Numerical and Compact Modeling of Embedded Flash Memory Devices Targeted for IC Design

THÈSE N^{O} 5274 (2012)

PRÉSENTÉE LE 23 JANVIER 2012

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE SYSTÈMES MICROÉLECTRONIQUES
PROGRAMME DOCTORAL EN MICROSYSTÈMES ET MICROÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Davide GARETTO

acceptée sur proposition du jury:

Dr C. Dehollain, présidente du jury
Prof. Y. Leblebici, Dr W. Clark, directeurs de thèse
Prof. T. Grasser, rapporteur
Prof. L. Larcher, rapporteur
Dr J.-M. Sallese, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2012

Abstract

In a semiconductor market dominated by portable consumer applications, embedded flash memory technology has experienced a rapid diffusion. It is now considered the preferred solid-state memory solution for its non–volatile characteristics, high read and write speeds and scalability properties. As technology scales down in the nanometer range, new accurate physical tools should be made available to circuit designers, to support the development and optimization of high–voltage circuit blocks.

A surface potential-based model for the flash memory cell has been developed with the purpose of providing a comprehensive physical understanding of the device operation, suitable for performance optimization in memory circuit design. An accurate validation methodology also takes into account charge balance effects on the isolated floating gate node and parasitic couplings inside and between the memory cells. The compact model supports DC, AC and transient analyses, including program/erase bias scalability, temperature effects, process corners and statistical variations. The results have been compared to Technology Computer-Aided Design (TCAD) simulations demonstrating that short channel effects, overlap capacitances and velocity saturation dominate over the intrinsic behaviour of the cell in ultrascaled devices. The approach includes drain disturb and memory endurance degradation models due to oxide aging. These effects are becoming dominant in ultrascaled devices. The model has been implemented using the Verilog-A language for portability into common circuit simulators. Validation has been performed on measurement results of test structures integrated in a 65nm derivative NOR CMOS technology.

The compact model development has been based on a rigorous modeling approach combining conventional TCAD simulation tools with physically-based analyses. A new TCAD tool has been proposed for the investigation of advanced quantum effects, band structure models, quantum tunneling and multiphonon-assisted charge trapping effects in dielectrics. The effects of charge trapping in oxide layers and Si/SiO_2 interfaces have been studied, specifically focusing on flash technology, where high voltage biases represent a major issue for dielectric degradation. A multiphonon-assisted model has been coupled with a Poisson-Schrödinger quantum solver. A novel impedance calculation method has been applied to the analysis of DC and AC MOS characteristics. This approach permits the physical modeling of trap filling, frequency response and device electrostatics. Transient effects of trap filling and trap-assisted tunneling through the gate have also been investigated. The adoption of such a multilevel approach permits to apply the methodology to flash memory cells. This enabled the investigation of the role of defects on electrostatics and program/erase efficiency reduction.

The flash compact model has been applied in a process development and IC memory design perspective. Technology development requires a profound understanding of trade-offs in flash devices, which affect DC, transient and long–term performances. The design, integration and characterization of a 40KB memory sector for smart card applications has been performed to demonstrate the capabilities of the compact model.

Keywords: semiconductor device modeling, compact modeling, embedded flash memory, Poisson-Schrödinger modeling, memory endurance, dielectric degradation, interface & oxide defects, multiphonon–assisted charge trapping.

Résumé

Dans un marché des semi-conducteurs dominé par les applications portables grand public, la technologie embarquée de mémoire flash a connu une diffusion rapide. Elle est devenue la solution de mémoire à état solide préférée pour ses propriétés de non-volatilité, ses temps d'accès et d'écriture très réduits et ses avantages en terme d'intégration. Alors que les dimensions des dispositifs électroniques approchent le seuil du nanomètre, des nouveaux outils des simulations doivent être mis à disposition des concepteurs de circuits pour le développement et l'optimisation de blocs à haute - tension environnants.

Un modèle compact à potentiel de surface de la cellule de mémoire a été développé afin de permettre une compréhension complète du fonctionnement physique du dispositif. Une méthodologie de validation précise inclut la prise en compte des effets d'équilibre de charge sur le nœud isolée, qui constitue la grille flottante, et des couplages parasites à l'intérieur et entre les cellules. Le nouveau modèle compact supporte les analyses DC, AC et transitoire, et comprent la dépendance avec les tensions, la durée et les algorithmes de programmation et d'effacement, les effets de température et ainsi que les variations statistiques. Les résultats du modèle compact ont été comparés à des simulations TCAD démontrant que dans les technologies avancées, les effets de canal court, des capacités parasites et de la vitesse de saturation dominent sur le comportement intrinsèque de la cellule. Le modèle offre également aux concepteurs la possibilité de déterminer des compromis importants, par l'inclusion des modèles de perturbation de l'etat de la cellule de type "drain disturb" et de la dégradation de l'endurance de la mémoire due au vieillissement de l'oxyde. Le modèle a été implémenté dans le langage portable Verilog-A et a ètè validé au moyen de résultats de mesures effectuées sur des structures de test intégrées en technologie CMOS 65nm.

Le développement du modèle compact a été basé sur une approche de modélisation rigoureuse, associant les outils TCAD conventionnels avec des modèles quantiques avancés. Pour ce faire, un nouvel outil a été développée pour l'étude des effets quantiques, des modèles de structure de bandes des matériaux, du tunnel quantique en régime direct et des effets de piégeage dans les diélectriques. Les effets de piégeage de charge dans les couches de l'oxyde et aux interfaces Si/SiO_2 ont été étudiés. Se concentrant spécifiquement sur la technologie flash, pour laquelle les signaux haute tension représent un enjeu majeur pour la dégradation des diélectriques. Un modèle rigoureu en prenant en compte les transitions multiphonon a été couplé à un solveur quantique de type Poisson-Schrödinger. De plus, un modèle de l'impédance du transistor a été élaboré et appliqué à l'analyse des caractéristiques du MOS en regimes DC et AC. Cette approche permet la modélisation physique du remplissage des pièges, de la réponse en fréquence et de l'électrostatique du composant. Les effets transitoires et l'effet tunnel assisté par pièges à travers l'oxyde ont également été étudiés. L'adoption d'une telle approche à plusieurs niveaux de modelisation offre la possibilité d'appliquer la méthodologie aux cellules de mémoire flash. Le développement d'un modèle compact pour la dégradation et l'analyse du rôle des défauts sur l'électrostatique et sur l'efficacité de programmation et d'effacement a été proposé.

Le modèle compact de la cellule flash a été appliqué à l'optimisation du développement de la technologie et à la conception de circuits intégrés de mémoire. Le développement du procédé technologique nécessite une compréhension profonde des compromis, qui peuvent affecter les fonctionnements en DC, en transitoire et à long terme de la cellule. La conception, l'intégration et la qualification d'un secteur de mémoire de 40KB destiné à les applications de carte à puce (Smart Card) ont été effectuées, afin de démontrer les capacités du modèle compact.

Résumé

Mots-clés : dispositif semi-conducteur, modélisation compacte, mémoire flash embarquée, modélisation Poisson-Schrödinger, endurance, dégradation des diélectrique, défauts d'interface & d'oxyde, piégeage des charges assisté par transition multiphonon.

Sommario

Nel mercato dei semiconduttori permeato da applicazioni portatili e di tipo consumer, la tecnologia delle memorie flash incorporate, offrendo vantaggi in termine di non-volatilità, ridotti tempi di accesso e scrittura e scalabilità, ha conosciuto una rapida diffusione, divenendo la soluzione di preferenza per le memorie a stato solido. Con l'addentrarsi della tecnologia nel dominio nanometrico, gli sviluppatori di circuiti analogici richiedono nuovi strumenti precisi per l'ottimizzazione dei blocchi ad alta tensione.

In questo contesto, un modello a potenziale di superficie per la cella di memoria flash é stato sviluppato, con l'obiettivo di offrire una migliore comprensione dei regimi di operazione del dispositivo durante l'ottimizzazione dei circuiti circostanti la matrice di memoria. Una procedura di validazione accurata richiede la considerazione di effetti di bilancio di carica sul nodo isolato e di effetti di componenti capacitive parassite nella cella. Il nuovo modello compatto permette l'analisi DC, AC e regime transitorio, includendo le dipendenze sulla durata e sulle tensioni di programmazione e di cancellazione, effetti della temperatura e variazioni statistiche dovute al processo di fabbricazione. I risultati del modello sono stati confrontati con simulazioni numeriche TCAD, dimostrando come gli effetti dovuti alla riduzione della lunghezza del canale, alla presenza di capacità parassite e alla velocità di saturazione delle cariche, dominino sul comportamento intrinseco della cella. Il modello offre la possibilità di valutare importanti compromessi specifici al design di memorie flash, attraverso l'aggiunta di modelli per la simulazione di effetti di disturbo sui terminali dei dispositivi, le componenti parassite dovute all'interazione con le celle circostanti nella matrice di memoria, e la resistenza della memoria al degrado del dispositivo. Il modello é stato implementato in linguaggio Verilog-A per facilitare la portabilità sui simulatori circuitali SPICE di tipo industriale, e validato mediante misure su specifiche strutture di test in tecnologia NOR 65nm CMOS.

Lo sviluppo del modello compatto é stato basato sull'analisi rigorosa del dispositivo con tecniche numeriche di tipo TCAD convenzionali e modelli fisici avanzati. Un nuovo software di modellizzazione é proposto per l'investigazione di effetti quantistici avanzati, effetti della struttura a bande dei materiali, effetto tunnel quantistico e intrappolamento di cariche nei dielettrici e alle loro interfacce attraverso transizioni assistite da fononi. Un modello quantistico per l'intrappolamento di cariche é stato implementato in un simulatore Poisson-Schrödinger ed un nuovo metodo per determinare l'impedenza sulla griglia del transistor é stato elaborato ed applicato per l'analisi di caratteristiche DC e AC. Questo approccio permette la modellizzazione fisica del riempimento dei difetti, la loro risposta in frequenza e il loro impatto sull'elettrostatica del sistema. Il modello permette l'analisi degli effetti in transitorio e delle componenti aggiuntive di tunneling assistito da difetti. I risultati del modello sono stati applicati alla discriminazione degli effetti del degrado dovuti al cambiamento dell'elettrostatica della cella e alla riduzione dell'efficacia di programmazione e cancellazione.

Il modello compatto della cella flash ha facilitato lo sviluppo del processo tecnologico e il design di un chip di test di 40KB di memoria per applicazioni smart card. Il design, l'integrazione e la caratterizzazione del chip dimostra le potenzialità del modello sviluppato.

Parole chiave: modellizzazione di dispositivi elettronici, modellizzazione compatta, memoria flash, modellizzazione Poisson-Schrödinger, endurance, degrado di ossidi, difetti in dielettrici e alle interfacce, intrappolamento di cariche.

Acknowledgements

Foremost, I would like to thank Mr. Neil Poulin, Dr. Hervé Jaouen, Dr. William Clark, Mr. Michael Kerbaugh and Prof. Yusuf Leblebici for giving me the opportunity to conduct this thesis project in their stimulating Research and Development groups in IBM STG, STMicroelectronics and EPFL. I deeply appreciated the excellent balance they granted me between freedom, support and advice, whenever the situation required.

I gratefully acknowledge for the collaboration between the different R&D groups in STMicroelectronics, in particular the eNVMRAM technology development group and the entire IBM STG team located in the Crolles facility.

I owe my most sincere gratitude to Dr. Denis Rideau, who acted as industrial technical advisor during my permanence in the TCAD team. I thank him for the very instructive scientific discussions, its modelling expertise and his untiring help during my difficult moments when "reaching saturation". His truly scientist intuition has made him a constant oasis of ideas and passions in science, which exceptionally inspired and enriched my growth as researcher and engineer.

I wish to thank Dr. Alexandre Schmid, who followed my PhD project since the very first beginning. I acknowledge him for his detailed and constructive comments, his important support throughout this work and his help revising the English of this manuscript and publications.

My sincere thanks are due to the entire TCAD modeling team, in particular to Dr. Erwan Dornel, Dr. Pascal Tannhof, Dr. Alban Zaka and Mr. Clement Tavernier. Their wide knowledge and logical way of thinking have been of great value for me and I extremely benefited from the integration into the group. Their ideals and concepts have had a remarkable influence on my entire career.

I wish to thank the entire SPICE modeling team, in particular Dr. Fabien Gilibert, Dr. David Souil, Dr. Michel Minondo, Mr. François Paolini and Dr. Birahim Diagne. I have benefited by advice and guidance from all of them. They kindly granted me vast portions of their precious time for answering many of my questions about compact modeling and extraction.

I would like to record my gratitude to Dr. Nicolas Degors, Dr. Sylvie Bruyere, Dr. Fausto Piazza and the whole M11 Process Integration team for their guidance as flash technology experts. Their understanding, encouraging and personal guidance have provided an excellent environment for the present thesis. In particular, I deeply enjoyed working with Dr. Degors on IC memory design and France-Italy cultural exchanges.

