this trade-off can be simply overcome considering an equivalent large device having very large width, with the disadvantage of losing 3D effects along W. Another critic can come from the ΔVth_{ox} extraction at the midgap state, since it can be affected by mobility lowering leading to an overestimation of this parameter. However, it is evident that considering IdVg curve in Logarithmic scale the extraction is exponentially more sensitive to electrostatic than transport drift, which extends the validity of the procedure till highly aged device.

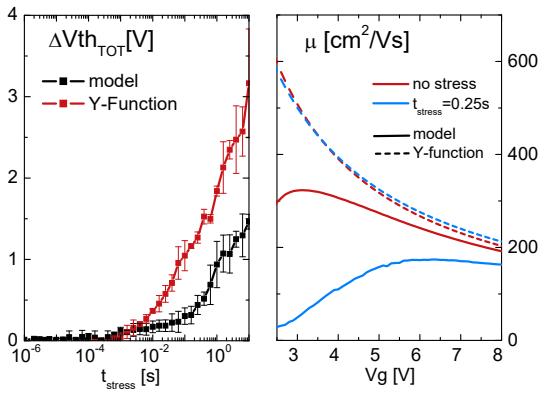
The proposed extraction technique has been applied on equivalent Flash transistors having experienced HCD, all the results are reported in Fig.2.11. In Fig.2.11(a) the results concerning the stress condition Vg=6V/Vd=4V are shown. Looking at the electrostatic degradation, Fig.2.11(a)(left), it is clear that the Vth drift is mainly driven by electrons trapped in the oxide bulk up to $t_{stress} \approx 100$ms. Afterwards, the interface traps start to play a role: SS^{-1} starts to decrease and the total ΔVth becomes higher than the electrostatic shift only due to oxide charges. It is important to notice that for $t_{stress} < 100$ms the fact that ΔDit does not play any role on Vth does not mean that they are negligible, since they can be close to the drain side and have a big impact on the mobility together with poor electrostatic signature, as simulations showed in Fig.2.6.

In Fig.2.11(a)(right) the results concerning transport degradation are shown, where μ(Vg)

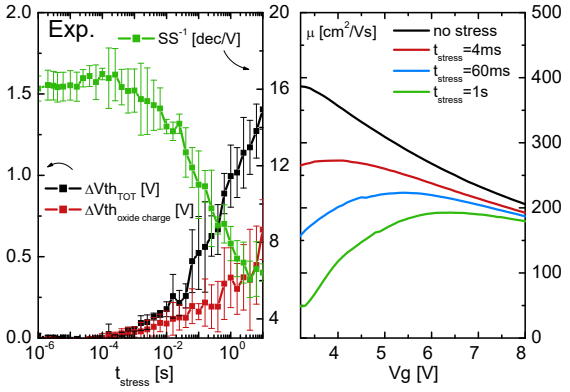
is observed to drift during stress. It is clear from these results that the mobility lowering occurs much strongly in the weak inversion as compared to the strong inversion region. This leads to an increase of the $Q_{inv,TH}$ necessary to screen the interface charges and so a drift of μ_{max} towards higher Vg-Vth is observed.



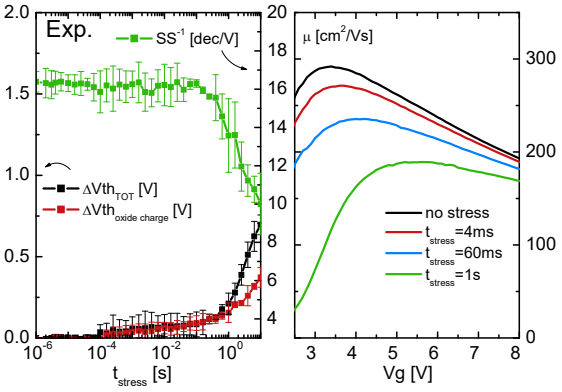
(a) Stress condition Vg=6V/Vd=4V.



(b) comparison between the *Method 2* and *Y-function*. Stress condition Vg=6V/Vd=4V.



(c) Stress condition Vg=8V/Vd=3.6V.



(d) Stress condition Vg=6V/Vd=3.6V.

Figure 2.11: Physical parameter extraction using *Method 2*. Error bars correspond the statistical dispersion of 3 nominal equal devices. The figures are divided in two sub-figures: (left) extracted electrostatic drift, (right) μ(Vg) evolution.

In Fig.2.11(b)(left) the extracted electrostatic degradations for *Y-function* method and for the model described here are compared. The important overestimation of ΔVth for the first methodology is evident whenever considering high stress time: it has been estimated up to $\approx$3V for $t_{stress} = 10s$. This is clearly not physical even just looking at the drift of the linear IdVg characteristics during the stress in Fig.2.2. Due to this bad electrostatic assessment, which in turn is caused by the lack of Coulomb Scattering description, the

extracted mobility is not accurate: μ(Vg) estimated by the *Y-function* procedure for high stress time (Fig.2.11(b)(right)) does not seem to suffer high degradation.

Results extracted from experimental data have shown how interface traps degrade significantly the low field mobility even if their electrostatic impact is visible just for high stress time. This occurs because they are mainly created close to the drain edge at the beginning and later, after 100ms, a significant presence near the center of the channel makes Vth increasing. This delay can be explained by a lower generation rate of interface traps, which is a result of lower average electron temperature whenever considering regions further from the drain side. This does not mean that the degradation rate at the drain side slows down during the stress, since interface traps can be created continuously there without having any electrical effect after a certain local concentration. Details of these dependences will be proposed in the next subsection, whereas we will deal with the HCD kinetics computation in the second part of the chapter.

In Fig.2.11(c) the results concerning the stress condition Vg=8V/Vd=3.6V are shown. It is worth noting that the Vth drift is mainly driven by electrons trapped in the oxide bulk up to $t_{stress} \approx 1$ms (till this time $SS^{-1} \approx$ const and $\Delta Vth_{TOT} \approx \Delta Vth_{OX}$), that means higher aging rate at the channel center than the previous stress condition. Despite the average electron energy has been doubtless decreased respect to before, especially at the drain side, the degradation rate at channel center has been increased thanks to a strong increase of the electron density (mainly driven by Vg).

Similar considerations can be done considering a different drain voltage during stress time. The results concerning the condition Vg=6V/Vd=3.6V are shown in Fig.2.11(d). The appearance of interface amphoteric traps close to the channel center occurs later. This is not surprising since the drain voltage defines the average electron energy at the drain edge, thus it sets the aging rate.

2.2.3 Degradation localization assessment

Using TCAD tools from Synopsis [12], we studied the relationship between the trap distribution at Si/SiO_2 interface and the drift of the trans-characteristic IdVg [17]. Looking at Fig.2.6(a), we may suppose that the electrostatic drift depends on the trap location: indeed, considering defects close to drain edge, no shift is observed. For this reason we performed simulations of IdVg trans-characteristics considering a fixed trap concentration at different positions along the interface. In Fig.2.12 the simulated trans-characteristics are shown in logarithmic scale. It is worth noting that traps closer to the drain edge have a negligible impact on the electrostatics, while traps closer to the channel center strongly shift the Vth, thus degrading the electrostatics.

It is well-known [18] that Short Channel Effects (SCE) lead the transistor interface not

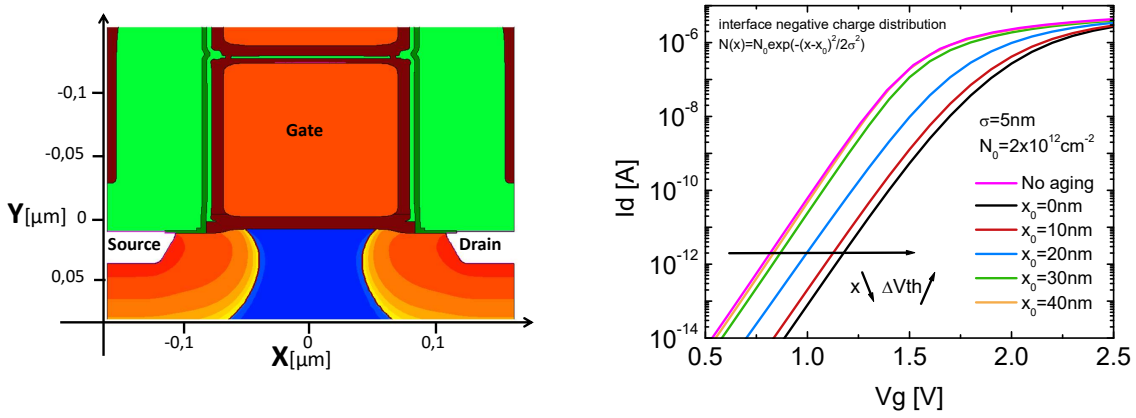


Figure 2.12: TCAD simulation: considering the calibrated structure in (left), IdVg has been simulated considering a negatively charged interface defect at different positions (right). "x" denoting the distance from the channel center as in (left)

to be uniform from an electrostatic standpoint. Indeed, considering the device in linear regime, regions closer to the channel center experience a lower surface potential, thus higher control on the overall device electrostatics since being the bottleneck for the "switching on" of the device. This means that charged defects created close to the channel center impact more the device threshold voltage [40]-[42]. Looking at Fig.2.13, the situation is clear: traps closer to the channel center impacts more significantly the minimum of the bell shape, thus the threshold voltage. Indeed, when simulating the Source-to-Drain current density around threshold, a clear decrease is observed whenever traps closer to the channel center are considered.

Since the drift of Vg at low current ΔVth_{CC} is strongly sensitive to electrostatic aging [13], this parameter has been simulated considering negatively charged interface defect at different positions and densities, as shown in Fig.2.14(a). As predicted, the impact of traps closer to the channel center (lower |x|) is higher [17].

With the same aim, the impact of an interface charged defect on the overall transport aging has been considered, simulating the device with calibrated coulomb scattering model for the mobility degradation [12],[36]. For this reason, on the same simulated IdVg characteristics, the drain current at $Vth_{CC} + 4V$, which is strongly sensitive to transport aging, has been extracted. Indeed, this parameter, denoted as Id_{ON}, represents the current flow at ≈fixed amount of inversion charge in strong inversion, thus reflecting the electron mobility at the same electrostatic condition [24]. Fig.2.14(b) shows the resulting drift of simulated Id_{ON} as a function of trap position and concentration and it is worth noting that an almost constant (slight increase) behavior is observed respect to the defect location x. However, it

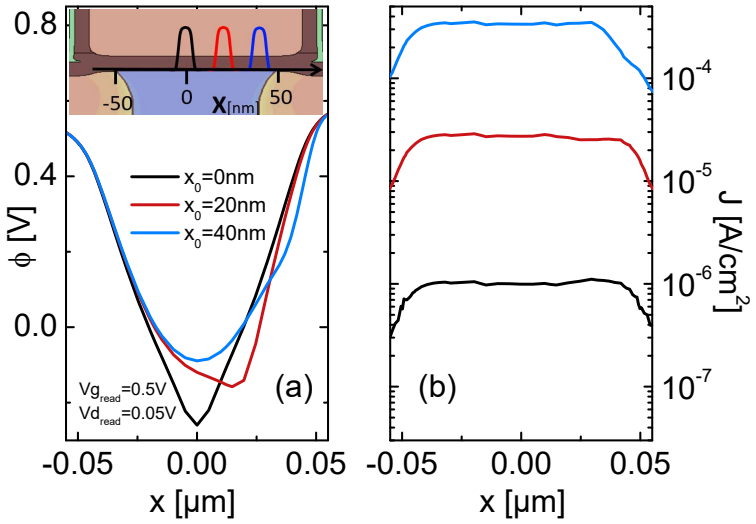


Figure 2.13: Simulations of the electrostatic potential (a) and of the related current density at Si/SiO_2 interface (b) considering three gaussian shape negatively charged defects at different interface locations. The considered defect experiences a concentration equal to $C(x) = C_{max}\exp\left(-(x - x_0)^2/2\sigma^2\right)$, where $\sigma = 5\text{nm}$, $C_{max} = 2 \cdot 10^{-12}\text{cm}^{-2}$, "x" denoting the distance from the channel center as in Fig2.12(left) and x_0 denoting the gaussian peak position.

has to be pointed out that when the trap moves inside the n^+ drain region, i.e. outside the active channel, its impact obviously decreases (not shown).

This almost independence on the interface position can be explained with simple considerations: in strong inversion and linear regime the device experiences essentially constant inversion charge along the channel. This means that, at first order, the impact on the mobility of a charged defect does not depend on the position since being screened by similar amount of inverted electrons.

Plotting the correlation between these two parameters, one sensitive to the electrostatics, thus to the trap position, and the other one sensitive to coulomb scattering in strong inversion, thus to the trap density, a clear separation of the trajectory is observed depending on the trap position (Fig.2.15). Taking advantage of this physical dependence, we can directly extract the barycenter of the trap distribution during HCD and the related evolution. Such a procedure gives important insights on the physical aging mechanism behind HCD and helps to calibrate the kinetic computation model, as we will see in the second part of the chapter.

In any case, coming back to extracted results previously seen in Fig.2.11, the parameter dependences with defect position explored in this subsection explain such evolutions. Indeed, up to a certain stress time, the mobility degrades due to Dit, while no effect is

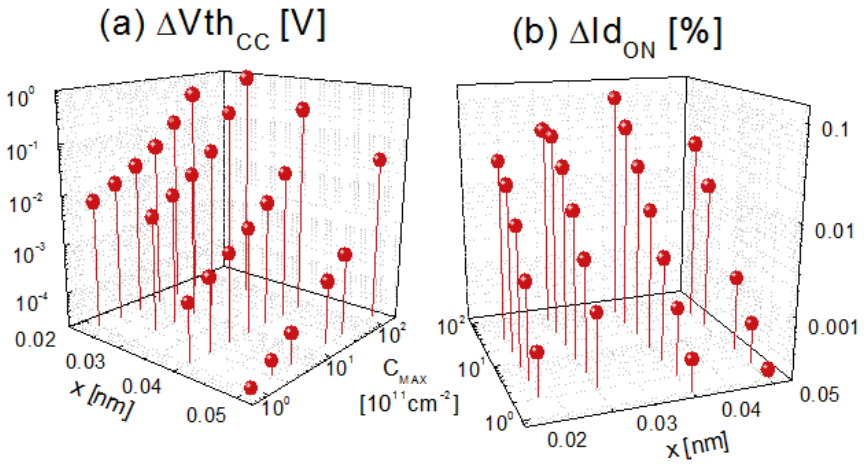


Figure 2.14: (a) Simulated drift of extracted Vg at low constant current ΔVth_{CC} (Idth=1nA) at Vd=50mV as a function of defect density and location ("x" denoting the distance from the channel center as in Fig.2.12(left)). The trap profile considered is a negatively charged gaussian defect distribution along the channel interface $C_{max}exp(-(x-x_0)2/2\sigma^2)$ having $\sigma = 5$nm. (b) Same as in (a), but concerning the simulated drift of Id_{ON}, i.e. extracted drain current at $Vg_{CC=1nA} + 4V$ in linear regime (Vd = 50mV).

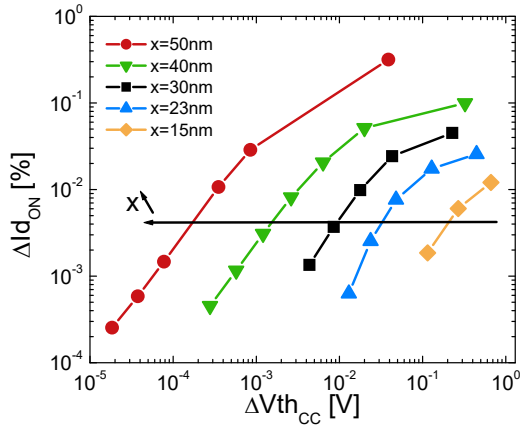


Figure 2.15: Correlation between ΔVth_{CC} and ΔId_{ON} shown in Fig.2.14. "x" denotes the distance of the defect from the channel center as in Fig.2.12(left). For a certain curve, each point represents an IdVg simulation considering a gaussian defect at the interface having a certain concentration. The trajectory is then built varying that concentration.

observed on the sub-threshold slope. Looking at the Fig.2.16, we clearly see this effect: till $\Delta Vth_{CC,high} = \Delta Vth_{CC,low}$, the sub-threshold slope does not change and the traps are located close to the drain edge. After a certain stress time, $\Delta Vth_{CC,high} > \Delta Vth_{CC,low}$ which means that the hot carriers start to degrade the channel center and thus the electrostatics

in linear regime.

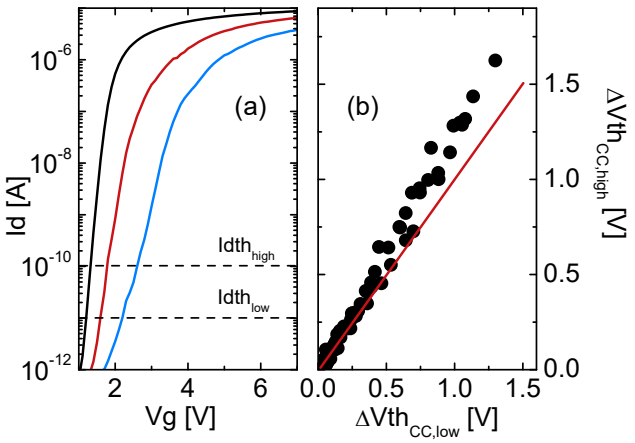


Figure 2.16: In (a) the extraction of $\Delta Vth_{CC,low}$ and $\Delta Vth_{CC,high}$ from the transcharacteristics (at Id_{low} and Id_{high} rispectively). In (b), the correlation between the two. Stress condition considered is Vg=6V/Vd=4V.

Since a fine extraction of Dit during HCD for short stress time (negligible SS drift) is impossible, we have to assume a simple model for the mobility aging to have an estimation of the total trap amount. Thus, in order to simplify the problem, we consider the simple analytical model for the drain current already seen in eq.2.1 and we assume that:

$$Vth = \overline{Vth} + p \cdot \frac{qDit\phi_B}{Cox} \tag{2.15}$$

$$\mu_0 = \frac{\overline{\mu_0}}{1 + \overline{\mu_0}\alpha qDit} \tag{2.16}$$

where $\overline{Vth}$ and $\overline{\mu_0}$ are the threshold voltage and low field mobility at fresh state, p and α are empirical factors which quantify the impact of a single trap on macroscopic electrostatic and transport parameters respectively. For example, the factor p links the local inversion charge variation to the overall threshold voltage. Thus, it is equal to ≈ 1 if the defect is close to channel center and ≈ 0 if very close to the drain edge. On the other hand, α links the local mobility variation due to coulomb scattering to the equivalent transport aging seen at drain contact. As previously seen, this relationship can be assumed to be independent on the channel position since extracted in strong inversion and linear regime. For this reason, this parameter has been extracted with the help of TCAD and fixed to a value of $2 \cdot 10^3$.

Then, in order to consider the IdVg curve evolution, we assume the simple model of the drain current in eq.2.1. Although we already remarked the drawbacks of such a model, we will consider it since it is simple and well describes the degradation evolution of the device. Indeed, in order to address an analytical expression linking macroscopic electrical (as

$\mathrm{Id_{ON}}$ and $\mathrm{Vth_{CC}}$) parameters and microscopic information (as trap distribution barycenter and concentration), strong simplifications which allow to get the order of magnitude of the degradation amount have to be done. Anyhow, the extraction precision of such an approach can be empirically increased considering a lower p factor in eq.2.15 accounting for the spurious additional Vth shift coming from mobility decay.

Now, considering θ_1 evolution, an assumption has to be done. Indeed, the drift of this parameter has been observed to be strongly linked with the drift of μ_0 (Fig.2.3(b)) since following the relationship $\theta_1 = \theta_{1,0} + \beta \mathrm{R_{sd}}$. Thus, $\mathrm{R_{sd}}$ has been extracted from this relationship and considered constant along stress time, as previously verified.

It is now possible to calculate analytically $\Delta\mathrm{Vth_{CC}}$, even though in strong inversion region, imposing $\mathrm{Id}(\mathrm{Vg_1}, \mathrm{Dit}=0) = \mathrm{Id}(\mathrm{Vg_2}, \mathrm{Dit})$. Computing this equality, $\mathrm{Vg_2} - \mathrm{Vg_1}$ function of Dit can be calculated:

$$\begin{aligned}\Delta\mathrm{Vth_{CC}}(\mathrm{Dit},\mathrm{p},\alpha) = & \frac{\overline{\mathrm{Vgth}}\left\{(\theta_{1,0}-\mathrm{k})\left(\frac{\mathrm{qDit}\cdot\phi_{\mathrm{SP}}}{\mathrm{Cox}}-\overline{\mathrm{Vgth}}\right)-1\right\}(\mathrm{qDit}\cdot\alpha\cdot\mu_0+1)}{\overline{\mathrm{Vgth}}\,(\mathrm{Cit}\cdot\alpha\cdot\mu_0+1)\,[\theta_{1,0}-\mathrm{k}]-\left[\theta_{1,0}\overline{\mathrm{Vgth}}+1\right]}+ \\ & +\frac{\overline{\mathrm{Vgth}}\left[\theta_{1,0}\overline{\mathrm{Vgth}}+1\right]-\frac{\mathrm{qDit}\cdot\phi_{\mathrm{BP}}}{\mathrm{Cox}}\left[\theta_{1,0}\overline{\mathrm{Vgth}}+1\right]}{\overline{\mathrm{Vgth}}\,(\mathrm{Cit}\cdot\alpha\cdot\mu_0+1)\,[\theta_{1,0}-\mathrm{k}]-\left[\theta_{1,0}\overline{\mathrm{Vgth}}+1\right]} \\ & \approx \mathrm{qD_{it}}\alpha\mu_0\overline{\mathrm{Vgth}}+\overline{\mathrm{Vgth}}^2\left(\theta_{1,0}\mathrm{qD_{it}}\alpha\mu_0-\frac{\mathrm{W}}{\mathrm{L}}\left(\mathrm{qD_{it}CoxR_{sd}}\alpha\mu_0^2\right)\right)+ \\ & +\frac{\mathrm{qD_{it}}\phi_{\mathrm{B}}\mathrm{p}}{\mathrm{Cox}}\end{aligned} \tag{2.17}$$

where $\overline{\mathrm{Vgth}}$ is the overdrive at fresh state and $\mathrm{k} = \mathrm{W/L}\cdot(\mathrm{qDit}\cdot\mathrm{Cox}\cdot\mathrm{R}\cdot\alpha\cdot\mu_0^2)/(\mathrm{qDit}\cdot\alpha\cdot\mu_0+1)$.

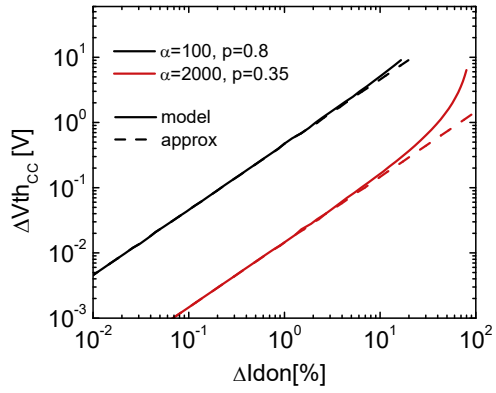
Concerning now the "transport" parameter, $\mathrm{Id_{ON}}$ is considered. The drift of this parameter is simply given by $\Delta\%\mathrm{Id_{lin}} = 1 - \mathrm{Id_{lin}}/\overline{\mathrm{Id_{lin}}}$, whose expression can be simplified as well:

$$\Delta\%\mathrm{Id_{ON}}(\mathrm{Dit},\mathrm{p},\alpha) \approx \mathrm{qD_{it}}\alpha\mu_0\left[1-\frac{\overline{\mathrm{Vgth}}}{\overline{\mathrm{Vgth}}\theta_{1,0}+1}\cdot\frac{\mathrm{W}}{\mathrm{L}}\left(\mathrm{CoxR}\mu_0\right)\right] \tag{2.18}$$

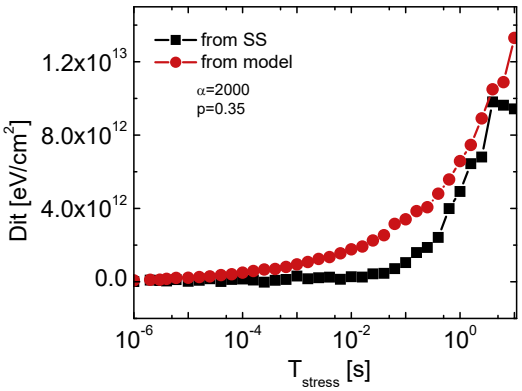
It is worth noting that $\overline{\mathrm{Vgth}}$ of eq.2.17 has to be different from the one in eq.2.18, since considering the transistor in deep strong inversion.

When plotting these two parameters function of each other, we obtain a curve that only depends on α and p. The simulation of this curve, using eqs.2.15,2.16, is shown in Fig.2.17(a) and, using eqs.2.17,2.18, it can be demonstrated that $\Delta\mathrm{Vth_{CC}}$ vs $\Delta\mathrm{Id_{ON}}$ can be approximated to a straight line in Log-Log scale having slope=1 and amplitude equal to $\mathrm{A} \approx \mathrm{K_1} + \mathrm{K_2}\cdot\mathrm{p}/\alpha$ where $\mathrm{K_1}$ and $\mathrm{K_2}$ are technology dependent parameters, i.e. function of Cox, $\mathrm{R_{sd}}$, ecc. This approximation fits well the simulations, as shown in Fig.2.17(a).

From these considerations, a simple methodology for extracting the total amount of amphoteric traps is proposed. This relies on the analytical expression of $\Delta\mathrm{Vth_{CC}}$ and $\Delta\%\mathrm{Id_{ON}}$ function of the total defect concentration previously seen. To address this point, the variation of p along the channel interface should be considered. However, in order to take into



(a) Simulation of $\Delta\%Id_{ON}$ vs ΔVth_{CC} for two random (p,α) couples of in eqs.2.15,2.16.



(b) Confront between Dit extracted from the SS decay and the one extracted from the model $\Delta Vth_{CC}(Dit, p, \alpha)/\Delta\%Id(Dit, p, \alpha)$. Stress condition considered is Vg=6V/Vd=4V.

Figure 2.17: A simple method for extracting the interface damage from macroscopic parameters, such as the drift of Vth_{CC} and Id_{ON}.

account this dependence, we should consider also that the aging kinetics is not constant along the channel interface, thus we should calculate the electron energy distribution function at each location, hence we should solve the Boltzmann equation ecc.. If we account for these effects, the difficulty of the extraction explodes, moving away from our purpose. For this reason, we consider an empirical overall p=0.35 factor, which well takes into account both drain edge fast kinetics and channel center slow rate. Fixing this value, the Dit can be easily extracted and compared respect to the ones previously obtained from SS decay. This is done in Fig.2.17(b) and we clearly see that for long stress time the two curves almost coincide, highlighting the validity of the approach. On the other hand, for short stress time, the new extraction achieves to estimate the amount of amphoteric traps not accessible by the SS decay since too much close to the drain region. However, the real advantage of this extraction relies on a drastically decrease of the noise, since the model is based on simple drifting parameters such as ΔVth_{CC} and $\Delta\%Id_{ON}$.

Concluding the section, it has to be noticed that the presence of interface traps and the related electrical signature have been treated separately. In particular, induced defects are shown to have or not an impact on electrical characteristics depending on the interface position. This is an important point achieved in the chapter, also because such a separation is rarely considered in scientific works. We will take advantage of all these results and considerations not only for the HCD model (presented in the next section), but also later whenever measuring electrical characteristics during Flash endurance.

2.3 Modeling of aging kinetics

Till now, we focused our attention on the extraction of physical parameters which gave us a full comprehension of HCD in 40nm NOR Flash technology and suggested us the parameters to track along stress time which are representative of the microscopic wear out. However, the real challenge is to model the aging kinetics computing mathematically the interaction between the single electron, described by its energy distribution, and the Si-H bond, described by a capture cross section that depends on the bond energy. In this section we go through the existing models in literature and we present our physical-based models for 40nm NOR Flash technology.

In a first rough approximation, the electron, accelerated by lateral field under hot carrier regime, gains an average energy equal to $\approx q \cdot Vd$ close to the drain edge. This means that MOSFET miniaturization, which involves a decrease of operating voltages as low as $\approx 1V$, would make hot electrons unlikely for scaled technologies. Indeed, HCD was expected to be totally removed or at least severely suppressed. However, this idea was initially discarded by Mizuno et al. [43], where an important drift of transistor characteristics was observed despite the drain voltage was less than 1V. For this reason, HCD theory has been extended in order to account for the microscopic description of the Si-H bond breakage mechanism depending on the carrier Energy Distribution Function (EDF). Indeed, the carriers have been shown to contribute to the Si-H dissociation process in two different ways. Firstly, the electron can gain energy through scattering mechanisms thus populate the high-energy tail of EDF even if the stress/operating voltage is low, thereby triggering HCD even for $q \cdot Vd < \epsilon_a$, where ϵ_a is the Si-H energy bond. Such a mechanism of degradation is called Single Vibrational Excitation (SVE) mode, since assuming that a single electron, accelerated by lateral field, breaks a Si-H bond, and thus generates an active defect. In order to explain the experimental results in this way, the role of Electron-Electron Scattering (EES) mechanism [44]-[50] is often emphasized. For this reason, EES mode is usually considered as a different degradation mode from SVE. Secondly, in scaled devices, the dominant mechanism factor for Si-H bond breakage is based on the Multiple Vibrational Excitation (MVE) of the bond, which is triggered by a bombardment of cold carriers [10],[51]-[54]. This is in contrast with long-channel devices, where the bond dissociation event can be induced by a solitary hot carrier in a single collision [10],[51]-[54].

At each point of the channel interface, the rates of both scattering and bond-breakage mechanisms strongly depend on the energy of the carriers. Mathematically, this means that a proper modeling of HCD needs to be based on the carrier EDF, which can be obtained solving the Boltzmann Transport Equation (BTE).

There are two different strategies to solve the BTE and thus calculate the energy occupation probability at each interface location: the stochastic Monte-Carlo method [55],[56] and

the deterministic approach based on the Spherical Harmonics Expansion (SHE) [57]. In the first one, the trajectories of each carrier (from a representative particle ensemble) has to be considered in the multidimensional space. As a result, computational resources needed for this task become cumbersome. An alternative approach is to represent the EDF as a series of spherical harmonics [57]. As opposed to the Monte Carlo method, this approach relies on a substantial amount of memory (required to store all the variables) rather than on CPU power.

However, even this deterministic SHE method has high computational demands. As such, it would be highly attractive if the solution of the BTE could be replaced by a simplified approach which represents the EDF by an analytic expression. There are many different approaches which could be employed in the context of HCD: the heated Maxwellian distribution [58], Cassi model [59], Hasnat model [60], Reggiani model [61],[62],[84], and Sharma approach [63]. Basically, they are based on guessing a certain shape of EDF, with few fitting parameters. Then, these parameters are computed imposing some conditions, as the electron density or the effective temperature, which are calculated with classical Drift-Diffusion simulations, that do not demand significant computational efforts.

Anyhow, most of these BTE-based approaches have the lack of being based on a high number of fitting parameters, so they can easily fit the experimental data speculating on the physics of the degradation and the nature of the limiting scattering mechanism factor. Indeed, it is easy to demonstrate that these models have infinite parameter sets which can perfectly reproduce the experimental data. In addition, the calibration is often done tracking the drift of a single parameter, which is not representative of the aging induced by hot carrier, as we demonstrated in the previous sections. Even worse, they are often calibrated on a general "Vth", whose extraction methodology is not even specified.

Due to all these simple considerations, these models are rightly never used in industry for product qualification, since being very far to be predictive and needing substantial computational resources. For this reason, alternative simplified approaches are considered which are based on effective degradation rates computed avoiding the BTE solution. One of the most successful HCD compact model has been developed by Bravaix group [6],[8],[67] and is based on the so-called Energy-Driven Paradigm (EDP), which has been previously proposed by Rauch and La Rosa [7],[68],[69]. According to it, the bond-breakage rate is driven by "knee" energies, which are related to operating conditions. Hence, the challenging evaluation of the EDF is skipped and the bond-breakage rates are modeled using some empirical parameters.

In this section, we show the existing hot carrier models in literature, from the compact empirical approaches to complex numerical simulations. Afterwards, our model is presented, which has been developed with the help of Synopsys TCAD tools [12] trying to simplify the

problem as far as we could in order to be as much as possible sensitive to microscopic defects avoiding any speculation. Then, an "had hoc" compact model is presented with the aim of addressing Flash cell modeling.

2.3.1 State of the Art

2.3.1.1 Lucky Electron Model

The Lucky Electron Model (LEM) of C.M. Hu [3],[65] remains a guiding principle of most industry standard hot carrier models and projection approaches. Free electrons are accelerated by lateral electric field till they collide with the gas atom. The interaction may ionize the atom, leading to two free electrons, which in turn are accelerated by the field. The process leads to "Impact Ionization" process (II). The probability of an electron traveling a distance d before suffering a collision is $\exp(-\mathrm{d}/\lambda)$, where λ is the "mean free path". Since the energy E is equal to dqF, where F is the electric field, the electron energy distribution becomes:

$$\mathrm{P(E)} = \exp\left(-\mathrm{E}/\mathrm{q}\lambda\mathrm{F}\right) \tag{2.19}$$

Since the electric field is not constant along the channel, F is replaced by $\mathrm{F_{max}}$, the maximum field. The impact ionization rate, roughly equal to the ratio between bulk and drain currents, becomes

$$\frac{\mathrm{Ib}}{\mathrm{Id}} = \mathrm{A} \cdot \exp\left(-\frac{\phi_{\mathrm{ii}}}{\mathrm{q}\lambda\mathrm{F_{max}}}\right) \tag{2.20}$$

where ϕ_{ii} is the threshold energy for impact ionization, assuming that the cross section is a step function triggering at this value.

Concerning the Hot Carrier damage, the interface trap concentration can be written as

$$\Delta\mathrm{N_{it}} \approx \left(\mathrm{t} \cdot \frac{\mathrm{Id}}{\mathrm{W}} \exp\left(-\frac{\phi_{\mathrm{it}}}{\mathrm{q}\lambda\mathrm{F_{max}}}\right)\right)^{\mathrm{n}} \tag{2.21}$$

where ϕ_{it} is the threshold energy for hot carrier damage. Now, considering that the lifetime duration τ is defined at a certain fixed damage $\overline{\Delta\mathrm{N_{it}}}$, the HCD rate becomes:

$$\tau^{-1} \propto \left(\frac{\mathrm{Id}}{\mathrm{W}}\right) \exp\left(-\frac{\phi_{\mathrm{it}}}{\mathrm{q}\lambda\mathrm{F_{max}}}\right) \approx \left(\frac{\mathrm{Id}}{\mathrm{W}}\right) \left(\frac{\mathrm{Ib}}{\mathrm{Id}}\right)^{\mathrm{m}} \tag{2.22}$$

where $\mathrm{m} = \phi_{\mathrm{it}}/\phi_{\mathrm{i}}$. Since m has been found experimentally around $2.5 - 2.9$ and knowing $\phi_{\mathrm{i}} \approx 1.3\mathrm{eV}$ [66], the threshold energy for HCD has been estimated around $\approx 3.5\mathrm{eV}$.

2.3.1.2 EDP and Bravaix approach

The starting point for the Energy-Driven Paradigm is a simplified expression for hot carrier rates due to an energy mediated process such as impact ionization or interface state

generation [69]. The rates are calculated by the following integral:

$$R = \int f(E)S(E)dE \tag{2.23}$$

where f is the energy distribution function (EDF) and S the cross section or scattering rate. The integrand of this equation generates peaks, which are referred to "dominant energies" because carriers close to these energies dominate the rate itself. This occurs when,

$$\frac{\ln(f)}{dE} = -\frac{\ln(S)}{E} \tag{2.24}$$

Mathematically, the dominant energies are controlled by 'knee' points (points of high curvature) of either ln(f) or ln(S). While the lucky electron model implicitly assumes that the knee points of ln(S) drives the dominant energies, the EDP is based on the idea that the dominant energies are controlled by the knee points of ln(f).

To illustrate the conditions under which the hot carrier behavior is energy driven, an idealized EDF has been used, $f_I(E)$, which collapses the knee to a single point.

$$\begin{aligned} f_I(E) &= \exp(-\chi E/qV_{eff}) && E \leq qV_{eff} \\ &= \exp(-\chi)\exp((qV_{eff} - E)/nkT) && E \geq qV_{eff} \end{aligned} \tag{2.25}$$

Concerning the scattering rate, $S(E) \approx (E - E_{TH})^p$ can be used. Hence, the energy driven regime can be defined as when the dominant energy is dominated by qV_{eff}, which means $E_{TH} + pnkT \leq qV_{eff} \leq E_{TH}/(1 - p/\chi)$ (for $\chi > p$).

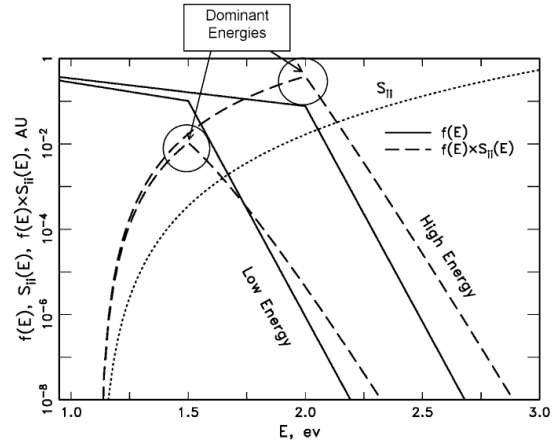
In Fig.2.18(a), an example for II mechanism is shown: $\chi = 3$ is considered and $S_{ii}(E)$ is taken from the model of Kamakura et al., which considers a $p \approx 4.6$ [70].

When the knee determines the dominant energies, the regime is called "energy driven": (1) dominant energies track with bias condition, (2) hot carrier bias dependences are primarily due to the energy dependence of the cross section S(E). The field dependence of the carrier EDF (value of χ) is secondary.

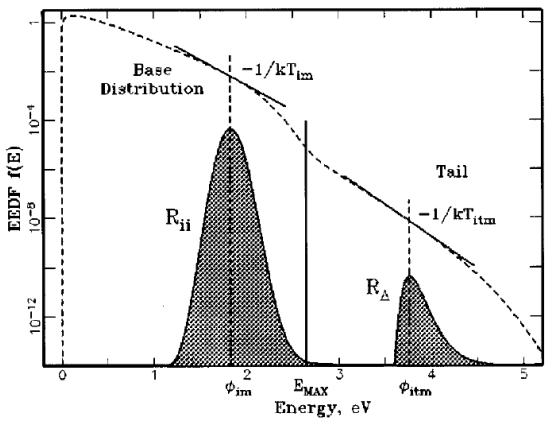
Rauch et al. have been the first to consider the effect of Electron-Electron Scattering (EES) and they inserted its description within EDP framework. The presence of EES brings a modification of the EDF at high energy, thus, it can be modeled as the sum of two distributions at different T_{eff}. In Fig.2.18(b), T_{ii} is associated to a population without EES, whereas T_{it} considers this mechanism.