Furthermore, I would like to thank Mr. Yoann Mamy Randriamihaja, Mr. Shawn Fetterolf, and Dr. Jean-Philippe Manceau for their technical assistance on reliability analysis, chip testing and device characterization, respectively.

During this work I have collaborated with many colleagues within the STMicroelectronics and IBM groups, for whom I have great regard and that I could not mention personally one by one. I wish to extend my warmest thanks to all those who have helped me with my work.

My thesis at EPFL was accompanied by the reliable administrative support of M^{me} Marie Halm from the EDMI Doctoral School – thank you.

I am indebted to many colleagues at EPFL for the assistance and motivational discussions during my summer permanence in Switzerland.

The financial support of IBM France and of the IBM PhD fellowship program is gratefully acknowledged.

Finally, my infinite gratitude goes to my family, for their unconditioned love, constant and patient support. They have lost a lot due to my research abroad. It would have been impossible for me to finish this work without their continuous encouragement and understanding.

"Whatever Nature has in store for mankind,
unpleasant as it may be, men must accept,
for ignorance is never better than knowledge."

Enrico Fermi - Nobel Prize in Physics (1938)

Contents

Introduction

Semiconductor memory devices and circuits represent key components of most of electronic systems due to the widespread presence of micro–controllers and software applications, to store fragments of executable code or large amount of data. This large diffusion has pervaded all the sectors of the semiconductor industry, experiencing a significant success following the emergence of recent applications into our day-to-day lives: from the storage of large amounts of information into portable electronic devices to the performance improvement in high-speed servers, from the storage of microcontroller's program code into any industrial equipment to simpler instruction sets (the so-called firmware) of DSP (Digital Signal Processors) circuits or any programmable unit.

The development of the semiconductor memory market has rapidly progressed over the years, as new applications and memory technologies have become available to industries and mature to be commercialized. All the presently available technologies can be classified into two main categories: RAMs (Random Access Memories), whose content can be changed within a period considered short and comparable to the clock period of the system, and ROMs (Read Only Memories), where the non-volatility aspect, i.e. the capability of retaining the information virtually for ever, enables storage when the power supply is disconnected. The predominance of data-centric software applications over computation-centric applications necessitates the development of the an ideal memory technology, combining both the high performances of RAMs with the high density and non-volatility properties of ROMs. However, a more accurate evaluation of new memory technologies should nowadays rely on a wide spectrum of criteria involving different domains and depending on product applications and customer needs, from low-power requirements to integration capabilities, from technology scalability to integration cost and complexity, from basic memory performance to device endurance.

A particular class of NVM (non-volatile memory) devices is represented by flash memory technology, whose market has grown to become a $20 billion/year giant in less than 3 decades [1–3]. Flash memory technology is a class of Electrically Erasable-Programmable ROM (EEPROM) originally developed by Masuoka in 1980 [4] and 1987 [5]. In such a device, the measurable attribute, i.e. the information, is represented by an electrical charge and the non-volatility aspect is achieved storing a given amount of charge on an isolated floating-gate (FG) node. As a benefit of its advantages, which include non–volatility properties, endurance and reliability, high access speed and scalability, the EEPROM Flash memory technology is nowadays the preferred storage memory in many portable consumer

and computer applications. Recently, the market demands for increasingly aggressive performances led to the development of advanced deep-submicron flash technologies, facing new scalability and integration issues.

Two mainstream Flash memory technologies have been developed to volume production:

- *NOR* flash is mostly used in embedded applications; featuring high speed and noise immunity, it is compatible with low-power applications and driven by the market of portable applications, offering random access capabilities;
- *NAND* flash are able to achieve higher packing density and consequently lower cost than NOR; this technology is suitable for large data storage, but not the ideal candidate for random access memory applications; the NAND architecture is indeed restricted to serial read/write operation.

The two technologies also differentiate in the memory matrix array organization and market demands and diffusion. In NOR architectures, the cells have common source lines, normally connected to the ground terminal of the array. Each cell can be addressed with its specific bit-line. Among the applications of NOR architectures one could mention the storage of parameters and instruction sets in DSPs' firmware in electronic automotive applications, as well as customer data in banking with Smart Card dedicated products. Circuitry complexity increases when the functionality to address a block of cells connected in series between the ground line and single bit line contacts is required. NAND technologies are primary dedicated to end-user products. In 2010, this latter technology was occupying 21% of the total market of semiconductor memories, while NOR flash architectures represented a portion equal to 12%. Nevertheless, both markets are in considerable increase with an average annual revenue growth of 12% [6].

Generally, CMOS device performances can be improved with the introduction of new materials, new device architectures and scaling paradigms, or new fabrication processes. As a consequence, the improvement in CMOS and post-CMOS technology has empowered the development of advanced models to predict the behavior of the device. For example, the oxide thickness of new nanoscale devices reaches few atomic layers and the influence of interface effects between the materials dominates over the bulk material properties. Moreover, fabrication processes in industrial CMOS technologies presently consist of more than 200 steps. Complex heterostructures, material stacks and technology boosters are applied. Advanced modeling techniques are thus required to predict the electrical behavior and the process variability, and to support the technology development and reference models formulation for simpler circuit-level approaches. The improvements of TCAD advanced models should proceed in parallel with the development of physical compact models to enable product development with short design tapeout time durations.

Embedded non–volatile RAM (eNVRAM) compact models that are presented in literature are generally limited to reproduce the behavior of a single flash cell only in read conditions, without considering the numerous physical phenomena that define the final overall performance of the device. While these models are suitable for end users, they are insufficient for IP designers implementing charge pumps or support circuits, and to

physically analyse the long-term properties and related issues (endurance, data retention, reliability, soft programming etc.) of the cell.

Compact models development should involve validation using physical TCAD simulations and hardware test-structures. From a methodology point of view, a rigorous approach combines conventional TCAD simulation tools with physically-based models. The investigation of physical phenomena and the calibration of empirical models implemented in commercial TCAD tools and compact models can be performed in this view. In response to the industrial need, new physically-based models have been developed in the academic world. However, their application to the industrial environment is rarely effective due to their limited versatility, portability, support and user-friendly interfaces.

TCAD analysis is specifically required in device reliability study. Specifically focusing on flash devices where high voltage biases represent a major issue for dielectric degradation, the effects of charge trapping in oxide layers and Si/SiO_2 interfaces require particular attention, as they cause uncontrolled device parameter variations and general device lifetime reduction. In flash devices, the highly-energetic carriers in programming by Channel Hot Electron Injection generate interface and bulk oxide states, affecting DC and AC performances, program/erase dynamics, memory endurance and retention.

This work develops within the Joint Development Agreement (JDA) between IBM Systems and Technology group and ST Microelectronics in Crolles (France) and is supported by the Laboratoire des Systemes Microelectroniques at EPF Lausanne. Its main objective relies on the development of a 65nm NOR eNVRAM derivative technology for automotive and smart card applications. A new physical, efficient and portable compact model suitable for SPICE circuit simulations has been developed, implemented in Verilog-A, validated on TCAD simulations and extracted from measurements. The development of such a physical model requires the establishment of a fully-comprehensive modeling methodology, investigating physical phenomena, improved convergence schemes and model applications in flash memory devices. This includes first principles simulations, TCAD analysis of the electrostatics of the cell, compact SPICE model development and extraction, and IC memory design of a testchip using the extracted model. This approach has been adopted in this work and a comprehensive description of its constituting parts is presented in this document.

The core of the document includes 5 chapters:

- In Chapter 1, we describe the general working principles of a flash memory device in NOR configuration. An overview of the technology aspects, integration challenges and cell performances is provided. By reviewing the modelling paradigms and state-of-the-art approaches, the reader is able to identify the advantages and limitations of modelling solutions presented in literature. Finally, circuit components in NVM memory sectors and issues emerging in IC memory design are described;

- in Chapter 2, a methodology to develop a surface potential-based model for the flash memory cell based on 2D and 3D TCAD simulations has been developed. This enabled the comprehensive physical understanding of the device operations and elec-

trostatics effects. We investigated the influence of short-channel effects, overlap capacitances and velocity saturation with respect to the intrinsic behaviour of the cell in ultrascaled devices by means of TCAD simulations. A robust algorithm is implemented to solve the charge balance on the isolated floating gate node, taking into account parasitic couplings in the cells. The new compact model supports DC, AC and transient analyses, including program/erase bias scalability, temperature effects, process corners and statistical variations. The model has been implemented using the Verilog-A language to guarantee compatibility with common circuit simulators and has been validated using 3D TCAD simulations and measurement results on designed test structures;

- in Chapter 3, the effects of oxide degradation in CMOS technologies have been analysed. A rigorous MultiPhonon-Assisted (MPA) charge trapping model has been coupled with a Poisson–Schrödinger solver and a novel impedance calculation method has been applied to the analysis of DC and AC MOS characteristics. This approach permits the physical modeling of trap filling, frequency response and device electrostatics, as it intrinsically takes into account self-consistence in the structure. The spatial/energy distributions of capture/emission time constants are modelled. Transient effects of trap filling are also investigated as Charge Pumping (CP) techniques can be adopted to extract interface trap concentration profiles. The characterization methods are compared extracting the energetic and spatial regions contributing to the response;

- in Chapter 4, secondary effects in flash memory devices are described and integrated in the aforementioned flash compact model. Towards a more analytical approach for the effects of defects of Chapter 3, a semi-analytical model for charge trapping and degradation has been implemented and the effects of degradation on endurance characteristics analysed. Compact solutions have been adopted to model disturb and cross-couplings in flash memory arrays, validating the approach on hardware and TCAD. Statistical analysis is a valuable functionality to permit designers to investigate the worst-case conditions of the device;

- in Chapter 5, the applications of the compact model to IC memory design are described: strategies for program/erase pulse analysis and optimization permit to determine the best conditions of operation of the memory cell. Finally, the design, integration and characterization of a 40KB memory sector for Smart Card (SC) applications has been performed using the compact model. Bit Line and Word Line parametric cells have been developed to simulate a complete set of devices in the memory sector. Characterization and functionality testing of the memory test-chip demonstrates the direct application of the physical SPICE model.

The formulation of the compact model has also conducted to the development of a new custom-made modeling tool, named UTOXPP, which has been adopted for the investigation of advanced quantum effects, band structure models, quantum tunneling and MPA

charge trapping effects in dielectrics. The new tool is described in Annex C where its functionalities and applications in CMOS technology are presented.

Introduction

Chapter 1

Flash memory technology

This Chapter exposes the basic working principles of a flash memory cell and the state-of-the-art of compact modeling of the device, suitable for the application to IC memory design.

1.1 Working principles and general definitions

The structure of a flash device is illustrated in Figure 1-1: a single cell is constituted by a floating gate (FG) metal-oxide semiconductor field-effect transistor (MOSFET), that is capable of retaining a previously-stored electrical charge for long periods of time even when power is removed. This type of transistor is a standard MOSFET device, whose gate is connected with one or more capacitors and is thus *floating*, i.e. completely electrically inaccessible and isolated from the other nodes. The dielectric between the channel and the polysilicon FG is also called tunnel oxide and its thickness T_{OX} results from a compromise between high program/erase (P/E) performances and endurance/retention capabilities. The layer between the FG and the control gate is commonly a trilayer stack formed of a Si_3N_4 insulator (nitride) sandwiched between two layers of SiO_2 and is also called oxide-nitride-oxide (ONO) layer. The capacitors are used to couple the FG node to the voltage biases that are applied to access or control the stored information. By appropriately biasing the four accessible terminals (drain - D, source - S, control gate - CG and substrate bulk - B), one is able to sense or vary the charge into the FG node. The process of removing the charge from a FG is called *erase* procedure, while the process of storing it *program* operation. Retrieving the information is performed in *read* operation. Triple-well CMOS technologies are commonly adopted to independently bias the isolated substrate of the memory device during read, program or erase operation.

Figure 1-2 shows the commonly-used operating modes of industrial flash devices, to sense, store and remove the electrical charge in the FG node. The amount of charge at the FG node can be sensed by relating it to electrical parameters, e.g. the drain/source current I_{DS} flowing into the channel of the MOSFET transistor constituted by terminals FG, D, S and B. In READ operation, the charge modifies the threshold voltage of the

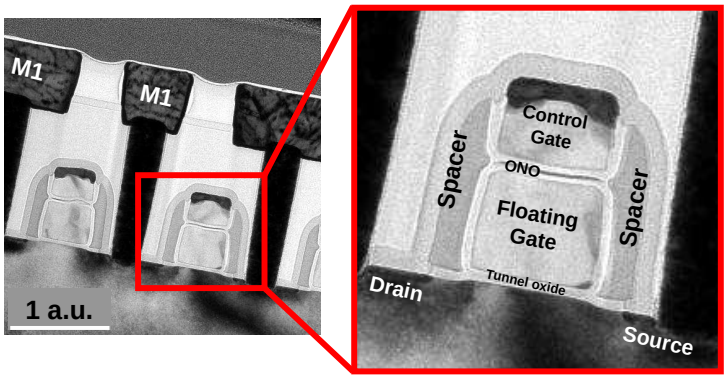


Figure 1-1 – Transmission Electron Microscope cross-section image along the channel length of a flash memory device in NOR configuration. In the zoom, the terminals of the device as well as the oxide layers can be identified. Courtesy of P. Guyader.

intrinsic MOSFET device, so that one can choose two stable reference amounts of charge to represent the 1 and 0 memory states. The threshold voltage of the cell V_{th} is consequently modified. Several definitions are present in literature for this parameter [7]; in this study the constant-current definition has been chosen, V_{th} being the voltage that has to be applied to the control gate to obtain an arbitrary reference I_{DS} through a memory cell in read operation. When a negative charge is present on the FG, V_{th} increases and the current flow is minimal (*programmed* state - 0); if the charge is removed, the cell MOSFET presents a low threshold voltage and I_{DS} increases (*erased* state - 1). Since the oxide completely surrounds the FG node, the injected charge is trapped in the floating gate and stores the information.

Among the techniques used to store or remove charges, Fowler–Nordheim (FN) tunneling and channel hot-electron injection (CHEI) are the most diffused solutions. Cell programming can be achieved using both the mechanisms. The tunneling of electrical charges through oxide layers has been applied since mid 70s on thin oxide layers. Fowler-Nordheim programming regime is achieved when a positive high voltage difference from 15V to 20V is applied between the control gate, i.e. the word line in a structured matrix array, and the source or the substrate of the cell[1]. This represents a more historical approach and is still applied in NAND, AND and DINOR (divided bit line NOR) technologies where high throughput is required sacrificing random-access memory capabilities. Indeed, one important advantage of such a method relies on the fact that multiple wordlines can be programmed in parallel using a very small programming current (less than 10nA/cell).

The Channel Hot Electron Injection (CHEI) operation consists in applying a lateral electric field along the channel to increase the energy of electrons and force a transversal electric field between the floating gate and the channel to enhance carrier injection. Contrary to FN tunneling, this low-efficiency mechanism requires large amounts of current (up to $\approx 500\mu$A/cell) and thus cannot be used in high P/E through-put applications. CHEI is currently the preferred approach for cell programming in NOR technologies where random-

[1]The electric field would give a better indication of the strong bias conditions and should be considered a more precise indicator as high voltage devices have high applied potential drops with low oxide electric fields [8].

access capabilities are available, and narrow threshold voltage distributions are required. Other drawbacks of CHEI include device degradation that reduce both memory endurance and retention.

The erase operation is usually achieved with reverse Fowler-Nordheim tunneling. In the considered high field regime, the control gate is biased at high negative voltage while a positive bias is applied on the source line and/or the isolated substrate of the device. The required erase current remains quite low, enabling the possibility to erase in parallel vast portions of the memory matrix. The low current consumption also contributes to relax requirements in embedded applications and permitting the integration of simpler circuit blocks for the generation of the required high voltage biases directly on chip. However, due to the exponential dependence of the tunneling current with oxide thickness and characteristics, accurate process control is required to achieve narrow V_{th} distributions.

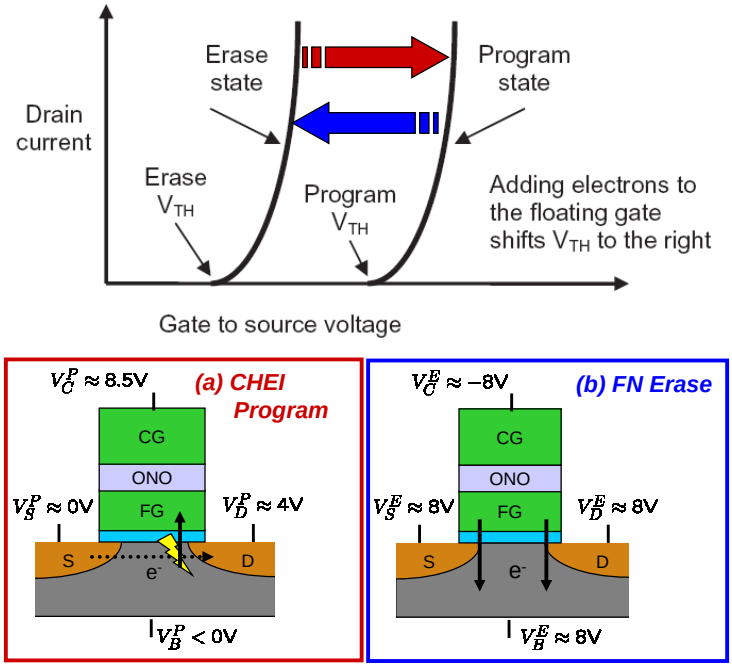


Figure 1-2 – Operating modes of a flash memory device. On the top of the figure, the drain current is shown as a function of the control gate voltage. The device switches between two or more states (program and erase states), represented by a different amount of electrical charge stored on the floating gate. In (a) program mode operation by Channel Hot-Electron Injection in the Lowly-Doped Drain (LDD) region is shown. Erase mode by Fowler-Nordheim tunneling across the Tunnel Oxide and uniformly along the entire channel is represented in (b).

P/E algorithms are able to inject/remove a given amount of charge into the floating gate creating a negative electrical potential, which then translates into a different threshold voltage of the equivalent MOSFET. As previously mentioned, the commonly adopted P/E mechanisms have several advantages and drawbacks that have to be carefully evaluated considering the total threshold voltage window, the spread of V_{th} distributions of the memory states, the device power consumption, the cell degradation and memory retention, and the possibility of exploiting P/E parallelization. Program/Erase efficiencies can be analysed evaluating the threshold voltage or the read current variation as a function of time. The latter approach is of particular importance in multistate memory architectures or in aged devices where charge leakage affects retention.

Variability between the cells of the matrix array causes the devices to have variable

characteristics, including variations of electrical properties of the layers composing the flash, geometrical asymmetries, different characteristics of material interfaces. The composition of all these variation results in a wide distribution of the cell performances, including the threshold voltage of the device. For this reason, modelling solutions should be able to discriminate such variations to offer designers the possibility to analyse worst case scenarios and analyse potential trade-offs.

Memory retention in a non-volatile memory technology is represented by the amount of elapsed time between data storage and the detection of a erroneous readout of the data [8]. Retention in flash memory technology is associated to the capability of the device to have reduced charge leakage from the floating gate. In modern Flash devices where high fields are required to switch the state of the device, interface and oxide defects are considered the main responsible of Stress-Induced low-field Leakage Currents (SILC), which decrease the quantity of stored electrons and alter the non-volatile information. It has been quantified that in deep-submicron technologies where the FG capacitance is approximately 1fF and the number of stored electrons in the FG does not exceed 20000, a loss of five electrons per day permits to achieve a commonly-accepted specification of 10 years in data retention [9]. Such requirements are even more aggressive in particular applications, e.g. in the automotive sector, where the device should be able to operate in a wide temperature range. Retention performances can also be affected by the properties of the materials, device geometries and architecture, P/E algorithms adopted in circuit design.

Another parameter altered by device degradation and electrical stress is represented by memory endurance, i.e. the capability of a memory device to withstand electrical stress conditions quantified in memory technologies by the number of P/E cycles [8]. Indeed, the device experiences V_{th} window closure or shift phenomena after a given amount of P/E operations which represent a serious concern for IP designers developing sense amplifier circuits for detecting the state of the device.

Word-line (WL) and bit-line (BL) disturbs represent other major issues in ultra-scaled technologies. Indeed, in NOR array configurations programming and erasing can be achieved only by biasing entire WLs and BLs in high voltage conditions. The programming disturb present on the unselected erased devices (attacked cells or *victims*) belonging to the same WLs (WL or gate disturb) may induce tunnelling of electrons from the channel to the FG. BL disturbs can generate from both program and read conditions. In such a case, the victim cells share the same BL of the device to be programmed or read. Also in this case the programmed condition is more sensitive to disturb as it is suspected that hot hole injection (HHI) phenomena occur from the channel and reduce the amount of electrons stored into the device [10, 11].

Although all the mechanisms detailed above have been investigated with compact solutions, technology scaling is facing an increasing number of issues related to quantum-effects, carrier confinement, reliability and oxide interfaces issues. As a consequence, more accurate and rigorous quantum models should be developed to act as a reference for ultra-scaled devices (see Chapter 3 and Annex C).

1.2 Modelling of flash memory devices

All compact DC models for flash devices described in literature divide the core of the device in two blocks: a capacitance network, mainly represented by the intra-poly capacitance, which isolates the floating gate node from the other terminals, and a MOS device, also called equivalent transistor, which represents the active component of the structure and whose gate is constituted by the floating node. The structure of the equivalent transistor can be obtained contacting the FG of the cell by shortening it with the control gate. This permits the characterization of the active region of the device.

1.2.1 Modelling methodology in industry

Compact or SPICE models in microelectronics reproduce the behaviour of electronic devices and find application in circuit design. They consist of analytical solutions with which IC designers can verify the electrical behaviour into modern circuit SPICE simulators. They provide a mathematical representation of the electronic device for the calculation of the terminal charges and currents depending on bias, geometry, temperature and other operating conditions. Compact models elegantly merge physics, simplicity, universality, accuracy and efficiency. With device scaling towards deep sub-micron, the development of compact approaches often requires trading off model performances with physicals accuracy and simplicity. The emergence of extrinsic phenomena requires new effects to be introduced, increasing the number of model parameters and thus the extraction complexity. In the case of MOSFET devices, compact models are usually structured with an intrinsic core model, that defines the transistor behaviour for the current and charges calculation, and a series of extrinsic models to simulate the resistive and capacitive components in the structure (access resistances, junction and overlap capacitances, interconnections, etc.) and the parasitic currents flowing in the structure (gate induced drain leakage - GIDL, junction and oxide leakage, etc.).

The state-of the-art of compact models of MOSFET devices adopted in industry [12] can be divided into three categories depending on the method adopted in the development of the intrinsic models: threshold voltage-based, charge-based and surface potential-based approaches. Threshold voltage-based models include the family of BSIM [13–15] developed by the University of Berkeley and MM9 from Philips [16], and are regionals models where currents and charges are determined analyzing the device in separate operating regions and using smoothing functions for the intermediate transitions. These models face issues when reproducing the physical behavior of the device as empirical parameters are introduced to control the smoothing of the transition. Consequently, other approaches are gradually replacing regional models. Charge-based approaches rely on the calculation of the inversion charge in the channel and are derived from the charge sheet approach originally adopted by Brews in 1978 [17, 18]. One of the most important representatives of this category is the EKV model, which founds important applications in IC analog design [19–22]. Although these approaches are presently adopted by a large number of research groups, the organism for the standardization of compact models (Compact Model Council) recently chose sur-

face potential-based models as future replacement of regional models. This last category includes HiSIM [23], MM11 [24], SP [25,26] and PSP [27–29] models, where analytical solutions permit an accurate and efficient determination of the surface potential in the MOS channel.

1.2.2 Modeling of the floating gate voltage in DC conditions

The specificity of the compact models for flash devices relies on the calculation of the floating gate potential. This section details the two paradigms that could be adopted to this purpose.

(i) Capacitive coupling coefficient models - CCM

The origin of the first compact floating-gate models is found in the mid 80s with the application to EEPROM cells. In such an approach, originally proposed by Bhattacharyya [30] and further adopted by Kolodny in [31], the base equations defining the floating-gate operation in read operation and the first concepts of intra-cell coupling are introduced. The charge Q_{F0} stored inside the FG node in DC conditions can be expressed as a function of applied biases and coupling capacitances:

$$Q_{F0} = C_{CF}(V_F - V_C) + C_{FS}(V_F - V_S) + C_{FD}(V_F - V_D) + C_{FB}(V_F - V_B), \tag{1.1}$$

where V_F is the potential on the FG, while C_{CF}, C_{FS}, C_{FD} and C_{FB} are the capacitances between the Floating Gate and the Control Gate, the Source, the Drain and the Bulk, respectively.

Assuming that one is able to determine the capacitances in the structure, the floating gate potential can be analytically determined using:

$$V_F = \alpha_C V_C + \alpha_D V_D + \alpha_S V_S + \alpha_B V_B + \frac{Q_{F0}}{C_T}, \tag{1.2}$$

with $C_T = C_{CF} + C_{FD} + C_{FB} + C_{FS}$ expressing the total floating gate capacitance.