Following the reasoning of LEM, with a number of carriers $\propto$ Ids for the base distribution and $\propto Ids^2$ for carriers which acquired high energy levels through EES mechanism, as Childs et al. mathematically demonstrated [46],[71],[72], they obtained the expressions for the life time durations τ_{SVE} and τ_{EES}:

$$\tau^{-1} = \tau_{SVE}^{-1} + \tau_{EES}^{-1} \approx C_1 Ids \left(\frac{Ibs}{Ids}\right)^{m_1} + C_2 Ids^2 \left(\frac{Ibs}{Ids}\right)^{m_2} \tag{2.26}$$



(a) A graphical representation [69] of the energy driven hot carrier paradigm applied to II. Two different values of Veff are used.



(b) Quantities involved in the generalized EETM [47]. The effective temperatures T_{im} and T_{itm} are related to the slopes of the EDF at ϕ_{im} and ϕ_{itm} as shown.

Figure 2.18: Energy Driven Paradigm graphical representations.

$m_1 = \phi_{it}/\phi_{ii}$, as found for LEM, and $m_2 = (\phi_{it}Tit)/(\phi_{ii}Tii)$ [47].

The group of Bravaix et al. took advantage of this attitude and improved the EDP approach. In accordance with Rauch model, the E_{dom} is associated with qV_{eff}, thus:

$$\frac{Ibs}{Ids}(V_{eff}) \approx S_{ii}(qV_{eff}) = (qV_{eff} - \phi_{ii})^{p_{ii}} \tag{2.27}$$

The eq.2.27 can be used to determine experimentally the II scattering cross section with energy, plotting the ratio Ibs/Ids function of V_{eff}. Guerin et al. [8] found $\phi_{ii} = 1.1$eV and $p_{ii} = 4.2$, values which have been used in several models for HCD.

In the energy driven regime, $E_{dom,ii}$ and $E_{dom,it}$ should be close since both determined by the tail of EDF. For this reason, assuming $f(E)S_{it}(E) \approx \delta(E - E_{dom,it})f(E_{dom,it})S(E_{dom,it})$, the eq.2.23 becomes:

$$R_{it} \approx f(E_{dom,it})S(E_{dom,it}) \approx f(qV_{eff})(qV_{eff} - \phi_{it})^{p_{it}} \approx Ids\left(\frac{Ibs}{Ids}\right)^{\frac{p_{it}}{p_{ii}}} \tag{2.28}$$

Respect to eq.2.22 and 2.26, the factor m is not ϕ_{it}/ϕ_{ii}, but is p_{it}/p_{ii}. As previously said, the value of ϕ_{it} has been fixed around 3.5eV. This hypothetical value of ϕ_{it} implies a strong reduction of HCD for voltages below 3V, since electrons can not reach the threshold energy necessary to break the Si-H bond. However, this is not verified and can be qualitatively explained by the new factor m in eq.2.28 which is not linked to the "threshold energy" anymore but to the scattering power law exponent. With similar observations, the EES mode becomes:

$$\tau_{EES}^{-1} \approx Ids^2\left(\frac{Ibs}{Ids}\right)^{m} \tag{2.29}$$

where m is a new exponent factor equal to p_{it}/p_{ii}.

In the following, the ratio $\frac{Ibs}{Ids}$ is named r_{ii}. Taking advantage of eq.2.28, an appropriate representation of lifetime duration is to track $\tau \cdot r_{ii}^{m} = \tau \cdot S_{it}$ function of Ids/W. This representation highlights the different degradation modes, as it has been done by Bravaix and Guerin [6],[8] for different oxide thicknesses and stress conditions in Fig.2.19. In this figure, the LEM lifetime plot $\tau \cdot S_{it}$ vs Ids/W shows the three different HC regimes. For high energy (mode 1), it is linearly dependent on $\propto Ids^{-1}$, for medium-energy range (mode 2) $\propto Ids^{-2}$, both predicted by the first version of EDP model. On the other hand, at low energy (mode 3) data largely deviate at high Ids/W values. This proves that damage due to cold carriers becomes driven by Multi-Vibrational Excitation (MVE) of the Si-H bond until breakage [73],[74],[75]. This had been already found using a scanning tunneling microscope (STM) under high current density [76],[77], where the dissociation energy paths of Si-H bonds under stretch mode has been found around 0.25eV [78], in relation to the long vibrational lifetime of Si-H bonds that can decay via the excitation in four phonons [77].

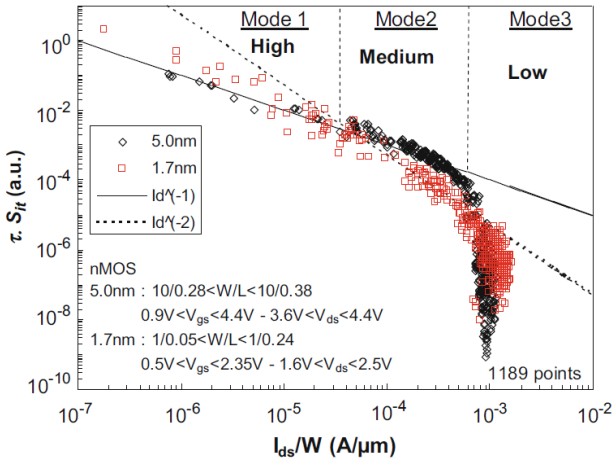


Figure 2.19: Transfer of the usual (LEM) lifetime plot using $\tau r_{ii}^{m} = \tau S_{it}$ vs Ids/W showing three different HC regimes (NMOSFETs with medium to thin Tox, 1,189 data points) [6]. The chosen lifetime criterion is 10% of the saturated Id_{sat} (Vgs=Vd=Vdd) in forward mode.

To describe the MVE process, Bravaix et al. used the truncated harmonic oscillator model for the Si-H bond [6],[79]. Considering $\hbar\omega$ being the distance between the oscillator levels in the corresponding quantum well, the respective MVE rate has been calculated in [6],[79] to be:

$$\tau_{MVE}^{-1} \approx \left[(qVd - \hbar\omega)^{1/2}\,(Ib/W)\right]^{E_B/\hbar\omega} \exp\left(-E_{emi}/kT\right) \approx Vd^{1/2}\left(\frac{Id}{W}\right)^{E_B/\hbar}\omega \qquad (2.30)$$

To translate this approach into kinetic age laws, the Lee method is usually applied [80]:

$$Age = \frac{t}{\tau} = t\left(\frac{1}{\tau_{SVE}} + \frac{1}{\tau_{EES}} + \frac{1}{\tau_{MVE}}\right) \qquad (2.31)$$

where Age is a parameter reflecting the quantity of interface traps. For this reason, the inverse of the maximum trans-conductance $1/\mathrm{Gm_{max}}$ is usually taken, in accordance with [33].

2.3.1.3 Physical models based on SHE solvers

The approaches seen in the previous section may suffer from some shortcomings. First of all, these models consider HCD mechanisms as independent processes. This implicit assumption is not physically reasonable because the charged traps distort the distribution function and hence the rates of the scattering mechanisms. In other words, the created interface defects impact the HCD kinetics itself. Thus, the energy exchange mechanisms and bond dissociation processes need to be considered self-consistently within the same simulation framework. In addition, such treatments are too simple and do not account for the spatial distribution of the rate along the channel interface and give just a general total amount of interface defects. For all these reasons, TCAD modeling has to be done to precisely simulate the degradation from a microscopic standpoint.

In order to simulate the HCD, the first step is to precisely acquire the EDF f(E) along the interface. Indeed, as the eq.2.23 underlines, the degradation rate strongly depends on this function. Starting from the principle that the probability to find a particle in the six-dimensional phase space of Cartesian coordinates r and momentum p is equal to $\mathrm{f(r,p)drdp}$ and assuming a net scattering rate $\mathrm{S_{net}(r,p)}$, the relationship $\mathrm{df(r,p,t)/dt = S_{net}(r,p)}$ is valid, thus:

$$\frac{\partial \mathrm{f}}{\partial \mathrm{t}} + \dot{\mathrm{r}} \cdot \nabla_{\mathrm{r}}\mathrm{f} + \dot{\mathrm{p}} \cdot \nabla_{\mathrm{p}}\mathrm{f} = \mathrm{S_{net}(r,p)} \tag{2.32}$$

Knowing the Lorentz's relationship between momentum and electric field in absence of magnetic field, developing $\mathrm{S_{net}(r,p)}$ making explicit the scattering term, and considering the Steady-State condition, we find the most common used expression for the Boltzmann Transport Equation (BTE):

$$\mathrm{v_g(p)} \cdot \nabla_{\mathrm{r}}\mathrm{f(r,p)} - \frac{\mathrm{q}}{\hbar}\mathrm{F(r)} \cdot \nabla_{\mathrm{p}}\mathrm{f(r,p)} = \int \mathrm{S(p',p)f(r,p')d^3p'} - \mathrm{f(r,p)} \int \mathrm{S(p,p')d^3p'} \tag{2.33}$$

where $\mathrm{v_g(p)}$ is the group velocity, $\mathrm{S(p_1,p_2)}$ is the scattering rate operator defining the momentum change from $\mathrm{p_1}$ to $\mathrm{p_2}$ and the second term represents the difference between in- and out-scattering. Now, it is evident that the solution of eq.2.33 needs massive computational efforts. For this reason, an approximated solution can be found expanding the EDF in spherical harmonic components as

$$\mathrm{f(r,p)} = \mathrm{f_0(r,p)} + \mathrm{Y_1^n}(\theta,\phi)\mathrm{f_1^n(r,p)} + \mathrm{Y_2^m}(\theta,\phi)\mathrm{f_2^m(r,p)}... \tag{2.34}$$

where (θ, ϕ) are the spherical coordinates and Y_i^j are the harmonic functions. Making explicit the carrier energy dependence, considering just the first-order component and including it into eq.2.33, an approximated version of the BTE can be found [81]:

$$-\nabla_r \cdot \left[\frac{v^2(r,E)}{3}\tau(r,E)g(r,E)\nabla_r f(r,E)\right] = g(r,E)S_{net}(r,E) \tag{2.35}$$

where τ^{-1} is the total scattering rate and g is the density-of-states per valley.

Once EDF has been acquired solving the SHE model of BTE, the hot carrier degradation rate can be calculated. The aging rate due to carriers between E and E+dE is simply given by $d\phi(E)\sigma(E)$, where $d\phi(E)$ is the electron flow in (E,E+dE) and $\sigma(E)$ is the cross section describing the Si-H bond. Since $d\phi(E) = dn(E) \cdot v(E) = f(E)g(E)dE \cdot v(E)$, the hot carrier rate becomes:

$$k_i = \int f(E)g(E)v(E)\sigma_i(E)dE \tag{2.36}$$

where i is for SVE or MVE, whereas the hot carrier cross section follows the Keldysh formulation [82]:

$$\sigma_i(E) = \sigma_{i,0}\left(\frac{E - E_i}{kT}\right)^{p_i} \tag{2.37}$$

It is worth noting that the rate equation in eq.2.23 used within EDP theory is a simplified expression of the more general eq.2.36.

In order to compute the aging kinetics, i.e. N_{it} vs $Time_{stress}$, using the rate coefficient in eq.2.36, the Reaction Diffusion (RD) model is usually considered [83]. Concerning such an approach, the passivation term, i.e. Si-H bond recovery kinetics, is commonly neglected in the formulation, since leading to a fast saturation of the aging dynamics, which never experimentally verified.

Now, considering separately the two breakage processes and using a modified overall kinetic equation for MVE mechanism, where the Si-H bond is modeled as N energy levels of a truncated-harmonic oscillator [85], it is simple to show that:

$$N_{it,SVE} = P_{SVE}N_0\left(1 - e^{-k_{SVE}t}\right) \tag{2.38}$$

$$N_{it,MVE} = P_{MVE}N_0\left[\frac{P_{emi}}{P_{pass}}\left(\frac{P_u}{P_d}\right)^N\left(1 - e^{-P_{emi}t}\right)\right]^{1/2} \tag{2.39}$$

where P_{emi} and P_{pass} are the emission/passivation probability rates having a form like $P_{emi/pass} = \nu \cdot \exp[-E/(kT)]$ [6], whereas P_u and P_d are the oscillator excitation/deexcitation probability rates given by [87]:

$$P_u = k_{ph}\exp(-E_{ph}/kT) + k_{MVE} \tag{2.40}$$

$$P_d = k_{ph} + k_{MVE} \tag{2.41}$$

with E_{ph} being the phonon energy corresponding to the distance between the oscillator levels.

It is worth noting that these kinetic equations are exponential laws with time, which lead to $\Delta Id_{lin} \approx 1 - \exp(t)$, whereas, as a matter of fact, a power-law is observed. This mismatch between the theory and the experimental observation has been explained with a dispersion of the Si-H bond energy, which, in this simulation framework, is E_i in eq.2.37. Varghese was the first showing how a superposition of weighted exponential kinetics can actually lead to an overall macroscopic power law [86]. Considering this assumption, a gaussian dispersion $g_A(E)$ of the bond energy is usually accounted for. Hence, the total kinetics becomes [84]:

$$N_{it,tot,i}(t) = \int_{E_i-3\sigma}^{E_i+3\sigma} g_A(E) N_{it,i}(t, E) dE \tag{2.42}$$

where i is SVE or MVE process. This model has been used by the Universities of Bologna and Vienna [54], [44], [85], [63], [62], [84], [61], [88].

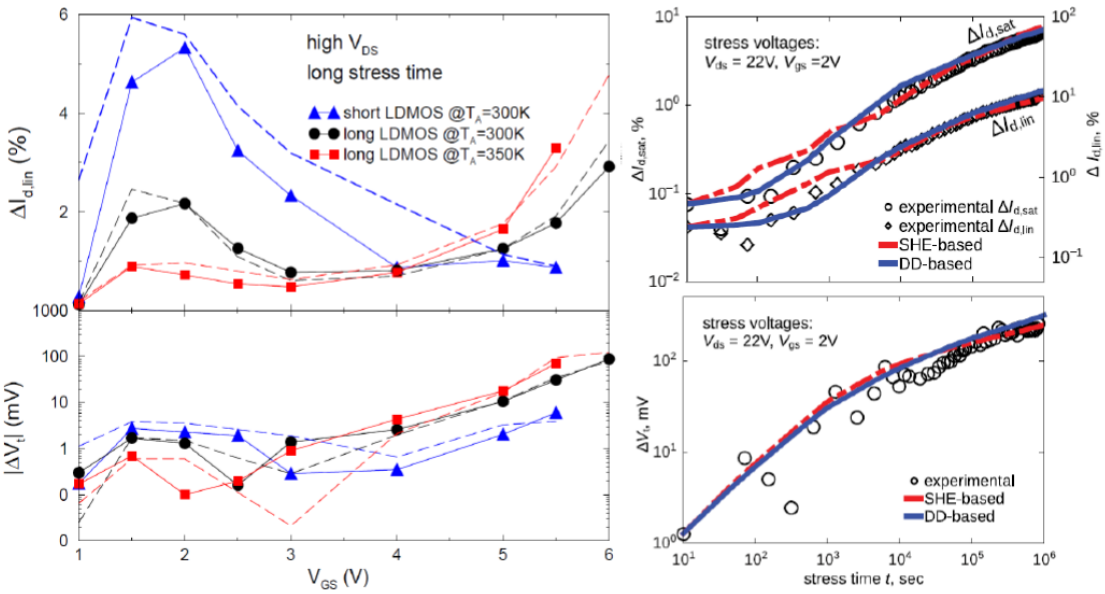


Figure 2.20: HCD results taken from [84] (left) and [63] (right). Experiments are compared with TCAD numerical simulations.

In addition, they tried to strongly reduce the computational time imposing an analytical non-Maxwellian formulation for the EDF. An example is given by Reggiani formulation [84]:

$$f(E) = \frac{1}{A} \exp\left[-\alpha \frac{\gamma(E)}{kT_e}\right] \tag{2.43}$$

where $\gamma(E)$ is the energy function accounting for band structure effects, T_e is the electron temperature, and A and α are fitting parameters which are determined forcing the EDF expression in eq.2.43 to respect the Drift-Diffusion parameters such as n and T_e. A similar formula has been proposed by Grasser group [63].

Once $N_{it}(\tilde{r}, t)$ is calculated at Silicon-Dielectric interface, using the complete TCAD model based on SHE solver or passing through the analytical expression of EDF, these data have to be loaded in the device model and linear parameters, such as Vth and Id_{lin} have to be simulated at different stress times. In this way, simulations and experimental results can be compared in terms of a certain Δ vs $Time_{stress}$. Some of these results are shown in Fig.2.20.

2.3.1.4 Limitations of existing models

In this section, the state of the art for HCD modeling has been presented. We saw how EDP-based compact models do not account for a correct description of the energy distribution function and the intrinsic two-dimensional nature of the problem. For this reason, BTE-based numerical models have been introduced. However, the complexity of these models do not allow to have a precise idea of what is really happening at microscopic scale, since accounting for all scattering and aging mechanisms without assuming any simplifications. Thus, they introduce an amount of fitting parameters that always allows to reproduce the experimental data. In order to highlight the effect of each microscopic parameter and thus directly capture the physics, they need to be simplified. In addition, such models are often calibrated on general Id_{lin} or Vth which are not representative of HCD. For example, as we saw in the first sub-chapter, a localized trap at the drain side has no impact on the electrostatics (Figs.2.14,2.15). Thus, why should we track Vth during stress time? On the other hand, Id_{lin} or Gm_{max} are mainly sensitive to the total amount of interface traps (Fig.2.15). Indeed, compact models based on EDP paradigm are rightly calibrated on these electrical characteristics, since their formulation is based on $N_{it,tot}$ vs $Time_{stress}$, but it is surely not enough for complex numerical TCAD models, since such a calibration loses the trap distribution localization.

In addition, Tyaginov rightly underlined that the energy exchange mechanisms and bond dissociation processes need to be considered self-consistently within the same simulation framework [89]. However, in most of TCAD models this is not accounted for anymore [62], [63]. Indeed, the EDF is calculated at the beginning of the stress, then the rate coefficient defined in eq.2.36 is computed at $Time_{stress} = 0$ and kept constant under the total hot carrier regime. This implicitly assumes that the aging kinetics stops when $N_{it} = N_0$, where N_0 represents the available total Si-H sites.

On the other hand, the energy bond dispersion representation adds fitting parameters and computational efforts. Considering directly a power law kinetics gives exactly the same results and lowers the complexity of the problem. Indeed, the fitting gaussian function $g_A(E)$ in eq.2.42 can be shown for representation purpose, as Varghese did in [86], but is not representative of the real energy bond cross section dispersion, unless this is directly addressed or measured.

For all these considerations, we will consider our own model and the related calibration methodology in the next section. This has been developed with the aim of overcoming all the limitations previously listed and of making the final model suitable for NOR Flash technology.

2.3.2 HCD for NOR Flash Technology

For all the observations previously done, we used the Synopsys tools [12] trying to reduce as maximum as possible the fitting parameters and considering a self-consistent SHE-solver based procedure. Moreover, we studied the effect of each model parameter at macroscopic scale and we increased the number of electrical characteristics to track in order to better address the aging evolution. Only the SVE mechanism has been considered in the degradation model, which we will widely demonstrate to well represent the HCD for such devices.

The idea of our works [16], [17] is to capture the trap information, such as total trap amount and localization, that are intrinsically contained in the evolution of the entire IdVg curve. For this reason, the trans-characteristic has been measured five times per stress time decade, as shown in Fig.2.21. In addition, different HC-stress conditions, i.e. (Vg,Vd) couples, have been applied up to 1s. Indeed, for such long stress time, corresponding to $\approx 10^6$ Program-only cycles, the Flash cell can be considered "failed", since 1/0 states are no more distinguishable.

It has to be pointed out that the experimental read operations have been performed 10s after the stress interruption. This has been done in order to relax the border trapped electrons and be sensitive mainly to HC-induced interface traps. Indeed, in the first part of the chapter, the presence of HC-induced oxide traps has been observed not to be negligible on device electrostatics.

2.3.2.1 TCAD model

The TCAD model being used in this work has been developed considering real process steps and is electrically calibrated [12]. First of all, although efforts are done by Hot Carrier community on studying reliability issues concerning Self Heating (SH) effects [90],[91], which have been demonstrated to increase the stress induced by HC [92],[93], especially for short channel transistors, in our work this is intentionally neglected. Indeed, as we will demonstrate, the driven degradation mechanism results to be SVE mode for NOR Flash technology. Thus, an increase of the local temperature leads to a (slight) improvement of the reliability [64]: phonon, impact ionization and surface roughness scatterings depopulate the high energy tail of the distribution function, thus with SH they are more efficient leading to a slowdown of the degradation rate. Hence, in the following, we will not consider SH -

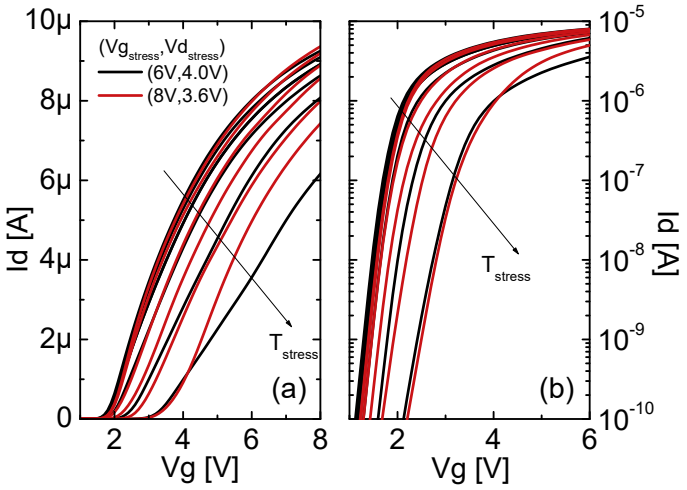


Figure 2.21: Experimental IdVg (Vd=50mV) in Lin (a) and in Log (b) for stress conditions (Vg,Vd) = (6V,4.0V) and (8V,3.6V). The measured curves refer to T_{stress}=10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1s.

HC interaction, since it is not likely to be at the origin of damage localization at drain side, whose detection analysis is an important aspect of our work.

Before considering hot carrier issues, the TCAD device deck has been developed considering the real process steps and has been electrically calibrated on the whole IdVg curve at different Vd_{read}. This procedure, although often neglected, is fundamental to have a quantitative analysis whenever HCD is considered. Concerning the computation of the aging kinetics, accurate knowledge of the non-equilibrium EDF is required. Thus, Spherical Harmonics Expansion (SHE) method has been considered [81], which computes it by solving the lowest-order SHE of BTE as in eq.2.35.

The Si-H bond depassivation rate k at each Si/SiO_2 interface location is given in accordance with the energy distribution function [12]:

$$k = \nu \cdot \left(1 + \delta_{SHE}\frac{qg_v}{2}\int_0^{+\infty} \min\left(\exp\left(\frac{\epsilon - \epsilon_a}{kT}\right), 1\right) g(\epsilon) f(\epsilon) v(\epsilon) d\epsilon\right) \tag{2.44}$$

where ϵ_a represents the activation energy of the Si-H bond [eV]. The integral in eq.2.44 sums at each energy level the electron flux, multiplied by a cross section featuring an exponential dependence on ϵ_a in accordance with [73]. Thus, the interface trap dynamics, i.e. N_{it} [cm^{-2}] vs T_{stress} [s], is computed in agreement with a power law kinetics reflecting both the stretching of the Si-H bond due to the local variation of the chemical potential [83] and the Si-H bond energy dispersion [86]:

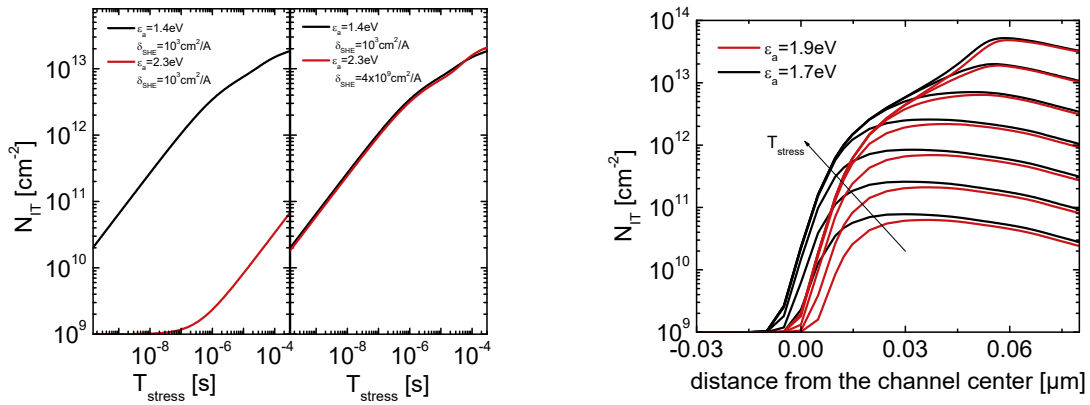
$$\frac{dN_{it}}{dt} = k\left(\frac{N_{it}}{N_{it,0}}\right)^{-\gamma} \tag{2.45}$$

The model has just 5 fitting parameters: ϵ_a, γ, δ_{SHE}, $N_{it,0}$, ν. It has been calibrated tracking the drift of parameters extracted in linear regime. However, looking at eqs.2.44,2.45, the parameters (δ_{SHE},$N_{it,0}$,ν) are redundant: all of them rigidly shift in time the aging kinetics. Thus, the number of fitting parameters is just three.

In our works [16],[17], we underlined the need of accurately studying the sensitivity of model parameters. There are two main reasons for that:

- First of all, to see and capture the macroscopic impact of these parameters. This leads to a deeper understanding of HCD itself.
- Secondly, to properly calibrate our physical-based TCAD model. Knowing the dependence of a certain model parameter, a proper calibration can be experimentally addressed.

As previously underlined, the model has three parameters: γ, ϵ_a and δ_{SHE}. The first is related to the activation energy dispersion thus being a fitting parameter related to oxide quality. The second one is the average Si-H bond energy and its effect is shown in Fig.2.22. Finally, δ_{SHE} is simply an empirical parameter used to shift in time the aging kinetics, similarly to P_{SVE} in eq.2.38.



(a) Simulated defect density N_{it} vs T_{stress} at drain edge (50nm from the channel center) with $\gamma = 0.5$ for two ϵ_a values in (left) with the same δ_{SHE} and in (right) with different δ_{SHE} for shifting in time purpose. Stress condition: (Vg,Vd)=(6V,4V).

(b) Simulated defect density N_{it} vs interface channel position for two ϵ_a at different $T_{stress} = 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1s$ for the case $\delta_{SHE} = 10^3 cm^2/A$ and $\gamma = 0.9$. Stress condition: (Vg,Vd)=(6V,4V).

Figure 2.22: The effect of changing the activation energy in the model on the trap distribution shape along the channel interface.

Concerning ϵ_a, this parameter does not influence the local degradation speed, in terms of $dLog(\Delta N_{it})/dLog(T_{stress})$; however it leads to a rigid shift in time, as shown in Fig.2.22(a),

that can be compensated by adjusting δ_{SHE}. It can be easily shown that this shift in time depends on the interface location. For example, lowering the activation energy, the aging kinetics close to the channel center is shifted at lower stress time more strongly respect to drain edge region since providing higher carrier concentration. This means that the activation energy controls the uniformity of the trap distribution along the interface, thus the defect localization, as clearly shown in Fig.2.22(b). For this reason, this parameter has been used for a quantitative calibration at different stress conditions (i.e. Vg,Vd couples). On the other hand, γ is used as a fitting parameter for the time exponent of the degradation kinetics and it is worth noting that a low value is needed for a proper power law kinetics description.

Taking advantage of all these considerations, the model has been properly calibrated fitting the drift linear characteristics experimentally measured. The parameter set used is shown in Tab.2.1.

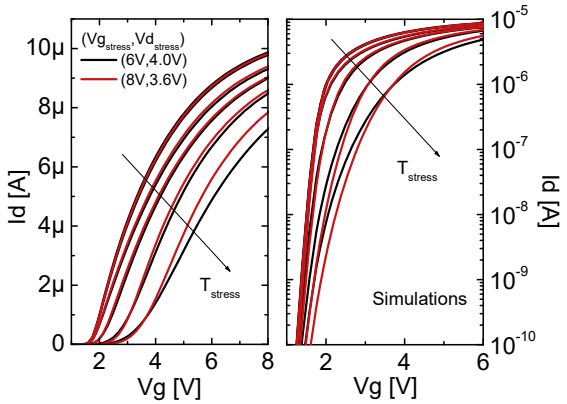
ϵ_a	γ	δ_{SHE}	ν	$N_{it,0}$
2.2eV	2.0	$10^{16}cm^2/A$	$10^7cm^{-2}s^{-1}$	10^9cm^{-2}

Table 2.1: Parameter set used in this work. These have been calibrated on experimental data at different stress conditions (i.e. Vg,Vd couples) and tracking different electrical parameters in linear regime [16]

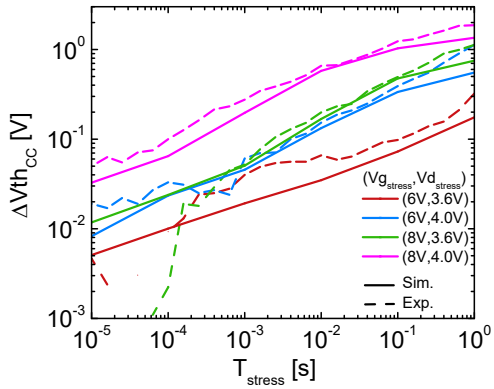
2.3.2.2 Results and discussion

In order to reproduce the experimental results, entire linear IdVg curves have been simulated at different stress times t_i corresponding to the calculated $N_{it}(t_i, x)$. The model parameters of Tab.2.1 have been used. The simulations corresponding to the experimental data of Fig.2.21 are shown in Fig.2.23(a). It is worth noting that the qualitative evolution of this curve is well caught, also considering different stress conditions.

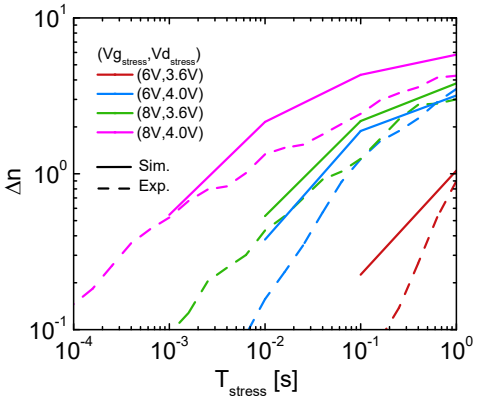
In order to quantify the degradation in terms of Δ vs T_{stress}, relevant parameters reflecting electrostatic and transport degradations have to be chosen. In accordance with [94],[13],[16], in the first case the drift of Vg at low current ΔVth_{CC} (Idth = 10nA) and of the sub-threshold slope Δn have been considered, while the drift of Id at high voltage ΔId_{CV} (Vgth=7V) has been tracked for the mobility aging. The figures 2.23(b), 2.23(c), 2.23(d) show the experimental and simulated results concerning these parameters. It is clear that a good fitting quality has been achieved. However it is worth noting that the sub-threshold slope drift is slightly overestimated, which means that a higher amount of interface traps close to the channel center is being considered in TCAD. Nevertheless, the simulation results qualitatively reproduce a clear Δn saturation as experimentally verified.



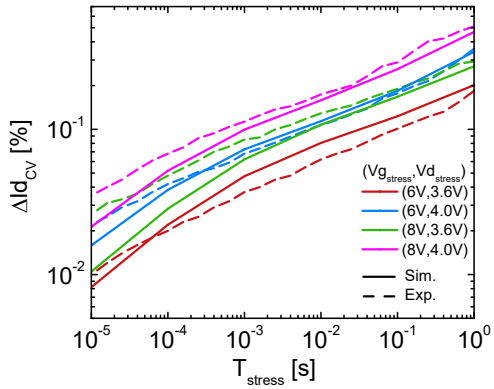
(a) Simulated IdVg (Vd=50mV) in Lin (left) and in Log (right) for stress conditions (Vg,Vd) = (6V,4.0V) and (8V,3.6V). The simulated curves refer to T_{stress}=10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1s.



(b) Vg at low current degradation ΔVth_{CC} (Idth=10nA) vs T_{stress} for simulation vs experimental results at different stress conditions.



(c) Sub-threshold slope degradation Δn vs T_{stress} for simulation vs experimental results at different stress conditions.

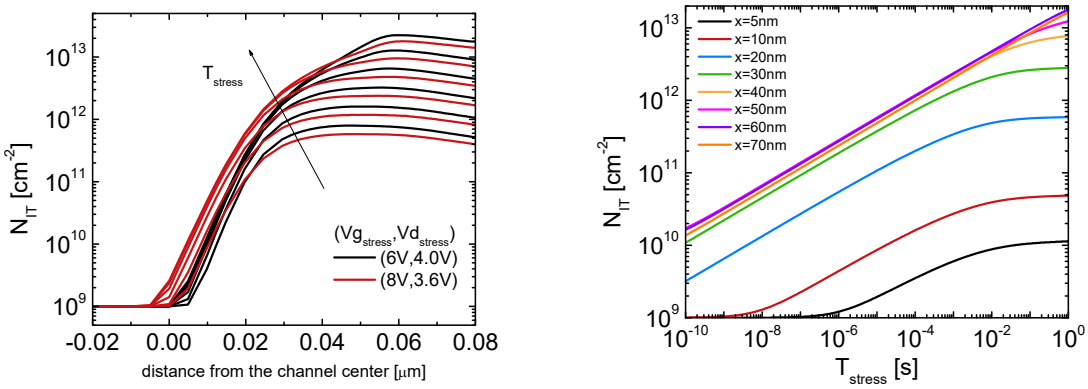


(d) Id at high voltage degradation ΔId_{CV} (Vgth=7V) vs T_{stress} for simulation vs experimental results at different stress conditions.

Figure 2.23: Simulation vs Experimental results of the Hot Carrier Aging model described in this section using the parameter set in Tab.2.1.

Looking at the simulations in Fig.2.24(a), the trap distribution expands in time towards the channel center because of different degradation rates along the interface. However, when the trap distribution in channel starts to significantly affect the electrostatics during hot carrier regime, Vth at Vd_{stress} increases rapidly. As a consequence, electrostatic aging (Δn and ΔVth_{CC}) quickly slows down while ΔId_{CV} slightly accelerates because of an increase of the average electron energy (bands pulled up, i.e. Vd_{sat} decrease). Microscopically, this means that locations further from the drain edge have an aging kinetics that saturates earlier,

as shown in Fig.2.24(b). Despite Id_{CV} drift is well-caught by the model, ΔVth_{CC} does not seem to slow down experimentally. A reasonable explanation of this slight disagreement relies on the presence of trapped electrons in the oxide. Indeed, their influence for this technology has been demonstrated to be significant [13], and a fast measurement setup would be required [95]. However, as already highlighted, for such long stress time, under nominal P/E biases, the Flash cell can be considered "failed", since the programming window is lower than the threshold dispersion. Thus, an accurate aging description for such highly stressed devices is not of interest.



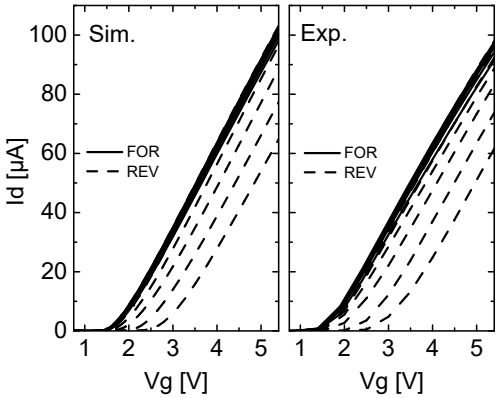
(a) Simulated defect density N_{IT} vs channel position for two stress conditions at different T_{stress}=10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1s

(b) Simulated defect density N_{IT} vs T_{stress} at different interface positions ("x" denoting the distance from the channel center) for the stress condition (Vg,Vd)=(8V,3.6V).

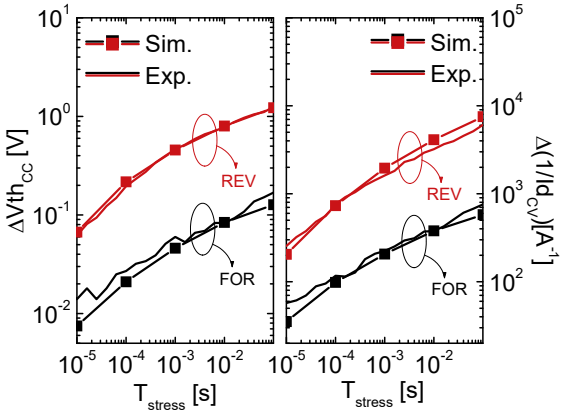
Figure 2.24: Aging kinetics for position further from the drain side occurs with a certain delay, whereas saturation occurs earlier. This is shown looking at the simulated interface defects along the interface at different stress times T_{stress} (a) and function of T_{stress} at different positions (b).

In order to extend the validity range of the model, drifting parameters in Forward (FOR) and Reverse (REV) mode in saturation regime are considered in order to be even more sensitive to the trap distribution location. Fig.2.25 shows the drift of the IdVg curves at Vd=3V and the corresponding extracted parameters in both simulated and experimental cases for the stress condition (Vg,Vd)=(8V,3.6V). It can be seen that the simulations well describe not only the parameter drifts, but also the degradation of the entire curve. It is worth noting that no transport aging is observed, since the scattering is mainly phonon-limited at these high energies. A good prediction is also achieved for the stress condition (Vg,Vd)=(6V,4V), as shown in Fig.4 in [17].

In addition, since HCS is mainly driven by lateral field, simulations are also compared with experiments performed on transistors with channel length ranging from 0.12μm to



(a) Simulated (left) and experimental (right) IdVg in saturation regime (Vd=3V) in Forward (FOR) and Reverse (REV) cases. The curves refer to T_{stress}=10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1s.



(b) Simulated and experimental parameter drifts in saturation regime (Vd=3V) in Forward (FOR) and Reverse (REV) cases. Vg at constant current drift (left) evaluated at Id = 30μA, Id at constant voltage drift (right) evaluated at Vg=4.5V.