In this approach, the coupling coefficients α_j, defined as the derivatives of the FG voltage with respect to the bias V_j applied to terminal j or equivalently as the coupling of the FG with respect to the electrode j, can be described by:

$$\alpha_j \equiv \frac{\partial V_F}{\partial V_j} = \frac{C_j}{C_T}. \tag{1.3}$$

The currents and charges in the active region of the device can be determined applying any analytical model to the MOSFET formed by the terminals F, D, S and B and replacing the gate voltage with the calculated floating gate potential. Several considerations may be done on the results provided by this method:

- the threshold voltage of the device V_{th} as determined from the control gate terminal

is linearly dependent on the charge stored in the floating gate node and the shift of the characteristics is given by $\Delta V_{th} = \frac{-Q_{F0}}{C_{CF}}$; in the same way the change in floating gate voltage is given by $\Delta V_F \approx \alpha_C \Delta V_{th} = \frac{-Q_{F0}}{C_T}$

- the model completely relies on the estimation of capacitive couplings; these can be considered model parameters but their accurate determination normally requires accessing the floating gate potential during characterization, a procedure that is usually impossible to perform with a standard FG cell;
- the control gate coupling α_C controls the capability of the control gate terminal to induce the floating gate voltage; the ONO capacitance is thus a critical structure parameter for the optimization of the device performance as an improved control of FG represents better program/erase performances; a small gate coupling coefficient corresponds to a small charge transfer on the FG;
- due to the drain coupling, drain turn-on effect can occur if one considers that the MOSFET active device can conduct even for control gate voltages $V_C < V_{th}$ [9];
- while the information can essentially be interpreted as binary, internally an analog information represented by the amount of stored charge can be considered; multi-state cells have also been proposed exploiting this property but they face serious retention and V_{th} distribution issues when the device is scaled down;
- the external threshold voltage in the absence of charge on the FG ($Q_{F0} = 0$) is also called ultraviolet threshold voltage V_{th}^{UV} [2] and it represents an important variable for compact model extraction;
- the channel coupling contribution is often neglected in Eq. 1.2 even though the large voltage dynamics between erase and program conditions can result in undesired approximations in V_F calculation during transient regime;
- when the accessible terminals are grounded, V_F can be different from zero, depending on the charge Q_{F0}.

The main drawback of the capacitive method is related to the accuracy of Eq. 1.2: due to the many extrinsic effects which are dominating at sub-micron scale, the determination of the coefficients can be a challenging task and the techniques used to experimentally extract them are often not accurate enough to be applied to all the regimes of operation. Generally, for the estimation of the α_i coefficients, the control gate coupling coefficient α_C is firstly determined while the other parameters are extracted by measuring the drain leakage currents or the difference between the erase voltages [32] [33]. The most common method to calculate α_C relies on the determination of the ratios between the threshold voltages, transconductances or subthreshold slopes of floating-gate cells and equivalent-transistor

[2] In older EEPROM technologies, the cell was erased by UV treatment completely removing the charge on the floating gate of the cell.

devices. It is evident that, since all the coefficients are derived from α_C, a minimal error on the estimation of this parameter affects all the other coupling coefficients.

A problem common to all the coupling coefficient extraction methods is the intrinsic difference between two structures generated by process variations, that can introduce random differences in identical devices on the same array. Additionally, all the cases where a comparison of equivalent-transistors and NVM characteristics is performed suffer from the fact that the two structures present differences in the morphology of the ONO capacitance due to the array configuration. Other issues in comparing equivalent-transistors and FG cells depend on the adopted extraction technique. The main limitation in the transconductance method described in [31] is its sensibility to short channel effects and mobility degradation. The subthreshold method [34] is the most accurate, but it suffers from the fact that the coupling coefficient α_C is determined in subthreshold regime of operation and its value cannot be considered constant for all the biases [9]. Bez [35] estimated the coupling coefficient α_C in the subthreshold regime of the transistor. In this region the dependence of the coupling with respect to the total capacitance C_T can be neglected and the value of the parameter can be evaluated from the programming characteristics of the cell, i.e. the variation of the threshold voltage shift ΔV_T as a function of time, and from the subthreshold current. Moison [36] considers a succession of write and erase operations varying the hold time of the pulse and calculates the charge and the threshold voltage shift at the end of the write operation.

The coupling ratios that have been defined as constants, are in reality strongly dependent on the biases [37, 38]. Larcher [38] proposed a new paradigm for the determination of the coupling coefficients dependency using a compact approach, showing an abrupt decrease of α_C when the device switches in weak inversion conditions. It is thus evident that the bias dependency cannot be neglected in an accurate and complete physical model. V_F calculation errors increase as the discrepancy between the applied biased in the operating conditions and in the regime where the coupling coefficients have been extracted increases [9]. Since the transient currents are exponentially dependent on V_F, these errors can compromise the accurate modeling of the transient dynamics.

Despite these limitations, the CCM also found application in several models for the simulation of the P/E transient dynamics of floating-gate cells. In particular, the classical formulation of the Fowler-Norhdeim current [39–41] was adopted for the calculation of the floating gate voltage variation as a function of time by [42, 43] in both program and erase conditions. This approach has also been applied to flash memory devices, adding the description of CHE injection currents [44] and the effects of degradation on endurance [45].

(ii) Charge balance models - CBM

Flash cell device modelling methodology has undergone a revolution during mid 90s when circuit simulators have been introduced for the optimization and the analysis of product performances. This need stimulated the development of more physical compact approaches, where the whole device is modelled. In particular, the common approach adopted nowadays is based on the charge balance method (CBM), originally proposed by Larcher and Pavan

in [9, 46]. It consists on simplifying the cell into a MOS transistor, whose gate is the FG of the flash cell, and a capacitor modeling the interpoly capacitance. This new paradigm has the advantage of avoiding the use of constant coupling coefficients that compromise the accuracy on the calculation of V_F.

A non–linear equation derived from the balance of charges is iteratively solved at the FG node. The total charge on this node is in fact the sum of the charge stored on the gate of the MOS transistor Q_G and of the charge stored on the bottom plate of the C_{CF} capacitor.

$$Q_G(V_F, V_S, V_D, V_B) + C_{CF}(V_F - V_C) = Q_{F0}. \tag{1.4}$$

Since the floating node is completely isolated from the other terminals, in DC conditions Q_{F0} is constant while its value is changed during an erase or programming operation. All the other parameters present in the equation are known or can be determined from geometrical considerations, as the charge Q_G can be calculated using standard compact models for MOS transistor such as BSIM3 [13], EKV [19–22], HiSIM or PSP [28]. The charge Q_G depends on all the biases in the intrisic MOSFET of the cell, and thus accurate models are required for its calculation. The accuracy of the model is strongly dependent on the precision of the adopted MOS compact model. Additionally, to assure model scalability, physical models are required for the calculation of the ONO capacitance C_{CF}. Even though its value can be estimated by means of characterization results on large area devices or TCAD simulations, compact scalable approaches are preferred for its calculation.

If we summarize the characteristics of the two approaches, we can conclude that the capacitive method based on the coupling coefficients is simple and efficient, thus it is commonly used in industry because the determination of the parameters on fully automated characterization benches is straightforward and fast. It offers a good estimation of the floating gate voltage but suffers from the extraction of the coupling coefficients in the device. Indeed, an accurate analytical calculation or experimental measurement of these parameters is problematic, and presently there is no complete compact model for their bias voltages dependence. When the coupling coefficients are considered constant, the accuracy is compromised, since, according to the conditions in which the coefficients are determined, an overestimation or underestimation of the FG voltage can occur, with a consequent exponential propagation of the error in the transient regime. In the charge balance method, the determination of the coupling coefficients is not required since the bias dependence of the charge is implicitly and self–consistently integrated inside the MOS gate charge model. The accuracy is improved but the computational time is slightly higher due to the solution of the non–linear equation on the FG node. Also in this case, the parameter extraction procedure is performed by applying the MOS parameter extraction procedure to the equivalent transistor cell. The potentiality of this paradigm however is even more significant, considering that all the models for scalability, mobility, saturation velocity, quantum effects, STI stress effects, etc.. are integrated into the MOS models, whose accuracies are highly consolidated in literature. Additionally, transient variations of the injected/extracted charges can be directly included in Eq. 1.4. In this work, we adopted the CBM paradigm in the development of the flash compact model.

1.2.3 Transient mechanisms modeling

This section deals with the physical phenomena leading to the variation of Q_{F0} during program/erase operation. The NVM cell being an analog device constantly operating in transient conditions, a particular critical point relies on the models of transient effects which determine the accuracy of program/erase simulation, disturb analysis and device ageing investigation. The most recent developments have permitted to define flash cell models based on the calculation of the surface potential of the device [47–49]. This facilitated an accurate calculation of the electric fields in the structure and largely improved the accuracy of the transient currents calculation. A wide spread of models for the different transient mechanisms in flash memory devices can be found in literature, covering both program/erase by Fowler-Nordheim tunneling and program by CHEI. Other models also include memory retention reduction with trap assisted tunneling currents or degradation models for the variation of the macroscopic performances of the cells during cycling and disturb mechanisms.

Any current calculated with the aforementioned models can be applied to one of the two paradigms for the calculation of the FG potential described in the previous subsection. Indeed, the injected charge calculated by integration of the total current I_{tot} from the bulk and the S/D terminals to the floating gate is given by:

$$\Delta Q_F = \int_t I_{tot} dt, \tag{1.5}$$

and can be added to Eq. 1.4 for the transient evaluation of V_F. Consequently, the most important effort should rely on the calculation of closed-form expressions of the tunneling currents in the different mechanisms.

(i) Tsu-Esaki model and compact approach for Fowler-Nordheim currents

A large number of compact models expressions in the different operating regimes derive from the Tsu-Esaki model for tunnelling phenomena evaluation through a potential barrier [50]. In such an approach, Electron-conduction-band and Holes-valence-band tunneling mechanisms are considered through an energy barrier representing a dielectric material between two semiconductor or metallic regions. Compact expressions have been derived for direct tunneling conditions (trapezoidal barriers) and Fowler-Norhdeim tunneling (triangular barriers) and have been introduced in most of the industrial compact models for gate leakage modeling [39]. The model will be described in detail in Chapter 2.

(ii) Program by CHEI

The preferred program mechanism in NOR technologies is based on Channel Hot-Electron Injection: the energy of electrons flowing into the channel in high drain bias conditions increases due to the high lateral electric field. Impact ionization occurs as the carrier approaches the low doped drain (LDD) region, generating holes and electrons couples.

While majority carriers are collected in the isolated P-well of the array, generating a weak avalanche substrate current, minority carriers flow towards the drain. However, due to the high energetic tail of the carrier distributions [51–53], some of the carriers have sufficient energy to overcome the tunnel oxide barrier and tunnel towards the FG. In contrast to quantum tunnelling, due to the complex 2D nature of the Channel Hot-Electron Injection phenomenon, only qualitative empirical models have been developed for compact modelling purposes. The physical and statistical complexity of the mechanism require numerical and Monte Carlo solutions to tackle this aspect. Certainly the most diffused approach in compact modelling is the lucky-electron model [54, 55], where the probabilities of energy increase, carrier scattering and injection are considered as independent events.

More advanced numerical models rely on taking into account non-local aspects of scattering mechanisms for the calculation of the non-Maxwellian energy distribution of carriers along the channel [56–58]. Such approaches are not suitable for the considered compact solutions and require input conditions from Monte Carlo simulations. In particular these models rely on the physical determination of the electron energy distribution in the channel as a function of the lateral field, the momentum distribution for the evaluation of scattering events, and the crossing probabilities. Models that describe the hot carriers distribution inside the channel have been studied by numerous authors: heated distribution functions and Non-Maxwellian energy distributions have been proposed by Abramo [52], Grasser [59–61] and Fiegna [62]. The distributions of cold carriers have to be determined as well to reproduce the transient dynamics in low-field conditions and at the end of the program pulse.

(iii) CHISEL operation

CHE injection fundamentally suffers from two main drawbacks: the drain voltage has to be kept high enough so that electrons acquire enough energy to cross the energetic barrier of the tunnel oxide, making the design of voltage multiplier circuits challenging in embedded applications; and the efficiency of the injection mechanism is very low with a considerable power consumption that makes ultra-scaled technologies less suitable for low-power applications. In the recent years, new mechanisms for cell programming have been investigated, *in primis* the CHannel Initiated Secondary ELectron (CHISEL) mechanism, where a negative bulk biasing of the flash array is used to boost impact ionization at the drain/bulk junction. In such a regime, after the creation of primary holes and electrons in the LDD region, the holes ionize again at the drain junction due to the high field applied, and generate other e^-/h^+ couples. While secondary h^+ are collected on the bulk, e^- have enough energy to contribute to the injection current as well.

Due to the large number of carrier interactions and scattering events, the formulation of compact models faces serious challenges when reproducing technology scalability of the phenomenon. The model proposed by Larcher in 2002 introduces different interaction probabilities and calculates the distribution of carriers from Monte Carlo results [63].