Figure 2.25: Experiments and simulations are compared looking at Forward and Reverse results in Saturation Regime for the stress condition (Vg,Vd)=(8V,3.6V).

0.16μm as a figure of merit of model robustness. Fig.2.26 shows that L-dependence is qualitatively well captured considering different stress conditions and drifting parameters.

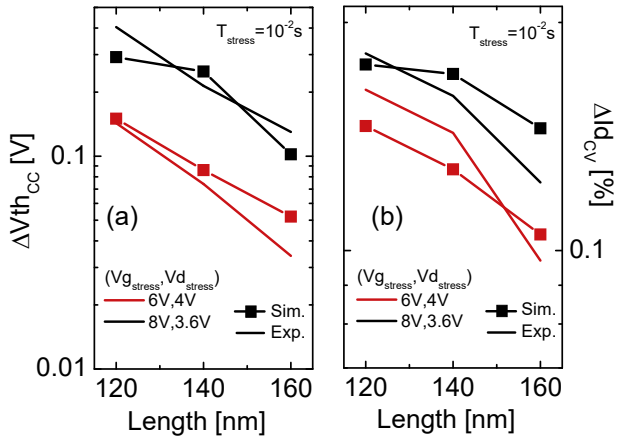


Figure 2.26: Drift of extracted Vg at low constant current (Idth = 140nA/L[nm]) (a) and extracted Id at constant voltage (Vgth=7V) (b), with Vd_{read} = 50mV, function of device length. The figure shows the comparison between experimental and simulated results for the stress conditions (Vg,Vd)=(6V,4V) and (8V,3.6V).

2.3.2.3 Analysis of defect location

Till now, efforts have been done on modeling the aging kinetics of the defect distribution through fitting of electrical parameter drifts. A complementary approach can be considered: the reconstruction of the trap distribution, from macroscopic electrical parameters, would allow a fine analysis of the defect shape evolution with the aim of addressing the Si-H bond depassivation process nature, as similarly done in [94]. This can be addressed looking at ΔVth_{CC} vs ΔId_{ON} graph, as shown in Figs.2.15,2.17(a). Plotting in this graph the trajectory drawn by a stress condition during time, the expansion of the trap distribution can be seen independently from its kinetics. This allows to only focus on the evolution of the defect shape uniformity along the Si/SiO_2 interface, which gives information about the Si-H bond activation energy and, in particular, about the driven HC-depassivation mode. It is worth noting that in Figs.2.14,2.15 a gaussian distribution with fixed standard deviation has been considered for simplification purpose.

In several works on Hot Carrier Degradation modeling the model calibration is done on one single parameter, generally the drift of the threshold voltage or the linear drain current, whose extraction methods are not always clear. Indeed, for calibration purpose, the correlation of two parameters addressing qualitatively the defect location (and concentration) must be captured, in order to catch its evolution independently from the aging kinetics, thus independently from the interface oxide quality (γ in eq.2.45). Afterward, the correct description law of the degradation rate has to be always empirically fitted: considering a power law [83],[86] or setting an "ad hoc" energy distribution of ϵ_a. Concerning the model described in the previous sub-section, the only parameter which can be tuned, in order to catch the trajectory drawn by ΔVth_{CC} vs ΔId_{ON} during the Hot Carrier Regime, is the activation energy: all the others, indeed, are rigid shift in time or affect the kinetic speed. As previously highlighted in Fig.2.22, the activation energy controls the uniformity of the trap distribution: when decreasing it, the defect expands towards the channel center, as also shown in 2.27(a), thus "moves" to higher ΔVth_{CC}, which is confirmed by simulation results in 2.27(b). For this reason, the correlation between these two parameters is a guideline for calibration purpose: its evolution helps to quantitatively extract a microscopic parameter such as the average activation energy of the Si-H bond. Concerning this work, the activation energy which well fits the experimental data is equal to 2.2eV. This high value reflects the single electron impact mode description, in accordance with the model being used.

The experimental results concerning six different stress conditions are shown in Fig.2.28(a). Comparing them with simulation data in Fig.2.28(b), obtained with the aging model described in this section, it is worth noting that a qualitative prediction of the trajectories is well achieved for several $(Vg_{stress}, Vd_{stress})$ couples. Having the simulated defect distribution formation, the trajectory drawn in ΔVth_{CC} vs ΔId_{ON} graph during stress time can be

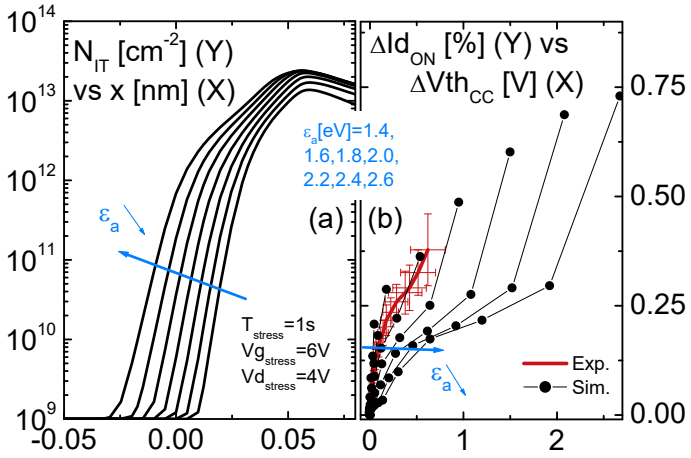


Figure 2.27: Simulated degradation for the stress condition Vg=6V, Vd=4V using the model described in the previous section and considering seven values of activation energy. (a) Trap concentration along Si/SiO_2 interface at fixed stress time, where "x" denotes the distance from the channel center in accordance with Fig.2.12(left). (b) Correlation between ΔVth_{CC} vs ΔId_{ON} as in Fig.2.15: each point represents one IdVg simulation. The red solid line corresponds to experimental results and the error bars are the standard deviations of the extracted values.

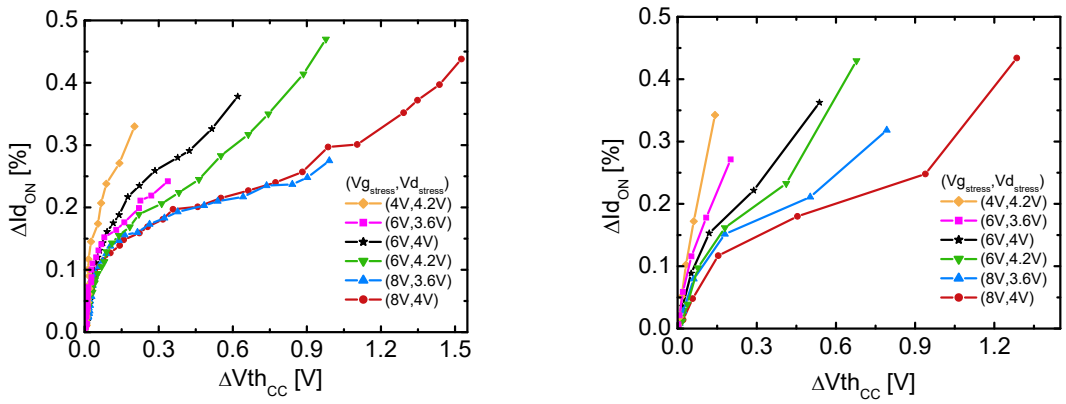
physically analyzed from a microscopic standpoint. Indeed, its evolution can be divided into three parts:

- Initially, almost vertical trajectories are observed, which means that the interfacial defects are mainly created in the first nanometers from the drain edge. This is in accordance with eqs.2.44,2.45: that region experiences higher aging rate, since having more energetic carriers.

- Then, the defect distribution expands towards the channel center and, in accordance with what said previously, the curve moves to higher ΔVth_{CC}. This is in accordance with the simple theory developed in the previous section (Fig.2.17): fixing α in eq.2.16 and extracting the effective p from experimental data using the relationship $\Delta Vth_{CC} \approx \Delta Id_{ON} \cdot (K_1 + K_2 \cdot p/\alpha)$, one would see an increase of this parameter with stress time (not shown here), indicating that the barycenter of the trap shape is shifting towards the channel center. In addition, a separation of the trajectories depending on the stress condition can also be observed. An increase of Vg_{stress} makes the trap distribution more uniform, i.e. closer to the channel center, since it accelerates the degradation in regions where aging rate is mainly driven by the quantity of electrons and not by their energy. On the other hand, considering a different level of Vd_{stress},

the aging kinetics barycenter is modified depending on the values of applied voltages, since an increase of it accelerates the carriers not only at the drain edge.

- Finally, an increase of $\partial\Delta Id_{ON}/\partial\Delta Vth_{CC}$ can be observed for long stress time. This can be explained with a slowdown of the aging kinetics in regions further from the drain edge. Indeed, for long stress time, the trap distribution in channel starts to significantly affect the electrostatics during Hot Carrier Regime, i.e. Vth at Vd_{stress} increases rapidly. We already saw that this leads to a slow down of the overall electrostatic wear out, whereas transport aging slightly accelerates thanks to an increase of the average electron energy. This means that locations further from the drain edge have an aging kinetics which significantly slows down earlier, as simulations confirm in Fig.2.24(b), thus the barycenter of the trap distribution moves back towards the drain edge. Within the theory framework developed in this section, this is translated into a change of trajectory in ΔVth_{CC} vs ΔId_{ON} graph: for long stress time $\partial\Delta Id_{ON}/\partial\Delta Vth_{CC}$ increases (i.e. effective p of eq.2.15 decreases). Looking at Figs.2.28(a),2.28(b), this effect is qualitatively well captured by simulation results, which highlights the validity of the physical interpretation.



(a) Experimental correlation.

(b) Simulated correlation using the Hot Carrier Aging model presented in this section.

Figure 2.28: Correlation between the extracted drifts of Vth_{CC} and Id_{ON} ($Vd = 50mV$) for different stress conditions.

However, considering low Vd_{stress} and long stress time, the position of the trap distribution might not be well caught by the model, since ΔVth_{CC} seems to be underestimated. The causes of this mismatch can be multiple. It can be due to the presence of negative oxide charges within the tunnel oxide, i.e. electrons trapped during the Hot Carrier Regime. Indeed, their influence for this technology has been demonstrated to be significant [95].

Anyhow, this mismatch can also be explained by an underestimation of traps close to the channel center, whose origin can be attributed to Multi Vibrational Excitation mode (MVE) and/or Electron-Electron Scattering (EES) [85],[8]. Indeed, their physical descriptions are missing within the model. In this case, the activation energy can be tuned in order to empirically capture the effect of low energy degradation modes in terms of an equivalent $\Delta\epsilon_a$. Nevertheless, the voltages actually involved in the Flash programming phase (Vd=4V-4.2V) are high enough to consider the process as SVE, making this HCS model of interest.

2.3.2.4 Compact modeling for NOR Flash Technology

In this chapter, it has been widely demonstrated that for 40nm NOR Flash Technology the aging mode driving HCD is SVE, which means that a single electron travelling from source to drain and exchanging energy due to scattering mechanisms (II, surface roughness and Coulomb scatterings) acquires enough energy to break a Si-H bond at Si/SiO_2 interface and create an active defect. This is not surprising, indeed the biases are high enough to consider the process as SVE. In particular, the drain voltage is set at 4-4.2V at product level in order to get fast programming operations. For such high values ($> \epsilon_a/q$) and long channel device, the degradation mode driving the device wear out results to clearly be SVE. It is worth noting that the absence of MVE is necessary, since it would lead to an additional degradation for the same quantity of injected charge in the Floating Gate, thus it would dramatically speed up the failure of the cell itsef.

Since MVE and EE modes are negligible for this technology (the EE scattering mechanism has not been considered in the SHE solver), our compact model has to refer to the simple LEM. For this reason, similarly to eq.2.22, the degradation rate can be written as:

$$\tau^{-1} \approx n \cdot \exp\left(-\frac{\phi_{it}}{q\lambda F_{max}}\right) \tag{2.46}$$

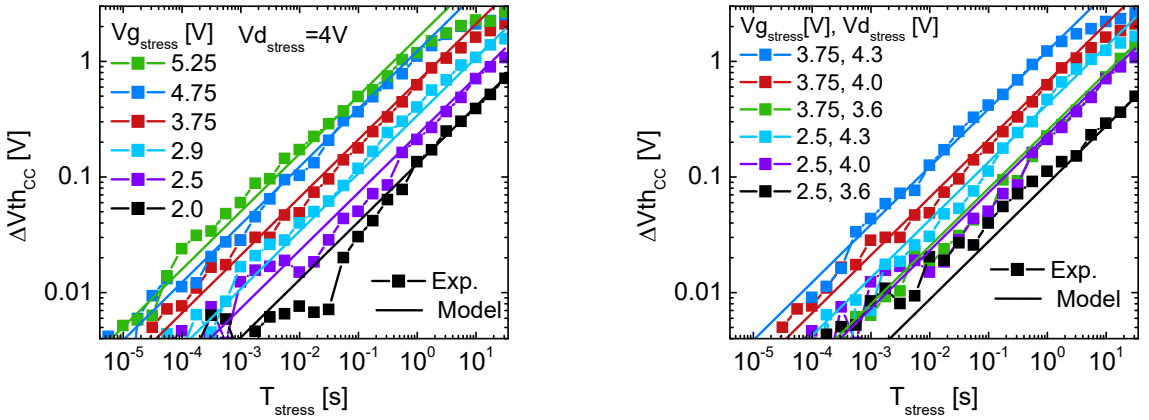
where n is the carrier density at $F = F_{max}$. In order to make explicit the bias dependences, the approximations $n \approx Vg^p$ and $F_{max} \approx Vd$ are considered. Then, the aging kinetics $\Delta(t)$ can be calculated as t/τ. However, in order to account for the activation energy (i.e. ϕ_{it}) dispersion, the following assumption is done:

$$\Delta \approx \left(\frac{t}{\tau}\right)^n \approx A \cdot Vg^p \cdot \exp\left(-\frac{\alpha}{Vd}\right) \cdot t^n \tag{2.47}$$

The compact model in eq.2.47 represents a simplified approach respect to the model developed by Arfaoui et al. in [96]. They considered a different Vd dependence and $n \approx (Vg - Vth)^p$ for the carrier concentration. However, it can be demonstrated that also this model well fits our experimental data (not shown here).

It has to be pointed out that the empirical parameter p in eq.2.47 not only accounts for the carrier concentration but also for the trap expansion dependence with the bias condition. Indeed, the kinetics along the interface is shifted in time depending on the position, because the electron flux f(E)g(E)v(E) is not constant along the channel, thus, the overall effect is an expansion of the created trap distribution towards the source. Simulating the HCD, this expansion clearly depends on Vg level, since modifying the electron flux at each interface location. In addiction, it has been demonstrated that the activation energy controls the uniformity of the aging kinetics along the interface. For this reason, the Vg dependence (i.e. p in eq.2.47) drastically varies when changing ϵ_a,. However, for simplification purpose, all these phenomena are not microscopically considered and are empirically incorporated in p.

In order to be consistent with Flash endurance results, the read condition has been changed into $Vd_{read} = 0.5V$, whereas ΔVth_{CC} at $Id = 8\mu A$ has been considered. Since this parameter has been extracted above the real Vth, it is representative for the overall hot carrier induced aging and it is perfectly correlated with Gm_{max} drift (not shown). The results are shown in Fig.2.29 and it is observed that the trends are well captured not only at different Vg, but also considering different drain biases. However, it is worth noting that for highly stressed devices the model overestimates the Vth_{CC} drift because the saturation appears. In any case, this condition occurs for long stress time, when the respective Flash cell is considered failed, thus it is not of our interest.



(a) ΔVth_{CC} vs T_{stress} at different Vg conditions, fixing Vd=4V.

(b) ΔVth_{CC} vs T_{stress} at different Vd conditions, fixing Vg=3.75V and Vg=2.5V.

Figure 2.29: HCD results concerning different stress conditions and tracking ΔVth_{CC} at $Id = 8\mu A$ ($Vd_{read} = 0.5V$). The model being used to fit the experimental data is eq.2.47 where A=135, p=2.6, $\alpha = 35$, n=0.5.

This simple model will be used to predict the NOR Flash endurance results in the last

chapter. However, in order to complete the study on HCD for this technology, the degradation of the gate current should also be considered. Unfortunately, a device allowing gate current measurement during Hot Carrier Regime is highly challenging. Indeed, a too small matrix of devices in parallel does not allow to measure the gate current since the signal would be too low (if increasing the measurement time, the measured gate current would be already degraded), whereas a large matrix would lead to a huge total drain current which is not manageable and dramatically increases the voltage drop at drain contact due to series resistances. In our case, it can be shown that a matrix of 10x10 transistors in parallel would be the perfect solution for our purpose. This device is the best compromise overcoming this trade-off and allowing the measurement of both gate current aging, i. e. efficiency decay of the cell, and the drift of linear parameters, i. e. the electrostatic evolution of the cell. In addition, the TCAD model can be calibrated considering both aspects. This can be crucial, since in all the works in which the BTE is numerically solved, the EDF is never verified and is considered as correct, despite the tremendous amount of fitting parameters and physical mechanisms considered. The reproduction of gate current and its relative decay would highly increase the accuracy and authenticity of such a model.

2.4 Conclusions

The chapter explores the Hot Carrier Degradation on equivalent MOSFET devices belonging to 40nm NOR Flash technology.

A deep investigation has been firstly done on HCD parameter extraction and on the related sensitivity respect to device aging at microscopic scale. After having analyzed the conventional extraction methodologies and underlined the difficulty to correctly separate pure electrostatic and transport degradations of linear characteristics occurring in the presence of interface charges, a novel technique that successfully addresses this point is presented and validated. In addition, the relationship between macroscopic drifting parameters and microscopic defects is studied. In particular, $\mathrm{Vth_{CC}}$ (extracted Vg at fixed low Idth) is shown to strongly depend on trap position, whereas $\mathrm{Id_{ON}}$ (extracted Id at $\mathrm{Vth_{CC}} + \Delta\mathrm{V}$) results to be just sensitive to defect concentration. Thus, the evolution of the related parameter correlation is shown to describe the microscopic expansion of the interface trap distribution.

Taking advantage of these considerations, a TCAD Spherical Harmonic Expansion (SHE) based HCD model has been calibrated at different stress conditions, read configurations and device lengths. The parameter sensitivity of the model has been analyzed, underling the need for a low exponent in the aging kinetics, simulated through a semi-empirical γ factor, and for a high Si-H bond activation energy, reflecting the single electron impact mode (SVE). It has been shown how the trap profiles expand towards the channel center till reaching zones controlling the electrostatics during HC regime. This modifies the degradation rates

along the interface leading the drain edge region to control back the overall device aging. Based on these simulation results and on the Id_{ON}-Vth_{CC} correlation, a simple technique for the analysis of trap distribution location has been proposed. The need to calibrate such a correlation with experimental data and its strong connection with Si-H bond activation energy have been highlighted. In addition, physical interpretations on device characteristic evolutions and insights at different stress conditions have been presented.

Finally, a compact model for device dynamics under HCD regime is proposed, with the aim of making it suitable for Flash cell aging evolution during P/E cycling. More generally, considerations and physical insights provided in this chapter for transistor degradation during HC process will be used for NOR Flash cell endurance understanding and modeling in the last part of the thesis.

Bibliography

[1] E.H. Nicollian, C.N. Berglund, P.F. Schmidt, J.M. Andrews, Electrochemical charging of thermal SiO2 films by injected electron currents. J. Appl. Phys. 42(12), 5654-5664 (1971)

[2] T.H., P.W. Cook, R.H. Dennard, C.M. Osburn, S.E. Schuster, H.N. Yu, 1m most VLSI technology Part IV: Hot-electron design constraints. IEEE Trans. Electron Dev. 26, 346-353 (1979)

[3] C. Hu, Lucky electron model for channel hot electron emission, in Proceedings of the International Electron Devices Meeting (IEDM), 1979, pp. 22-25

[4] G. Groeseneken, R. Bellens, G. Van den bosch, and H. E. Maes. Hot carrier degradation in submicrometre MOSFETs: from uniform injection towards the real operating conditions. Semiconductor Science and Technology, v. 10, n. 9, p 1208-1220, September 1995

[5] M. Cho, P. Roussel, B. Kaczer, R. Degraeve, J. Franco, M. Aoulaiche, et al. Channel Hot Carrier Degradation Mechanism in Long/Short Channel n-FinFETs. IEEE Transactions on Electron Devices, v. 60, n. 12, p 4002-7, December 2013

[6] A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, Hot-carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature, in Proceedings of the International Reliability Physics Symposium (IRPS), 2009, pp. 531-546

[7] S. E. Rauch, G. La Rosa. The Energy-Driven Paradigm of NMOSFET Hot-Carrier Effects. IEEE Transactions on Device and Materials Reliability, v. 5, n. 4, p 701-5, December 2005

[8] C. GuÃ©rin, V. Huard, A. Bravaix. The Energy-Driven Hot-Carrier Degradation Modes of nMOSFETs. IEEE Transactions on Device and Materials Reliability, v. 7, n. 2, p 225-235, June 2007

[9] B. Tuttle, C.G. Van de Walle, Structure, energetics, and vibrational properties of Si-H bond dissociation in silicon. Phys. Rev. B 59(20), 12884-12889 (1999)

[10] W. McMahon, K. Matsuda, J. Lee, K. Hess, J. Lyding, The effects of a multiple carrier model of interface states generation of lifetime extraction for MOSFETs, in Proceedings of the International Conference on Modelling and Simulation Micro, vol. 1, 2002, pp. 576-579

[11] Y. Nishi, Study of silicon-silicon dioxide structure by electron spin resonance. Japanese Journal of Applied Physics, v. 10, n. 1, p 52-62, January 1971

[12] Synopsys, Zurich, Switzerland, Sentaurus device user guide, J-2014.09

[13] G. Torrente, J. Coignus, S. Renard, A. Vernhet, G. Reimbold, D. Roy, G. Ghibaudo, Physically-based extraction methodology for accurate MOSFET degradation assessment, Microelectronics Reliability 55 (9-10) (August 2015) 1417-21

[14] G. Verma and N. Mielke. Reliability performance of ETOX based Flash memories. 26th Annual Proceedings, Reliability Physics, p 158-166, 1988

[15] S. Aritome, R. Shirota, G. Hemink, T. Endoh, F. Masuoka, Reliability issues of flash memory cells, Proc. IEEE 81 (5) (May 1993) 776-88

[16] G. Torrente, X.Federspiel, D.Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo, Hot Carrier Stress modeling: from degradation to kinetics to trap distribution evolution, Proc. IEEE International Integrated Reliability Workshop (IIRW) (Oct. 2015)

[17] G. Torrente, X.Federspiel, D.Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo, Hot Carrier Stress: aging modeling and analysis of defect location, IEEE International Reliability Physics Symposium (IRPS), p 5A-4 (6 pp.), 2016.

[18] Y. Taur, T. H. Ning. Fundamental of Modern VLSI Devices. 2nd Edition, June 2013

[19] J.S. Brugler; P.G.A. Jespers, Charge pumping in MOS devices, IEEE Transactions on Electron Devices, v ED-16, n 3, p 297-302, March 1969

[20] G. Groeseneken, H. E. Maes, N. Beltran, R. F. de Keersmaecker. A Reliable Approach to Charge-Pumping Measurements in MOS Transistors. IEEE Transitions on Electron Devices, v. ed. 31, n. 1, p 42-53, January 1984

[21] G. Van den bosch; G.V. Groeseneken; P. Heremans; H.E. Maes, Spectroscopic charge pumping: A new procedure for measuring interface trap distributions on MOS transistors, IEEE Transactions on Electron Devices, v 38, n 8, p 1820-31, Aug. 1991

[22] S. Mahapatra; C.D. Parikh ; V. Ramgopal Rao; C.R. Viswanathan; J. Vasi, A comprehensive study of hot-carrier induced interface and oxide trap distributions in MOSFETs using a novel charge pumping technique, IEEE Transactions on Electron Devices, v 47, n 1, p 171-7, Jan. 2000

[23] G. Van den bosch; G.V. Groeseneken; H.E. Maes, On the geometric component of charge-pumping current in MOSFETs, IEEE Electron Device Letters, v 14, n 3, p 107-9, March 1993

[24] G. Ghibaudo. New method for the extraction of MOSFET parameters. Electronics Letters, v. 24, n. 9, p 543-545, April 1988

[25] M.F. Hamer. First-order parameter extraction on enhancement silicon MOS transistors. v 33, n 2, p 49-54, April 1986

[26] A. Tsormpatzoglou; K. Papathanasiou; N. Fasarakis; D.H. Tassis; G. Ghibaudo; C.A. Dimitriadis, A Lambert-function charge-based methodology for extracting electrical parameters of nanoscale FinFETs, IEEE Transactions on Electron Devices, v 59, n 12, p 3299-305, Dec. 2012

[27] N. Fasarakis; A. Tsormpatzoglou; D.H. Tassis; I. Pappas; K. Papathanasiou; M. Bucher; G. Ghibaudo; C.A. Dimitriadis, Compact Model of Drain Current in Short-Channel Triple-Gate FinFETs, IEEE Transactions on Electron Devices, v 59, n 7, p 1891-8, July 2012

[28] J.B. Henry; Q. Rafhay; A. Cros; G. Ghibaudo, New Y-function based MOSFET parameter extraction method from weak to strong inversion range, Solid-State Electronics, v 123, p 84-8, Sept. 2016

[29] D. Fleury; A. Cros; H. Brut; G. Ghibaudo, New Y-function-based methodology for accurate extraction of electrical parameters on nano-scaled MOSFETs, IEEE Conference on Microelectronic Test Structures, p 160-5, 2008

[30] Y. Pan; K.K. Ng; C.C. Wei, Hot-carrier induced electron mobility and series resistance degradation in LDD NMOSFET's, IEEE Electron Device Letters, v 15, n 12, p 499-501, Dec. 1994

[31] S.K. Manhas; D. Chandra Sehkar; A.S. Oates; M.M. De Souza, Characterisation of series resistance degradation through charge pumping technique, Microelectronics Reliability, v 43, n 4, p 617-24, April 2003

[32] P.I. Suciu; R.L. Johnston, Experimental derivation of the source and drain resistance of MOS transistors, IEEE Transactions on Electron Devices, v ED-27, n 9, p 1846-8, Sept. 1980

[33] R. Woltjer, G. Paulzen. Universal description of hot-carrier induced interface states in NMOSFETs. International Electron Devices Meeting, p 535-8, 1992

[34] Y.H. Chang; Y.F. Wu; C.S. Ho; A simple method to extract source/drain series resistance for advanced MOSFETs, IEEE International Conference on Electron Devices and Solid-State Circuits - EDSSC '07, p 87-90, 2007

[35] D. Fleury; A. Cros; G. Bidal; J. Rosa; G. Ghibaudo, A new technique to extract the source/drain series resistance of MOSFETs, IEEE Electron Device Letters, v 30, n 9, p 975-7, Sept. 2009

[36] C. Lombardi, S. Manzini, A. Saporito, M. Vanzi, A physically based mobility model for numerical simulation of nonplanar devices, Proc. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 7 (11) (November 1988) 1164-1171

[37] A. Subirats, X. Garros, J. Mazurier, J. El Husseini, O. Rozeau, G. Reimbold, et al. Impact of single charge trapping on the variability of ultrascaled planar and trigate FD-SOI MOSFETs: Experiment versus simulation. IEEE International Reliability Physics Symposium, v. 60, n. 8, p 2604-10, August 2013

[38] C. G. Sodini, T. W. Ekstedt, and J. L. Moll. Charge accumulation and mobility in thin dielectric MOS transistors. Solid-State Electronics, v. 25, n. 9, p 833-841, September 1982

[39] P. J. McWhorter, P. S. Winokur. Simple technique for separating the effects of interface traps and trapped-oxide charge in metal-oxide-semiconductor transistors. Applied Physics Letters, v. 48, n. 2, p 133-5, January 1986

[40] C. Guerin, V. Huard, A. Bravaix, M. Denais, Impact of hot carrier degradation modes on I/O nMOSFETS aging prediction, Proc. IEEE International Integrated Reliability Workshop (IIRW) (2006), pp63-7

[41] X. Wang, A.R. Brown, Binjie Cheng, A. Asenov, Statistical distribution of RTS amplitudes in 20nm SOI FinFETs, Silicon Nanoelectronics Workshop (2012) 1-2

[42] A. Bekaddour, M.G. Pala, N. Chabane-Sari, G. Ghibaudo, Deterministic method to evaluate the threshold voltage variability induced by discrete trap charges in Si-nanowire FETs, Transaction on Electron Devices 59 (5) (May 2012) 1462-7

[43] T. Mizuno, A. Toriumi, M. Iwase, M. Takanashi, H. Niiyama, M. Fukmoto, M. Yoshimi, Hot-carrier effects in 0:1 m gate length CMOS devices, in Proceedings of the International Electron Devices Meeting (IEDM), 1992, pp. 695-698

[44] S. Tyaginov, I. Starkov, H. Enichlmair, J.M. Park, C. Jungemann, T. Grasser, Physics-based hot-carrier degradation models (invited). ECS Trans. 35(4), 321-352 (2011)

[45] A. Zaka et al., An Efficient Nonlocal Hot Electron Model Accounting for Electron-Electron Scattering, IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 59, NO. 4, APRIL 2012

[46] P.A. Childs, C.C. Leung, New mechanism of hot carrier generation in very short channel MOSFETs. Electron. Lett. 31(2), 139-141 (1995)

[47] S.E. Rauch, G. La Rosa, F.J. Guarin, Role of E-E scattering in the enhancement of channel hot carrier degradation of deep-submicron NMOSFETs at high Vgs conditions. IEEE Trans. Device Mater. Reliab. 1(2), 113-119 (2001)

[48] J.D. Bude, Gate-current by impact ionization feedback in submicron MOSFET technologies, in Proceedings of the VLSI Symposium on Technical Digest, 1995, pp. 101-102

[49] F. Venturi, E. Sangiorgi, B. Ricco, The impact of voltage scaling on electron heating and device performance of submicrometer MOSFET's. IEEE Trans. Electron Devices 38(8), 1895-1904 (1991)

[50] J.E. Chung, M.C. Jeng, J.E. Moon, P.K. Ko, C. Hu, Low-voltage hot-electron currents and degradation in deep-submicrometer MOSFET's. IEEE Trans. Electron Devices 37, 1651-1657 (1990)

[51] W. McMahon, A. Haggag, K. Hess, Reliability scaling issues for nanoscale devices. IEEE Trans. Nanotechnol. 2(1), 33-38 (2003)

[52] A. Bravaix, V. Huard, Hot-carrier degradation issues in advanced CMOS nodes, in Proceedings of the European Symposium on Reliability of Electron Devices Failure Physics and Analysis (ESREF), tutorial, 2010

[53] S. Rauch, G. La Rosa, CMOS hot carrier: From physics to end of life projections, and qualification, in Proceedings of the International Reliability Physics Symposium (IRPS), tutorial, 2010

[54] S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmail, C. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric, T. Grasser,

Interface traps density-of-states as a vital component for hot-carrier degradation modeling. Microelectron. Reliab. 50, 1267-1272 (2010)

[55] C. Jungemann, B. Meinerzhagen, Hierarchical Device Simulation (Springer, Wien/New York, 2003)

[56] M. Fischetti, S. Laux, Monte-Carlo study of sub-band-gap impact ionization in small silicon field-effect transistors, Proc. International Electron Devices Meeting (IEDM) 1995, pp. 305-308.

[57] S.-M. Hong, A. Pham, C. Jungemann, Deterministic solvers for the Boltzmann transport equation. 1 em plus 0.5 em minus 0.4 em, Springer, 2011.

[58] T. Grasser, T.-W. Tang, H. Kosina, S. Selberherr, A review of hydrodynamic and energy-transport models for semiconductor device simulation, Proc. IEEE 91 (2) (2003) 251-273.

[59] D. Cassi, B. Ricco, An analytical model of the energy distribution of hot electrons, IEEE Trans. Electron Dev. 37 (6) (1990) 1514-1521.

[60] K. Hasnat, C.-F. Yeap, S. Jallepalli, S.A. Hareland, W.-K. Shih, V.M. Agostinelli, A.F. Tasch, C.M. Maziar, Thermionic emission model of electron gate current in submicron NMOSFETs, IEEE Trans. Electron Dev. 44 (1) (1997) 129-138.

[61] S. Reggiani, G. Barone, E. Gnani, A. Gnudi, G. Baccarani, S. Poli, R.Wise, M.-Y. Chuang, W. Tian, S. Pendharkar, M. Denison, Characterization and modeling of electrical stress degradation in STI-based integrated power devices, Solid-State Electron. 102 (12) (2014) 25-41.

[62] S. Reggiani, S. Poli, M. Denison, E. Gnani, A. Gnudi, G. Baccarani, S. Pendharkar, R. Wise, Physics-based analytical model for HCS degradation in STI-LDMOS transistors, IEEE Trans. Electron Dev. 58 (2011) 3072-3080.

[63] P. Sharma, S. Tyaginov, Y. Wimmer, F. Rudolf, K. Rupp, M. Bina, H. Enichlmair, J.-M. Park, R. Minixhofer, H. Ceric, T. Grasser, Modeling of hot-carrier degradation in nLDMOS devices: different approaches to the solution of the Boltzmann transport equation, IEEE Trans. Electron Devices. 62 (6) (2015) 1-8.

[64] S. Tyaginov, M. Jech, P. Sharma, J. Franco, B. Kaczer, T. Grasser, On the Temperature behavior of Hot-Carrier Degradation Proc. IEEE International Integrated Reliability Workshop (IIRW) (Oct. 2015)

[65] C. Hu et al., IEEE Trans. Electron Devices 32, 375 (1985)

[66] S. Tam et al., Lucky-Electron Model of Channel Hot-Electon Injection in MOSFET's, IEEE Trans. Elec? Dev. 31, 1116 (1984)

[67] A. Bravaix, V. Huard, F. Cacho, X. Federspiel, D. Roy et al., Hot-carrier degradation in decananometer CMOS nodes: From an energy driven to a unified current degradation modeling by multiple carrier degradation process, in Hot-Carrier Degradation, ed. by T. Grasser (Springer, Wien/New York, 2015)

[68] S. Rauch, G. La Rosa, The energy driven paradigm of NMOSFET hot carrier effects, in Proceedings of the International Reliability Physics Symposium (IRPS), 2005

[69] S. Rauch, F. Guarin, The energy driven hot carrier model , in Hot-Carrier Degradation, ed. by T. Grasser (Springer, Wien/New York, 2015)

[70] Y. Kamakura et al., J. Appl. Phys. 75, 3500 (1994)

[71] P. A. Childs and C. C. Leung, A one-dimensional solution of the Boltzmann transport equation including electron-electron interactions, J. Appl. Phys., vol. 79, no. 1, pp. 222-227, 1996.

[72] M. Y. Chang, D.W. Dyke, C. C. Leung, and P. A. Childs, High-energy electron-electron interactions in silicon and their effect on hot carrier energy distributions, J. Appl. Phys., vol. 82, no. 6, pp. 2974-2979, 1997.

[73] K. Hess, L.F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J.W. Lyding, I.C. Kizilyalli, Theory of channel hot-carrier degradation in MOSFETs, Physica B 272 (1-4) (December 1999) 527-31

[74] A. Haggag, W. McMahon, K. Hess, K. Cheng, J. Lee, J. Lyding, in Proceedings of the International Reliability Physics Symposium (IRPS), vol 271 (2001)

[75] A. Haggag, M. Lemanski, G. Anderson, P. Abramovitz, M. Moosa, in Proceedings of the International Reliability Physics Symposium (IRPS), vol 93 (2007)

[76] T.C. Shen, C. Wang, G.C. Abeln, J.R. Tucker, J.W. Lyding, P. Avouris, R.E. Walkup, Science 268, 1590 (1995)

[77] 97. P. Avouris, R.E.Walkup, A.R. Rossi, T.-C. Shen, G.C. Abeln, J.R. Tucker, J.W. Lyding, Chem. Phys. Lett. 257, 148 (1996)

[78] K. Hess, B. Tuttle, F. Register, D.K. Ferry, Appl. Phys. Lett. 75, 3147 (1999)

[79] C. Guerin, V. Huard, A. Bravaix, General framework about defect creation at the Si=SiO2 interface. J. Appl. Phys. 105, 114513-1-114513-12 (2009)

[80] Lee et al., Circuit aging simulator (CAS), IEEE Int. Elec. Dev. Mee. 134, (1988)

[81] Gnudi et al., Two dimensional MOSFET Simulation by Means of Multidimensional Spherical Harmonics Expansion of the Voltzmann Transport Equation, Solid-State Electronics, v 36, n 4, p 575-81, 1993

[82] J. Bude and K. Hess, Thresholds of impact ionization in semiconductors, J. Appl. Phys., vol. 72, no. 8, pp. 3554-3561, Oct. 1992.

[83] O.Penzin, et al., MOSFET degradation kinetics and its simulation, IEEE Trans. Electron Dev. 50 (2003) 1445-1450.

[84] S. Reggiani et al., TCAD Simulation of Hot-Carrier and Thermal Degradation in STI-LDMOS Transistors, IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 60, NO. 2, p.691-8, FEBRUARY 2013

[85] S. Tyaginov et al., A Predictive Physical Model for Hot-Carrier Degradation in Ultra-Scaled MOSFETs, SISPAD, 2014

[86] D.Varghese, et al., OFF-State Degradation in Drain-Extended NMOS Transistors: Interface Damage and Correlation to Dielectric Breakdown, IEEE Trans. Electron Dev.54(2007)2669-2677.

[87] I. Starkov, S. Tyaginov, H. Enichlmair, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Ceric, and T. Grasser, Hot-carrier degradation caused interface state profile-Simulations versus experiments, J. Vac. Sci. Technol., vol. 29, no. 1, pp. 01AB09-1-01AB09-8, Jan. 2011.