(iv) Trap assisted tunnelling, SILC and endurance

A serious concern in deep sub-micron flash memory technologies relies on the gate leakage currents which eventually reduce the amount of stored charge and cause information loss. After the device undergoes a certain amount of cycles and electrical stress across the oxide layer, band-to-band and trap assisted tunnelling currents linked with oxide degradation phenomena emerge. These stress induced leakage current (SILC) components dominate in high temperature applications and are also considered responsible of increased disturb effects [64]. SILC is usually observed in highly stressed devices as a carrier flow present below the onset of direct/FN tunneling regimes and strongly limits the gate oxide thickness scalability and P/E voltages. Clear correlation has been found between this parasitic current and the oxide defect concentration, which increases as the device undergoes additional stress [65]. Two components have been identified in inelastic trap assisted tunneling at the origin of SILC: the steady state component contributing to the oxide leakage is caused by the capture and emission of carriers with multiphonon absorption/emission events [66–69]; transient SILC is represented by the displacement current caused by the dynamic filling of traps [68, 70, 71].

While multiple characterization techniques based on charge pumping, admittance measurements and DCIV analyses have been recently developed for extracting the trap concentrations in the oxide layers, a large number of empirical or compact independent models for the different trapping phenomena have been developed [72–74]. These compact solutions are used to extract analytical defect distributions in electrically stressed oxide layers but their accuracy can be strongly limited due to the approximations in the models [75].

One of the most diffused and complete model for SILC has been formulated by Ielmini in 2001 where capture/emission events of carriers from gate and channel reservoirs are calculated using an approach derived from the expressions of Tsu-Esaki tunnelling and taking into account the capture cross-sections of the defects. In such a way, the steady-state SILC current density is a function of the defect distribution N_T and of an effective capture cross-section, treated as an empirical parameter or extracted from more rigorous simulations [69]. The current also depends on the carrier occupation at the defect site and in the substrate. A generalized Tsu-Esaki approach extended to charge trapping can be adopted [69, 76, 77]. On the other hand, the calculation of transient SILC components requires the resolution of the transient rate equation with a numerical finite difference approach [68].

The critical issue in SILC modelling is represented by the accurate and physical calculation of the capture/emission time constants. Although analytical solutions are usually adopted for modelling the capture cross-sections studies performed with advanced physical models show that multi-phonon inelastic tunnelling events dominate and thus the barrier transmission and energy dependence should be taken into account in the calculation.

Other models assimilate the entire current contribution to a single defect placed at a single energy and in the middle of the oxide stack where the TAT current contribution is maximum [78,79]. In such cases, the total TAT current can be expressed as a function of the trap capture cross-section, the trap concentration N_T and a uniform tunnel current density

J_T through an effective energetic barrier. More recently statistical TAT current models through multiple trap sites have been applied to the determination of SILC leakage [80–84] and oxide leakage in novel oxide stacks [85–87].

(v) Disturb modeling

The optimization of P/E conditions for scalability plays a critical role in the tradeoff between tunnel-oxide reliability and device performance. During asymmetrical program operation by CHEI, it can be shown that oxide degradation is localized at the drain overlap [88]. Additionally, bitline biasing induces degradation by program disturb mechanisms caused by hot-hole injection (HHI) at the drain [11]. Indeed, any cell sharing the same wordline (bitline, respectively) with the cell being programmed, erased or read is subject to gate (drain, respectively) disturbs. These can enhance charge loss and degradation performances. In such regimes, the electrical fields in the tunnel oxide are smaller than the nominal program condition, but the average cumulative disturb time is considerably longer than a single program pulse [11]. For this reason, the complete analysis of cell reliability should take disturb currents into account. However, a physical investigation of HHI mechanisms requires advanced 2D statistical simulations for the calculation of band-to-band tunneling currents, impact ionization effects and holes injection [11], resulting in an approach not suitable for application to compact models. Nevertheless, compact solutions exist and rely on variations from the classical expressions of FN or CHEI currents [89], where the empirical parameters need to be extracted from characterization and suffer from bias and technology scalability.

Models described in Sections 1.2.3(i) and (ii) are developed in Chapter 2, while Chapter 3 discusses the effects mentioned in Section 1.2.3(iv). Chapter 4 includes the modeling of disturb dynamics.

1.3 IC design for flash eNVM

Analog ICs of embedded flash memory circuits are complex architectures where the different functional blocks not only provide access and modification of the stored information, but also testability functions, repair/correction circuitry and optimization of P/E algorithms. The architecture of a flash memory chip includes at least the following blocks:

- the flash memory array appropriately partitioned in small blocks or sectors and organized in NOR or NAND configuration;
- row/column decoding circuits to provide the read/program/erase path from the analog high voltage blocks to the flash cells;
- the analog management system constituted by a group of high voltage multipliers, regulators and switches to generate and control the voltage biases during the different operating modes;

- a series of sense amplifiers to read the information with high throughput;
- a control logic to generate the control signals for the memory array and the analog blocks.

A peculiarity of the design of flash memory circuits resides in the organization of the memory matrix, which strongly depends on the product applications, the P/E specifications and the addressing scheme. In NOR architectures, the array is divided in sectors in which the cells generally share a common source line and an isolated p-well for the bulk contact. This configuration permits to adopt erase algorithms to restore logic "1" over an entire sector. Sectors are separated and surrounded by dummy flash devices to reduce lithographic proximity effects. In large memory arrays, parasitic capacitances and access resistances on the word and bit lines of the sectors can play a role on the access times and P/E dynamics. For this reason, IC designers employ compact models of entire WLs and BLs to properly size the decoding drivers and meet the product specifications. The accurate estimation of the cross-coupling capacitances between the cells in the array should be realized with physical accurate models. Historically, memory circuit designers have adopted static compact models for flash devices, where simple static corners provide an estimation of the worst-case cell electrostatics. These consist in separate extracted model parameter cards that the designer can select depending on the state under analysis. Such an approach however does not provide any insight about power consumption or P/E efficiency as transient currents cannot be evaluated. Other simpler approaches consist in modelling entire bit and word lines with distributed RC networks, whose parameters are extracted from TCAD simulations of the complete matrix interconnection network. Equivalent circuits are also used for estimating the load from the parasitic cross-coupling capacitances, while the ONO capacitance plays an important role as it represents a large portion of the WL load.

Decoding circuits are one of the most critical blocks to be designed. Indeed, in ultra-scaled embedded technologies where the high voltage biases are internally generated, a high voltage conduction path has to be created from the voltage multiplying circuits to the flash devices. For this reason, the derivative NVM technology includes particular thick-oxide FETs, which are able to sustain high voltage biases on the oxide stack and drive flash cells in a dynamic range from -10V up to +10V. The row decoder circuit is usually the first block to be designed as soon as the array sector configuration is defined [90]. Indeed, the design is challenged by the final driver sizing that has to deliver sufficient current to drive a long WL or BL, while it also has to fit into the row pitch of the matrix. Additionally, the row decoder should be able to drive the device in different bias conditions, from high positive voltages in program and read modes, to large negative biases in erase mode. This constitutes a serious issue in low-power technologies, where the control stimuli provided by the logic block are low voltage signals and generally not sufficient to reach the threshold voltage of HV transistors and drive HV blocks [91, 92]. As a consequence, one or more stages of level shifting circuits are designed to increase the dynamics of the control signals to intermediate values or to the final desired voltage [93]. Hierarchical approaches in

decoder design permit to reduce the occupied area and partially solve footprinting issues[3], due to the fact that HV devices only slightly scale down with technology and the overhead imposed by these devices can dominate.

Specific circuits are present in flash memory technology to handle the large variety of voltage conditions to drive the device. In an embedded low-power technology where a single supply bias is available, several DC voltages need to be generated on-chip. Moreover, the market requirements driving the technologies to low-power operation contrasts with the high-voltage operation of flash devices, which cannot be subdued to preserve cell performance specifications. As a consequence, voltage-level shifting circuits provide an interface between the low-voltage digital signals and the analog memory array. Level-shifters permit to increase the dynamic range of the low-voltage signal to the required high-voltage signal. Critical trade-offs exists between the switching speed of levels shifters, contention mechanisms that slow down and reduce the efficacy of the switching and the dissipated power. With the scaling down of the digital supply level, new solutions have been proposed. Some multi-stage approaches use intermediate voltage to gradually increase the dynamic of the signal [91]. This approach is relatively straightforward to implement, but speed and area are negatively affected. One of the most diffused approach relies on the contention-reduction scheme, where the switching speed is improved by driving a couple of low-voltage drivers also on the up-level network [94, 95]. Other solutions for power optimization have been developed for specific applications operating the circuit in particular modes depending on the output voltage required (Bypassing Enabled Level Shifter [96]).

Switched capacitor circuits, or charge pumps (CP), are used to generate the HV biases and eliminate the need of integrating transformers and inductors, which may cause potential electromagnetic-interference concerns and increase the cost. Additionally, in flash technology high and well controlled internal programming/erase voltages are required, regardless of the decreasing power supply trend. In general charge pumps are closed loop systems, where the output is regulated at a determined level [97]. Periodically charging and discharging a network of capacitors by reconfiguring the circuit topology permits to achieve voltage boosting.

Different charge pump architectures are used in flash design to provide the high voltages needed both in read, program and erase operation [97–103]. In read mode simple boost circuits or voltage triplers can be sufficient as the voltage on the control gate voltage needs to be ramped only to values within the range of 4 - 5V to detect the state of the cell by measuring its current. The drain voltage is maintained significantly below the maximum V_{DD} supply and thus no CP circuit is needed in this phase. The specifications of the charge pumps for the program mode strongly vary depending on the adopted program technique. In CHE injection, voltages as high as 9V need to be applied to the control gate (word line) of the device. Since no current flows from the control gate to the floating gate due to the thick isolation layer, the charge pump specifications on the load current are relatively relaxed. On the other hand, the bit-line drain voltage in program mode

[3]Footprinting consists in the issue one faces when designing and matching the large size of WL/BL drivers with the reduced dimension of the device at layout level.

needs to be increased to 4V-5V and in this case, due to the low efficiency of CHE injection, the load current requirements can be quite high. Additionally, designers have to consider carefully-estimated loads for both the entire bit lines and word lines. This task can be simplified by the use of complete models for flash devices, which can provide not only the correct estimation of coupling effects and the load, but also the maximum injection current when programming the device. Also in erase regime, high voltages are required to force the FN regime both on the control gate and on the common isolated substrate of the devices. In particular since high negative voltages need to be applied to the WL, negative voltage multipliers are needed in this operating mode.

In Chapter 5 the importance in the development of an accurate flash model for some of the circuit blocks will be discussed with a direct application to chip design in a 65nm technology.

1.4 Conclusion

In this chapter, the technological details of semiconductor flash memory technology have been presented. This non-volatile memory technology relies on stocking the information in an electrically-isolated floating terminal integrated in a double-gate structure. The structure of the cell, the operating regimes and the definitions that determine nominal and worst-case performances have been defined in the view of modelling the device. The two common modelling paradigms (CCM and CBM) in use for several decades have been presented, with an overview of supplemental models for the determination of the transient currents in the device. Finally, insights into IC design for flash circuits have been given, introducing the most common IP blocks to be optimized and stressing why compact models of flash cells are important in determining tape-out time and optimizing the design efficiency.

Chapter 2

From TCAD to compact modeling of flash devices

2.1 Introduction

In this chapter, semi-analytical and compact modeling approaches applied to flash NVM devices are described using TCAD models as reference. The flash structure and modeling methodology are introduced in Section 2.2. DC/AC models have been developed around a surface potential-based compact model for the intrinsic MOSFET part of the device. A novel scalable and physical ONO capacitance model is proposed and applied to simplify compact model extraction and provide geometrical scalability. It has been chosen to build the model adopting the charge balance paradigm introduced in Chapter 1. 2D and 3D TCAD simulations have been performed to validate the approach. The role of couplings and the importance of their bias dependencies is illustrated and modelled.

Afterwards, semi-analytical solutions are adopted for program and erase transient mechanisms with non-local and compact approaches for CHEI and quantum tunneling for erase current. Compact solutions are integrated to develop a SPICE model, called NVM-SPICE, and based on a surface potential solution for the core device. In such an approach, transient currents are calculated using closed-form expressions to support efficient circuit simulation.

The complete validation of the DC and transient model on measurement results performed in an embedded 65nm derivative technology is illustrated in Section 2.3. A description of the model verification procedure from test structures definition to characterization and verification on equivalent transistors and flash cells is provided.