[88] S. Reggiani et al., TCAD degradation modeling for LDMOS transistors, ESSDERC, p 185-8, 2012

[89] S. Tyaginov, Physics-Based Modeling of Hot-Carrier Degradation ed. by T. Grasser (Springer, Wien/New York, 2015)

[90] C. Prasad, L. Jiang, D. Singh, M. Agostinelli, C. Auth, P. Bai, T. Eiles, J. Hicks, C.H. Jan, K. Mistry, S. Natarajan, et al., Self-heat reliability considerations on Intel's 22nm Tri-Gate technology, Proc. IEEE International Reliability Physics Symposium (IRPS) (April 2013) 5D.1.1-5

[91] F. Stellari, K.A. Jenkins, A.J. Weger, B. Linder, P. Song, Self-heating characterization of FinFET SOI devices using 2D time resolved emission measurements, Proc. IEEE International Reliability Physics Symposium (IRPS) (April 2015) 2B.1.1-6

[92] S. Mittl, F. Guarin, Self-heating and its implications on hot carrier reliability evaluations, Proc. IEEE International Reliability Physics Symposium (IRPS) (April 2015) 4A.4.1-6

[93] H. Jiang, Y. Longxiang, L. Yun, X. Nuo, Z. Kai, H. Yandong, D. Gang, L. Xiaoyan, Z. Xing, Comprehensive understanding of hot carrier degradation in multiple-fin SOI FinFETs, Proc. IEEE International Reliability Physics Symposium (IRPS) (April 2015) XT.6.1-4

[94] W. Arfaoui, X. Federspiel, P. Mora, M. Rafik, D. Roy, A. Bravaix, Experimental analysis of defect nature and localization under hot-carrier and bias temperature damage in advanced CMOS nodes, Proc. IEEE International Integrated Reliability Workshop (IIRW) (Oct. 2013) 78-83

[95] J. Coignus, A. Vernhet, G. Torrente, G. Reimbold, S. Renard, D. Roy, Relaxation-free Characterization of Flash Programming Dynamics along P-E Cycling, to be published in IEEE IIRW 2015 report

[96] W.Arfaoui, X. Federspiel, P. Mora, F. Monsieur, F.Cacho, D. Roy, Energy-driven Hot-Carrier model in advanced nodes, IRPS 2014

Chapter 3

Fowler-Nordheim Degradation

The degradation of a gate oxide under high-field stress is of great importance for MOS devices, as wear out and failure are limited by charge buildup on defect sites in the SiO_2 layer [1],[2]. Indeed, it is well-known that high field stress in SiO_2 layer gives rise to aging mechanisms as interface trap generation [3], [4] and charge trapping [1]. They cause the gradual drift of oxide electrical properties leading to the decay of device performance thus accelerating its failure and breakdown [5],[6].

In this reliability context, both for nonvolatile memories [6],[7] and standard CMOS applications [8], it is generally believed that the charge buildup at bulk-oxide sites is governed by pure trapping process of electrons travelling towards the anode through Fowler-Nordheim mechanism (FN) [9] and by holes travelling towards the cathode through Anode Hole Injection (AHI) mechanism [10],[11].

On the other hand, there was a long debate about the origin of the bulk-oxide traps responsible for Stress-Induced Leakage current (SILC) [12] and for the eventual breakdown [13]. However, Esseni [14], [15] and Wu [16] experimentally demonstrated that the limiting factor for both aging mechanisms is AHI, discarding the Hydrogen Release (HR) model [12] which postulated that hydrogen species released at Si/SiO_2 interface would produce the oxide damage responsible for those macroscopic aging mechanisms.

For all these reasons, the evaluation of wear out properties of gate dielectrics is a key issue for the development of ultra-scaled integration Silicon technologies. Thus, reliable techniques for the assessment of trapping kinetics [17], as well as methodologies for the extraction of drifting parameters representing the different aging mechanisms [3] and for the evaluation of oxide defect distribution position, as for example through the centroid assessment [18], [19], are often adopted.

Concerning NOR Flash technology, whose Erase operation is performed via Fowler-Nordheim (FN) mechanism, the integrity and reliability of the SiO_2-based tunnel oxide is essential for Program/Erase (P/E) operation efficiencies [2],[6]. Thus, proper cell opti-

mization requires fine characterization and modeling of the Erase FN-induced defects and of the electrical impact they have on Flash characteristics and P/E operations. Indeed, aging-induced defects are known to impact both device electrostatics (static drift) and cell operation efficiencies (performance decay). We will see that both contributions make Vth read at Erase state drifting towards higher values, till becoming undistinguishable from the Program state. For that condition, the Flash memory cell can be considered failed.

Although FN aging mechanism has been widely studied in the past [1]-[23], a clear separation of these two contributions, their microscopic analysis and the respective impact on the overall cell performance are still lacking.

In this chapter and in our work published in *Transaction on Device and Materials Reliability* [24], a deep investigation of Fowler-Nordheim Stress (FNS) is done on devices belonging to 40nm NOR Flash technology. A complete set of experiments is presented, which aims to capture the signatures of the different Erase-induced defects and quantitatively address the overall SiO_2 wear out. Therefore, for the first time, the effect of charged defects within the tunnel oxide is studied from both Si/SiO_2 and $Poly/SiO_2$ interfaces with delay-free experiments considering different test structures, which allow to address both electrostatic aging and Erase efficiency decrease.

On the other hand, we will limit our analysis on the evaluation of oxide wear out and on the modeling of the respective aging mechanisms, without linking them with SILC and breakdown processes. Indeed, this chapter is focused on the effect of FN-induced defects on Flash characteristics just after the FN stress itself, without accounting for relaxation and data retention effects and without pushing the endurance till the breakdown to detect the Time-Dependent gate Dielectric Breakdown (TDDB), because out of our purpose.

Since FN degradation cannot be studied directly on Flash cell independently from Program induced aging, equivalent Flash transistors have to be considered to properly characterize and model the stress induced just by the Erase operation. In particular, the equivalent Erase-only stress will be applied on these structures.

The chapter is divided in two parts:

- In the first section, the role of each FN-induced aging contribution on the respective Flash cell and the related impacts on PW evolution are highlighted.
- In the second part, FN-induced electrostatic aging and FN current decay are separately studied with specific transistor structures. During the equivalent Erase-only stress, accurate physical-based parameters, representing the different defects, are properly extracted in order to quantify the Erase-related degradation mechanisms in simple kinetic laws which can be applied on Flash cell modeling during P/E cycling.

3.1 FN aging contributions in Flash endurance

Three different test structures have been considered in this chapter: Flash cell, equivalent Flash transistor, built by shorting the Control Gate (CG) and the Floating Gate (FG) of the cell, and CAST structure, which is a matrix of $5 \cdot 10^5$ equivalent transistors connected in parallel. This last device has been considered in order to directly measure the FN current and it has been preferred respect to a single device having $W = 5 \cdot 10^5 \cdot W_{single}$ since preserving the 3D border effects of the single transistor.

In order to validate the consistency of the test structures, tunnel oxide transport in FN regime has been measured and compared. In Fig.3.1 IfgVfg characteristic is shown for fresh devices, i.e. not degraded case. This curve has been acquired both directly measuring the gate current on CAST structure and calculating each Ifg and Vfg level from Vth(t) evolution during the Step Pulse Erase (SPE) mode for the Flash cell, as similarly done for Prog case in [25],[26] and in the previous chapters. It is worth noting the good accordance between the results, which well fits with the standard 1D-FN model (see extracted FN parameter in Fig.3.1).

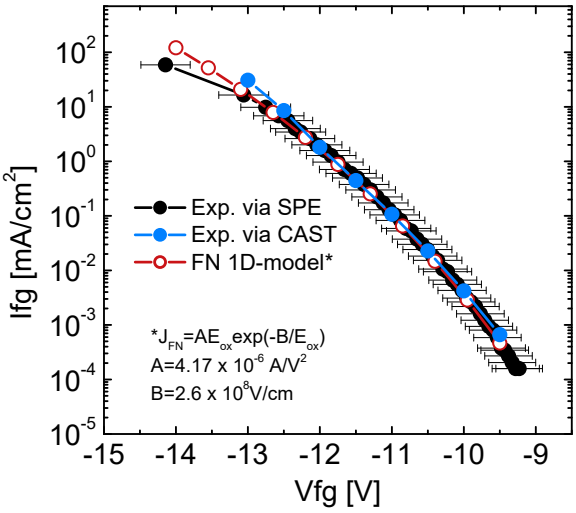


Figure 3.1: Fresh IfgVfg characteristic in Fowler-Nordheim mode. The current directly measured on the CAST structure, the IfgVfg indirectly acquired via SPE (average of 6 samples) and the 1D FN-simulation are shown.

In order to physically address the role of Fowler-Nordheim aging and contextualize it within Flash endurance framework, we take advantage of the methodology developed in the first chapter of this thesis, which insures that Flash tunnel oxide stress during Erase operation is perfectly reproduced on equivalent transistors [27]. In particular, ramp-optimized patterns are considered on the CG electrode, which lead to a constant FG potential (Vfg) during Flash P/E operations. The electrical stress suffered by tunnel oxide is reproduced on equivalent

transistors applying the respective square patterns, i.e. Vg_{stress} = Vfg level of the Flash. In Fig.3.2(a), the drift of Id_{ON}, i.e. extracted drain current at Vfg = Vth_{mos} + 4V[1], known to be mainly sensitive to channel mobility degradation [28], is compared for the two devices which are supposed to be subjected to the same wear out. The results clearly indicate that the stress suffered by them is identical. Further details and results of this methodology will be shown in the next chapter.

In this experimental context, we will see in the next chapter that, using simple conventional Flash equations [29], the drift of cell Vth at Erase level, i.e. Vth_E calculated at Constant drain Current (CC) during the Read phase, reduces to:

$$\Delta Vth_E = \frac{1}{\alpha_G} (\Delta Vth_{mos} + \Delta Vfg_\infty) \tag{3.1}$$

where ΔVth_{mos} is the Vfg shift in read mode (Vd=0.5V) in order to provide a certain fixed low drain current level Idth, whereas ΔVfg_∞ is the Vfg shift, in absolute value, during Erase phase (Vd=Vs=0V) in order to provide a fixed FN tunneling gate current equal to $b \cdot C_{ono}$, where b is the ramp speed in V/s and C_{ono} is the capacitance between FG and CG electrodes. The first term represents the aging of the electrostatic, i. e. of the transistor "inside" the cell, while the second one represents the degradation of the tunnel oxide transport because linked to the gate current decay.

Since ΔVth_{mos} is directly measured during equivalent stress on Flash transistors, it is possible to separate the electrostatic aging from the loss of Erase efficiency at Vth_E level. Looking at Fig.3.2(b), the impact of this second contribution on Vth_E drift is quite significant, indeed it is known to play an important role on the Program Window (PW) closure [2], since directly linked to electron trapping process. In the following, these two degradation contributions are studied separately with the help of equivalent transistors in order to focus only on the aging induced by FN mechanism.

The sequent experimental strategy is adopted:

- By stressing the single equivalent Flash transistor and reading its drifts in linear regime, it is possible to address Si/SiO_2 interface and border traps, which directly impact the electrostatics of the Flash cell during P/E endurance, i.e. ΔVth_{mos} in eq.3.1. For this reason, AC FNS has been performed on single MOSFET alternating Flash-equivalent Erase pulses and read operations, i.e. acquisitions of the linear characteristic (Vd=50mV).

[1] as defined in the first chapter, Vth_{mos} is FG potential which induces the read condition, i.e. Id = 8μA at Vd=0.5V. Thus, it can be read directly on MOSFET structures as long as such a device experiences exactly the same degradation of the respective Flash cell. On the other hand, Id_{ON} can be measured on the Flash cell at Vcg = Vth + 4V/α_G, where Vth is the threshold voltage of the cell itself, in accordance with the terminology used so far.

- On the other hand, by stressing the CAST structure and reading the gate current evolution, it is possible to capture the effects of defects close to Poly/SiO_2 interface, which strongly impact the current and thus the Erase efficiency operation of the memory cell, represented by ΔVfg_∞ in eq.3.1. For this reason, Constant Current Stress (CCS) and Constant Voltage Stress (CVS) have been carried out on the CAST structure.

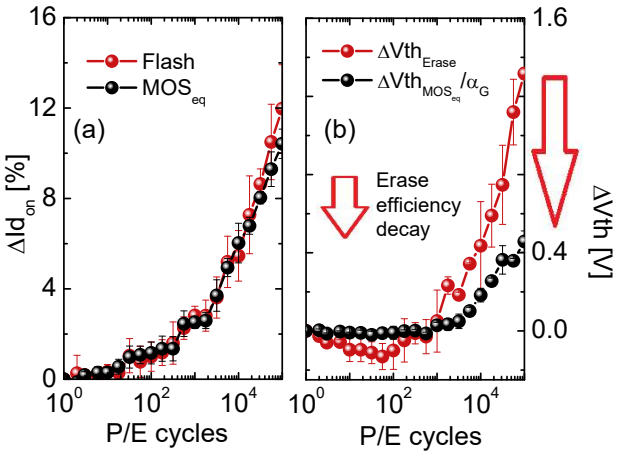


Figure 3.2: In (a) the comparison of ΔId_{ON} (i.e. mobility aging [28]) between Flash cell and equivalent transistor, stressed respectively with ramp and square patterns is shown along P/E cycling (see the next chapter for details). In (b) the drift of Vth at Erase state is shown together with the electrostatic aging of the equivalent transistor divided by the coupling coefficient. The loss of Erase efficiency at Vth_E level is highlighted. Error bars refer to the standard deviations due to statistical dispersion (6 samples considered). All the data here refer to standard read operation (Vd=0.5V, Idth = 8μA), considering Vfg=-12V/T_{pulse} = 3.2ms for the Erase pulse.

Although CCS and CVS have been performed using standard SMUs belonging to Keysight B1500, measurement setup controlling the delays between the different operations has to be considered for AC stress, since trapped charges in SiO_2 may relax during the sweep between the stress and the read phase. For this reason, the measurement setup described in the first chapter of this thesis and in [30] has been considered. In addition, it has been verified that AC stress gives exactly the same results, in terms of drifting parameters, as DC case (not shown here), thus it does not suffer from any extra-relaxation effect.

Since we will uniquely deal with equivalent transistors, from now on all electrical parameters present in this chapter will just refer to MOSFET structures[2] (unless specified) for simplification purpose in terminology. In particular, Vth refers to the threshold voltage

[2]as for example Id, Vg, but also the extracted Vth, Id_{ON} and so on

calculated at CC in read mode ($Vd_{read} = 50mV$ and $Id = 1\mu A$, which corresponds to the "real" threshold condition at fresh state).

3.2 FN aging understanding and modeling

Concerning the device aging, a sketch of the band diagram during FN regime is shown in Fig.3.3. As depicted, the electron flow from the gate gives rise to charge trapping within the tunnel oxide stack in pre-existing defect sites [22]. The injected electrons in Si-substrate have enough energy to break the strong covalent bonds via Impact Ionization mechanism (II) and thus create hot holes which are then re-injected back towards the PolySilicon: this phenomenon is called Anode Hole Injection (AHI) [10],[11]. Similarly, the injected holes have a certain probability to be trapped in the tunnel oxide stack, which acts as a competitor, from an electrostatic standpoint, respect to negative oxide charges.

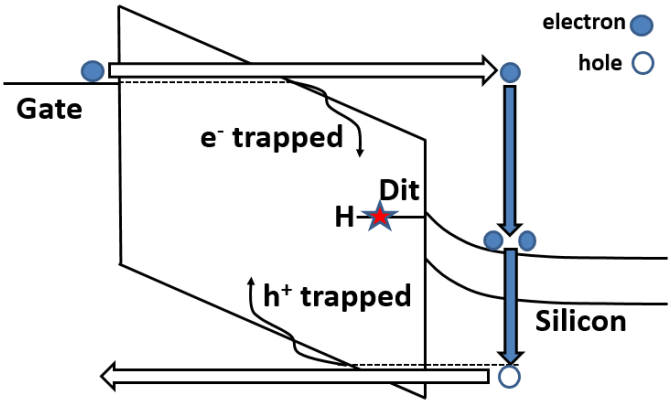


Figure 3.3: Energy band diagram for 1D-MOS structure under FNS. Three types of defects are created: trapped electrons via FN, trapped holes via AHI and Si/SiO_2 interface amphoteric states (Dit), which are created breaking the Si-H bonds.

Whenever FN regime is set, Si/SiO_2 damage process is always observed [3], which gives rise to amphoteric interface traps Dit [31]. However, this phenomenon has been rarely studied in the past and its physical origin is still doubtful. Shiue et al. [4] claimed that this aging process is driven by hot holes, thus by AHI mechanism. However, in this chapter we will link the interface state generation with the injected electrons in order to more easily transfer the model on the respective Flash cell endurance.

In the following, the signatures of all these defects are separately measured and their impacts on electrostatic and Erase efficiency degradations of the respective Flash memory cell are analyzed and quantified in kinetic laws. It is worth noting that we will implicitly consider the degradation to be uniform along the channel. Indeed, as highlighted in the

first chapter, very high Vg values (in amplitude) make the Si/SiO_2 interface strictly uniform from an electrostatic standpoint, which leads the FN current, and the induced aging, to be constant along the channel. This allows to approximate the FN as a 1D mechanism.

By measuring and modeling the degradation induced on Flash equivalent transistors, as done in this chapter for Fowler-Nordheim case, it is possible to study the aging evolution of the memory cell during P/E cycling and, in particular, to properly separate the wear out mechanisms within the PW drift. Indeed, in the last chapter, using the appropriate extracted parameters, their kinetic laws and their impact on the aging contributions, we will deal with the application of this framework on the respective memory cells and a complete physical-based model for Flash endurance will be presented.

3.2.1 Electrostatic degradation

AC FNS on single equivalent transistor has been performed for different Vg levels. In Fig.3.4 the degradation of the IdVg trans-characteristic in read mode (Vd=50mV) is shown at different stress time for the condition $Vg_{stress} = -12V$ (Vd=Vs=0V). Looking at the curve in linear scale in Fig.3.4(a), the effect of interface states is evident: the mobility in strong inversion degrades due to Coulomb scattering with charged amphoteric defects. On the other hand, looking at Fig.3.4(b) at low current levels, the electrostatic shift due to the damage in oxide bulk is negative, which is a clear evidence of hole trapping process, occurring through AHI mechanism.

In order to quantify the impact of these defects, physical drifting parameters have been extracted from IdVg curves. In Fig.3.5 the related electrostatic effects are shown: the threshold voltage drift ΔVth is due to both trapped oxide charges ΔVth_{ox} and to acceptor interface states ΔVth_{Dit}.

Concerning the electrostatic effect of trapped charges in oxide bulk, the related ΔVth_{ox} has been extracted at Mid-Gap (MG) current level [32], which has been calculated for the fresh device fitting the whole IdVg curve, as done in [33] and in the second chapter of the thesis. Finally, ΔVth_{Dit} has been simply computed from the difference $\Delta Vth - \Delta Vth_{ox}$.

It is worth noting that the electrostatic effects of Dit and Qox^+ are similar in absolute value, which induce a total ΔVth shift close to zero. Indeed, we will see in the next chapter that the overall threshold voltage drift during the respective endurance on the Flash memory cell is limited by the Program operation, thus by Hot Carrier aging mechanism.

However, although being qualitatively clear, the results in Fig.3.5 come with high standard deviation values, thus making hard a fine quantitative analysis. For this reason, a different approach is considered. Indeed, the degradations of the sub-threshold slope (SS) and ΔVth_{Dit} are strictly linked by the well-known relationships [3]:

$$Dit \approx \frac{Cox}{2.3kT} \cdot \Delta SS \tag{3.2}$$

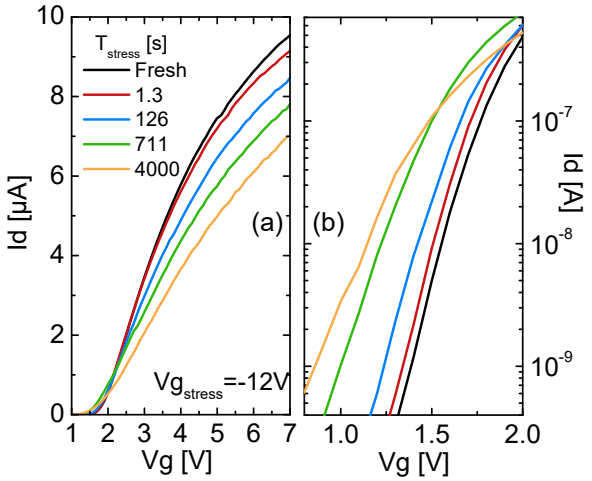


Figure 3.4: AC stress on single transistor. The evolution of the entire IdVg curve in linear regime (Vd=50mV) is shown in Lin (a) and in Log (b) scales.

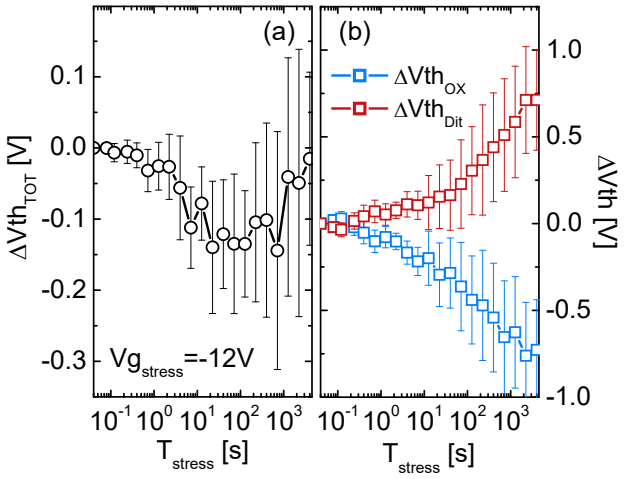


Figure 3.5: Degradation of extracted parameters from IdVg. ΔVth is shown in (a), whereas its parts due to oxide charges and interface traps are plotted in (b). ΔVth_{ox} has been extracted at MG level [32],[33] and ΔVth_{Dit} from the difference $\Delta Vth - \Delta Vth_{ox}$. Error bars refer to standard deviations due to statistical dispersion (6 samples considered).

$$\Delta Vth_{Dit} \approx \frac{q \cdot Dit}{Cox} \cdot \phi_B \cdot K \tag{3.3}$$

where K is a fitting parameter accounting for the energy distribution shape within the Silicon band-gap of the active interface traps: if they are uniformly distributed between intrinsic Fermi level and the conduction band, K would be equal to 1. In Fig.3.6(a) the comparison between ΔVth_{Dit} calculated with eqs.3.2,3.3 and $\Delta Vth - \Delta Vth_{ox}$ previously computed is shown. A good agreement has been found with K=0.6 and it is worth noting that the extraction noise has been strongly reduced calculating this parameter from SS decay.

The problem can be even more simplified, in terms of extraction noise, considering the parameter Id_{ON}, whose decay has been demonstrated to be only sensitive to Dit amount. Differently from what shown for Hot Carrier case in the previous chapter and in [28], the drift of Id_{ON} is strongly correlated with the electrostatic effect of aging-induced interface traps, since the channel degradation is surely uniform. Thus, a certain amount of degradation at Si/SiO_2 interface, due to the created Dit, has a unique impact on ΔVth_{Dit} (or ΔSS) and on ΔId_{ON}. This is experimentally verified looking at the correlation of these two parameters in Fig.3.6(b) considering different stress conditions: the drifts of SS^{-1} and Id_{ON} are aligned by a linear relationship. Thus, this parameter, which does not suffer any extraction noise, can be used to directly calculate ΔVth_{Dit}; then, ΔVth_{ox} can be extracted from the difference $\Delta Vth - \Delta Vth_{Dit}$. Concerning the Flash cell endurance, these relationships are very important, since ΔId_{ON}, which is easily measurable, is observed to be FN-limited.

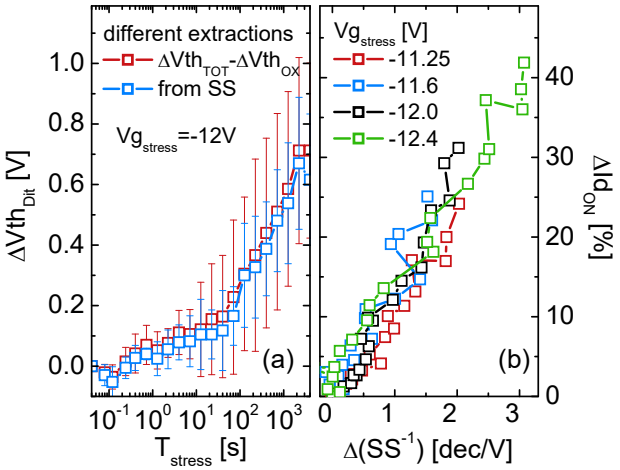


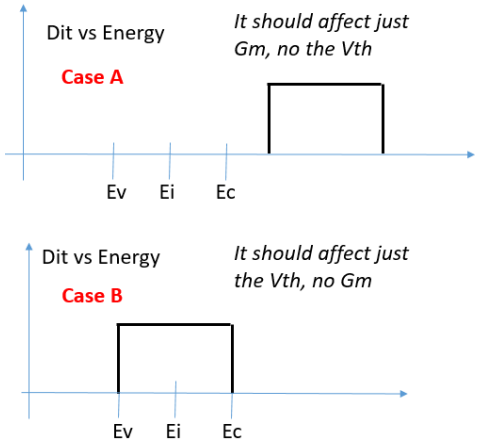
Figure 3.6: In (a) the comparison between the extracted ΔVth_{Dit} as in Fig.3.5(b) and the direct extraction from SS eqs.3.2,3.3 is shown. In (b) the correlation between Id_{ON} and SS^{-1} degradations is highlighted. Error bars refer to standard deviations due to statistical dispersion (6 samples considered).

Till now, we have dealt exclusively with traps located within the Silicon bandgap, as most of the works did in the past [3],[4]. Indeed, defects above the conduction band are rarely considered since being usually uncharged in linear regime and their effect is negligible. However, our devices are heavily doped in the Silicon substrate ($N_A \approx 10^{18} cm^{-3}$) and the Fermi level easily gets inside the conduction band charging those traps.

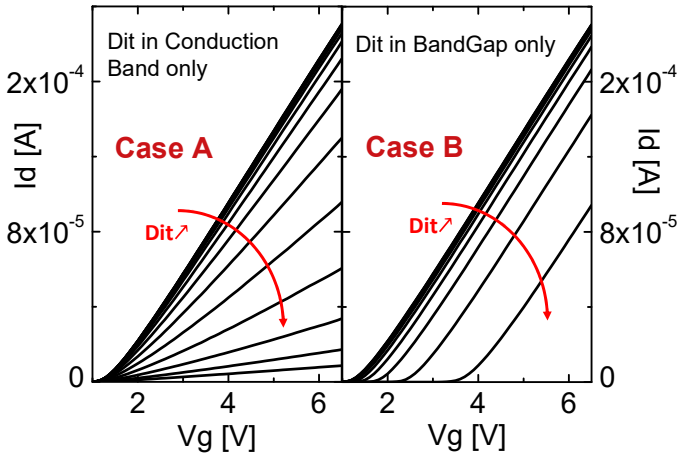
The effect of defects over the conduction band is not easy to capture since the linear trans-characteristic over the threshold condition suffers from both mobility and electrostatic degradation. For this reason, the saturated trans-characteristic has been considered. The current in saturation regime is strictly proportional to $Qinv(Vg) \cdot \mu_{sat}$, thus it does not suffer

from any transport degradation since the mobility is always limited by phonon scattering [9]. During the Vg-sweep above threshold, i.e. $E_F > E_C$, the interface traps are more and more charged thus giving a cumulative electrostatic drift which degrades the trans-conductance Gm. In Fig.3.7(a) the effect of interface defects depending on the energy position is schematically shown: if they are located above E_C, the Vth drift is negligible and the Gm degradation is observed, whereas the impact of traps within the bandgap is exactly the opposite. Simulating the IdVg trans-characteristics in saturation regime with Synopsys tools [34], these effects have been verified and are shown in Fig.3.7(b). It is worth noting that this analysis only accounts for traps close to source contact since the read is performed at high Vd values. However, as previously discussed, the FN-aged devices experience a constant defect concentration along the channel, thus this analysis addresses the overall wear out.

It can be of interest to correlate ΔVth with ΔId_{ON}, which represents the degradation of IdVg slope in inversion, during stress time, since giving insights on the energy position of the created active interface defects (as similarly done in the previous chapter for the spatial localization of the trap distribution after Hot Carrier stress). Then, knowing the barycenter in energy, the trap concentration can be extracted and the total trap amount can be accurately estimated. However, this is not done in the thesis since not being of our interest.



(a) Schematic representation of the effect on Vth and Gm of interface traps located over the conduction band (Case A) and within the Band-Gap (Case B).



(b) TCAD simulations of the saturated IdVg curve for the Case A and B represented in (a) varying the trap concentration. Concerning the Case A, the trap distributions have been considered from $E_{min} = E_C + 0.2eV$

Figure 3.7: Simulation of IdVg trans-characteristics in saturation region for different Dit energy-distributions.

In Fig.3.8 the related experimental results are shown for the condition $Vg_{stress} = -12V$. In the figure, the drift of IdVg slope, that has been quantified in terms of ΔId_{ON}, is evi-

dent, which is a clear signature of interface traps above the conduction band, as previously explained. This is in contrast with Hot Carrier Degradation results seen in the previous chapter. In that case, the slope of the trans-characteristic does not degrade in saturation region, whereas the Vth drifts, which is signature of defects within the Silicon bandgap. This means that HCD and FNS create traps at different energy location, that probably comes from the difference in the Si-H breakage mechanism.

Looking at low current levels in Fig.3.8(a), the effect of hole trapping process can be observed: it negatively shifts the threshold voltage. However, after a certain stress time, the kinetics of Dit generation becomes important moving back the threshold voltage towards higher values. The competition of these two different aging mechanisms can be also observed looking at the linear results in Fig.3.5(a).

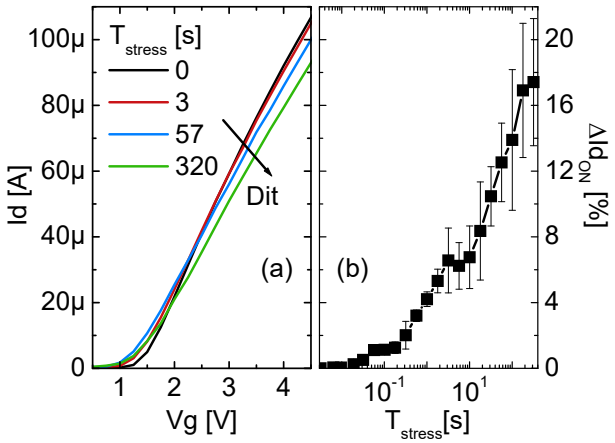


Figure 3.8: Experimental data concerning the evolution of the saturated IdVg characteristics: in (a) the curve at different stress time; in (b) the drift of the extracted Id_{ON}. Stress condition considered is $Vg_{stress} = -12V$.

3.2.2 Tunnel oxide transport aging

In order to study the degradation of the Erase efficiency of the memory cell, the aging of FN current is explored applying CCS and CVS on the CAST structure.

In Fig.3.9(a) experimental CCS results are shown at different gate current levels. An increase of the gate voltage during stress time in order to maintain the same current level is observed (denoted ΔVg_{FN} in the following), which is a clear signature of trapped electrons during FN injection.

In literature, a power law correlation between injected and trapped electrons during FNS, i.e. Qinj and Qox respectively, has been highlighted [17]. This relation has been

physically explained with the superposition of exponential laws, which come from the first-order trapping kinetics [35]:

$$\frac{dQox}{dt} = \sigma \cdot J_{FN} \cdot (qN_{tot} - Qox) \tag{3.4}$$

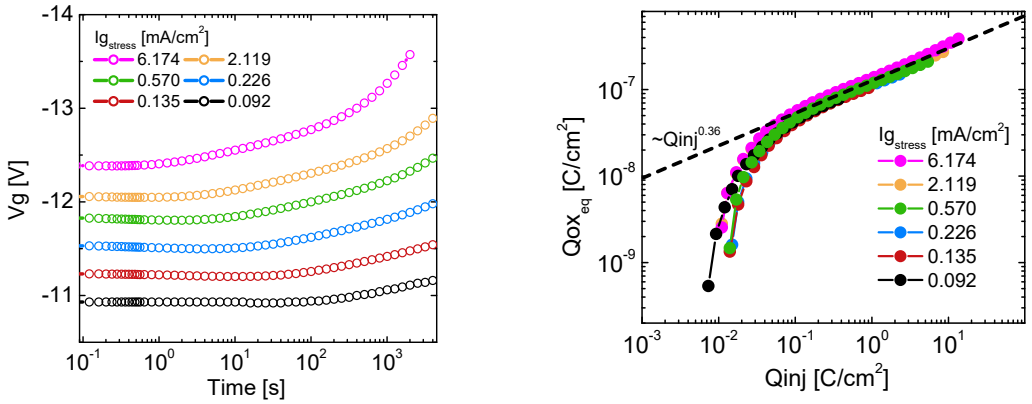
where σ is the capture cross-section and N_{tot} is the total oxide defect concentration. Knowing that $Qinj(t) = \int_0^t J_{FN}(t')dt'$, an analytical solution can be easily found for trapped electrons:

$$Qox(t) = qN_{tot} \cdot [1 - \exp(\sigma \cdot Qinj(t))] \tag{3.5}$$

This relation turns into a power law considering N_{tot} and σ exponential functions of the oxide position, as suggested in [17]. Hence, the Vg shift measured at the gate electrode can be linked to the oxide damage $Qox_{eq} = \Delta Vg_{FN} \cdot Cox$, which is equal to:

$$Qox_{eq} = \int_0^{Tox} Qox'_{eq}(x)dx = \int_0^{Tox} \left(1 - \frac{x}{Tox}\right) \cdot qN'_{tot}(x) \cdot [1 - \exp(\sigma(x) \cdot Qinj)]\, dx \approx A \cdot Qinj^{\nu} \tag{3.6}$$

where x is the distance from Poly/SiO_2 interface and ν is an oxide quality empirical parameter. An example of such a simulation is shown in Fig.3.10.



(a) Evolution of the gate voltage directly measured.

(b) Extracted negative Qox_{eq} plotted function of injected charges.

Figure 3.9: Experimental results of Constant Current Stress experiment applied on CAST structure at different Vg.

This power law correlation has been experimentally shown on positive FNS [17], i.e. Fowler-Nordheim mechanism occurring under $Vg > 0V$ condition. However, looking at the results in Fig.3.9(b), it is also verified in our case. The small mismatch from the power law at the beginning of the stress is likely due to higher rate for trapped holes via AHI, as already observed by Papadas in [1].

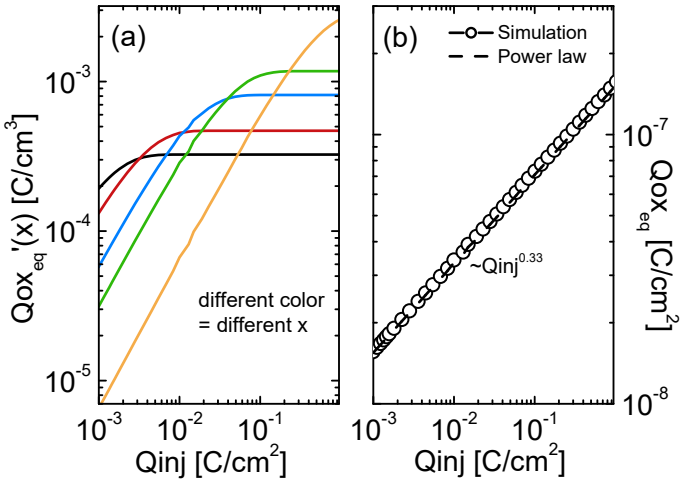


Figure 3.10: Simulation of oxide trapping kinetics. The exponential dynamics at different oxide positions (a) and the resulting overall power law kinetics (b) are shown.

The simple dependence of FN-induced aging with injected charges gives a very important guideline for Flash endurance understanding. Indeed, the Erase efficiency reduction of the memory cell can be directly linked with the injected charge, which is simply calculated as:

$$\mathrm{Qinj(i)} = \mathrm{C_{ono}} \cdot \sum_{\mathrm{k=0}}^{\mathrm{k=i}} \mathrm{PW_k} \tag{3.7}$$

at cycle=i, where $\mathrm{C_{ono}}$ is the capacitance between FG and CG electrodes and PW is equal to $\mathrm{Vth_P} - \mathrm{Vth_E}$.

Using the simple relationships in eqs.3.6,3.7, the estimation of $\Delta \mathrm{Vfg}_\infty$ in eq.3.1 can be addressed for the cell endurance. However, it has to be pointed out that the Program-injected charges and the effect of Program operation on Erase degradation kinetics are not taken into account in this estimation. In the next chapter, we will explore these issues.

Coming back to the extracted parameters, it is worth noting the strong contraposition between the results in Fig.3.9(a) with the ones previously seen in Figs.3.4,3.5. Indeed, stressing the CAST structure, the effect of oxide charges suggests a majority of trapped electrons in the tunnel oxide stack, whereas looking at $\Delta \mathrm{Vth_{ox}}$ of the equivalent transistor, the signature is clearly opposite. This difference is explained considering the position of the charge distribution $\rho(\mathrm{x})$ within the oxide:

$$\Delta \mathrm{Vg_{FN}} = -\int_0^{\mathrm{Tox}} \left(1 - \frac{\mathrm{x}}{\mathrm{Tox}}\right) \cdot \frac{\rho(\mathrm{x})}{\mathrm{Cox}} \mathrm{dx} \tag{3.8}$$

$$\Delta \mathrm{Vth_{ox}} = -\int_0^{\mathrm{Tox}} \frac{\mathrm{x}}{\mathrm{Tox}} \cdot \frac{\rho(\mathrm{x})}{\mathrm{Cox}} \mathrm{dx} \tag{3.9}$$

where x is the distance from Poly/SiO_2 interface. Thus, it is demonstrated that the electrons are mainly trapped close to PolySilicon, whereas the holes, injected from the Si-substrate via AHI mechanism, are trapped close to the bottom Si/SiO_2 interface, since having very low mobility in the oxide, as remarked by Park in [3]. Hence, from now on, ΔVth_{ox} is assumed to represent uniquely the positive charges in the oxide, i.e. trapped electrons have a negligible impact on Vth.

3.2.3 Insights on degradation kinetics

As similarly done for the extracted electron trapping, it is of prime interest to calculate the kinetics of Dit creation and hole trapping process. This can be achieved establishing the link between the drifting electrostatic parameters, extracted during AC FNS, and the gate current evolution during the respective stress at constant voltage, which gives information on injected electrons and holes. For this reason, CVS on CAST structure has been considered too.

Experimental results concerning Ig-decay during CVS are shown in Fig.3.11 for different stress conditions. Looking at the gate current evolution in linear scale (Fig.3.11(b)), the signature of excess holes at the beginning of stress is evident: an increase of FN-current is observed. These extra holes for short stress time explain the over-Erase at the beginning of the endurance on the memory cell, i.e. $\Delta Vth_E < 0$ in the first cycles in Fig.3.2(b). However, concerning the trapped electrons, the equivalent Qox seen at Poly/SiO_2 interface is not directly measured, since the gate voltage is fixed. Thus, using the relationship

$$J_{FN} = A \cdot \left(E_{ox} - \frac{Qox_{eq}}{\epsilon_{ox}} \right)^2 \cdot \exp\left(- \frac{B}{E_{ox} - \frac{Qox_{eq}}{\epsilon_{ox}}} \right) \tag{3.10}$$

it is possible to extract Qox_{eq} from the current directly measured. A and B constants used are the same as in Fig.3.1.

The results regarding the correlation between Qox_{eq} and Qinj are shown in Fig.3.12(a) and it is worth noting that a good prediction of the trajectory has been achieved as compared to the one directly measured in CCS case, which highlights the accuracy of the parameters used.

Concerning the interface damage at Si/SiO_2 interface, the related Dit aging kinetics can be easily addressed. In Fig.3.12(b) the correlation between the degradation of Id_{ON}, measured during AC stress, and the injected charges, measured during CVS, is shown. Since the first parameter represents the quantity of interface states, the relationship represents the aging kinetics of the Si-H bond depassivation process. However, in this relation, the physical descriptions of the energy distribution for the electrons and the Si-H bond activation energy

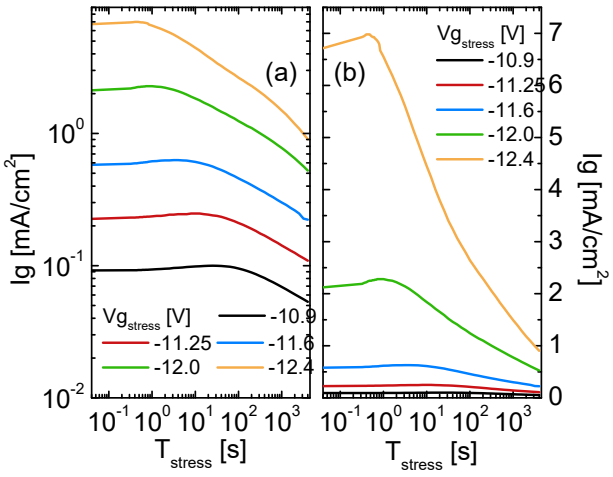


Figure 3.11: Ig evolution with time directly measured on CAST during CVS in Log (a) and in Lin (b) scales.

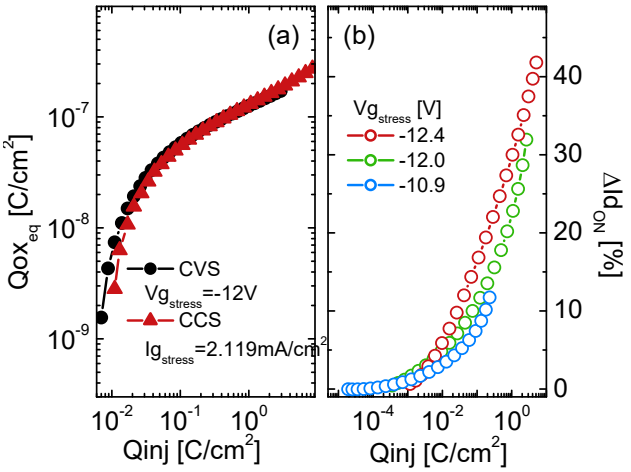


Figure 3.12: (a) Qinj vs Qox, i.e. injected vs trapped electrons, calculated during CVS experiment compared with the respective CCS results (of Fig.3.9(b)). In (b) the degradation of Id_{ON} is plotted function of the corresponding injected electrons calculated during CVS experiment on CAST structure.

dispersion [36] are clearly missing. Indeed, for higher Vg_{stress} in amplitude the electrons are more energetic when reaching the Si/SiO_2 interface.

As previously said, the physical origin of the Si-H bond depassivation process is not clear nowadays. Indeed, the bond dissociation mechanism can be linked to both carriers: considering the electrons, they have higher concentration, whereas the hole breakage process experiences a lower Si-H capture cross section. However, in both cases, the increase of the gate voltage during FN regime leads to higher interface aging. Translating it into macroscopic parameters, this means that Id_{ON} degradation should be higher at fixed Qinj,

as experimentally verified looking at the results in Fig.3.12(b). Nevertheless, the dependence on Vg_{stress} level is not so strong as expected. This can be explained considering that the electrons may partially lose the initial potential energy dependence, since having travelled from the cathode towards the Silicon substrate for almost 10nm of oxide thickness. Taking advantage of this experimental observation, a unique law describing the Dit aging kinetics can be assumed for reliability engineering purpose. Considering this simplification during the respective Flash cell endurance, the error done is surely negligible since equivalent Vfg=const patterns are considered.

This simple law has the big advantage to be in opposition with the Hot Carrier Dit kinetics seen in the previous chapter. This difference may help us to distinguish the physical origin of some degradation mechanisms whenever the two P/E operations are considered together in Flash endurance context.

Concerning the kinetics of hole trapping process, the link between ΔVth_{ox} of the transistor, which represents the equivalent trapped holes seen from the Si/SiO_2 interface, with the amount of injected holes via AHI, calculable from Ig-decay during CVS, has to be addressed. Indeed, the direct measurement of AHI current via CCS or CVS is not possible, since it is part of the gate current which is dominated by tunneling electrons. For this reason, Schuegraf's model has been used [11] for the calculation of the AHI current. The quantum efficiency considered is:

$$\frac{J_{AHI}}{J_{FN}} = \alpha_p \cdot \exp\left(-\frac{\widehat{B}}{E_{ox}}\left[\phi_p(V_{ox})\right]^{3/2}\right) \tag{3.11}$$

where J_{FN} is the gate current directly measured during CVS on CAST, $\widehat{B} = \frac{8\pi\sqrt{2m_{p,ox}}}{3hq}$ and $q\phi_p = Eg_{SiO_2} - q\phi_b - E_{gain}$. The energy gained from the oxide field by a single electron before arriving at the anode is [37]:

$$E_{gain} = q\phi_b + qE_{ox}\lambda \cdot \left[1 - \exp\left(-\frac{1}{\lambda}\left(X_{ox} - \frac{\phi_b}{E_{ox}}\right)\right)\right] \tag{3.12}$$

where λ is the electron mean-free path in the oxide conduction band. All parameter values used for the current calculation have been taken from [11].

For the computation of the AHI-current degradation, $Q_{TOT,eq} = \Delta Vth \cdot Cox$ has been considered in the model, since the hole flow starts at the anode, thus all the defects seen at Si/SiO_2 interface affect the AHI current. Fig.3.13 shows the extracted results concerning AHI parameters, i.e. the correlation between $Qinj_{AHI}(t) = \int_0^t J_{AHI}(t')dt'$ and $Qox_{eq}(t) = \Delta Vth_{ox}(t) \cdot Cox$. As similarly observed in Fig.3.9(b), a universal power law relationship is found to govern the correlation between injected and trapped holes, which highlights the trapping nature of the process. It is worth noting that the power exponent for holes is

lower than for electrons. This qualitatively explains the slight increase of FN-current at the beginning of the stress and thus the over-Erase of the memory cell at the beginning of the endurance in Fig.3.2(b), although the total amount of injected electrons is evidently higher.

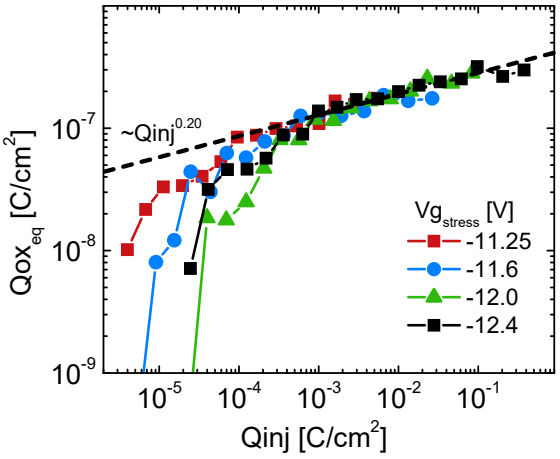


Figure 3.13: Extracted positive Qox_{eq} during AC stress plotted function of injected charges via AHI. Qinj, i.e. amount of injected holes, has been calculated from the gate current, measured via CVS on the CAST structure, using Schuegraf's model [11].

3.3 Conclusions

The chapter explores the Fowler-Nordheim Stress on equivalent MOSFET devices belonging to 40nm NOR Flash technology.

After having shown how to properly apply realistic Flash cell stress on equivalent transistors and thus isolate Erase-induced degradation, the different wear out contributions have been studied separately. Concerning the aging of Erase efficiency, the gate current evolution has been measured via CCS and CVS on big arrays capturing the impact of trapped charges close to Poly/SiO_2 interface. On the other hand, with the aim of addressing the electrostatic drift of the memory cell, thus capturing the effect of defects close to the bottom oxide interface, AC-stress alternated with Id-Vg measurements has been performed on single device.

The respective signatures of trapped holes and electrons and amphoteric interface states have been shown. Concerning the induced interface traps, it has been pointed out how such defects are distributed also over the conduction band, which makes the extraction more difficult. On the other hand, it has been highlighted how the electrons are mainly trapped close to Poly/SiO_2 interface, whereas the holes, injected via AHI mechanism, are

mostly located close to Si/SiO_2 interface. For the first time, the power law correlation between trapped and injected charges has been highlighted for both types of carriers under negative FNS. A lower power exponent has been found for the hole trapping kinetics, which qualitatively explains the slight increase of FN current at the beginning of the stress and thus the over-Erase of the memory cell at the beginning of the endurance. In addition, a fine study on the extraction of drifting electrostatic parameters has been done on experimental results in order to quantify the oxide wear out and reduce the extraction noise. The strong correlation between Id_{ON}, SS and Vth_{Dit} drifts has been highlighted, in agreement with uniform degradation along the channel interface. Finally, the interface trap aging kinetics has been experimentally addressed, emphasizing the weak electron-energy dependence.

Knowing the degradation kinetic laws for each defect generation during FN process in equivalent transistors, it is possible to transfer such a knowledge on Flash cell aging evolution during P/E cycling. In particular, all considerations and physical insights provided in this chapter will be used for NOR Flash cell endurance understanding and modeling in the last part of the thesis.

Bibliography

[1] C. Papadas, G. Ghibaudo, G. Pananakakis, C. Riva, P. Ghezzi, Model for programming window degradation in FLOTOX EEPROM cells, IEEE Electron Device Letters, v 13, n 2, p 89-91, Feb. 1992.

[2] A. Chimenton, P. Olivo, Reliability of erasing operation in NOR-Flash memories, Microelectronics Reliability, v 45, n 7-8, p 1094-108, 2005.

[3] Young-Bog Park, D.K. Schroder, Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM, IEEE Transactions on Electron Devices, v 45, n 6, p 1361-8, June 1998.

[4] Jao-Hsian Shiue, Joseph Ya-min Lee, and Tien-Sheng Chao, A Study of Interface Trap Generation by Fowler-Nordheim and Substrate-Hot-Carrier Stresses for 4-nm Thick Gate Oxides, IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 46, NO. 8, AUGUST 1999 1705

[5] Hong Yang, Hyunjae Kim, Sung-il Park et al., Reliability issues and models of sub-90nm NAND flash memory cells, 2006 8th International Conference on Solid-State and Integrated Circuit Technology, 3 pp., 2006.

[6] A. Aritome, R. Shirota, G. Hemink, T. Endoh and F. Masuoka, Reliability issues of flash memory cells, in Proc. IEEE, v 81, n 5, p 776-88, May 1993.

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

[8] J. Stathis and D. J. DiMaria, Reliability projection for ultrathin oxides at low voltage, in IEDM Tech. Dig., 1998, pp. 167-170.

[9] Y. Taur, T. H. Ning. Fundamental of Modern VLSI Devices. 2nd Edition, June 2013

[10] D.J. DiMaria, J.H. Stathis, Anode hole injection, defect generation, and breakdown in ultrathin silicon dioxide films, Journal of Applied Physics, v 89, n 9, p 5015-24, May 2001.

[11] K.F. Schuegraf, C. Hu, Hole injection oxide breakdown model for very low voltage lifetime extrapolation, 31st Annual Proceedings. Reliability Physics 1993, p 7-12, 1993.

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

[13] M. A. Alam, B. E. Weir, J. D. Bude, P. J. Silverman, and D. Monroe, Explanation of soft and hard breakdown and its consequences for area scaling, in IEDM Tech. Dig., 1999, pp. 449-452.

[14] D. Esseni, J. D. Bude, and L. Selmi, On interface and oxide degradation in VLSI MOSFETs-Part I: Deuterium effect in CHE stress regime, IEEE Trans. Electron Devices, vol. 49, pp. 247-253, Feb. 2002.

[15] D. Esseni, J. D. Bude, and L. Selmi, On interface and oxide degradation in VLSI MOSFETs-Part II: Deuterium effect in CHE stress regime, IEEE Trans. Electron Devices, vol. 49, pp. 254-263, Feb. 2002.

[16] J. Wu, E. Rosembaum, B. MacDonald, E. Li, J. Tao, B. Tracy, and P. Fang, Anode hole injection versus hydrogen release: The mechanism for gate oxide breakdown, in Proc. Int. Reliability Physics Symp., 2000, pp. 27-32.

[17] G. Pananakakis, G. Ghibaudo, C. Papadas, E. Vincent, R. Kies, Generalized trapping kinetic model for the oxide degradation after Fowler-Nordheim uniform gate stress, Journal of Applied Physics, v 82, n 5, p 2548-57, September 1997.

[18] D. DiMaria, J. Appl. Phys. 47, 4073 1976.

[19] R. Kies, T. Egilsson, G. Ghibaudo, and G. Pananakakis, A method for the assessment of oxide charge density and centroid in metal-oxide-semiconductor structures after uniform gate stress, Appl. Phys. Lett. 68 (26), 24 June 1996

[20] D.J. DiMaria, E. Cartier, D. Arnold, Impact ionization, trap creation, degradation, and breakdown in silicon dioxide films on silicon, Journal of Applied Physics, v 73, n 7, p 3367-84, April 1993.

[21] Y. Nissan-Cohen, J. Shappir, D. Frohman-Bentchkowsky, Measurement of Fowler-Nordheim tunneling currents in MOS structures under charge trapping conditions, Solid-State Electronics, v 28, n 7, p 717-20, July 1985.

[22] C. Papadas, P. Morfouli, G. Ghibaudo, G. Pananakakis, Analysis of the trapping characteristics of silicon dioxide after Fowler-Nordheim degradation, Solid-State Electronics, v 34, n 12, p 1375-9, December 1991.

[23] C. K. Williams, Kinetics of trapping, detrapping, and trap generation, Journal of Electronic Materials, v 21, n 7, p 711-20, July 1992.

[24] G. Torrente, J. Coignus, A. Vernhet, J.L. Ogier, D. Roy, G. Ghibaudo, Microscopic analysis of Erase-induced degradation in 40nm NOR Flash Technology, IEEE Transaction on Device and Materials Reliability (TDMR), v 16, n 4, p 597-603, Dec. 2016.

[25] A. Modelli, F. Gilardoni, D. Ielmini, and A. S. Spinelli, A new conduction mechanism for the anomalous cells in thin oxide flash EEPROMs, IEEE International Reliability Physics Symposium (IRPS), p 61-66, April-May 2001.

[26] B. Eitan and D. Frohman-Bentchkowsky, Hot-electron injection into the oxide in n-channel MOS devices, IEEE Transaction on. Electron Devices, vol. ED-28, pp. 328-340, March 1981.

[27] J. Coignus, G. Torrente, A. Vernhet, S. Renard, D. Roy, G. Reimbold, Modelling of 1T-NOR Flash Operations for Consumption Optimization and Reliability Investigation, IEEE International Reliability Physics Symposium (IRPS), p PR-1 (4 pp.), 2016.

[28] G. Torrente, X.Federspiel, D.Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo Hot Carrier Stress: aging modeling and analysis of defect location, IEEE International Reliability Physics Symposium (IRPS), p 5A-4 (6 pp.), 2016.

[29] P. Pavan, R. Bez, P. Olivo, E. Zanoni, Flash memory cells-an overview, Proceedings of IEEE, v 85, n 8, p 1248-71, August 1997.

[30] J. Coignus, A. Vernhet, G. Torrente, S. Renard, D. Roy, G. Reimbold, Relaxation-free Characterization of Flash Programming Dynamics along P-E Cycling, IEEE International Integrated Reliability Workshop Final Report (IIRW), p 119-21, October 2015.

[31] Y. Nishi, Study of silicon-silicon dioxide structure by electron spin resonance, Japanese Journal of Applied Physics, v 10, n 1, p 52-62, January 1971.

[32] P.J. McWhorter, P.S. Winokur, Simple technique for separating the effects of interface traps and trapped-oxide charge in metal-oxide-semiconductor transistors, Applied Physics Letters, v 48, n 2, p 133-5, January 1986.

[33] G. Torrente, J. Coignus, S. Renard, A. Vernhet, G. Reimbold, D. Roy, G. Ghibaudo, Physically-based extraction methodology for accurate MOSFET degradation assessment, Microelectronics Reliability Journal, v 55, n 9, p 1417-21, August-September 2015.

[34] Synopsys, Zurich, Switzerland, Sentaurus device user guide, J-2014.09

[35] P. Samanta, M. Chan, Effects of gate material on Fowler-Nordheim stress induced thin silicon dioxide degradation under negative gate bias, Journal of Applied Physics, v 96, n 3, p 1547-55, August 2004.

[36] D.Varghese, et al., OFF-State Degradation in Drain-Extended NMOS Transistors: Interface Damage and Correlation to Dielectric Breakdown, IEEE Trans. Electron Dev.54(2007)2669-2677.

[37] C. Chang, C. Hu, R.W. Brodersen, Quantum yield of electron impact ionization in silicon, Journal of Applied Physics, v 57, n 2, p 302-9, January 1985.

Chapter 4

Flash endurance understanding and modeling

NOR Flash memory technology has been scaled for years, leading to high integration densities [1]. This technology significantly evolved over the years: from the standard 1T-NOR Floating Gate cells, non-conventional structures, such as 2T-NOR cells [2],[3], charge trapping layer based cells [4],[5], MirrorBit cells [6] and Split 1T-NOR [7]-[9], have been developed and industrialized in the recent past. However, device lifetime is always limited by charge buildup on defect sites in the tunnel oxide layer together with induced amphoteric traps at Si/SiO_2 interface [10]. This directly leads to the closure of Program Window (PW) due to the loss of Program/Erase (P/E) efficiencies and the drift of cell electrostatics [11]. Although several studies deal with Flash reliability issues, especially for NAND technology [12],[13], concerning SILC mechanism [14]-[16] and Vth statistical dispersions [17],[18], one cannot find works which accurately explored the evolution of the PW during P/E cycling from a microscopic physical standpoint in Hot Carrier programmed Flash technology.

In general, during P/E cycling, the overall cell aging can be decoupled into two contributions: Electrostatic Degradation (ED), i.e. the static Vth drift due to the induced oxide wear out, and Operation Efficiency Loss (OEL), i.e. Vth shift brought on by tunnel oxide current decay during P/E phase. It has to be underlined that a complete understanding of these cell characteristic drifts during P/E endurance is fundamental for such a technology. Indeed, an accurate physical comprehension of their evolutions and dependences helps not only to find new technological optimizations, but also to push the memory cell towards its maximum intrinsic performance by correctly managing the P/E electrical operations, as recently shown in [19],[20].

Considering conventional Floating Gate 1T-NOR Flash technology, cells are programmed by Hot Carrier (HC) mechanism and erased by Fowler-Nordheim (FN) process. As we saw in the previous chapters, the degradation induced by these two mechanisms is different: a

schematic picture of the two wear out processes is shown in Fig.4.1 for a MOSFET device. Although Hot Carrier Degradation (HCD) [21],[22] and Fowler-Nordheim Stress (FNS) [23],[24] have been widely studied on transistors, their role on PW evolution during P/E cycling and how both mechanisms interact with each other have not been addressed in literature. In particular, their respective impact on Flash cell characteristic drifts, i.e. ED and OEL, has not been studied yet. In addition, a clear separation of these cell characteristics has never been properly addressed within the Flash cell dynamics theory.

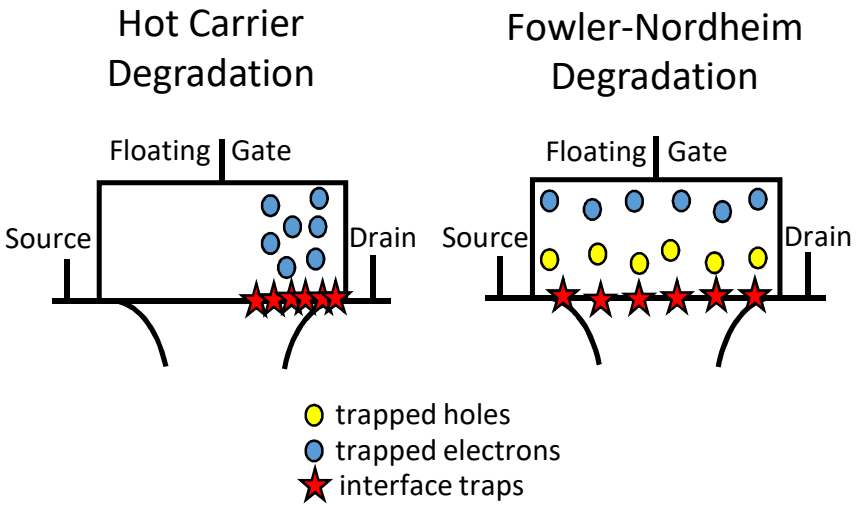


Figure 4.1: Schematic illustration of HC and FN degradations suffered by a Flash cell.

This significant lack of knowledge mainly comes from experimental limitations. Indeed, in order to correctly separate the cell characteristic degradations, the reproduction of the stress suffered by the cell on ad-hoc Flash transistor structures has to be performed. Taking advantage of these devices, the Program-induced aging can be studied independently from Erase operation (and vice versa) by turning off the Erase phase, which is simply not possible on the memory cell itself. In particular, the degradation suffered by the transistor allows to access the memory cell wear out purely induced by electrostatics (ED), making possible to separate such characteristic from the one just linked to the injection current degradation (OEL). In any case, in order to compare equivalent transistor degradation results with the ones directly measured on memory cell during P/E cycling, such an approach needs an advance experimental setup which allows customizable shape waveforms for CG pattern during P/E phases. In other words, since on Flash cell the FG node is not directly accessible, we are obliged to customize the external voltages in order to induce a specific FG potential which is then applied on the respective equivalent transistor. In addition, such an experimental setup demands a delay-free system for controlling all possible delays between the different electrical operations and keeping P/E cycling tests within reasonable durations. For all these reasons, i.e. the need of advanced experimental setup and ad-hoc MOSFET structures, such

an approach is rarely pursued.

In this chapter and in our work published in *Microelectronics Reliability Journal* [25], we take advantage of the physical comprehension previously achieved on Program- and Erase-induced degradations, i.e. HC and FN aging respectively, on equivalent transistor structures, we make use of the advanced experimental setup previously described and we finally explore the overall evolution of the Flash cell during endurance experiments. The philosophy is to decouple the drift of Vth read at Program and Erase states, i.e. $\mathrm{Vth_{P/E}}$, into the evolution of the electrostatics and the loss of P/E operation efficiency. Then, each contribution is analyzed, highlighting the driving microscopic aging mechanism, such as for example Program-induced interface defect generation, Erase-induced hole trapping and so on. Finally, the proper aging kinetics describing this process is established and included in the Flash endurance framework.

It is important to underline that the approach used in this chapter has the advantage to be general, hence it can be applied to all non-conventional Flash cells as well. This would allow to deeply analyze the difference between the cell structures, giving the possibility to quantitatively explore cell degradation variations.

The chapter is divided into three parts:

- In the first section, an accurate separation of the macroscopic aging contributions, such as the loss of P/E efficiency and the static drift of the cell during Flash P/E endurance experiments, is carried out at $\mathrm{Vth_{P/E}}$ states.
- In the second part, the interplay between Program- and Erase-induced degradations is analyzed and addressed from an experimental standpoint with the help of equivalent Flash transistors, which have been used in order to switch on/off one of the two cell operations. This approach will help to properly quantify the aging kinetic laws that determine the Flash cell evolution during P/E cycling.
- In the last section, a complete physical-based model will be presented computing the drifts of $\mathrm{Vth_{P/E}}$ during Flash endurance experiments considering different P/E patterns. The respective role of Program-efficiency loss, Erase-efficiency loss and static cell aging on the overall PW evolution are explored. A proper characterization and understanding of each contribution is performed, and guidelines towards better reliability performance are proposed.

4.1 Separation of cell aging contributions

The experiments presented in this chapter have been done at 25 Celsius degrees on devices issued from 40nm embedded NOR Flash technology developed at STMicroelectronics

(tox = 9.7nm, L = 140nm, W = 60nm). Similarly to the previous chapters, we take advantage of equivalent Flash transistors, which are built by shorting Control Gate (CG) and Floating Gate (FG) of the cell, in order to study the stress suffered by the tunnel oxide and reproduce P or E induced aging only. Two different transistor/cell couples have been considered in this chapter: having $\alpha_{G,lin} = 0.648$ (group 1) or $\alpha_{G,lin} = 0.635$ (group 2), where $\alpha_{G,lin}$ is the cell coupling coefficient acquired in linear regime.

To reproduce the stress on equivalent transistors, we take advantage of the methodology presented in [20] and in the first chapter of this thesis. It has been demonstrated that an accurate knowledge of Vth_{mos}[1] and of IfgVfg during P/E operations allows to design a proper ramp pattern for the CG potential (Vcg), which perfectly induces a constant FG level (Vfg) during P/E phase. The validity of such an approach has been also verified through the measurement of the drain current during the Program operation, which results to be strictly constant (Fig.4 in [20]). Such a "constant Vfg" endurance can be easily reproduced on equivalent transistors with the respective square patterns, i.e. Vg_{stress} = Vfg level of the Flash cell, as schematically shown in Fig.4.2(a). In Fig.4.2(b), the drift of Id_{ON}, i.e. extracted drain current at $Vfg = Vth_{mos} + 4V$[2], that has been demonstrated to be only sensitive to channel mobility degradation [26], is compared for the two devices and clearly indicates that the stress suffered by them is identical, which highlights the accuracy of the equivalent transistor stress. In Tabs.4.1, 4.2 all the endurance conditions are listed for groups 1 and 2 respectively considering a nominal PW=5.75V.

In Fig.4.3, the drifts of P/E threshold voltages ($Vth_{P/E}$) are shown during Flash endurance experiment for the condition A of Tab.4.1. In the figure, the roles of the static degradation (ED), represented by $\Delta Vth_{mos}/\alpha_{G,lin}$ acquired on the equivalent transistor endurance, and of P/E efficiency losses (OEL) are highlighted. The first assesses the memory cell wear out purely induced by electrostatics, i.e. due to the created/filled defects within the oxide and at the Si/SiO_2 interface. On the other hand, the second one is linked to the injection current degradation, which may have a different dependence from the same defect configuration. As we will demonstrate, these two Flash characteristics fully define the overall cell wear out read at the two memory states, i.e. Program and Erase levels. Anyhow, a priori, such characteristics are not completely independent, since the same single trapped electron can affect one of the two, none of the two, or both characteristics. Taking advantage of these considerations, we will see that this approach allows to microscopically explore the

[1] as defined in the first chapter, the read condition has been set at Id = 8μA, Vd=0.5V. For such a condition, Vcg=Vth (Flash cell threshold voltage) and Vfg=Vth_{mos} (MOSFET threshold voltage). The last parameter can be directly read on the equivalent transistor structure even when the device is aged, as long as it experiences exactly the same degradation of the respective Flash cell.

[2] accordingly to Flash cell electrostatic system, such a parameter is measured on the Flash cell at $Vcg = Vth + 4V/\alpha_G$

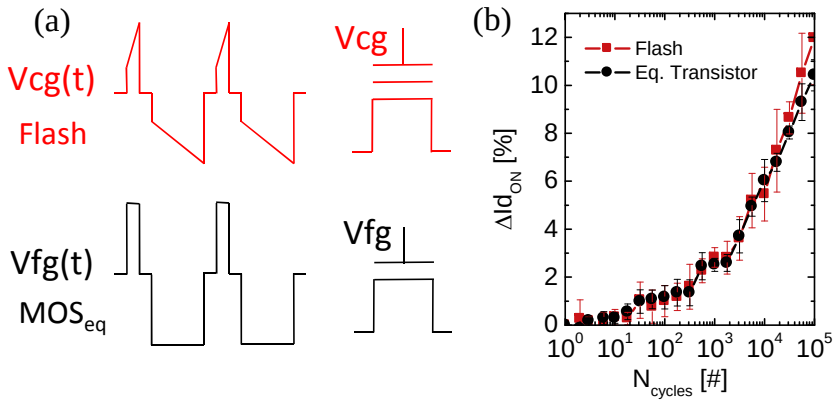


Figure 4.2: Flash cell and equivalent transistor endurance. In (a) the illustration of P/E cycling carried out on Flash cells using optimized CG ramp patterns and of the respective square pulses applied on the equivalent transistors in order to induce the same stress [20] is shown. The representation of Vd=const pattern is skipped for simplification. In (b) the comparison of Id_{ON} (≈mobility aging [26]) between the two devices is shown for the condition A of Tab.4.1.

device aging, identify the weak points of cell architecture and address the main physical mechanism factors limiting the correct device working condition.

Coming back to the Fig.4.3, it is worth noting that at the Vth_P state the two contributions are in opposite phase, whereas at Vth_E they are summed and contribute to a positive ΔVth_E. In the following, the correct separation of these aging contributions is addressed within the Flash theory framework and each characteristic drift is accurately analyzed.

Stress	$Vfg_{eq,E}$	b_E	$Vfg_{eq,P}$	b_P
A	-12V	1.78kV/s	4V	2,98MV/s
B	-11.35V	308V/s	4V	2,98MV/s
C	-10.5V	28.6/s	4V	2,98MV/s
D	-12V	1.78kV/s	2.9V	250kV/s
E	-12V	1.78kV/s	2.5V	49kV/s

Table 4.1: Endurance conditions for PW=5.75V applied on both Flash and equivalent transistors belonging to group 1 ($\alpha_{G,lin} = 0.648$). For each one, a statistics of 3 equal devices has been considered. $Vfg_{eq,P/E}$ indicates the level of the stress during P/E pulse at FG potential, which, for the Flash case, is induced by a ramp on the CG electrode with a speed of $b_{P/E}$ (Fig.4.2). The drain potential (Vd) considered is equal to 4V, 0.5V and 0V during Program, Read and Erase phases respectively, whereas Vb=Vs=0V all time.

The advanced measurement setup described in [27] and in the first chapter of the thesis

Stress	$Vfg_{eq,E}$	b_E	$Vfg_{eq,P}$	b_P	Vd_P
A	-11.465V	640V/s	3.75V	1.93MV/s	4V
B	-11.0V	178V/s	3.75V	1.93MV/s	4V
C	-10.5V	40V/s	3.75V	1.93MV/s	4V
D	-11.465V	640V/s	2.5V	15kV/s	4V
E	-11.0V	178V/s	2.5V	15kV/s	4V
F	-10.5V	40V/s	2.5V	15kV/s	4V
G	-11.465V	640V/s	3.52V	746kV/s	4V
H	-11.465V	640V/s	3.3V	306kV/s	4V
I	-11.465V	640V/s	2.0V	1.76kV/s	4V
J	-11.465V	640V/s	3.75V	669kV/s	3.6V
K	-11.465V	640V/s	3.75V	48kV/s	2.8V
L	-11.0V	178V/s	3.3V	306kV/s	4V

Table 4.2: Same as in Tab.4.1, but for the group 2 ($\alpha_{G,lin} = 0.635$), considering also the Vd dependence and a statistics of 5 samples.

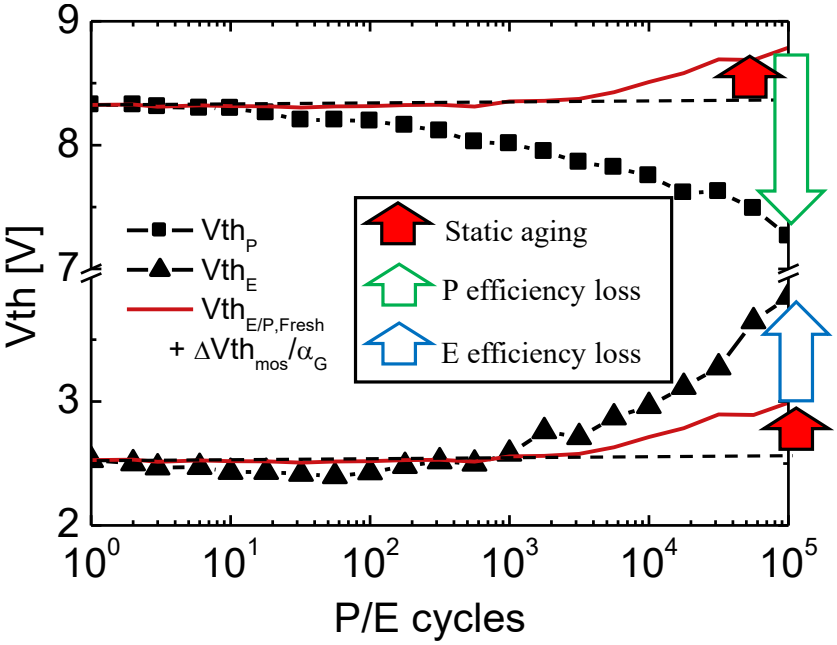


Figure 4.3: Illustration of the evolution of aging contributions during P/E cell endurance performed with optimized ramp patterns. The impact of static (where Vth_{mos} is measured during the equivalent P/E endurance on Flash transistors) and P/E efficiency degradations are highlighted at both states. The condition A of Tab.4.1 is considered.

has been considered, which features 4 synchronized Keysight B1530 Fast-Measurement Units (FMUs), for Program operation and Read phase (i.e. fast IdVg acquisitions), a Pulse Gener-

ator (PG), for Erase operation, and a CMOS switch that connects the gate with FMUs and PG offering ns-switching duration between device operations. As we saw, this setup allows a systematic measurement of Program dynamics (Id(t) with a 5ns sampling period), together with fast and delay-free read operations for capturing intrinsic device characteristics.

In order to study the drift of $Vth_{P/E}$ and properly separate electrostatic and cell performance decays, the device dynamics within P/E operations have to be studied. The standard equation of Flash cell [28] is considered during the P/E phase:

$$Vfg(t) = \alpha_{G,P/E} Vcg_{P/E}(t) + \alpha_{D,P/E} Vd_{P/E} + \frac{Q_T(t)}{C_T} \tag{4.1}$$

where C_T is the total capacitance seen from FG and Q_T is the sum of FG and tunnel oxide charges (Qfg(t) and Qox respectively). In eq.4.1, the drain potential is constant (=4V during Program phase and =0V within the Erase pulse) and the coupling coefficients do not significantly vary. Thus, differentiating the equation in time, we get:

$$\frac{dVfg}{dt} = \alpha_{G,P/E} \cdot \frac{dVcg}{dt} + \frac{Ifg(Vfg)}{C_T} \tag{4.2}$$

where Ifg(Vfg) is the HC/FN characteristic (depending on the operation being considered). Considering ramp patterns on the CG electrode, i.e. $Vcg_{P/E}(t) \propto b_{P/E} \cdot t$, where $b_{P/E}$ is the ramp speed in V/s, the cell dynamics is described by the equation:

$$\frac{dVfg}{dt} = \alpha_{G,P/E} \cdot b_{P/E} + \frac{Ifg(Vfg)}{C_T} \tag{4.3}$$

Solving eq.4.3, Vfg(t) within P/E cycle is computed. It is worth noting that the Steady-State (SS) condition, i.e. Vfg(t)=const, reduces to:

$$|Ifg_\infty| = \alpha_{G,P/E} \cdot b_{P/E} \cdot C_T = b_{P/E} \cdot C_{ono} \tag{4.4}$$

where C_{ono} is the capacitance between FG and CG electrodes. This condition is reached when the |Vfg| increase induced by the ramp pattern, i.e. $b_{P/E} \cdot \Delta t \cdot \alpha_{G,P/E}$, perfectly compensates the |Vfg| decrease due to the charge variation in FG caused by HC/FN tunnel oxide current. In addition, the SS condition of eq.4.4 is reached independently from the aging, thus the final FG potential $|Vfg_\infty|$ is always the one which induces a tunnel oxide current equal to $b_{P/E} \cdot C_{ono}$, as simulations confirm in Fig.4.4(b) for Program case.

In this figure, Ifg vs time during the Program ramp pattern has been simulated from eq.4.3 using the Ifg(Vfg) curve of Fig.4.4(a) separately extracted from the Step Pulse Program (SPP) experiment [29]. As already discussed, this experiment consists in the acquisition of Vth evolution during a train of short Program CG constant pulses (100ns), as schematically shown in the inset of Fig.4.4(a). From the experimental Vth_{SPP} vs time, Ifg(Vfg)

characteristic can be extracted [29], since $Ifg \propto dVth_{SPP}/dt$ and $Vfg \propto Vth_{mos} - Vth_{SPP}$. The advantage of performing the "constant Vfg" cycling is that Vth_{mos} can be estimated from the equivalent transistor endurance. Thus, Ifg(Vfg) curve can be extracted also for an aged cell as long as SPP experiment is performed. Later, we will take advantage of such an approach in order to analyze the Program-efficiency degradation since able to experimentally extract the Ifg(Vfg) evolution along P/E cycling.

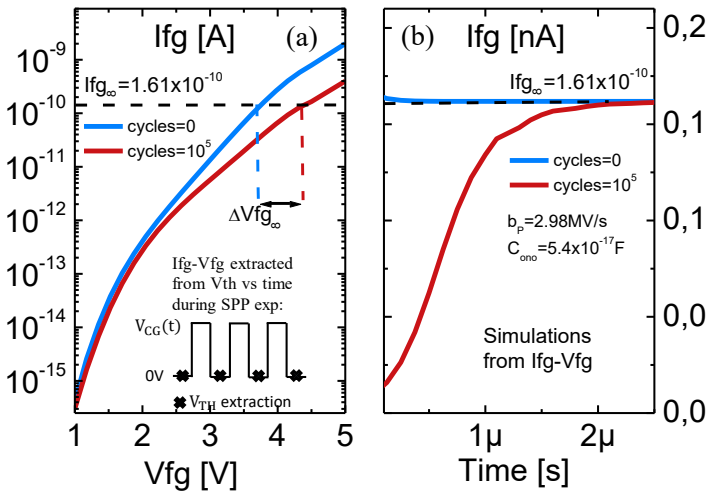


Figure 4.4: FG current characteristics. In (a) the Program IfgVfg curve, acquired from Step Pulse Program (SPP) experiment [29], is shown for fresh and aged cell. In (b) the dynamics of Ifg within the Program ramp pulse is simulated for the two cases using eq.4.3. The condition A of Tab.4.1 is considered.