2.2 Structure description and model methodology

The reference flash structure in NOR configuration that has been modelled in this work is presented in Figure 2-1. 2D TCAD process simulations have been performed with a commercial simulation package [104]. Calibration on the 65nm node manufacturing process flow based on measurements has been performed using the most advanced dopant/defect

pair diffusion models. The 3D structure is built extruding the 2D cross-sections along the width of the device. It has been supposed that the doping profile is uniform along the width of the device; this first order solution is acceptable when assuming that the influence of parasitic transistors in proximity of the Shallow Trench Isolation (STI) region in the width direction can be neglected. In the left part of Figure 2-1, a 3D cross-section of the cell in the middle of the channel is presented. The three 2D cross-sections following the planes AA', BB' and CC' are shown in (a), (b) and (c), respectively. The characteristic geometrical dimensions width W, length L and floating gate wing W_{fg} are introduced. This latter parameter represents the extension of the inter–poly structure over the channel region in the width direction and is of critical importance for improving the control of the FG node with the CG.

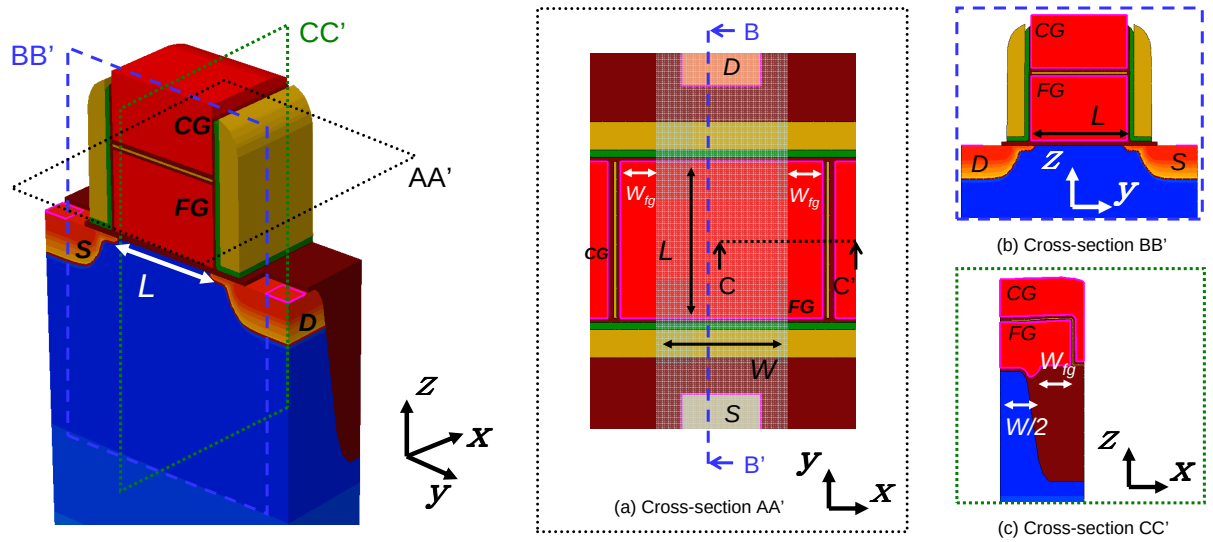


Figure 2-1 – Left: 3D cross-section along the length of the simulated flash structure. Right: 2D cross-sections, (a) top view along AA' plane, (b) lateral view along the length, (c) transversal view along the width. The doping profile of (b) is extruded along this direction.

From the compact modeling point of view, the structure can be divided into three main distinct parts: (a) a MOS device, whose gate is constituted by the floating gate of the cell; (b) a network of capacitances, which electrically isolate the floating gate from the other terminals; (c) a set of current generators for considering the transient currents into the device.

A successful modelling of floating gate devices for circuit simulations requires considering several key issues which include:

- the efficient and physical modelling of the intrinsic MOSFET characteristics using surface-potential or charge-sheet models for the analysis of the active part of the device;

- the accurate calculation of the floating gate voltage V_{FB} in DC, AC and transient conditions following the charge balance paradigm;
- the determination of intra and extra-cell couplings for taking into account 3D effects into the cell structure;
- the evaluation of transient effects, including program/erase currents, controlling the nominal transient dynamics, device degradation and disturbs;
- a physically-based and scalable extraction methodology to permit the model application to next-generation technology nodes.

Figure 2-2 illustrates the features and effects taken into account into the analytical model [105]. The majority of them is investigated in this chapter starting from numerical TCAD simulations for the development of compact approaches. In Chapter 4, complementary effects will be also described, including an endurance model to reproduce the behaviour of degraded devices, disturb parasitic currents, and statistical variations that together with static process corners permit designers to address random spreading of the electrical characteristics and worst-case design conditions.

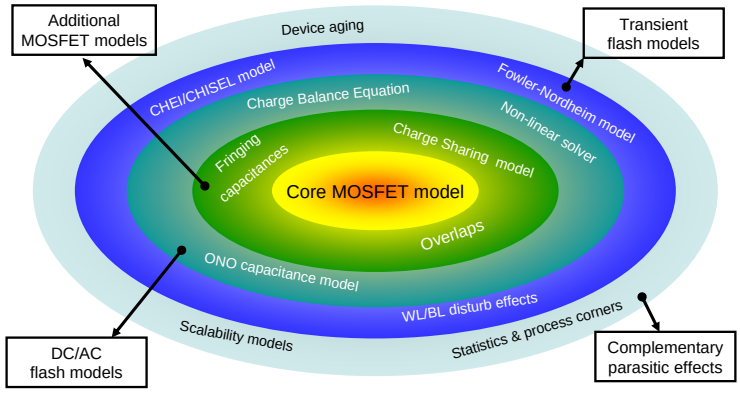


Figure 2-2 – Building blocks and physical effects considered in the proposed flash model.

2.2.1 DC/AC model

(i) Surface potential calculation

In the present approach, it has been chosen to adopt a surface potential model for a more accurate estimation of the potential distribution in the MOS device. Indeed, these models are presently the standard approaches used in industry and recognized by the CMC organism [106]. Furthermore, these approaches permit the direct determination of the potentials in the channel and across the oxide for tunneling current calculation. A charge-based or surface-potential based technique is also required for the implementation of charge sharing model for short channel and overlap effects.

A first critical issue relies on the evaluation of the potential distribution across the channel of the MOS device that represents the active part of the flash cell. Analytical and compact models are usually derived from Pao-Sah approximation [18], where the surface potential along the channel $\psi_S(y)$ at the Si/SiO_2 of the MOS is determined. The development of the 2D Poisson equation and the application of the gradual channel approximation [17, 107] leads to the surface potential equation (SPE) to be solved at each position y along the channel:

$$\begin{aligned}(V_{FB} - V_{fb} - \psi_S(y))^2 &= \gamma^2 \psi_T \left(e^{-\frac{\psi_S(y)}{\psi_T}} + \frac{\psi_S(y)}{\psi_T} - 1 + \right. \\ &\left. + e^{-\frac{2\psi_F}{\psi_T}} \left(e^{\frac{\psi_S(y) - V_C(y)}{\psi_T}} - \frac{\psi_S(y)}{\psi_T} - 1 \right) \right), \end{aligned} \tag{2.1}$$

where V_{FB} denotes the floating gate voltage, V_{fb} is the flat band voltage, $\psi_T = k_B T/q_0$ is the thermal voltage calculated from the device temperature T; $\psi_F = -(E_F - E_I)/q_0$ indicates the potential difference between the semiconductor Fermi level E_F and the intrinsic Fermi level E_I. The channel voltage $V_C(y)$ in the channel varies from the source at $y = 0$ to the drain $y = L$ from $V_C(0) = 0$ to $V_C(L) = V_{DS}$, respectively. The body factor γ is expressed as:

$$\gamma = \frac{\sqrt{2 q_0 \epsilon_0 \epsilon_{Si} N_{ch} \psi_T}}{C_{OX}}, \tag{2.2}$$

being N_{ch} the channel doping and $C_{OX} = \epsilon_0 \epsilon_{ox}/T_{OX}$ the tunnel-oxide surface capacitance and $\epsilon_{Si}/\epsilon_{ox}$ the Si/SiO_2 relative permittivities.

Eq. 2.1 is a non-linear expression, whose solution requires an iterative scheme. An analytical solution based on Taylor series approximation is commonly adopted in compact surface potential-based models [28, 108] determining ψ_S in the source and in the drain and applying the Symmetric Linearisation Method (SLM) for the linearisation of the bulk and inversion charges (see Annex A and [109] for further details on the approximations introduced). The application of the charge sheet approximation introduced by Brews [17], permits to derive closed-form expressions for the terminal charges and for the current I_{DS} in the channel.

(ii) ONO capacitance model

Having calculated the electrostatics of the active part of the cell, a novel physical model has been developed and implemented to accurately determine the capacitance between the control gate and the FG, separated by an oxide-nitride-oxide (ONO) dielectric layer. This model has been detailed in [110]. Most compact models used in industry consider fringing or corner capacitances as fitting parameters. This is not appropriate when technology scalability must be taken into account. The proposed model is based on a structure-decomposition approach and on the principle of Gauss law integration along electrical field lines between the CG and the FG [111, 112].

Figure 2-3 shows the cross-sections along the length and width of the device where the structure decomposition principle is applied. The ONO capacitance C_{CF} determined by:

$$C_{CF} = C_{pt} + 2(C_{pl} + C_{fl} + C_{ft} + 2C_{crn}), \tag{2.3}$$

results from the sum of: (a) two parallel–plate capacitances (C_{pt} and C_{pl}); (b) two parasitic fringing capacitances (C_{ft} and C_{fl}) and (c) the capacitance C_{crn} in the edge corner separating C_{pt} from C_{pl}. The capacitances C_{pt} and C_{pl} represented in Figure 2-3(a-b) are calculated as $C_{pt} = \epsilon_0\epsilon_{ono}\frac{(W_{fg}+W)L}{T_{ONO}}$ and $C_{pl} = \epsilon_0\epsilon_{ono}\frac{(T_{PO1}-T_{ONO})L}{T_{ONO}}$, respectively. T_{PO1} and T_{ONO} indicate the polysilicon gate and ONO stack thicknesses, respectively. An effective ONO permittivity ϵ_{ono}, estimated from electrical and physical measurements on large capacitor structures, has been adopted[1]. The corner component of C_{CF} represented in Figure 2-3(c) has been calculated using the Gauss law in polar coordinates $C_{crn} = \frac{\pi\epsilon_0\epsilon_{ono}}{2}L$.

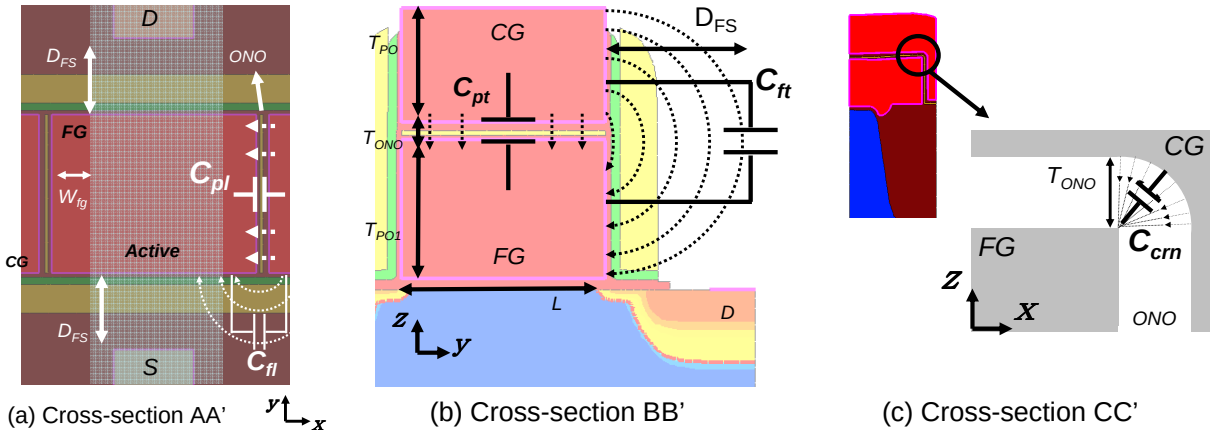


Figure 2-3 – Structure decomposition principle applied to the top view (a) and the cross-section *AA'* (b) of the cell. In (a), a part of the CG has been removed to show the FG. The two components visible in (a) are the lateral parallel–plate capacitance C_{pl} and the lateral fringing capacitance C_{fl}. In (b), a top parallel–plate capacitance C_{pt} and the related fringing capacitance C_{ft} between the sides of the two polysilicon gates can be identified. The field lines are approximated with semi–ellipses. For symmetry reasons, the same capacitances C_{fl} and C_{ft} are present on the source side. In (c), scheme of the cross–section *CC'* showing the corner edge capacitance C_{crn} and how the electrical field lines distribution is approximated in this region.

The fringing capacitances C_{ft} and C_{fl} represent the coupling between the sides of the FG and CG and have been calculated integrating the Gauss law along the electric field lines. A first application of this approach is found in [113] for the determination of static capacitances, while it has been adopted by Goren in [111, 112] for the calculation of the fringing components in transmission lines. Such a method is based on the assumption that the shape of the field lines is known and can be modelled by mathematical expressions.