From now on, we assume that the condition in eq.4.4 is always reached in the whole Flash cell endurance. It is worth mentioning that this approach can be extended considering any CG pattern. Indeed, it can be demonstrated that the cell dynamics always reaches a condition $Vfg_{\infty}(t)$ which depends uniquely on C_{ono}, Ifg(Vfg) and applied Vcg(t). Thus, it does not depend on the initial state (i.e. stored charge in FG before the pulse) and on the electrostatic degradation in linear regime.

A similar analysis has been done in [30] by Chimenton et al., who considered a sequence of box pulses with increasing voltage amplitude. Analogously, the equilibrium condition, i.e. constant tunnel oxide current of eq.4.4, had been analytically calculated with technological parameters. However, their work did not go further exploring the impact of the degradation on cell dynamics for such a pulse scheme, thus the respective role of ED and OEL, and, moreover, it just dealt with the Erase phase. Indeed, as we will show in the following, the use of such an approach for the Program phase represents a real advantage for properly optimizing the cell drift during P/E cycling.

Taking advantage of all the considerations previously done, the cell aging evolution at

$Vth_{P/E}$ levels can now be easily studied. The standard equation of Flash cell [28] is used both in the end of P/E operation (@ $t = t_{P/E}$) and in the following read condition (@ Id = 8μA, Vd=0.5V):

$$Vfg_{\infty} = \alpha_{G,P/E} Vcg(t_{P/E}) + \alpha_{D,P/E} Vd_{P/E} + \frac{\overline{Q_T}}{C_T} \quad @\ t = t_{P/E} \tag{4.5}$$

$$Vth_{mos} = \alpha_{G,lin} Vth_{P/E} + \alpha_{D,lin} Vd_{lin} + \frac{\overline{Q_T}}{C_T} \quad \text{read condition} \tag{4.6}$$

where $\overline{Q_T}$ is the total charge in the end of P/E pulse. The same $\overline{Q_T}$ has been considered in eq.4.5 and eq.4.6, assuming that during the IdVcg sweep, which is performed in 90μs, the FG charge does not vary. Computing the difference between these equations in order to be charge insensitive, we get $\alpha_{G,lin} Vth_{P/E} = Vth_{mos} - Vfg_{\infty} + K$, where K is constant along P/E endurance. Thus, the Vth shift at P/E states along cycling can be easily calculated as:

$$\Delta Vth_{P/E} = \frac{1}{\alpha_{G,lin}} \left(\Delta Vth_{mos} \mp |\Delta Vfg_{\infty}| \right) \tag{4.7}$$

where Δ has to be considered from cycle i to cycle 0. The sign $\mp$ comes from the different calculation at the two states, as already remarked in Fig.4.3. At the Program level, the two aging contributions are in opposite phase (-), whereas at Erase state they are summed (+).

The eq.4.7 represents the decouple of Vth drift at P/E states into two contributions: one component is linked to the static degradation, quantified in terms of ΔVth_{mos}, i.e. ΔVfg in linear regime in order to keep the constant current read condition, i.e. 8μA at $Vd_{read} = 0.5V$, whereas the second one represents the loss of P/E efficiency during the P/E operation, quantified in terms of ΔVfg_{∞}, i.e. ΔVfg in HC/FN regime in order to keep the "ramp" condition $b_{P/E} \cdot C_{ono}$ as shown in Fig.4.4.

Once ΔVth_{mos} is acquired on equivalent transistors, the two aging contributions at $Vth_{P/E}$ states can be quantified as shown on the PW evolution in Fig.4.3. However, the assessment of ΔVfg_{∞} has to be addressed for methodology validation and physical understanding.

Before starting to model the kinetics of these parameters, preliminary studies have to be done in order to have an overview of the cell drift during P/E cycling. Thus, for a given parameter, i.e. ΔVth_{mos} or $\Delta Vfg_{\infty,E}$ or $\Delta Vfg_{\infty,P}$, which represents a macroscopic aging contribution, the interplay between the different microscopic aging mechanisms and the respective driving force have to be addressed. Then, proper physical based laws are considered in order to reproduce the drift of Vth_E and Vth_P using the eq.4.7(a). For this reason, in the next section, the interplay between HC and FN induced degradations are considered on equivalent Flash transistors.

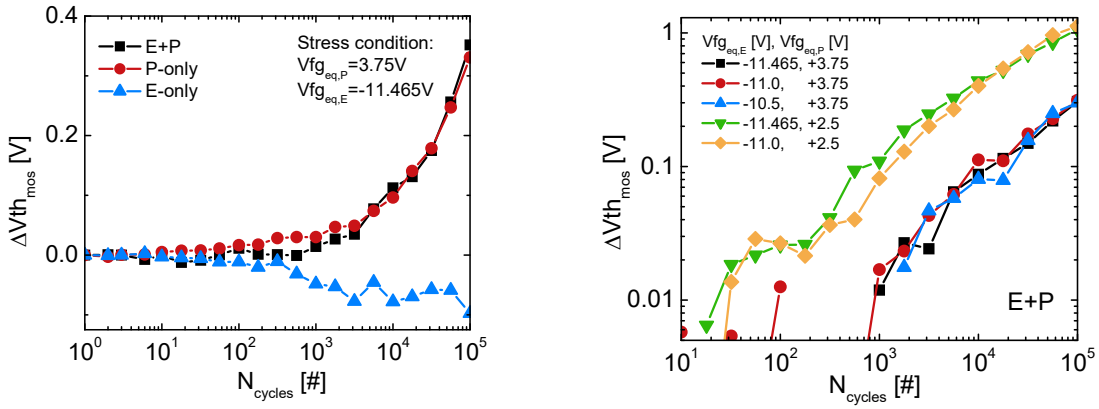
4.2 E+P static aging interplay

In the previous two chapters, HC and FN stresses have been separately studied. However, during Flash P/E endurance, the two degradation mechanisms are present in the same time: within one single cycle, one pulse of Erase and one of Program are successively applied. In other words, the two different induced wear out schematically shown in Fig.4.1 are mixed together following a degradation kinetics that may be different since influenced by the other operation.

For this reason, the device behavior has to be studied taking into account the interaction between these two aging processes. Having as a reference the eq.4.7, we have three macroscopic drifting parameters, i.e. ΔVth_{mos}, $\Delta Vfg_{\infty,P}$ and $\Delta Vfg_{\infty,E}$, and, for each of them, the driving aging mechanism has to be addressed. In this section, only Flash transistors are considered, on which the equivalent endurance, schematically drawn in Fig.4.2(a), is considered.

4.2.1 Erase vs Program wear out

One simple way to study the role of HC and FN degradations within the Flash cell endurance is simply to consider them separately and see the evolution of the device in the three different situations: Program-only (P-only), Erase-only (E-only), Program+Erase (P+E). In Figs.4.5,4.6,4.7 these experimental results are shown.



(a) ΔVth_{mos} for the condition A of Tab.4.2, switching on/off one of the two operations.

(b) ΔVth_{mos} during equivalent P/E endurance for the conditions A,B,C,D,E of Tab.4.2.

Figure 4.5: ΔVth_{mos} during equivalent P/E endurance applied on Flash transistors. The measurement has been done after the Program operation.

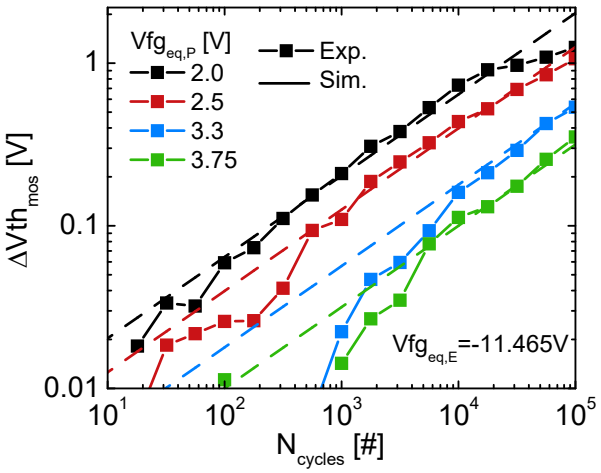
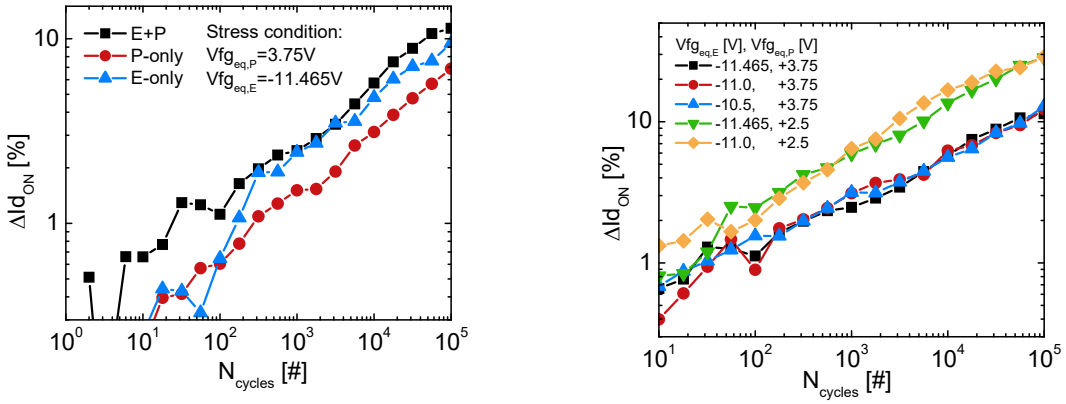


Figure 4.6: ΔVth_{mos} evolution during equivalent endurance on Flash transistors for the conditions A,D,H,I in Tab.4.2. Differently from Fig.4.5, the extraction has been performed after the Erase operation.



(a) ΔId_{ON} for the condition A of Tab.4.2, switching on/off one of the two operations.

(b) ΔId_{ON} during equivalent P/E endurance for the conditions A,B,C,D,E of Tab.4.2.

Figure 4.7: ΔId_{ON} measured during equivalent P/E endurance applied on Flash transistors. The measurement has been done after the Erase operation and no different is found if extracting the same parameters after the Program phase (not shown).

Considering the evolution of ΔVth_{mos}, which represents the static drift of the cell, in Figs.4.5,4.6, the situation is clear: HC-induced interface defect generation is the mechanism which drives the electrostatic aging. Indeed, as we saw in the previous chapter, the FN mechanism induces two different defects: interface traps, which increase Vth_{mos}, and trapped holes close to the Si/SiO_2 interface, which decrease Vth_{mos}. The net threshold

voltage shift induced by FN has been observed to be close to zero. Since the total drift can be roughly estimated, using the superposition principle in electrostatic framework, as $\Delta Vth_{mos} \approx \Delta Vth_{HC} + \Delta Vth_{FN}$, the FN mechanism is not expected to significantly vary the cell electrostatics. Looking at the experimental results in Fig.4.5(a), this assumption results to be correct. Thus, following the compact HCD model presented in the second chapter, we can assume that at cycle=i:

$$\Delta Vth_{mos}(i) \approx \Delta Vth_{mos,HC}(i) \approx A \cdot Vfg_{eq,P}^{p} \cdot \exp\left(-\frac{\alpha}{Vd_P}\right) \cdot (t_P \cdot i)^n \tag{4.8}$$

In accordance with these considerations, the static degradation does not vary considering different Erase patterns, as shown in Fig.4.5(b), whereas the overall increase/decrease of ΔVth_{mos} strictly follows the increase/decrease of $\Delta Vth_{mos,HC}$.

Considering different $Vfg_{eq,P}$ levels for the equivalent endurance, the experimental results are nicely reproduced by the law in eq.4.8, as shown in Fig.4.6. However, an empirical correction factor has to be accounted for whenever Vth_{mos} is acquired after the Erase operation, i.e. A in eq.4.8 lowers of 10%. Indeed, a significant electron detrapping towards Silicon substrate occurs during the single Erase pulse since supported by a high oxide electric field.

Looking at the Fig.4.6, it is worth noting that lower $Vfg_{eq,P}$ levels induce higher aging. This is explained considering the dependences of the Program-efficiency, i.e. $Ifg(Vfg_{eq,P})$, and the Program-induced degradation, i.e. $\Delta Vth_{mos}(Vfg_{eq,P}, t)$. Indeed, ΔVth_{mos} has a weak $Vfg_{eq,P}$ dependence and follows a power law in time [31], which comes from the superposition of exponential laws of different interface bond activation energies [32], whereas Ifg has an exponential dependence on $Vfg_{eq,P}$. Thus, considering lower $Vfg_{eq,P}$ patterns, the Program pulse duration, i.e. t_P in eq.4.8, has to be exponentially increased in order to maintain the same PW and, consequently, Vth_{mos} drifts more. In particular, considering $Vfg_{eq,P,2} = Vfg_{eq,P,1} - \Delta Vfg$, it is easy to demonstrate that, for $Ifg(Vfg_{eq,P}) = B \cdot 10^{K \cdot Vfg_{eq,P}}$, the acceleration rate factor of the static aging becomes:

$$\frac{\Delta Vth_{mos,2}}{\Delta Vth_{mos,1}} = \left(1 - \frac{\Delta Vfg}{Vfg_{eq,P,1}}\right)^{p} \cdot 10^{K \cdot n \cdot \Delta Vfg} \tag{4.9}$$

which is always greater than 1 in our case.

Coming back to the eq.4.8, two points have to be highlighted. Firstly, we implicity assumed that the HC-induced aging rate is unaffected by Erase-induced degradation. In the next subsection, we will explore the variation of the aging kinetics whenever the two operations are considered together. We will see that the assumption behind the eq.4.8 can be considered a good approximation. Secondly, whenever considering very low $Vfg_{eq,P}$ levels, the model fails for high cycles since not accounting for the saturation effect, as evident looking at Fig.4.6. As we saw in the second chapter, the saturation does not occur when all available Si-H interface states are depassivated, but it occurs for a self-limitation effect, which here

may be accelerated by the presence of Erase-induced trapped holes. Since this process is hard to be translated in a simple mathematical expression and does not significantly improve the precision, it is totally neglected in our model.

Concerning the drift of Id_{ON}, the experimental results are shown in Fig.4.7. Considering this parameter, the superposition principle is not valid anymore, since it cannot be applied on the mobility aging, i.e. $\Delta\%\mu(\text{Dit}) \not\propto \text{Dit}$. However, qualitative considerations can be done. Looking at Fig.4.7(a), it seems that the driving aging mechanism is the Fowler-Nordheim, which is not so surprising since affecting all the channel interface. However, as we saw in the previous chapter, the amount of interface defects created by FN mechanism depends mainly on the injected charges, and not on the Vfg stress level. Thus, considering different Erase patterns keeping fixed the PW, the degradation should not vary. Looking at Fig.4.7(b), this assumption results to be correct. In addition, it is worth noting that the HC degradation has a clear influence on the total mobility aging: considering lower $Vfg_{eq,P}$ patterns, thus degrading more, the mobility wear out is accelerated. In other words, in order to calculate the total degradation, we should take into account the partial superposition of the created interface trap distributions, thus a law like $\Delta Id_{ON} \approx \left(\Delta Id_{ON,HC}^{1/n} + \Delta Id_{ON,FN}^{1/n}\right)^n$ should be adopted, as similarly done in [33] by Federspiel. However, this has not been done since not of our interest.

4.2.2 FN/HC static aging interaction

After having determined which is the driving aging mechanism for static drifting parameters, the wear out process interactions have to be explored. Indeed, it is of our interest to get the behavior of HC-aging kinetics whenever FN-induced defects are present and vice versa. In order to experimentally address this point, the P/E aging rates have to be analyzed independently from E/P-induced degradation respectively. This can be done taking advantage of the read operation in saturation regime in Forward (FOR) or Reverse (REV) modes, i.e. Vd/Vs=4V, which addresses uniquely defects at source or drain side respectively [34]. It is worth noting that the trans-characteristics in saturation regime have been acquired extremely fast, i.e. within 16μs, in order to avoid any additional HC stress.

4.2.2.1 P influence on E-related aging rate

As already pointed out, it is of our interest to study the degradation uniquely induced by the Erase operation with and without the Program phase. In other words, does the Program operation have an effect on the Erase aging rate? To answer this question we take advantage of the methodology already developed in the previous chapter concerning the FN-induced aging, which has been detected with parameters acquired in saturation FOR

mode. Indeed, the saturated trans-characteristics gives important information on the wear out located at source side: Id_{ON} represents the interface state amount over the conduction band (Ec), whereas ΔVth_{mos} is related to the sum of Qox and Dit roughly below Ec (see the chapter 3 for details). Since the Program phase does not introduce any additional defects at that region, we can acquire exclusively the Erase-induced defects independently from the Program operation.

In Fig.4.8, experimental results are shown. It is evident that the drift of each parameter is slowed down whenever the Program pulse is considered, which seems counterintuitive. These results suggest that the Erase-induced degradation rate, in terms of both Dit generation and hole trapping, is slowed down by Program operation. This can be explained only if we accept that during the Program operation there is a strong hole detrapping process, which reduces the electron energy whenever the carrier reaches the Si/SiO_2 interface during FN regime. In this case, the AHI mechanism, together with the related Dit generation and hole trapping processes, would be less efficient. This is justified by the fact that during the Program phase the bands are pulled down in energy enhancing the hole relaxation process.

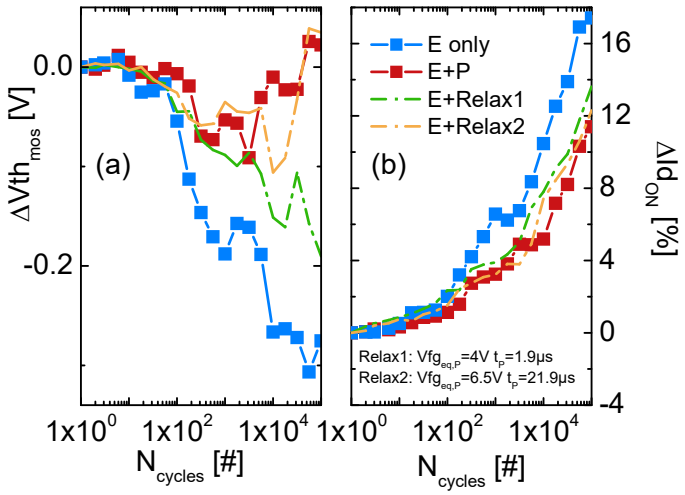
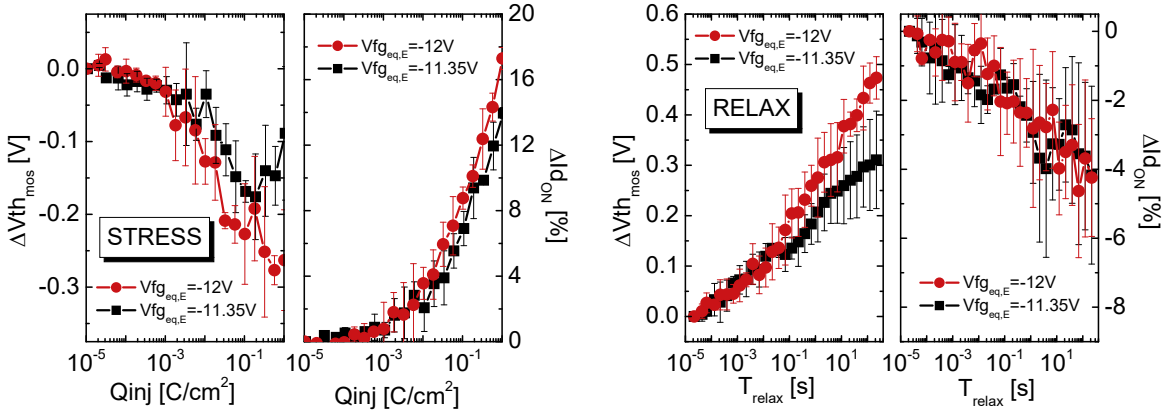


Figure 4.8: Experimental results of parameters in saturation regime during the equivalent endurance applied on Flash transistor. The condition A of Tab.4.1 is considered (red curve), together with E-only case (blue curve). The other two curves correspond to the same situation replacing the Program operation with a Relax phase (Vd=0V): $Vfg_{eq,P} = 4V$, $t_P = 1.9\mu s$ (green curve), which corresponds to the condition A but with Vd=0V, and $Vfg_{eq,P} = 6.5V$, $t_P = 21.9\mu s$ (yellow curve).

In order to confirm the hypothesis, this phenomena has to be directly addressed: stressing E-only the device and then looking at the relaxation process enhanced by a certain Vfg>>0V. This has been done for the conditions A,B of Tab.4.1 and the related experimental results

are shown in Fig.4.9. In Fig.4.9(a) the parameter drifts are shown during the E-only stress condition and it is worth noting the slight dependence on $Vfg_{eq,E}$ as expected from AHI theory. On the other hand, important evolutions can be observed during the Relax phase. Indeed, as predicted, the hole detrapping process strongly increases the threshold voltage and partially recovers the mobility degradation since border holes are detrapped.

Although the hole relaxation process has been demonstrated to be significant during the Program phase, it is not enough to fully explain the experimental data. Indeed, if this is the case, considering $Vd_{eq,P}$=0V (instead of 4V) during the equivalent endurance for the condition A of Tab.4.1, no significant variations should be observed, i.e. similar evolutions respect to the red curves of Fig.4.8 are expected. However, looking at the green curve in Fig.4.8, this does not entirely happen; thus, the interpretations previously done do not fully explain the situation. In addition, setting $Vd_{eq,P}$=0V during the Program operation, the hole relaxation process should be even increased since occurring all along the channel and not just at the source side. In order to match the results with $Vd_{eq,P}$=4V, $Vfg_{eq,P}$ has to be increased up to 6.5V and the time pulse t_P up to 21.9μs during the relaxation period (yellow curve). It is worth mentionting that this Relax condition has been verified not to add any additional degradation (not shown here). The only way to fully explain the experimental results in Fig.4.8 is to consider the acceleration of electron trapping process during the Program case due to the presence of Erase-induced trapped holes. However, this effect does not strongly affect the overall drifts during the cell endurance; thus, it will be not henceforth accounted for.



(a) Drifting parameters acquired during the E-only stress corresponding to the conditions A and B of Tab.4.1

(b) Evolution of the parameters acquired after stress and during the hole relaxation enhanced by Vfg=6.5V.

Figure 4.9: Experimental results measured on equivalent transistors during E-only stress (a) and during the following Relax phase (b).

4.2.2.2 E influence on P-related aging rate

In the previous subsection, the slow down of Erase aging kinetics has been explored whenever the Program operation is considered. Similarly, the opposite situation can be analyzed by reading the parameter drifts in saturation regime in REV mode, i.e. Vs=4V and Vd=0V. Basically, the created damage at drain side is mainly due to HC regime, thus, reading the IdVg trans-characteristics in REV saturation mode, the Program-induced interface traps are addressed independently from Erase operation. In Fig.4.10, the related results are shown for the conditions A,E of Tab.4.1 considering E+P and P-only cases. It is worth noting that the degradation induced at drain side in terms of ΔVth_{mos} is just slightly decelerated by the Erase phase since the Erase-induced interface states at source side may decrease the aging rate during HC regime. For this reason, it can be assumed, at first order, that the HC aging kinetics is unaffected by the Erase operation, as implicitly done in eq.4.8. However, accurate interpretations are hardly achievable looking at Fig.4.10 since Erase-induced Dit at the drain side are not negligible for high cycles as evident looking at ΔId_{ON} evolution.

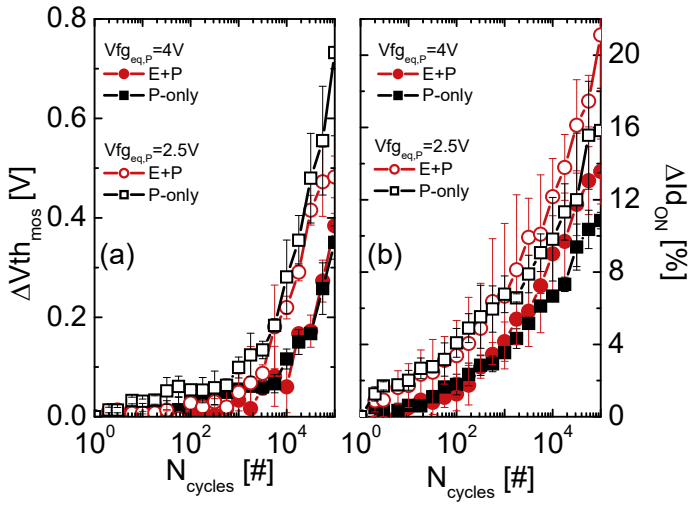
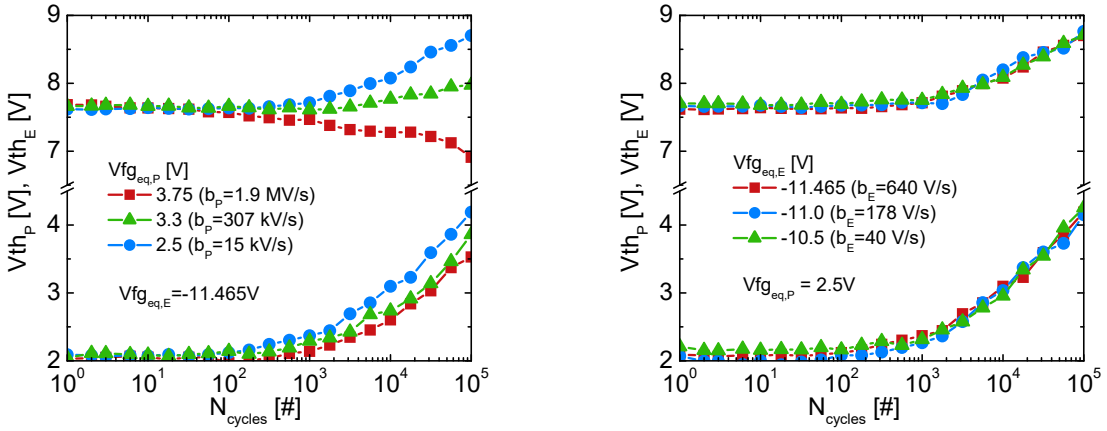


Figure 4.10: Parameter drifts extracted in REV saturation regime (Vs=4V) during equivalent P/E endurance applied on Flash transistors. Conditions A,E of Tab.4.1 are considered for E+P and P-only cases.

4.3 Flash modeling endurance

As already pointed out, for a correct technology understanding and optimization, it is necessary to acquire and separate the aging contributions within the PW evolution dur-

ing endurance experiments. Indeed, throughout P/E cycling, the drift of $Vth_{P/E}$ has to be decoupled into static cell degradation and P/E operation efficiency loss. This can be addressed with the help of equivalent Flash transistors and by accurately modeling the different contributions. For instance, the physical reasons behind cell evolution when modifying the optimized pattern during Program or Erase operation, as in Fig.4.11, has to be addressed. Following eq.4.7 and Fig.4.3, this can be achieved by knowing the aging laws for ΔVth_{mos}, $\Delta Vfg_{\infty,E}$, $\Delta Vfg_{\infty,P}$ function of stress condition, i.e. $Vfg_{eq,E}$, $Vfg_{eq,P}$, Vd_P, and of stress time or injected charges. In order to get the kinetic laws for such parameters, the driving degradation mechanism factor has to be properly addressed. This has been partially done in the last section looking at the interplay between HC and FN aging processes for equivalent transistor structures.



(a) Endurance results for different Program patterns. Conditions A,H,D of Tab.4.2 are shown.

(b) Endurance results for different Erase patterns. Conditions D,E,F of Tab.4.2 are shown.

Figure 4.11: $Vth_{P/E}$ evolutions during endurance experiments considering different Program/Erase optimized ramp patterns.

Although the working framework has been set, a clarification has to be done on the meaning of "static" aging. Indeed, it has to be pointed out that a certain defect distribution within the oxide gives an electrostatic shift larger than $\Delta Vth_{mos}/\alpha_G$, since charges affecting the FG potential, i.e. traps close to FG electrode, increase ΔVth_{Flash} in spite of a negligible effect on Vth_{mos}[3]. This is true in general: for a fixed Qfg, the cell threshold voltage degradation is higher than $\Delta Vth_{mos}/\alpha_G$. However, during endurance experiments, we read Vth_{Flash} at P/E state, which experiences different FG charge amount along cycling. Indeed, as we already demonstrated, in the end of P/E phases the Vfg potential depends uniquely on the Ifg

[3]in accordance with the terminology used so far, Vth_{mos} shift corresponds to Vfg shift at threshold condition. On the other hand, Vth (or Vth_{Flash}) shift is measured at CG node, thus it refers to Vcg shift at threshold condition.

degradation, thus, it does not directly depend on traps influencing the electrostatics in read condition. In other words, assuming that SS condition is always reached within the pulse, $\overline{Q_T} = \overline{Qfg} + Qox$ is fixed by the degradation of P/E tunnel oxide current dynamics, thus, during the following read condition, the electrostatic drift is strictly equal to $\Delta Vth_{mos}/\alpha_G$.

In this section, the aging contributions are analyzed separately. Simple compact models are introduced for each of them in order to reproduce the PW evolution using different P/E patterns as in Fig.4.11. As we will see, this approach gives important insights on the nature of cell degradation and on driving aging mechanisms.

4.3.1 Static aging

In the previous section, we explored the physics behind the electrostatic drift of the cell, i.e. ΔVth_{mos}. Applying separately Erase and Program induced degradations on equivalent transistors, HCD has been demonstrated to be the limiting mechanism factor. Thus, using the eq.4.8, a good prediction of ΔVth_{mos} under equivalent endurance on Flash transistors has been achieved, as shown in Fig.4.6. However, the stress induced by Flash transistors with equivalent square patterns may be different from the real wear out suffered by the MOS structure "inside" the cell. In particular, we expect similar wear out for the two devices in the beginning of the endurance, then, for long cycling duration, a difference between them is expected. In order to maintain the condition in eq.4.4, the cell should increase Vfg potential in absolute value. For Erase induced degradation, nothing significantly changes since the aging depends mainly on the charge flow and not on Vfg level. On the other hand, for Program-induced aging, the situation is different. We will see in the next subsection that higher Vfg levels suffer from higher tunnel oxide current wear out. Thus, increasing the $Vfg_{eq,P}$ potential, a larger cell aging is expected respect to the respective equivalent transistor endurance. Looking at the results in Fig.4.12, this is the case: considering low $Vfg_{eq,P}$, the two devices drift similarly, whereas for high $Vfg_{eq,P}$ levels the Flash cell degradation deviates from transistor results for high cycling. It has to be pointed that this extra degradation can also be skipped in the calculation since not significant. However, the error in the computation of ΔVth_{mos} may be not negligible whenever is transferred on Flash cell endurance since divided by $\alpha_G \approx 0.6 - 0.7$.

In order to calculate ΔVth_{mos} really experienced by the Flash cell, we take advantage of its relationship with ΔId_{ON}. As already seen in Chapter 2, the relationship between these two parameters depends on the trap distribution location. However, in a first approximation, during the Flash endurance, the aging along channel interface is fairly uniform, as already remarked by Verma in [35]. For this reason, changing the gate pattern, the correlation ΔVth_{mos} vs ΔId_{ON} should not significantly vary. Looking at the results obtained with

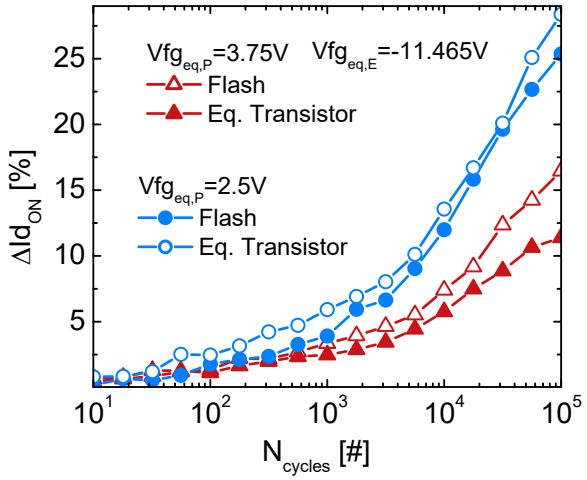
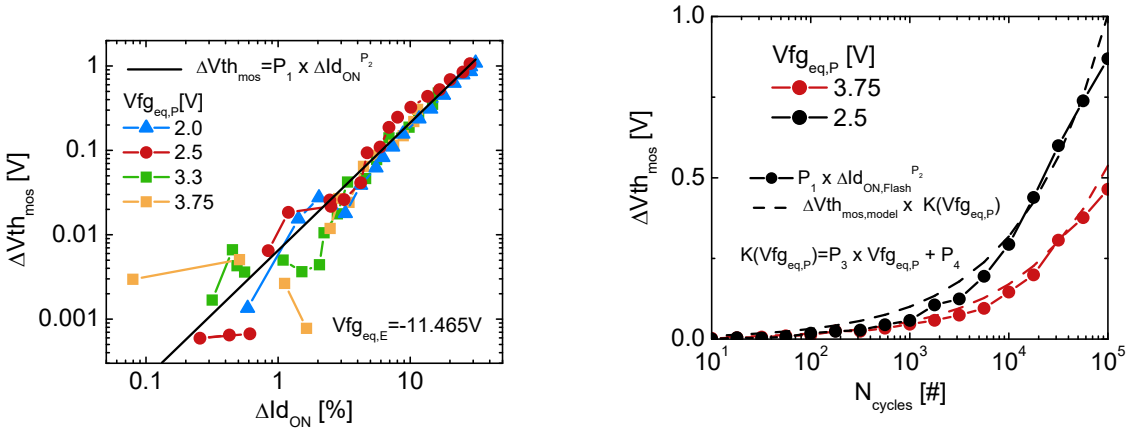


Figure 4.12: Degradation of Id_{ON} for equivalent transistor and Flash cell under the endurance conditions A,D of Tab.4.2.

equivalent transistor endurance in Fig.4.13(a), this is the case.



(a) Correlation between ΔVth_{mos} and ΔId_{ON} during equivalent transistor endurance for the conditions A,D,H,I of Tab.4.2. The real ΔVth_{mos} during Flash cell endurance can be estimated.

(b) Comparison between ΔVth_{mos} extracted from Flash Id_{ON} drift (using (a)) and the model, i.e. eq.4.8, accounting for the acceleration factor $K(Vfg_{eq,P}) \propto Vfg_{eq,P}$.

Figure 4.13: Model of static degradation acceleration whenever passing from equivalent to real Flash endurance with ramp optimized patterns. Extraction in (a) and fitting in (b).

A linear relationship is observed in Log-Log scale between the two parameters, thus a power law well describes the correlation. Since ΔId_{ON} is directly measurable during the cell endurance, ΔVth_{mos} can be estimated using this law. Finally, these results can be compared with the model of Vth_{mos} degradation accelerated by a factor $K(Vfg_{eq,P}) \propto Vfg_{eq,P}$,

in accordance with what previously said. A linear acceleration law has been chosen for simplification purpose. The results for two different $Vfg_{eq,P}$ levels (fixed $Vfg_{eq,E}$) are shown in Fig.4.13(b) and it is worth noting the good accuracy of the model.

4.3.2 P/E efficiency losses

4.3.2.1 E-efficiency loss

In the previous chapter, the degradation of tunnel oxide current under FN regime has been explored. In accordance with a pure trapping process, we experimentally found the relationship [36],[37]:

$$\Delta Vfg_{\infty,E} = A \cdot Qinj^{\nu} \tag{4.10}$$

where Qinj represents the injected charges and ν is an oxide quality parameter. Since this law can be direcly extracted during CCS experiment on CAST structure, as previously done, and Qinj at cycle=i can be simply computed as

$$Qinj(i) = C_{ono} \cdot \sum_{k=0}^{k=i} PW_k \tag{4.11}$$

the calculation of $\Delta Vfg_{\infty,E}$ is straightforward.

Using this approach together with the model for the static drift previously shown, the extracted aging contributions at Vth_E state can be obtained, as done in Fig.4.14(a) for the endurance condition A of Tab.4.1. Then, the respective estimated total drift ΔVth_E can be finally compared with experimental results, as shown in Fig.4.14(b) for the same stress condition.

The mismatch with experiments at the beginning of the endurance in Fig.4.14(b) comes from the well-known turnaround of Vth_E [10] caused by hole trapping process, which is not taken into account in the calculation of $\Delta Vfg_{\infty,E}$. In addition, it is worth noting an underestimation of ΔVth_E for high cycles. Since the computation of Vfg dynamics, which has been performed using Ifg(Vfg) extracted from Step Pulse Erase (SPE) [29], confirms that SS condition is reached (not shown here), the mismatch must be due to HC-induced aging. Indeed, the Erase-efficiency decay has been attributed only to FN-induced wear out. However, the presence of trapped electrons in the oxide injected close to FG during the Program pulse [38] must significantly impact the FN current. In addition, as remarked in the previous section, a strong hole detrapping is present during the Program operation, which may slow down the Erase-efficiency as well. For all these reasons, in order to account for this mismatch for high cycles, an empirical multiplicative factor is added in eq.4.10. However, knowing that the electron injection during the Program phase occurs from the Si-substrate

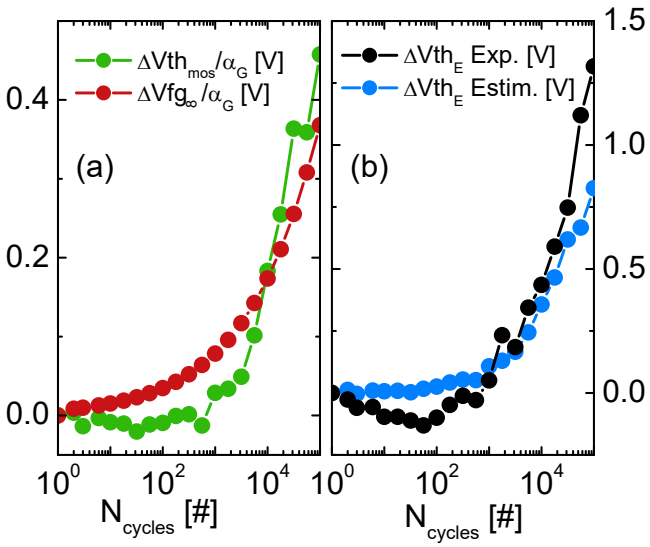


Figure 4.14: Endurance condition A of Tab.4.1. In (a) the aging contributions, in accordance with eq.4.7, are shown. In (b) the estimated drift of Vth_E is compared with the experimental one.

towards the Polysilicon, this factor must be between 1 and 2. In our case, we estimated a value of 1.8.

4.3.2.2 P-efficiency loss

Concerning the Program-efficiency wear out, the degradation of HC tunnel oxide current for NOR Flash technology has not been the subject of several studies in the past. Generally, it is attributed to the Program-induced interface traps at the drain side. In [39] this is justified by a strong coulomb scattering mechanism, which reduces the electron energy, thus the tunnel oxide current. However, this mechanism results to be completely negligible in the energy tail of the electron energy distribution function [40], since phonon and surface scatterings are dominating. In any case, these traps may affect the Program-efficiency by lowering the tunnel oxide transparency, as we will see later.