[1]Similar considerations could be applied separating the ONO stack in the different layers.

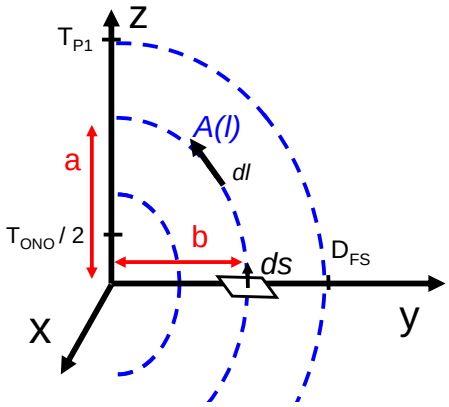


Figure 2-4 – Scheme illustrating the principles for the calculation of the fringe capacitance components by integration along the field lines.

TCAD electrical simulations on both the 2D cross–sections (a) and (b) have been performed to justify the approximation of field lines in semi–elliptical shape. Using Ramanujan's approximation [114] their length is given by:

$$\int A(l)dl \approx \frac{\pi}{2}\left[3(a+b)-\sqrt{(3a+b)(a+3b)}\right], \tag{2.4}$$

where a and b indicate the semiaxes of the ellipses. Following Figure 2-3 and 2-4, these can be correlated by a linear relation. For C_{fl}, one can assume $a = y$ and $b = \frac{W+W_{fg}}{2D_{FS}}y + \frac{T_{ONO}}{2}$, while for C_{ft} one has $a = y$ and $b = \frac{T_{PO1}}{D_{FS}}y + \frac{T_{ONO}}{2}$.

Assuming that the electric field F is constant along the line $A(l)$ and that the total length of each line can be computed, we can write:

$$F = \frac{V_{CF}}{\int_l A(l)dl}, \tag{2.5}$$

where V_{CF} is the voltage between the two considered terminals. We now calculate the charge on the plates using Gauss law: $Q_{CF} = \int_S S(s)\epsilon_0\epsilon_{ox}Fds$, considering a surface S at the polysilicon/oxide interface collecting the field lines and supposing that the field line is perpendicular to the surface (constant potential in the terminal). The fringing capacitances $C_{fr} = \{C_{fl}, C_{ft}\}$ per unit length are thus given by:

$$C_{fr} = \frac{Q_{CF}}{V_{CF}} = \epsilon_0\epsilon_{ox}\int_S \frac{S(s)ds}{\int A(l)dl}. \tag{2.6}$$

Using $ds = dxdy$ and knowing that the field lines are parallel in the width direction, the previous equation can be rewritten with only the integration along y.

After substitution of $A(l)$ in Eq. 2.4, one finds:

$$C_{fr} = \epsilon_0\epsilon_{ox}\kappa_{fr}\int_0^{D_{FS}} \frac{dy}{\pi(\lambda_{fr}y + \eta_{fr} - \sqrt{\alpha_{fr}y^2 + \beta_{fr}y + \gamma_{fr}})}, \tag{2.7}$$

in which the coefficients are associated with geometrical parameters and calculated as in Table 2.1 for the two fringing capacitances. A more compact approach considers straight lines for modelling the field lines:

$$\int_l A(l)dl \;\approx\; \sqrt{a^2+b^2}. \tag{2.8}$$

Consequently, one finds:

$$C_{fr} = p_{frC}\frac{\kappa_{frC}\epsilon_{ox}\epsilon_0}{2\sqrt{\alpha_{frC}}}\ln\left(1+\frac{4\alpha_{frC}D_{FS}+2\sqrt{\alpha_{frC}\beta_{frC}D_{FS}}}{\beta_{frC}+2\sqrt{\alpha_{frC}\gamma_{frC}}}\right), \tag{2.9}$$

where the prefactor $p_{frC}=1/\sqrt{2}$ has been introduced to take into account the curvature of the field lines and attenuate the overestimation of the capacitances.

Parameter	**Value for C_{ft}**	**Value for C_{fl}**
α_{fr}	$(3+\frac{T_{PO1}}{D_{FS}})(1+\frac{T_{PO1}}{D_{FS}})$	$(3+\frac{W+W_{fg}}{2D_{FS}})(1+\frac{W+W_{fg}}{2D_{FS}})$
β_{fr}	$5T_{ONO}+3\frac{T_{ONO}T_{PO1}}{D_{FS}}$	$5T_{ONO}+3\frac{T_{ONO}(W+W_{fg})}{2D_{FS}}$
γ_{fr}	$\frac{3}{4}T_{ONO}^2$	$\frac{3}{4}T_{ONO}^2$
η_{fr}	$\frac{3}{2}T_{ONO}$	$\frac{3}{2}T_{ONO}$
λ_{fr}	$3(1+\frac{T_{PO1}}{D_{FS}})$	$3(1+\frac{W+W_{fg}}{2D_{FS}})$
κ_{fr}	$2(W+W_{fg}-T_{ONO}/2)$	$2(T_{PO1}-T_{ONO}/2)$
α_{frC}	$\frac{T_{PO1}^2}{D_{FS}^2}+1$	$\frac{(W+W_{fg})^2}{4D_{FS}^2}+1$
β_{frC}	$\frac{T_{ONO}T_{PO1}}{D_{FS}}$	$\frac{T_{ONO}(W+W_{fg})}{2D_{FS}}$
γ_{frC}	$\frac{1}{4}T_{ONO}^2$	$\frac{1}{4}T_{ONO}^2$

Table 2.1 – Model parameters used for the calculation of the fringe capacitance components of the ONO capacitance. The geometrical variables adopted are indicated in Figure 2-3.

To validate the model, 3D TCAD numerical simulations are required as the value of the ONO capacitance cannot be measured due to the morphology of the structure. Indeed, due to the complete isolation of the FG, the terminal cannot be contacted for standard AC analysis. Measurements on equivalent transistor devices are not useful to extract the accurate value of C_{CF} due to the mismatch between the device structures. Consequently, the value of the ONO capacitance has also been extracted from 3D TCAD AC simulations and used for validating both the analytical and compact ONO capacitance models, for various lengths L ($0.12\mu\text{m} < L < 1.00\mu\text{m}$) and widths W ($0.08\mu\text{m} < W < 1.00\mu\text{m}$). Figure 2-5 shows the role of the different components in the calculation of the total capacitance, as a function of both L and W of the cell. Figure 2-6 shows the predictions of the ONO capacitance model for all the simulated dimensions of the device, including the FG wing W_{fg}, varying from 45nm to 115nm. The distance to contact D_{FS} is less than

90nm and $T_{PO1} = T_{PO0} < 100$nm. Good agreement is found, demonstrating the capability of the model to accurately estimate the FG–CG coupling capacitance, even when the cell is aggressively scaled down to nanoscale dimensions, e.g. for embedded and high-density applications.

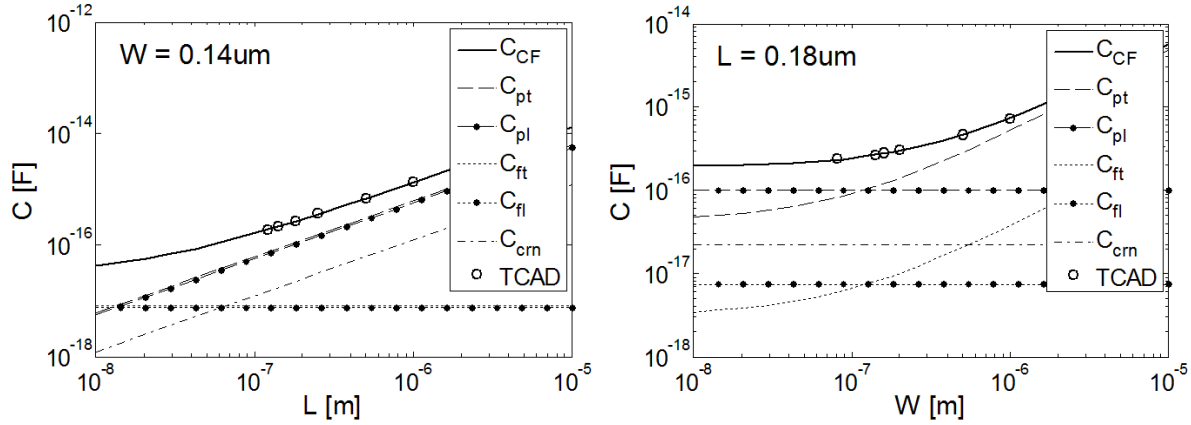


Figure 2-5 – Scaling of the components of the ONO capacitance as a function of length L (left) and width W (right) of the device. 3D TCAD simulations are reported in symbols.

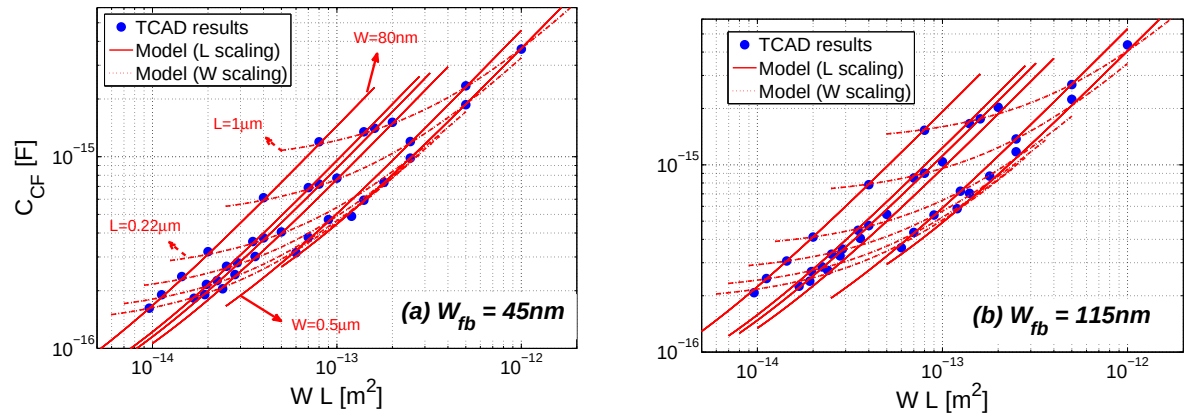


Figure 2-6 – C_{CF} capacitance versus area of the active device $W \cdot L$ for a FG wing W_{fg} of 45nm (a) and 115nm (b). Results calculated with the compact model are shown in red lines while 3D TCAD simulations are indicated in blue symbols. Dashed isolines indicate devices having the same L where only W is scaled, while devices on the same solid lines illustrate the scaling in L.

(iii) Charge balance and coupling coefficients

Using the charge balance paradigm introduced in Chapter 1, an iterative root-finding algorithm to determine the floating gate potential V_{FS} has been applied. Considering also

transient effects, the charge balance equation is expressed as:

$$Q_G(V_{FS}, V_{DS}, V_{BS}) + C_{CF}(V_{FS} - V_{CS}) = Q_{F0} + \int (I_{FS} + I_{FB} + I_{FD})dt, \tag{2.10}$$

where the C_{CF} capacitance is determined using the aforementioned ONO capacitance model, the initial net charge Q_{F0} on the floating gate remains constant in DC simulations and the MOSFET charge Q_G is determined at each iteration from the value of ψ_S (see Section A.2 in Annex A). The currents I_{FS}, I_{FB} and I_{FD} between the floating gate and the other terminals of the structure are determined with one of the transient models described afterwards in Subsection 2.2.2. Since an analytical expression of the derivative of the gate charge with respect to V_{FB} is not available, the equation is solved with a *regula falsi* method [115], which guarantees both convergence and speed (convergence order ≈ 1.6). The model has been implemented in Verilog-A and integrated into an industrial SPICE simulator; this solutions offers a good compromise between portability for SPICE simulators, computational speed and industrial application.

Figure 2-7(a) shows the variation of the floating gate voltage in DC with the control gate voltage V_{CS} and the charge Q_{F0} in the FG node for a long and wide device. An uniform doping profile with constant value $N_{ch} = 1.45 \cdot 10^{18}\text{cm}^{-3}$, while an oxide thickness T_{OX}=98Å has been used. The slope of the $V_{FS}(V_{CS})$ curve shows non-linearities in weak inversion and represents the coupling coefficient α_C. In (c), the dependence of α_C on N_{ch}, T_{OX} and W_{fg} is shown. The value of α_C peaks in proximity of the V_{th} of the equivalent transistor device, where inversion occurs, to become relatively flat at higher voltages.