Before modeling the tunnel oxide current degradation and exploring the driven aging mechanism factor, we take advantage of the direct measurement of the drain current Id(t) within the Program pulse (Fig.4.15(a)) [27] in order to directly address $\Delta Vfg_{\infty,P}$ and validate the methodology which separates the aging contributions. First of all, it is worth noting that the drain current always reaches the SS condition Id_∞ in Fig.4.15(a), even for the last cycle considered, which means that $Vfg = Vfg_\infty$ within the pulse as previously assumed. In order to extract Vfg_∞ from the measured Id_∞, Flash transistors have been stressed with equivalent

square patterns and IdVfg has been acquired @Vd=4V during the read phase. In order to avoid any extra HC-stress, the whole IdVfg curve has been measured within 16μs. Using these results, as schematically shown in Fig.4.15(b), $\Delta Vfg_{\infty,P}$ can be easily extracted. It is worth noting that this technique is also valid when the SS condition (eq.4.4) is not reached, because the cell state in the end of the Program operation is directly captured.

As already shown, the degradation of the trans-characteristic in saturation regime (Fig.4.15(b)) depends uniquely on the Erase induced aging, since directly linked with the generated interface traps at source side [34]. Thus, the evolution of this curve can be acquired just once for different Program patterns. In addition, a supplementary simplification could be done: the degradation of IdVfg @Vd=4V can be considered negligible. In this case, the Vth_P drift is underestimated since not taking into account the effect of Erase-induced aging on Program-efficiency decay, as similarly observed for the opposite case in Fig.4.14(b).

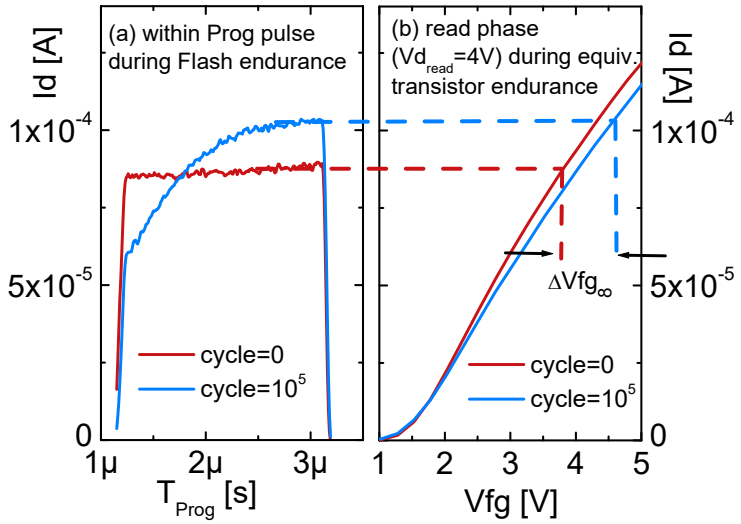


Figure 4.15: Direct measurement of Id(t) within the Program pulse during the Flash endurance (a) is compared with the degradation of IdVfg of the equivalent transistor measured @Vd=4V (b) in order to extract $\Delta Vfg_{\infty,P}$ at Vth_P. The stress condition B of Tab.4.1 is considered in the figure

For methodology validation, we can sum $\Delta Vfg_{\infty,P}/\alpha_G$, extracted as in Fig.4.15, and $\Delta Vth_{mos}/\alpha_G$, acquired with equivalent transistor stress, in order to reproduce Vth_P evolution. In Fig.4.16 this is shown for three different Program conditions, i.e. $Vfg_{eq,P}$ levels, and it is worth noting the perfect agreement with the experimental data measured during Flash cell endurance. This highlights the accuracy and pertinence of the methodology used.

It has to be pointed out that the three considered Program conditions might create different trap distributions, since a change of $Vfg_{eq,P}$ potential at fixed Vd_P induces a shift

of the defect position along Si/SiO_2 interface [26]. However, this does not affect the analysis done so far since being independent from the trap distribution barycenter.

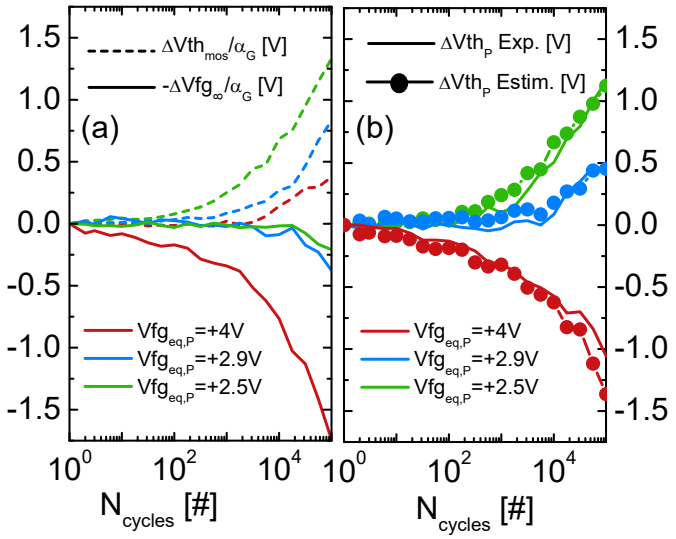


Figure 4.16: In (a) the two contributions of the total Vth_P shift (eq.4.7) for the conditions A,D,E of Tab.4.1 are extracted. In (b) the comparisons between the estimated ΔVth_P evolutions and the ones experimentally measured are shown

In order to address the real degradation of Program-efficiency, we should track the drift of the entire IfgVfg curve @Vd=4V along cycling. Thus, SPP experiment [29] has been carried out along P/E cycling. Looking at the experimental results Vth_{SPP} vs Program time in Fig.4.17, the roles of the two aging contributions are evident: the slow down of tunnel oxide current reduces the injection efficiency especially at the beginning of the Program time, whereas the degradation of the cell electrostatics increases Vth_{SPP} for long time since $Vth_{SPP} \propto Vth_{mos}$. In particular, as we already saw in the first chapter, using eqs.4.5,4.6, we get:

$$Vfg_{SPP}(t) = \alpha_{G,P} Vcg_{SPP} + \alpha_{D,P} Vd_P - \alpha_{D,lin} Vd_{lin} - \alpha_{G,lin} Vth_{SPP}(t) + Vth_{mos} \tag{4.12}$$

$$|Ifg_{SPP}(t)| = C_{ono} \frac{dVth_{SPP}(t)}{dt} \tag{4.13}$$

where the first three terms of eq.4.12 do not vary during SPP experiment (i.e. Vcg_{SPP}=const during this experiment). Looking at the Fig.4.17 for long Program time, in a first approximation dVth/dt does not seem to significantly vary with cycling. Indeed, assuming that Ifg degradation is negligible at low Vfg levels, as experimentally observed in Fig.4.4(a), dynamic simulations of SPP experiment in Fig.4.18(b) show that Vfg potential evolution is unchanged after a certain Program time. These simulations have been performed solving

the eq.4.2 with Vcg_{SPP}=const and considering $Ifg(Vfg) = -A \cdot \exp(B \cdot Vfg)$ (Fig.4.18(a)). It is easy to demonstrate that, after a certain time, the Vfg dynamics follows the law:

$$Vfg_{SPP}(t) = -\frac{1}{B} \cdot \ln\left(t_{SPP} \cdot \frac{A \cdot B}{C_T}\right) \tag{4.14}$$

which depends uniquely on IfgVfg characteristic, whose degradation is observed negligible for low Vfg values. Thus, after a certain time, $Vfg_{SPP}(t)$ evolution is unchanged and, in accordance with eq.4.12, $Vth_{SPP} \propto Vth_{mos}$, which justifies the cross point in Fig.4.17 and what previously said.

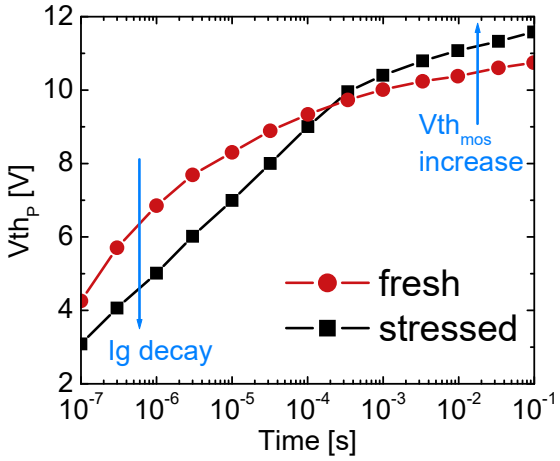


Figure 4.17: Step Pulse Program (SPP) performed on a fresh device and on an aged one ($N_{cycling} = 10^5$) with $Vth_{ini} = 2V$ and Vcg_{SPP}=8.5V.

Using the two eqs.4.12,4.13, and thus Vth_{mos} previously estimated, the IfgVfg characteristic can be acquired from the evolution in time of Vth_P during SPP experiment. This has been carried out during P/E cycling with an automatic experimental setup. The results of the two different Program stress conditions are shown in Fig.4.19. First of all, it is worth noting that for low Vfg levels the tunnel oxide current degradation is negligible, which is in accordance with the assumption previously done. Secondly, looking at the figure, the IfgVfg curve is observed to drift more at higher Vfg levels (i.e. Ifg), as already remarked in literature [27],[39]. This occurs since more Dit are filled as FG potential is increased. These traps impact the tunnel oxide current by degrading either the oxide transparency either the lateral transport during HC regime. In particular, the interface defects which effectively slow down the Program-efficiency can be divided into two types:

- Program-induced Dit at drain side which affect the barrier transparency at the injection region.

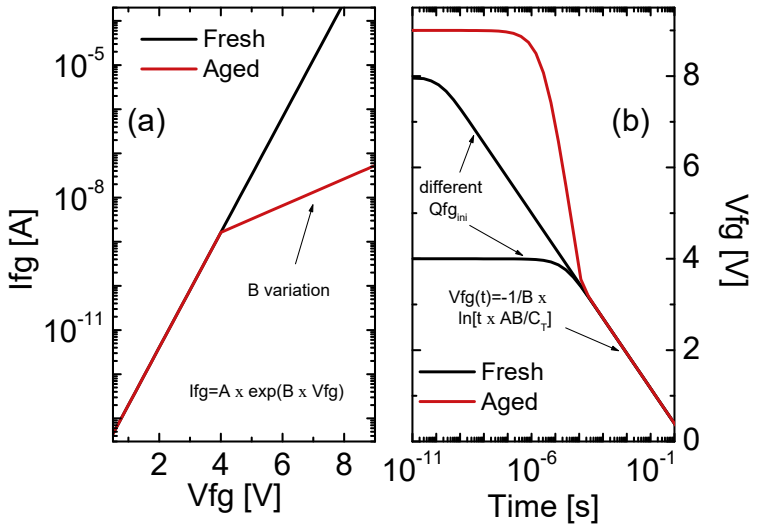


Figure 4.18: Simulations of the cell dynamics within the pulse. Starting with a general IfgVfg as in (a), Vfg(t) can be computed from eq.4.3 (b). It is worth noting that if assuming negligible degradation for low Ifg levels, as observed in Fig.4.4(a), the dynamics follows a universal law during SPP experiment, which does not depend on the initial Qfg charge.

- Erase-induced Dit at source side which impact the electrostatics, i.e. the inversion charge, of the device during the Program phase. Thus, the available charges are reduced and Ifg, as well as Id, degrades. As previously pointed out and observed in Fig.4.15(b), these traps reduce the trans-conductance dId/dVfg in saturation regime for an incremental electrostatic shift. Hence, a similar behavior is expected for IfgVfg curve.

In any case, considering lower $Vfg_{eq,P}$ patterns, i.e. lower Ifg_{∞} levels, $\Delta Vfg_{\infty,P}$ is surely inferior in absolute value, as experimental verified looking at Fig.4.19.

Anyway, it is fundamental to address the limiting factor of Program-efficiency slow down for modeling purpose and physical understanding. Indeed, knowing the relationship $Ifg = A \cdot 10^{B \cdot Vfg}$, we can estimate the $\Delta Vfg_{\infty,P}$ as:

$$\Delta Vfg_{\infty,P}(i) = \Delta\%B(i) \cdot Vfg_{eq,P} + K(i) \tag{4.15}$$

where $\Delta\%B(i)$ is the slope decay for IfgVfg at cycle=i, whereas K(i) represents the drift in amplitude. Now, if the driving aging mechanism relies on the Program-induced Dit, we should write $\Delta\%B \propto \Delta Vth_{mos,HC}$, whereas, if the electrostatic degradation is the limiting factor, we expect $\Delta\%B \propto Qinj^{d}$, since the generation of Erase-induced Dit at source side has been experimentally demonstrated not to depend on the oxide electric field. However, looking at Fig.4.19, it is evident that the second contribution must be the dominant one,

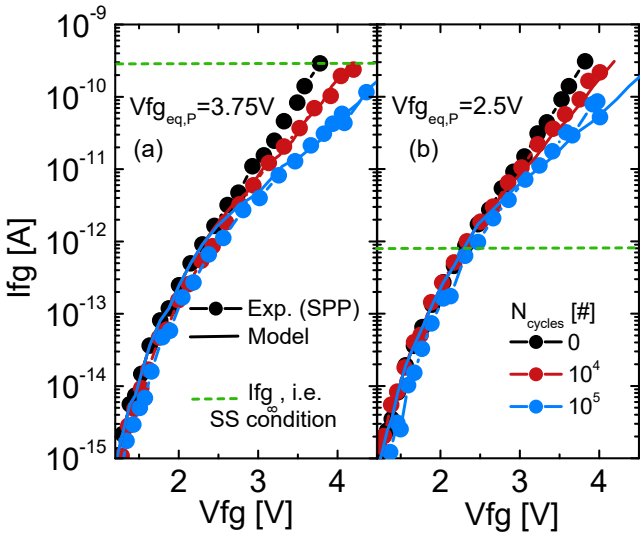


Figure 4.19: IfgVfg characteristic experimentally acquired during SPP experiment using eqs.4.12,4.13 along cycling for the conditions A,D in Tab.4.2. The SS condition (eq.4.4) is highlighted. The drift of IfgVfg curve has been modeled using the eq.4.16.

otherwise we should see an acceleration of the slope decay equal to eq.4.9 whenever $Vfg_{eq,P}$ is reduced from 3.755V to 2.5V. Indeed, the slope decay for the two conditions seem to be independent on $Vfg_{eq,P}$ level, thus, approximating the eq.4.15 in a Vfg drift from a fixed electrostatic point as in Fig.4.18(a), $\Delta Vfg_{\infty,P}$ can be computed as:

$$\Delta Vfg_{\infty,P}(Vfg_{eq,P}) = A \cdot Qinj^{d}(Vfg_{eq,P} - Vfg_0) \cdot (Vfg_{eq,P} > Vfg_0) \tag{4.16}$$

Applying this model on the IfgVfg curve degradation starting from the fresh state, the experimental results are well reproduced as shown in Fig.4.19, which highlights the validity of the model and of the physical interpretation. These results showed that, contrary to the common beliefs, the slow-down of Program-efficiency along cycling is not dominated by HC induced aging, but is mainly driven by interface traps generated by the Erase operation at source side. This emphasizes the role of the Program-Erase interplay, which strongly accelerates the PW closure, thus reducing the cell lifetime. On the other hand, these important results open the way to new possible technological solutions, which can be targeted at process level. For instance, a physical separation of FN/HC injection regions through a new cell architecture can envisaged.

4.3.3 Complete Flash cell aging modeling

After having explored the role and the physical origin of the different aging contributions within the PW evolution, the overall cell drift can now be addressed. In this section, we

analyze the Flash cell degradation under different endurance conditions in order to have a general overview of the device drift keeping a microscopic standpoint. This will help to properly optimize the ramp patterns for given technology restrictions.

Modeling parameters used in this subsection are listed in Tab.4.3. They have been extracted and refined as previously explained.

Static Degradation ΔVth_{mos}	eq.4.8	$A = 135V$	$p = 2.6$	$\alpha = 35$
		$n = 0.5$	$P_3 = 0.8$	$P_4 = -1.24$
E-efficiency decay $\Delta Vfg_{\infty,E}$	eq.4.10	$A = 0.45V$	$\nu = 0.36$	
P-efficiency decay $\Delta Vfg_{\infty,P}$	eq.4.16	$A = 0.85V$	$d = 0.4$	$Vfg_0 = 2.3$

Table 4.3: Parameters used for the endurance model proposed in this thesis for 40nm NOR Flash cell and based on eq.4.7. In eqs.4.8, 4.10, 4.16 all Vfg/Vd values considered are divided by 1V, Qinj by $1C/cm^2$ and t by 1s. The considered experiments are listed in Tab.4.2.

4.3.3.1 Effect of Program pattern

First of all, the effect of Program phase is explored. In Figs.4.20,4.21 the impact of the aging contributions is shown respectively on ΔVth_E and ΔVth_P considering different Program patterns, i.e. $Vfg_{eq,P}$. These results refer to the PW evolution of Fig.4.11(a). Using the models previously developed, the evolution of Vth at both states is reproduced with high accuracy, as shown in Figs.4.20(b),4.21(b).

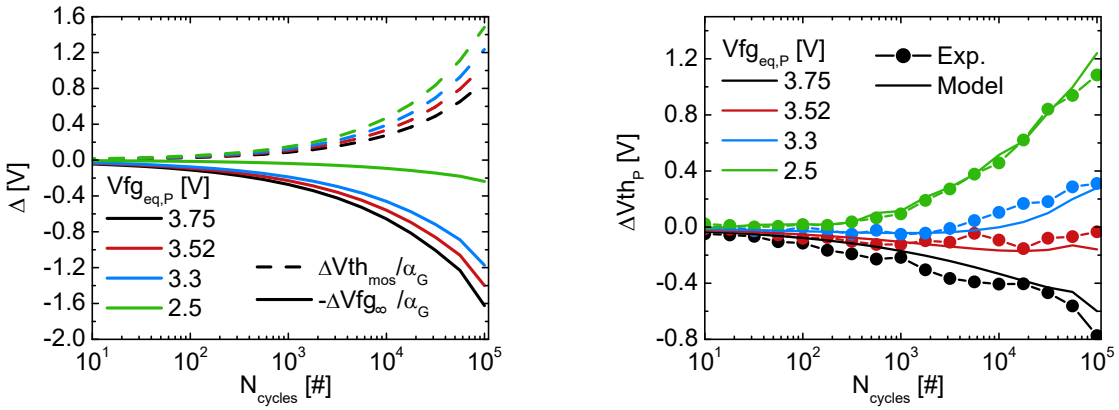
Firstly, looking at the threshold voltage drift at Program state, both contributions are observed to make ΔVth_P shifting towards positive values, i.e. $\Delta Vth_{mos} \nearrow$ and $|\Delta Vfg_{\infty,P}| \searrow$, whenever lower $Vfg_{eq,P}$ patterns are considered (Fig.4.20(a)). This is explained within the theory framework developed in the previous subsections:

- $\Delta Vth_{mos} \nearrow$ since it follows a power law in time, in accordance with HC-induced degradation and the Si-H bond activation energy dispersion, and with the fact that t_P has to be exponentially increased in order to maintain PW fixed. Whenever considering a $\Delta Vfg_{eq,P}$ reduction within the Program phase for a fixed injection charge, the acceleration factor in the static aging follows the law in eq.4.9.

- $|\Delta Vfg_{\infty,P}| \searrow$ since a lower concentration of Erase-induced interface defects at source side is integrated. Thus, although the degradation is almost the same looking at the IfgVfg curve, a reduction in $Vfg_{eq,P}$ level leads to lower aging sensitivity.

On the other hand, concerning Vth_E evolution, the situation can be simply analyzed looking at the aging contributions in Fig.4.21(a). Both static and Erase-efficiency degradations are accelerated whenever considering lower $Vfg_{eq,P}$ levels. However, the major factor

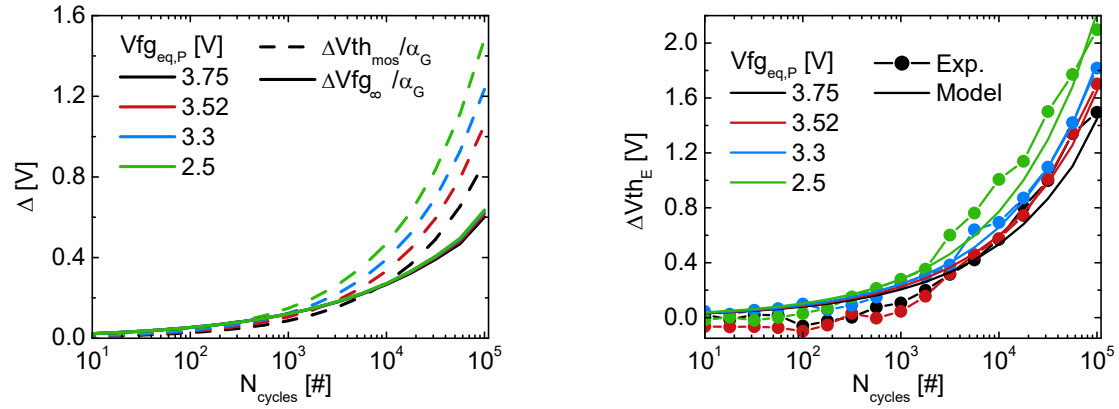
relies on the increase of ΔVth_{mos}, since the difference in the second contribution comes from a negligible increase of the injected charges.



(a) ΔVth_P contributions.

(b) ΔVth_P evolution: experiments vs model.

Figure 4.20: Vth_P evolution during P/E cycling changing the Program pattern: conditions A,G,H,D are considered for Tab.4.2. In (a) the static and E-efficiency decays are modeled following the eqs.4.8,4.16 respectively, whereas in (b) their sum is compared with the experimental data.



(a) ΔVth_E contributions.

(b) ΔVth_E evolution: experiments vs model.

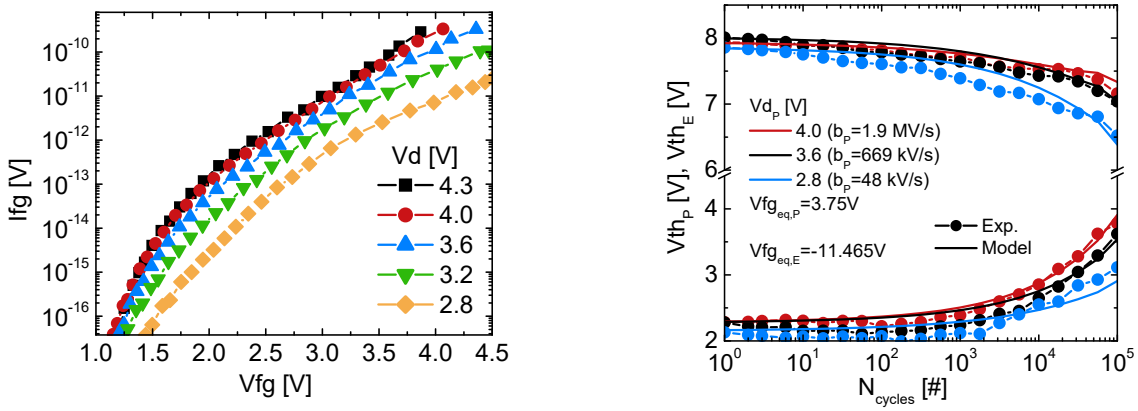
Figure 4.21: Vth_E evolution during P/E cycling changing the Program pattern: conditions A,G,H,D are considered for Tab.4.2. In (a) the static and E-efficiency decays are modeled following the eqs.4.8,4.10 respectively, whereas in (b) their sum is compared with the experimental data.

In addition, it has to be underlined that longer Program patterns, i.e. lower $Vfg_{eq,P}$ levels, improve the net PW drift, as shown in Fig.4.23(a). Indeed, in a first approximation, the

static degradation affects both $Vth_{E,P}$ levels in the same way, thus the PW, which is equal to $Vth_P - Vth_E$, is insensitive on Vth_{mos} aging. For this reason, the PW drift improves when lower $Vfg_{eq,P}$ levels are considered, since $|\Delta Vfg_{\infty,P}| \propto Vfg_{eq,P}$. However, it is worth noting that the cell lifetime, defined as the cycle number i where $Vth_{E/P,i} = (Vth_{E,0} + Vth_{P,0})/2 \mp \delta V$ for a certain δV, may be reduced since Vth_E drifts more. We will see that a proper pattern optimization, achievable within the theory framework developed in this thesis, helps to push the correct Flash cell working condition towards its physical limit.

One more degree of freedom has to be considered in the modeling of Program ramp pattern: the drain potential. Changing Vd_P level, a difference in the static aging is expected exclusively, i.e. both P/E efficiency losses are unchanged for given $Vfg_{eq,P}$ and Qinj.

First of all, the SPP technique has been performed at different Vd_P values, as shown in Fig.4.22(a). This experiment has been carried out at different Vd_P levels on the same device, thus the statistical dispersion is simply reduced to zero. Respect to $Vfg_{eq,P}$ case, the roles of IfgVfg and ΔVth_{mos} are inverted as Vd_P is modified. Indeed, an increase of Vd_P does not significantly vary the injection tunnel oxide current, whereas the degradation explodes, in accordance with eq.4.8. Looking at the $Vth_{E/P}$ drifts for different drain voltage levels in Fig.4.22(b), the model well reproduces the experimental data. This means that the physical interpretation is correct: the PW evolution is unchanged, as shown in Fig.4.23(b), whereas both $Vth_{E/P}$ are shifted towards higher values whenever considering higher Vd_P levels, in accordance with an acceleration of the static degradation.

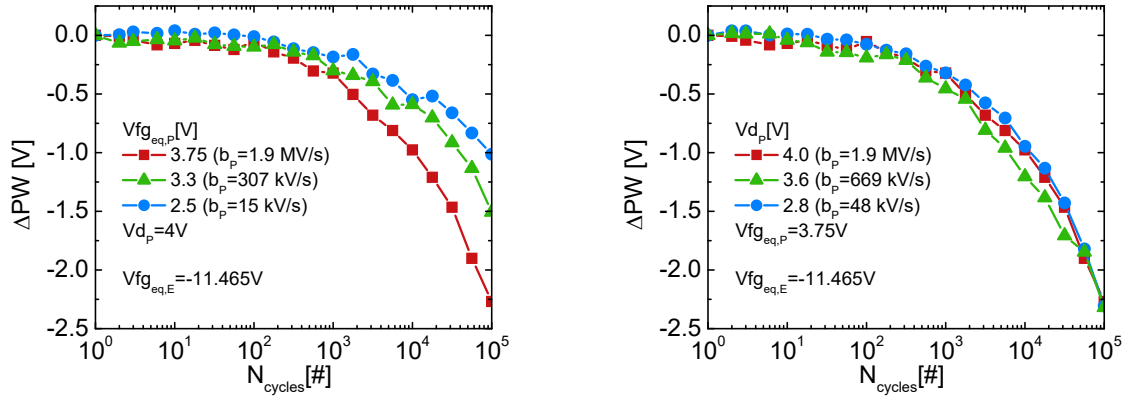


(a) IfgVfg acquired through SPP experiment [29] at different drain voltages.

(b) $Vth_{E/P}$ drifts for the conditions A,J,K of Tab.4.2.

Figure 4.22: Effect of the drain potential during the Program operation on the injection current (a) and on the PW evolution (b) during cell endurance experiment.

We can take advantage of the PW drift dependences shown in Fig.4.23, together with the considerations previously done, in order to improve the cell endurance performance. As



(a) Considering different $Vfg_{eq,P}$ levels.

(b) Considering different Vd_P levels.

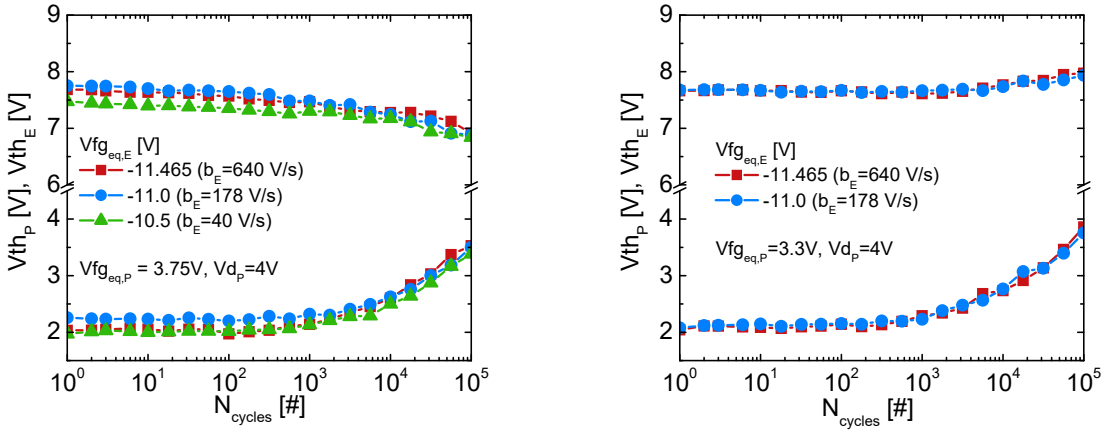
Figure 4.23: PW degradation evolution during P/E cycling changing the Program pattern: conditions A,H,D,J,K are considered for Tab.4.2.

remarked, when decreasing $Vfg_{eq,P}$ level, the net PW degradation lowers and Vth at Erase state drifts more, which can be the limiting factor for the cell lifetime. This issue can be overcome by lowering the drain potential: the PW drift is unchanged and ΔVth_E kinetics slows down. Although it is clear that reducing both $Vfg_{eq,P}$ and Vd_P voltages improves the cell lifetime, it is worth noting that their levels cannot be lowered as much as we want. Indeed, the Program phase duration and the energy dissipation exponentially increase with this reduction. For this reason, additional technology limitations are applied in order to guarantee also fast operations and/or low consumption.

4.3.3.2 Effect of Erase pattern

Concerning the cell evolution dependence on the Erase pattern, the situation is more simple. In Fig.4.11(b), the drifts of $Vth_{E/P}$ are shown for different Erase ramps. It is worth noting that considering lower $Vfg_{eq,E}$ patterns (so longer pulses in time), no difference is observed in terms of Vth_E and Vth_P drifts. This is simply explained considering the two contributions of eq.4.7. Applying separately Erase and Program patterns on equivalent transistors, ΔVth_{mos} has been observed to be Program-limited. Thus, modifying the Erase pattern, the electrostatic aging does not significantly vary, as verified in Fig.4.5(b). Concerning the Erase-efficiency aging, $\Delta Vfg_{\infty,E}$ is observed to just depend on the injected charges (see the previous chapter for further details). Hence, fixing the PW at the beginning of the endurance, Vth_E must evolve similarly, as experimentally verified in Fig.4.11(b). On the other hand, Vth_P cannot vary since depending uniquely on injected charges and static aging, which are unchanged. These clear observations are confirmed also at different Program

patterns, as shown in Fig.4.24.



(a) Conditions A,B,C of Tab.4.2 are shown.

(b) Conditions H,L of Tab.4.2 are shown.

Figure 4.24: $\mathrm{Vth_{P/E}}$ evolutions during endurance experiments at different Erase optimized ramp patterns.

However, as previously remarked, different $\mathrm{Vfg_{eq,E}}$ levels may affect TDDB and SILC related results. Indeed, they have been demonstrated to be physically linked with AHI mechanism, thus to the electric field of the cell during FN operation. Hence, lower $\mathrm{Vfg_{eq,E}}$ values should improve these reliability issues at fixed injection charges. Anyhow, this has not been faced in this thesis since out of our purpose.

4.3.3.3 Pattern optimization

The approach developed in this chapter allows not only to address the cell degradation from a physical perspective, but also gives clear guidelines for a correct optimization of P/E phases. This is fundamental for technology setting up, since a proper management of P/E operation pushes the memory cell towards its maximum performance gaining few decades of correct working condition in P/E cycling process. In order to quantify the improvement in device reliability, we define the cell failure condition as:

$$\min(\mathrm{Vth_{mid}} - \mathrm{Vth_E}, \mathrm{Vth_P} - \mathrm{Vth_{mid}}) = \delta\mathrm{V} \tag{4.17}$$

where $\mathrm{Vth_{mid}} = (\mathrm{Vth_E} + \mathrm{Vth_P})/2$ at fresh state and δV is the minimum safety margin condition which has to be proportional to the Vth statistical dispersion. The equation means that the cell state cannot be determined if $\mathrm{Vth} \in (\mathrm{Vth_{mid}} - \delta\mathrm{V}, \mathrm{Vth_{mid}} + \delta\mathrm{V})$. Taking advantage of the model developed in this section, the best optimized pattern can now be chosen. Indeed, we can write the margin, i.e. $\min(\mathrm{Vth_{mid}} - \mathrm{Vth_E}, \mathrm{Vth_P} - \mathrm{Vth_{mid}})$, function of $\mathrm{Vfg_{eq,P}}$

and Vd_P for a fixed PW at fresh condition. Simulations of lifetime are shown in Fig.4.25. It is worth noting the significant improvement that can be achieved if the Program pattern is correctly optimized. In particular, this can be addressed lowering both $Vfg_{eq,P}$ and Vd_P, as previously predicted.

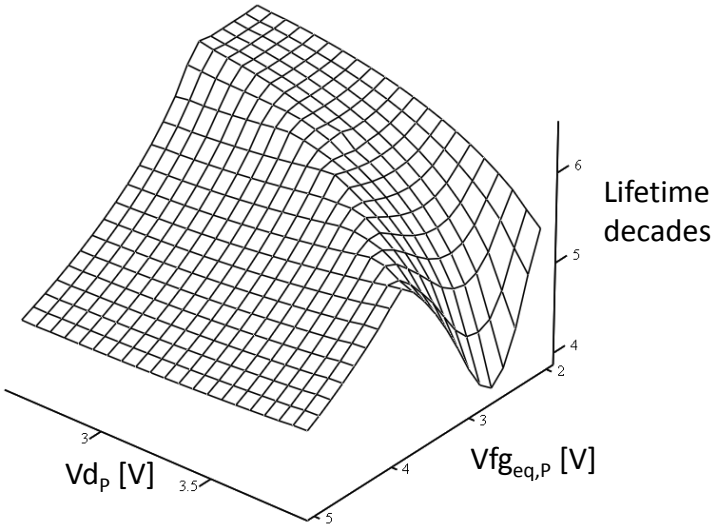


Figure 4.25: Lifetime simulation function of ($Vfg_{eq,P}$, Vd_P) using the model developed in this section and the parameters in Tab.4.3. An initial PW=5.75V and a minimum margin, i.e. δV in eq.4.17, equal to 0.7V have been considered in the model.

However, technology qualifications are usually made at fixed cycling number and, in addition, further restrictions are adopted, as for example Id_{max}, maximum energy dissipation, $t_{P,max}$ and so on. The related simulation results are shown in Fig.4.26 in a color chromatic scale for the cycle number 10^5. It is worth noting that for low $Vfg_{eq,P}$ values, a decrease of Vd_P improves the margin since reducing the static drift and consequently Vth_E. On the other hand, for higher $Vfg_{eq,P}$, the situation is opposite since the limiting cell state becomes the Program one, i.e. Vth_P. Concerning the applied restriction, we considered the maximum applicable Program pulse time. Imposing $t_{P,max} = 0.1$ms or $t_{P,max} = 10$ms, a trajectory is drawn on the graph and represents the minimum ($Vfg_{eq,P}$, Vd_P) that can be applied, since lowering one of the two voltage levels would increase this parameter. On that curve, the optimum ($Vfg_{eq,P}$, Vd_P) couple, which guarantees the maximum margin between $Vth_{E/P}$ and Vth_{mid}, can be chosen. This pattern is always the one which respects the following condition:

$$Vth_{mid} - Vth_E = Vth_P - Vth_{mid} \tag{4.18}$$

Then, taking advantage of the color levels, the major restriction in eq.4.17 has to be considered. If, for example, a minimum margin $\delta V = 2V$ is chosen, no ($Vfg_{eq,P}$, Vd_P) couples ensure $t_{P,max} = 0.1$ms too, whereas solutions can be easily found if $t_{P,max} = 10$ms.

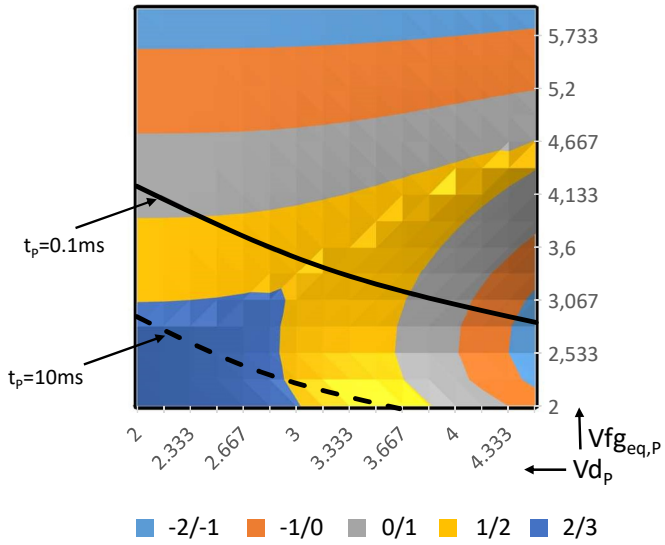


Figure 4.26: Chromatic scale for $\min(Vth_{mid} - Vth_E, Vth_P - Vth_{mid})$ function of ($Vfg_{eq,P}$, Vd_P) for a fixed PW=5.75V and cycle number 10^5. Each colour represents this parameter varying in a range defined as in the legend. The black lines represent an example of technology restriction: $t_{P,max} = 0.1$ms or $t_{P,max} = 10$ms. Lower ($Vfg_{eq,P}$, Vd_P) levels do not satisfy such constraints.

In addition, the model can be improved taking into account the increase of Vth statistical dispersion along P/E cycling [18]. Thus, the minimum margin can be written, in a first approximation, function of Qinj.

This simple example demonstrates that a proper modeling of the cell degradation leads to a correct P/E phase optimization which guarantees the technology constraints. Similar examples can be done imposing conditions as Id_{max} and maximum energy dissipation E_p. This has been done in [20] at the fresh state (see the Chapter 1 for details). However, both energy dissipation and maximum drain current have been demonstrated not to significantly vary along P/E endurance (Fig.6 of [20]). Thus, the same restrictions can be applied in Fig.4.26, as previously done for $t_{P,max}$. The intersections between the different trajectories drawn in the ($Vfg_{eq,P}$, Vd_P) space will determine the range of working conditions for the memory cell which guarantees the correct operation.

In order to extend the cell lifetime, a different approach can be considered. Instead of optimizing the P/E patterns at "time zero", it is possible to use an algorithm that modifies the P/E pattern shapes along cycling in order to increase the PW when necessary. In particular, if $Vth_P < Vm_P$ / $Vth_E > Vm_E$, for certain $Vm_{P/E}$ values, the Program / Erase phase can be extended respectively, as shown in Fig.4.27, maintaining the same ramp speed. Indeed, assuming that the SS condition is reached within the pulse, i.e. $Ifg = b_{P/E} \cdot C_{ono}$,

we force a $\Delta Qfg = \Delta Vth_{P/E} \cdot C_{ono}$ in order to push back $Vth_{P/E}$ to the fresh value. This is translated into an increase of the P/E pulse duration equal to:

$$\Delta t_{P/E} = \frac{\Delta Vth_{P/E}}{b_{P/E}} \tag{4.19}$$

This relationship translates the fact that, in SS condition, the increase in CG potential within the P/E pulse is directly transferred to an increment of $Vth_{P/E}$ evolution. If the SS condition is not satisfied, the cell is under-programmed/erased. However, as already remarked, the SS condition can always be considered reached as far as the P/E ramp pulse is properly modeled.

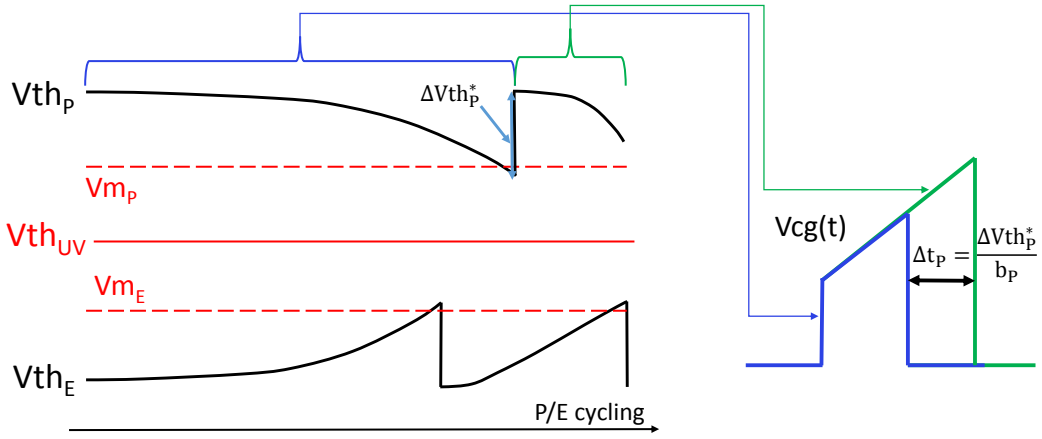


Figure 4.27: Schematic illustration of a different approach for improving the cell lifetime. The pulse duration is increased whenever the $Vth_{P/E}$ drift is significant, i.e. $Vth_P < Vm_P$ or $Vth_E > Vm_E$, as shown in the figure for the Program case.

Coming back to the optimization schematically shown in Fig.4.27, it is worth noting that the degradation kinetics increases along cycling whenever the safety condition is not respected, since the P/E phase duration is extended. However, the main problem of such an approach is that the Vcg level significantly increases and may reach values which are not sustainable, since the ramp speed has been kept fixed. In that case, the pulse can be simply extended in time fixing the CG level to the maximum available voltage. Thus, in order to guarantee that $Vth_{P/E}$ gets back to the fresh value, the degradation of the entire IfgVfg curve has to be taken into account. Since the evolution of such a curve has been previously modeled, it is possible to directly calculate the new $t_{P/E}$ from the simulation of Qfg(t) evolution within the pulse (solving the eq.4.2).

To conclude, all the work done in this chapter on Flash cell modeling for "constant Vfg" endurance can be extended for any pattern shape. Indeed, knowing the degradation

evolution of IfgVfg, the drift of Vth_{mos} and their relative dependences, the cell behavior along cycling can always be predicted. In particular, since the HC aging rate has been observed to be unaffected by the channel degradation itself, the static wear out at cycle=i becomes:

$$\Delta Vth_{mos}(i) = \sum_{k=0}^{k=i} \int_0^{t_{P,k}} A \cdot Vfg_{eq,P,k}(t)^p \cdot \exp\left(-\frac{\alpha}{Vd_{P,k}(t)}\right) \cdot n \cdot t^{n-1} dt \tag{4.20}$$

where $Vfg_{eq,P,k}(t)$ depends on the CG pattern and on the couple (IfgVfg, Vth_{mos}) at cycle k. Similar observations can be done for P/E efficiency losses. It is worth noting that, changing the pattern shape, the SS condition is different and depends on the degradation of IfgVfg characteristic. However, calculating the drift of this curve, the equivalent $\Delta Vth_{eff-loss}$ can be computed from the simulation of the cell dynamics within the pulse (eq.4.2).

4.4 Conclusions

The chapter explores the 40nm NOR Flash cell endurance under P/E cycling taking advantage of the studies previously done on HCD and FNS on equivalent transistors.

Firstly, the separation of cell aging contributions has been addressed during P/E cycling performed with CG optimized ramp patterns. The cell dynamics within Flash operations has been theoretically studied, highlighting that the steady-state condition reached by the tunnel oxide current depends uniquely on technology and ramp speed. Based on this study, the electrostatic and cell performance decays have been properly decoupled.

Anyhow, before starting to analyze the related experimental results, some insights on HCD-FNS interplay had to be done. It has been underlined that the static aging is mainly driven by HCD, whose kinetics does not significantly vary with the presence of FN-induced traps. On the other hand, the FNS kinetics resulted to be strongly reduced with the application of Program pulses, because of the additional HC-induced electron trapping and of the enhanced hole detrapping during the Program phase.

Then, taking advantage of the equivalent transistor results and separately extracting the impact of P/E efficiency losses, the assessment of aging contribution evolutions has been provided for different ramp speeds at fixed initial PW.

It has been observed that a slower ramp, i.e. longer pulse duration in order to provide same PW, during Erase does not influence the PW drift. Indeed, the Erase efficiency decay depends uniquely on the injected charge, in accordance with FN trapping mechanism. On the other hand, the electrostatics is not significantly modified since being mainly HC-induced.

Conversely, a significant increase of Vth in both states has been observed when lowering the ramp speed of CG pattern during the Program operation. This has been physically

explained considering separately the evolution of the degradation contributions. The electrostatics gets worse since the pulse time has been exponentially increased in order to maintain the same PW. On the other hand, the Program efficiency wear out improves because a lower amount of interface amphoteric traps impacts the tunnel oxide current whenever smaller FG potential levels are considered. Further insights have been presented considering also the Vd dependence during the Program operation, which results to interfare only with the cell electrostatics.

In addition, it has been underlined how the Program-induced defects strongly influence the Erase efficiency degradation and vice versa. In particular, HC-induced trapping strongly slows down the Erase efficiency operation, whereas the FN-induced interface defects at Source side results to be the main responsible of tunnel oxide current reduction during the Program phase.

Based on all these experimental extractions and considerations, together with the kinetics laws previously obtained for HCD and FNS processes, a complete physical-based model reproducing the PW evolution for different P/E patterns has been provided. This allows to properly optimize the Flash cell endurance finding the best trade-off, depending on application restrictions, between performance and reliability. Some examples of optimizations have been provided in the end of the chapter. In particular, it has been highlighted that a correct management of P/E operations can significantly increase the cell lifetime.

Bibliography

[1] S. Tehrani, J. Pak, M. Randolph, Y. Sun, S. Haddad, E. Maayan, Y. Betser, Advancement in Charge-Trap Flash Memory Technology, IEEE International Memory Workshop (IMW), p 9-12, May 2013.

[2] T. Guoqiao, H. Chauveau, Do Dormans, R. Verhaar, A quantitative study of endurance characteristics and its temperature dependence of embedded flash memories with 2T-FNFN NOR device architecture, IEEE Transactions on Device and Materials Reliability, v 7, n 2, p 304-9, June 2007.

[3] A. Scarpa, G. Tao, J. Dijkstra, F.G. Kuper, Tail bit implications in advanced 2 transistors-flash memory device reliability, Microelectronic Engineering, v 59, n 1-4, p 183-8, Nov. 2001.

[4] T. Ping-Hung, C.L. Kuei-Shu, L. Te-Chiang et al., Charge-trapping-type flash memory device with stacked high-k charge-trapping layer, IEEE Electron Device Letters, v 30, n 7, p 775-7, July 2009.

[5] Y. Hsin-Chiang, H. Tze-Hsiang, K. Fu-Hsiang et al., SONOS-type flash memory using an HfO2 as a charge trapping layer deposited by the sol-gel spin-coating method, IEEE Electron Device Letters, v 27, n 8, p 653-5, Aug. 2006.

[6] MirrorBit technology past, present and future: the on-going scaling of nitride-based flash memory, 21st Non-Volatile Semiconductor Memory Workshop. (IEEE Cat. No. 06EX1246), p 8, 2006.

[7] A. T. Wu, T. Y. Chan, P. K. Ko, and C. Hu, A novel high-speed, 5-volt programming EPROM structure, in IEDM Tech. Dig., p 584-7, 1986.

[8] W. Yu-Hsiung, W. Meng-Chyi, L. Chrong-Jong et al., An analytical programming model for the drain-coupling source-side injection split gate flash EEPROM, IEEE Transactions on Electron Devices, v 52, n 3, p 385-91, March 2005.

[9] C.Y.-S.Cho, M.-J.Chen, C.-F.Chen, P.Tuntasood, D.T.Fan, and T.Y.Liu, A novel self-aligned highly reliable sidewall split-gate flash memory, IEEE Trans. Electron Devices, v 53, n 3, p 465-73, March 2006.

[10] C. Papadas, G. Ghibaudo, G. Pananakakis, C. Riva, P. Ghezzi, Model for programming window degradation in FLOTOX EEPROM cells, IEEE Electron Device Letters, v 13, n 2, p 89-91, Feb. 1992.

[11] N. Mielke, H. Belgal, I. Kalastirsky, P. Kalavade, A. Kurtz, M. Qingru, N. Righos, J. Wu, Flash EEPROM Threshold Instabilities due to Charge Trapping During Program/Erase Cycling, IEEE Transactions on Device and Materials Reliability, v 4, n 3, p 335-44, September 2004.

[12] S. Aritome, R. Shirota, G. Hemink, T. Endoh, F. Masuoka, Reliability issues of flash memory cells, Proceedings of the IEEE, v 81, n 5, p 776-88, May 1993.

[13] Hong Yang, Hyunjae Kim, Sung-il Park, Jongseob Kim, et al. Reliability issues and models of sub-90nm NAND flash memory cells, 2006 8th International Conference on Solid-State and Integrated Circuit Technology (IEEE Cat. No. 06EX1294), 3 pp., 2006.

[14] D. Ielmini, A.S. Spinelli, M.A. Rigamonti, L. Lacaita, Modeling of SILC based on electron and hole tunneling. I. Transient effects, IEEE Transactions on Electron Devices, v 47, n 6, p 1258-65, June 2000.

[15] D. Ielmini, A.S. Spinelli, M.A. Rigamonti, L. Lacaita, Modeling of SILC based on electron and hole tunneling. II. Steady-state, IEEE Transactions on Electron Devices, v 47, n 6, p 1266-72, June 2000.

[16] E. Vianello, F. Driussi, D. Esseni, L. Selmi, et al, Explanation of SILC probability density distributions with nonuniform generation of traps in the tunnel oxide of flash memory arrays, Transactions on Electron Devices, v 54, n 8, p 1953-62, Aug. 2007.

[17] A. Chimenton, F. Irrera, and P. Olivo, Impact of Pulsed Operation on Performance and Reliability of Flash Memories, IEEE Transactions on Electron Devices, vol. 54, no. 6, June 2007.

[18] A. Chimenton, P. Olivo, Reliability of erasing operation in NOR-Flash memories, Microelectronics Reliability 45 (2005) 1094-1108.

[19] V. Della Marca, G. Just, A. Regnier, J.-L. Ogier, R. Simola, S. Niel, J. Postel-Pellerin, F. Lalande, L. Masoero, G. Molas, Push the flash floating gate memories toward the future low energy application, Solid-State Electronics, v 79, p 210-17, January 2013.

[20] J. Coignus, G. Torrente, A. Vernhet, S. Renard, D. Roy, G. Reimbold, Modelling of 1T-NOR Flash Operations for Consumption Optimization and Reliability Investigation, IEEE International Reliability Physics Symposium (IRPS), p PR-1 (4 pp.), 2016.

[21] A. Bravaix, C. Guerin, V. Huard, D. Roy, J. Roux, E. Vincent, Hot carrier acceleration factors for low power management in DC-AC stressed 40nm NMOS node at high temperature, International Reliability Physics Symposium (IRPS), 2009, pp. 531-546.

[22] S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, et al., Interface traps density-of-states as a vital component for hot-carrier degradation modeling, Microelectron. Reliab. 50, 1267-1272 (2010).

[23] C. Papadas, P. Morfouli, G. Ghibaudo, G. Pananakakis, Analysis of the trapping characteristics of silicon dioxide after Fowler-Nordheim degradation, Solid-State Electronics, v 34, n 12, p 1375-9, December 1991.

[24] Young-Bog Park, D.K. Schroder, Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM, IEEE Transactions on Electron Devices, v 45, n 6, p 1361-8, June 1998.

[25] G. Torrente, J. Coignus, A. Vernhet, J.L. Ogier, D. Roy, G. Ghibaudo, Physically-based evaluation of aging contributions in HC/FN-programmed 40nm NOR Flash technology, Microelectronics Reliability Journal, 2017, in press.

[26] G. Torrente, X.Federspiel, D.Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo, Hot Carrier Stress: aging modeling and analysis of defect location, IEEE International Reliability Physics Symposium (IRPS), p 5A-4 (6 pp.), 2016.

[27] J. Coignus, A. Vernhet, G. Torrente, S. Renard, D. Roy, , G. Reimbold, Relaxation-free Characterization of Flash Programming Dynamics along P-E Cycling, IEEE International Integrated Reliability Workshop Final Report (IIRW), p 119-21, October 2015.

[28] P. Pavan, R. Bez, P. Olivo, E. Zanoni, Flash memory cells-an overview, Proceedings of IEEE, v 85, n 8, p 1248-71, August 1997.

[29] A. Modelli, F. Gilardoni, D. Ielmini, and A. S. Spinelli, A new conduction mechanism for the anomalous cells in thin oxide flash EEPROMs, IEEE International Reliability Physics Symposium (IRPS), p 61-6, April-May 2001.

[30] A. Chimenton, P. Pellati, and P. Olivo, Constant Charge Erasing Scheme for Flash Memories, IEEE Transactions on Electron Devices, v 49, n 4, p 613-8, 2002.

[31] G. Torrente, X.Federspiel, D.Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo, Hot Carrier Stress modeling: from degradation kinetics to trap distribution evolution, IEEE International Integrated Reliability Workshop Final Report (IIRW), p 134-7, October 2015.

[32] D. Varghese, M.A. Alam, B. Weir, A generalized, IB-independent, physical HCI lifetime projection methodology based on universality of hot-carrier degradation, IEEE International Reliability Physics Symposium (IRPS), p 1091-4, May 2010.

[33] X. Federspiel, F. Cacho, D. Roy, Experimental characterization of the interactions between HCI, off-state and BTI degradation modes, IIRW p 133-6, 2011

[34] L. Perniola, S. Bernardini, G. Iannaccone, P. Masson, B. De Salvo, G. Ghibaudo, C. Gerardi, Analytical Model of the Effects of a Nonuniform Distribution of Stored Charge on the Electrical Characteristics of Discrete-Trap Nonvolatile Memories, IEEE Transactions on Nanotechnology, v 4, n 3, May 2005.

[35] G.Verma and N.Mielke, Reliability performance of ETOX based Flash memories, IEEE IIRPS Proc., pp.158-166, 1988

[36] G. Pananakakis, G. Ghibaudo, C. Papadas, E. Vincent, R. Kies, Generalized trapping kinetic model for the oxide degradation after Fowler-Nordheim uniform gate stress, Journal of Applied Physics, v 82, n 5, p 2548-57, September 1997.

[37] G. Torrente, J. Coignus, A. Vernhet, J.L. Ogier, D. Roy, G. Ghibaudo, Microscopic analysis of Erase-induced degradation in 40nm NOR Flash Technology, IEEE Transaction on Device and Materials Reliability (TDMR), v 16, n 4, p 597-603, Dec. 2016.

[38] Z. Sun, M. Zhang, Z. Huo, S. Li, Y. Yang, S. Qiu, H. Wu, M. Liu, Effect of Damage in Source and Drain on the Endurance of a 65-nm-Node NOR Flash Memory, IEEE Transactions on Electron Devices, v 60, n 12, p 3989-95, December 2013.

[39] S. S. Chung, C.-M. Yih, S.-M. Cheng, M.-S. Liang, A New Technique for Hot Carrier Reliability Evaluations of Flash Memory Cell After Long-Term Program/Erase Cycles, IEEE Transactions on Electron Devices, v 46, n 9, p 1883-9, September 1999.

[40] Synopsys, Zurich, Switzerland, Sentaurus device user guide, J-2014.09

General conclusions and perspectives

In this thesis we faced the reliability concerns for 40nm NOR Flash technology with particular attention to cell endurance, thus to the Programming Window (PW) evolution during Program/Erase (P/E) cycling. We highlighted the difficulty of accurately analyzing the $Vth_{P/E}$ drifts from a microscopic standpoint, since coming from different cell aging contributions, i.e. electrostatic drift and loss of P/E operation efficiency, and from different degradation mechanisms, such as Hot Carrier Degradation (HCD) and Fowler-Nordheim Stress (FNS). We overcame such difficulties thanks to an advanced experimental setup and ad-hoc equivalent transistor structures. The complexity of the problem has been faced, highlighting the need of building a simple compact model having the capability to be suitable for industry without losing the physical meaning and thus the accuracy.

To achieve this goal, we firstly showed how to properly characterize and model the Flash cell underlining the importance in the extraction accuracy of cell characteristics and the novelty of these approaches. Then, we presented the advanced experimental setup used in this thesis, which allows customizable electrical pattern shapes for Control Gate (CG) electrode during P/E phases, together with delay-free system for keeping P/E cycling tests within reasonable durations. Taking advantage of the cell model and of such an experimental setup, we optimized P/E operations in order to have constant equivalent Floating Gate (FG) patterns. Profiting from this approach, the cell degradation during P/E can be easily studied on Flash transistors, which are equivalent structures built by shorting CG and FG of the cell at process level. In particular, we showed how to reproduce the cell wear out on these devices with the respective square patterns. This allows to address both the total cell electrostatic aging, decoupling it from P/E efficiency losses, and the HCD/FNS impacts distinctly. Based on these considerations, these two aging mechanisms have been studied separately on equivalent MOSFET structures.

Concerning HCD, we firstly optimized the parameter extraction methodology and studied the parameter sensitivity respect to the microscopic degradation. This gave the possibility to assess the wear out localization and thus to properly calibrate our kinetic model. Such a TCAD Spherical Harmonic Expansion (SHE) based model computes the aging kinetics at each interface location considering uniquely the Single Vibrational Excitation (SVE) mode,

which has been widely demonstrated to well describe the device wear out. In addition, physical insights on trap distribution evolution at different stress conditions have been proposed.

Concerning FNS, a complete electrical characterization and aging extraction have been performed, taking advantage of different transistor structures. The kinetics of interface traps, trapped electrons and trapped holes have been extracted and modeled. It has been underlined the effects of such degradation mechanisms on Flash cell characteristics. In particular, the electron trapping slows down the Erase efficiency, whereas the overall electrostatic drift can be considered negligible.

Finally, considering the Flash cell endurance, HCD and FNS have been considered together, highlighting the interplay between them. The cell static aging is shown to be HCD-driven, whereas the P/E efficiency losses are enhanced as E/P-induced defects are considered respectively. In particular, Erase-induced defects at Source side are shown to be the limiting mechanism factor for the Hot Carrier gate current decay. On the other hand, the electrons injected in the oxide during Program operation are shown to represent around 40% of the total Erase efficiency loss. Using all this information, together with the aging kinetic laws accurately refined for Flash cells, a complete model for the cell endurance during P/E cycling is proposed, giving important guidelines for technological qualification and optimization.

Such a work represents a solid brick on which future studies can be inspired. Indeed, the approach proposed in this thesis allows to perfectly control general cell performance, such as power consumption, P/E pulse durations, maximum drain current and so on, together with reliability concerns having a microscopic standpoint. In particular, taking advantage of an advanced experimental setup and an accurate cell model, it has been shown how to separate the cell electrostatic drift from the P/E efficient loss and, on the other hand, to quantify the impacts of HCD and FNS on each contribution. This represents an important result for the comprehension of cell endurance and allows a physical-based technology optimization.

It has to be pointed out that all the work done in this thesis has been performed referring to "constant Vfg" endurance. However, it can be easily extended for any CG pattern shape and duration. Indeed, knowing the degradation evolution of P/E Ifg-Vfg characteristics, the drift of Vth_{mos} and their relative dependences, the cell behavior along cycling can be always predicted by simulating the device dynamics within P/E phases.

Another perspective based on this work may be the application of such a methodology on different device geometry characteristics, such as length and oxide thickness. This would allow to optimize the cell geometry, also finding new technological solutions. Similarly, the doping concentration and profile may completely change the device wear out picture, since mainly impacting HC-induced degradation. Indeed, the relationship between the Program gate current and the related HCD can be completely different. Thus, such a dependence represents a very interesting feature to be explored.

An additional insight is the extension of the cell endurance model also at different temperatures. This represents a difficult and important challenging task for NOR Flash Technology, since a temperature increase may completely change the physical degradation regime during P/E cycling, especially for HCD, which may be enhanced by Self-Heating (SH) effect and Multi Vibrational Excitation (MVE) mode.

Moreover, the physical-based approach developed in this thesis can be completed with the complementary study on data retention and device breakdown. In particular, during the thesis, we clearly saw that the effect of AHI-induced traps is negligible, a part of the slight effect on Vth_E in the beginning of the endurance. However, as already remarked, such a degradation mechanism is known to be the main responsible for the generation of defects affecting SILC and Time-Dependent gate Dielectric Breakdown (TDDB) processes. Thanks to the FN aging analysis performed in this thesis, it is possible to link AHI kinetic laws with the characteristics of such processes, in order to complete the physical based model for device degradation. Also in this case, the temperature dependence is fundamental, since accelerated tests may be necessary. In any case, such an approach would allow to handle a unique model describing Flash cell characteristics, $Vth_{E/P}$ evolutions during P/E cycling and induced SILC always having a microscopic perspective. This gives the possibility to optimize the device operations finding the best trade-off between performance, endurance and data retention.

List of publications

In this section, the publications related to this thesis and published during the PhD are listed.

Conference proceedings

- **G. Torrente**, J. Coignus, S. Renard, A. Vernhet, G. Reimbold, D. Roy, G. Ghibaudo

  Physically-based extraction methodology for accurate MOSFET degradation assessment

  Proc. European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF) 2015.

- **G. Torrente**, X. Federspiel, D. Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo

  Hot Carrier Stress modeling: from degradation to kinetics to trap distribution evolution

  Proc. IEEE International Integrated Reliability Workshop (IIRW), p 134-7 (Oct. 2015).

- J. Coignus, A. Vernhet, **G. Torrente**, G. Reimbold, S. Renard, D. Roy

  Relaxation-free Characterization of Flash Programming Dynamics along P-E Cycling

  Proc. IEEE International Integrated Reliability Workshop (IIRW), 119-21 (Oct. 2015).

- C. Tavernier, F. G. Pereira, O. Nier, D. Rideau, F. Monsieur, **G. Torrente**, M. Haond, et al.

  TCAD modeling challenges for 14nm Fully Depleted SOI technology performance assessment

  Proc. IEEE International Conference on Simulation of Semiconductor Processes and Devices (SISPAD-2015), p 4-7.

- **G. Torrente**, X. Federspiel, D. Rideau, F. Monsieur, C. Tavernier, J. Coignus, D. Roy, G. Ghibaudo

  Hot Carrier Stress: aging modeling and analysis of defect location

  Proc. IEEE International Reliability Physics Symposium (IRPS), p 5A-4 (6 pp.), 2016.

- J. Coignus, **G. Torrente**, A. Vernhet, S. Renard, D. Roy, G. Reimbold

  Modelling of 1T-NOR Flash Operations for Consumption Optimization and Reliability Investigation

  Proc. IEEE International Reliability Physics Symposium (IRPS), p PR-1 (4 pp.), 2016.

- X. Federspiel, **G. Torrente**, W. Arfaoui, F. Cacho, V. Huard

  Temperature sense effect in HCI self-heating deconvolution

  Proc. IEEE International Reliability Physics Symposium (IRPS), p XT-09-1 (4 pp.), 2016.

Journal papers

- **G. Torrente**, J. Coignus, S. Renard, A. Vernhet, G. Reimbold, D. Roy, G. Ghibaudo,

  Physically-based extraction methodology for accurate MOSFET degradation assessment

  Microelectronics Reliability 55 (9-10) (August 2015) p 1417-21.

- **G. Torrente**, J. Coignus, A. Vernhet, J. L. Ogier, D. Roy, G. Ghibaudo

  Microscopic analysis of Erase-induced degradation in 40nm NOR Flash Technology

  IEEE Transaction on Device and Materials Reliability (TDMR), v 16, n 4, p 597-603, Dec. 2016.

- **G. Torrente**, J. Coignus, A. Vernhet, J. L. Ogier, D. Roy, G. Ghibaudo

  Physically-based evaluation of aging contributions in HC/FN-programmed 40nm NOR Flash technology

  Microelectronics Reliability Journal, 2017, in press.

Annex

Résumé en Français

La technologie Flash représente aujourd'hui la mémoire non-volatile de référence dans plusieurs applications électroniques. Néanmoins, la réduction d'échelle (ou "scaling") des cellules Flash conventionnelles fait aujourd'hui face à plusieurs limitations et un effort d'optimisation accru est nécessaire pour atteindre de meilleures performances, notamment en termes de fiabilité (rétention de données et tenue en endurance). La rétention de l'information stockée a ainsi fait l'objet de nombreuses études dans la littérature, aboutissant à une bonne compréhension et une modélisation précise des phénomènes de Stress Induced Leakage Current (SILC). En revanche, une description précise et microscopique des mécanismes de dégradation de cellules Flash en cours d'endurance Programmation/Effacement (P/E) reste manquante. Notamment dans le cas des technologies Flash de type NOR, dont la nature 2D des mécanismes de programmation, et donc de dégradation, complexifie l'analyse, la compréhension et la modélisation de la perte de performances en cours d'endurance. En effet, la cellule NOR est programmée en exploitant le phénomène de porteurs chauds (Hot Carrier) et elle est effacée grâce à l'effet tunnel exercé à travers l'oxyde de grille (régime Fowler-Nordheim). Ces différents régimes d'opération conduisent à des mécanismes de dégradation différant par la nature des défauts ainsi induits et par leur localisation.

Cette thèse se propose d'investiguer le vieillissement en endurance de la technologie embarquée NOR Flash 40nm produite à STMicroelectronics et de la modéliser en prenant en compte l'ensemble des mécanismes physiques microscopiques impliqués.

Pour atteindre ce résultat, il a été nécessaire, dans un premier temps, d'extraire les caractéristiques fondamentales de la cellule Flash. Grâce à l'utilisation de Transistors Equivalents (TrEq) et de méthodes de caractérisation électrique spécifiquement développées, les facteurs de couplage et les caractéristiques courant tunnel - tension de la grille flottante (i.e. Floating Gate, FG) ont été obtenus avec précision. En outre, un modèle TCAD dédié a été calibré dans le but de fournir une compréhension physique poussée des mécanismes de vieillissement.

De même, un banc de test spécifique a été mis en place pour étudier et mesurer la dégradation des dispositifs sans être impacté par d'éventuels effets de relaxation électrostatique (mesures rapides sans délais entre les opérations P/E). Grâce à cette approche expérimentale innovante et les caractéristiques de la cellule précédemment acquises, nous avons développé un protocole de Programmation optimisé, qui permet d'obtenir un potentiel de grille flottante constant pendant les phases d'écriture et d'effacement grâce à l'application d'impulsions de tension spécifiques, et modélisées *a priori*, sur la grille de contrôle (i.e. Control Gate, CG). Cette méthode permet ainsi de reproduire le stress électrique des opérations Programmation/Effacement directement sur TrEq à l'aide d'impulsions de tension constante (i.e. reproduisant ainsi le stress subi par la grille flottante de cellules Flash soumises à impulsions optimisées sur la grille de contrôle). En conséquence, cette approche permet de décorréler l'impact sur le vieillissement en endurance des deux opérations : dégradation par Fowler-Nordheim (i.e. Fowler-Nordheim Stress, FNS) et par porteurs chauds (i.e. Hot Carrier Degradation, HCD).

Grâce à ces approches expérimentales innovantes, l'essentiel du travail de thèse a consisté à étudier au niveau microscopique les deux phénomènes physiques de dégradation, i.e. HCD et FNS, sur des structures TrEq et les contextualiser ensuite dans un environnement Flash, en incluant les possibles interactions entre les mécanismes.

Une analyse poussée du vieillissement HCD a tout d'abord été réalisée, en s'attachant à extraire les paramètres les plus significatifs décrivant correctement le vieillissement du TrEq. La relation entre la dégradation microscopique et la dérive des paramètres extraits a été mise en lumière grâce à des simulations TCAD. En particulier, ces simulations ont permis de modéliser la sensibilité de ces paramètres par rapport à la concentration et localisation des défauts d'interface créés lors du processus « porteurs chauds ». Avec cette approche, nous avons également souligné que les techniques conventionnelles d'extraction ne sont pas satisfaisantes si appliquées à un dispositif

dégradé par Hot-Carrier. En particulier, nous avons souligné la difficulté de découpler correctement la dégradation du transport des électrons et la dégradation électrostatique lors de la phase de lecture.

Grâce à cette analyse, nous avons développé une nouvelle méthodologie qui permet d'extraire précisément les deux composantes, grâce à deux métriques aisément extractibles : Id_{ON} et Vth_{CC}. Nous avons également montré comme l'évolution de ces deux paramètres renseigne précisément la quantité d'états d'interface et leur localisation.

La deuxième partie relative au HCD a été dédiée à la modélisation de la cinétique de rupture des liaisons d'interface Si-H. Après avoir précisément décrit l'état de l'art, nous avons utilisé le model « Spherical Armonic Expansion » (SHE), implémenté dans le logiciel TCAD de Synopsys, qui résout l'équation du transport de Boltzmann en prenant en compte uniquement le mode « classique » de dégradation SVE (i.e. Single Vibrational Mode). Le modèle a été calibré en considérant différentes conditions de stress (i.e. différentes tensions de grille et de drain), différentes configurations de lecture et plusieurs géométries. Nous avons également analysé les paramètres du modèle en soulignant la nécessité d'une valeur plutôt élevée pour l'énergie de liaison Si-H, qui nous confirme l'hypothèse faite sur le mode de dégradation considéré. En outre, grâce à l'évolution du rapport Id_{ON} et Vth_{CC}, nous avons pu décrire qualitativement la dégradation le long du canal durant le stress et nous avons montré une perte de cinétique de dégradation en commençant par les régions du canal éloignée du drain, en cohérence avec les simulations.

Grâce à l'ensemble de ces résultats et à la compréhension physique associée, un modèle compact simplifié a été proposé.

La troisième partie du manuscrit est dédiée à la dégradation par Fowler-Nordheim (FNS) qui a été appliquée sur les structures TrEq. L'analyse a été conduite en séparant la dégradation statique et celle du transport tunnel, qui correspondent respectivement à la dégradation électrostatique de la cellule et la perte d'efficacité lors de la phase d'effacement. Le premier volet a été adressé à l'aide de stress électriques appliqués à des dispositifs TrEqs et d'analyse de la dérive des caractéristiques en régime linéaire. Concernant l'évolution du courant tunnel, des matrices de 0.5 millions de TrEqs en parallèle ont été considérées pour que le courant de grille soit directement lu durant le FNS.

Les signatures des différents défauts crées pendant le régime FNS ont été soulignées : électrons piégés dans l'oxyde de grille, trous également piégés dans l'oxyde tunnel (à travers le mécanisme de Anode Hole Injection, i.e. AHI) et états d'interface. Les cinétiques associées ont été extraites avec le support de modèles physiques dédiés en soulignant l'adéquation d'une loi en puissance pour le piégeage des porteurs, en cohérence avec la littérature. Nous avons également montré que l'exposant de puissance pour les trous est plus faible, expliquant ainsi la croissance du courant tunnel au début du régime FNS et donc le sur-effacement observable sur cellule Flash à bas cycles. L'approche précédemment citée a également montré qu'une majorité des trous sont piégés proche de l'interface Si/SiO_2 et, donc, sont responsables de la dégradation électrostatique du TrEq, conjointement aux états d'interface. Par ailleurs, les électrons sont piégés proches du PolySilicium (grille du TrEq, grille flottante Flash) et, par conséquence, représentent la cause de la chute du courant de tunnel et de l'efficacité d'effacement de la cellule.

Grâce aux études réalisées sur la dégradation par porteurs chauds et par FN du TrEq, la dernière partie du manuscrit explore l'endurance de la cellule NOR 40nm et un modèle de vieillissement en cours d'endurance est proposé.

Avec cet objectif, il a été d'abord nécessaire étudier théoriquement la dynamique de la cellule pendant les phases d'écriture et d'effacement. Cela a permis de séparer les composants de la dégradation pour un niveau de Vth (tension de seuil de la cellule) donné. Avec la supposition d'atteindre toujours la condition de "Steady-State" du dispositif, on a montré que la dérive du Vth, à l'état de programmation ou d'effacement, peut être mathématiquement écrite comme la somme de la dégradation électrostatique et d'un composant qui est strictement lié à la perte d'efficacité de l'opération en question. La première partie est décrite par la dérive du TrEq et la deuxième par la dégradation du courant tunnel de la cellule (i.e. Ifg-Vfg).

Pour fournir un modèle fiable d'endurance, l'interaction des deux mécanismes de vieillissement HCD et FNS a été étudiée. Nous avons ainsi montré que la dégradation statique de la cellule Flash a pour origine le processus de HCD (donc de Programmation) et que le rôle du FNS s'avère être négligeable.

Grâce aux résultats obtenus sur TrEq et à l'extraction des pertes d'efficacité des opérations P/E, les évolutions des différents mécanismes ont été extraites avec précision en considérant différentes vitesses de rampe de CG (i.e. différents niveaux constants de FG) et à fenêtre de programmation

PW (i.e. Vth_P-Vth_E) fixée. Nous avons montré que le niveau de FG pendant l'effacement n'a aucune influence sur la dégradation en endurance, en cohérence avec notre étude : la perte d'efficacité pendant l'effacement dépend des charges injectées et la dégradation statique est uniquement liée au HCD, donc à la phase de programmation. Par ailleurs, une augmentation significative a été observée pour les deux états de Vth lorsqu'une rampe lente (i.e. bas niveau de FG) est considérée. Tout cela a pu être facilement expliqué grâce à la caractérisation électrique et modélisation dédiées développées le long de la thèse. En effet, l'électrostatique dégrade plus car le temps doit être exponentiellement augmenté pour garder la même PW, alors que la dépendance de la dégradation est une puissance de Vfg. Concernant la perte d'efficacité d'écriture, il y a une nette amélioration car une quantité mineure d'état d'interface impacte le courant de tunnel. Observations supplémentaires ont été également faites pour ce qui concerne la dépendance à Vd pendant la phase d'écriture.

Au cours de ce chapitre, nous avons souligné que les deux mécanismes de dégradation interagissent fortement entre eux. Les électrons piégés proche du PolySilicium pendant la phase d'écriture provoquent une perte d'efficacité supplémentaire pendant la phase d'effacement d'environ 40%. De l'autre côté, les états d'interface crées proche de la Source pendant l'effacement ont été identifiés comme responsables de la perte d'efficacité de programmation.

Grâce à la description microscopique et exhaustive établie, un modèle physique dédié a été développé pour prédire l'évolution des performances des mémoires Flash NOR en cours d'endurance, et ce quels que soient les schémas de programmation et d'effacement. Enfin, nous avons montré en quoi l'application de ce modèle permet de définir des conditions de programmation optimales conduisant en une amélioration de la durée de vie des cellules NOR Flash considérées.

Acknowledgements

First and foremost I want to thank my chefs David Roy (STMicroelectronics, Crolles) and Jean-Coignus (CEA, Grenoble). The first one gave me the direction to follow and drove me towards the "truth" as a real mentor, especially in the beginning of my PhD, when I really needed such a support. The second one taught me everything like a second mother. I am grateful for all his contributions, ideas, supports he gave me along all my PhD period in both bad and good times. I will never forget all those days we spent drawing, writing equations, scrawling energy bands everywhere, sketching on paper all kind of ideas!! It was just amazing. I miss all those moments. He supported and helped me as real friend not only from a technical point of view, but also humanely. I must express my thanks also to my "orchestra director" Gérard Ghibaudo (IMEP-LAHC, Grenoble). It was an honor to have a thesis director like him. He was just there; ready to help me, whenever and whatever I needed.

I am grateful for the support of all reliability team in ST: Xavier Federspiel, Jean-Luc Ogier, David Ney, Sophie Renard, Jerome Goy, and all the PhD students, Damien, Tsha, JB, Ajith and in particular Giulio Marti. He showed me the fantastic world of games: it doesn't matter which kind of game, he already played it!! I must have to acknowledge all the other PhD colleagues in STMicroelectronics: Bastien, Darayus, Boris, Andrej, Nils, Romu etc.. How many good moments we spent in ST restaurant ah?

I must thank also the entire electrical characterization group in CEA-Leti, in particular Alexandre (I owe you everything man!!), Antoine, Push, Clement, Thomas and all the others. How can I forget all the soccer 5 matches we played? Especially Italy vs France matches!!! (NB: we are stronger btw).

A special “grazie amico” to Blend Mohamad, who became a second brother to me and made me discover the best experience a human being can make in his life: eating kebab.

Then, I have to thank my parents, who support me in each moment of my life, whatever I do, wherever I am. And I have also to congratulate them, since they finally learned how to use Skype and Whatsapp!!!

I cannot forget in this list Svetlana, my love. She is the first person I met in France and, then, she became the most important person in my entire life. Thank you for everything my “piccola”.

Finally, I cannot exclude one last important and fundamental person to mention. Myself. I came here in France without knowing anybody, I could not speak French, I could barely speak English. Here in Grenoble I met just amazing people; I found a home and a second family.

Grazie a tutti ragazzi, grazie davvero di tutto.