Two-dimensional TCAD simulations have been performed on the 2D cross-section in Figure 2-1(b) for various device dimensions, avoiding the influence of width and STI effects. The absolute value of α_C is lower compared to 3D simulations, but the shape of the curve is preserved when varying W or W_{fg}. Figure 2-8 illustrates the results over a wide range of bias voltages, revealing the dependence of the coupling coefficients α_C and α_D for a long device where extrinsic effects are not considered. Both the coupling coefficients present a strong variation with the control gate and drain bias voltages. An investigation of the impact of short channel effects on flash devices using a dynamic charge sharing approach is provided in the following section and in [116].

(iv) Charge sharing and short channel effects

Short-channel effects (SCE) are mostly originated by the limitation imposed on electron drift characteristics in the channel and by the modification of the threshold voltage due to the shortening of the channel length [117]. Additionally advanced devices and flash memory cells make use of back-bias polarization to tune the threshold voltage or improve program efficiency, and thus an accurate modeling of the V_{BS} dependence in short devices has to be included.

An adaptation of the static charge-sharing model originally proposed by Wu [118] for V_{th}-based models and reprised in [119] with a dynamic version suitable for a charge-sheet approach has been developed and integrated in the analytical model to investigate SCE

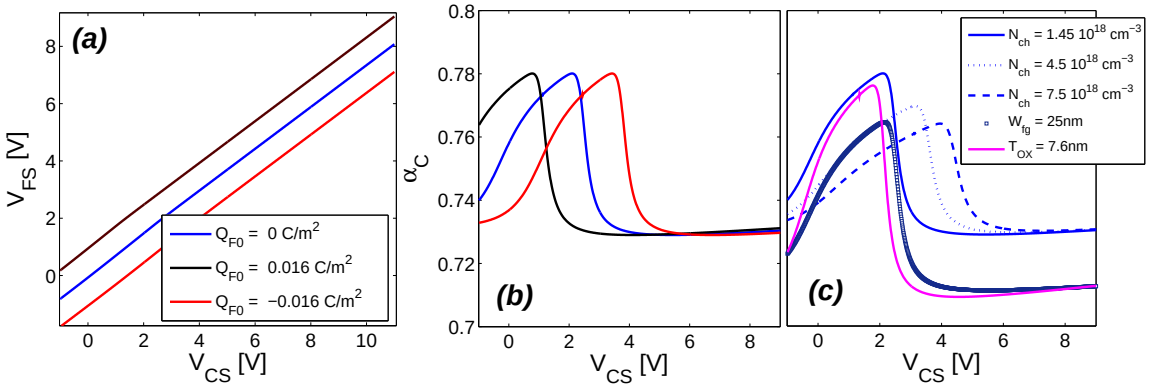


Figure 2-7 – (a) Floating gate voltage versus control gate voltage at $V_{DS} = 0V$ for different values of the charge stored on the floating gate Q_{F0}. When Q_{F0} and V_{CS} are null a negative V_{FS} is found due to the compensation of the positive charges on the bulk. Non-linearities can be found when comparing the control gate coupling coefficient α_C versus V_{CS} in (b). The coupling is affected by both process variations and layout parameters as in (c) and determines most of the cell performances.

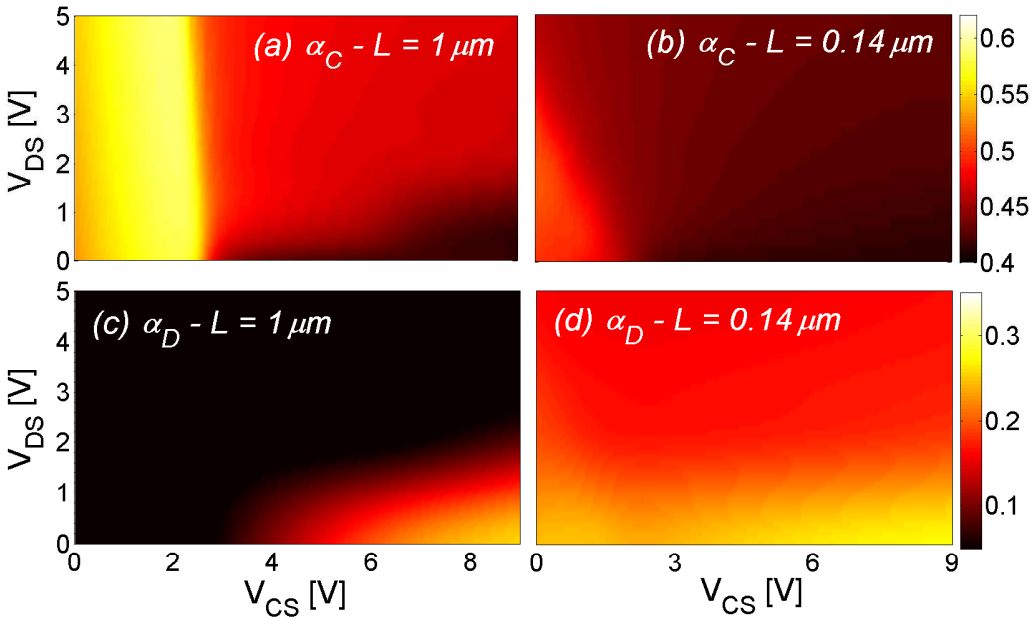


Figure 2-8 – Simulated control gate α_C(a,b) and drain α_D (c,d) coupling coefficients dependencies versus V_{CS} and V_{DS} biases, obtained using their definition in Eq. 1.3 for long ((a,c) $L = 1\mu$m) and short ((b,d) $L = 0.14\mu$m) flash cells.

on the flash electrostatics. The complete treatment could be found in [116]. The charge sharing effect is due to a gate voltage-dependent reduction of the depletion charge. This reduction can be estimated considering geometrical aspects as in [119], where the variation

of the source and drain depletion lengths X_S and X_D in the bulk junctions and the one near the interface x_S and x_D depend on the bias voltages and on ψ_S. Consequently, one has:

$$\begin{aligned} X_{\{S,D\}} &= \sqrt{\frac{2\epsilon_0\epsilon_{Si}}{q_0 N_{LDD}}(V_{bi}+V_{\{S,D\}B})}, \\ x_{\{S,D\}} &= \sqrt{\frac{2\epsilon_0\epsilon_{Si}}{q_0 N_{LDD}}(V_{bi}+V_{\{S,D\}B}-\psi_S)}. \end{aligned} \tag{2.11}$$

The geometrical quantities are indicated in the schematic of Figure 2-9. The depletion charge contribution proportional to ψ_S in long devices, is corrected using:

$$q_{dep}(\psi_S) = \gamma\sqrt{\psi_S}C_{OX}\cdot F(\psi_S), \tag{2.12}$$

where $F(\psi_S)$ is a geometrical charge sharing coefficient resulting from the trapezoidal and voltage–dependent shape of the charge distributions in the channel region:

$$F(\psi_S) = 1 - \frac{x_s + x_d + X_s + X_d}{2L}. \tag{2.13}$$

This factor also represents the percentage of bulk charge controlled by the gate. Eq 2.1 can be thus modified to take into account SCE and the depletion charge modification according to the charge sharing principles:

$$\begin{aligned} (V_{FB}-V_{fb}-\psi_S)^2 &= \gamma^2\psi_T\left(e^{-\frac{\psi_S}{\psi_T}} + F^2(\psi_S)\frac{\psi_S}{\psi_T} - 1 + \right. \\ &\quad \left. + e^{-\frac{2\psi_F}{\psi_T}}\left(e^{\frac{\psi_S-V_C}{\psi_T}} - \frac{\psi_S}{\psi_T} - 1\right)\right). \end{aligned} \tag{2.14}$$

This effect is equivalent to considering a surface potential–dependent body factor (or equivalently doping), which traduces the reduction of the control of the channel in short devices. Rigorously, Eq. 2.14 should be evaluated at each position y in the channel. However, this would compromise the symmetric linearisation assumption and the symmetry of the device. To avoid this issue, we computed only a single factor F using an avarage of the value computed at the drain and at the source of the device.

In Figure 2-10, the model predictions have been compared to 2D TCAD simulation results illustrated in Figure 2-8 for different device lengths (0.14μm, 0.37μm and 1μm). In this comparison, the surface potential equation has been solved with the implicit semi-analytical approach as in Eq. 2.1. Excellent agreement between TCAD and model simulations has been found on the bias dependency of α_C. This results from the inclusion of both the dynamic charge sharing and overlap capacitance models (see the following Section). Indeed, the SCE model adopted in the analytical model and based on the dynamic charge sharing [116], not only accurately reproduces the V_{th} roll–off, but has also a strong impact on the charges in short devices altering the role of Q_G in the charge balance equation.

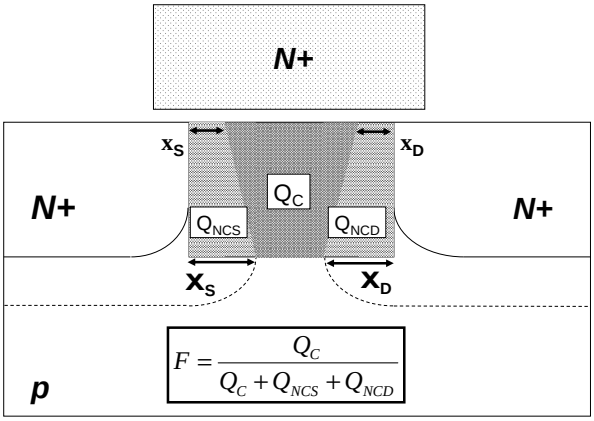


Figure 2-9 – Schematic of the dynamic charge sharing approach with the represented geometrical quantities. In the static model of Wu [118] only the variations of source and drain depletion widths X_S and X_D in the deep junction have been considered.

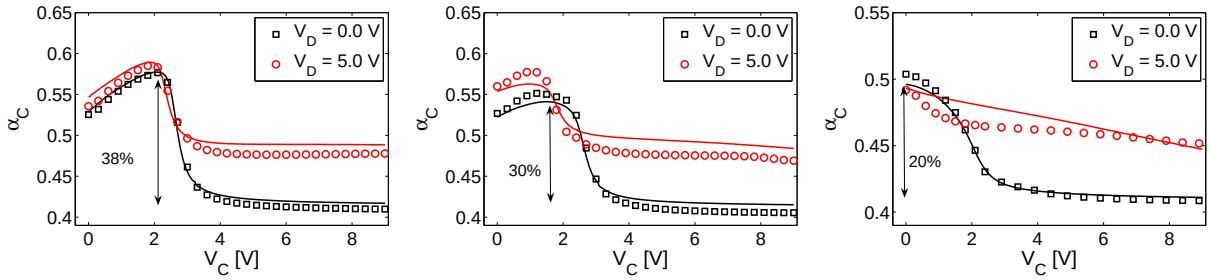


Figure 2-10 – Matching of TCAD (lines) and semi-analytical model (symbols) simulations of the gate coupling coefficient α_C as a function of the control gate voltage V_{CS}, for different drain biases V_D and different cell dimensions: (left) L=1μm, (middle) L=0.375μm, (right) L=0.14μm. In long devices the CG has a better control over the FG but the variation in V_{CS} can reach even 38% of its magnitude and consequently it needs to be considered for an accurate modeling. For decreasing lengths the effects of parasitic capacitances dominate, causing a global decrease of α_C and an increase of α_D. Additionally charge sharing reduces the control of the gate over the channel and the variation in V_{CS} is smoothed. Threshold voltage roll-off also occurs.

While the integration of such a dynamic charge sharing is relatively simple in semi-analytical charge sheet models that solve the SPE with an iterative scheme, its implementation in compact approaches when the surface potential is solved by means of an analytical approximation remains problematic [28]. Certain modern industrial approaches solve the SPE twice [28], adopting a different set of parameters for the two regimes to model the effects of charge sharing or non-uniform doping and improve the model agreement with DC and AC results. While this technique can lead to good results in both the regimes, it remains disputable due to intrinsic inconsistencies in the different regimes of operation and challenges the model parameter extraction methodology. It can be seen that the application of static charge-sharing approaches to the closed form SPE in industrial compact models is feasible, since in such a case the charge sharing coefficient S is constant for a given operating point and does not depend on ψ_S. Moreover, capacitance reciprocity is

not guaranteed in all bias conditions when adopting such an approach [120].

The effects of narrow channel and full geometrical scaling of the cell have been studied by means of 3D TCAD simulations. Results are reported in Figure 2-11. In (a), the decrease of the width strongly increases α_C due to the linear reduction of the overlap capacitance between the floating gate and the channel. Additionally, due to the morphology of the cell in the STI region, parasitic transistors presenting different mobility and V_{th} can alter the shape of the $\alpha_C(V_{CS})$ curve in weak inversion regime. In (b), the cell is scaled along the length from $1\mu m$ to 120nm. Charge sharing phenomena and the increase of the importance of the overlap capacitances decrease the coupling. In (c), two floating-gate wing dimensions are simulated. The decrease of W_{fg} reduces the ONO capacitance and α_C. Two drain voltage conditions are also shown to highlight the effects of DIBL.

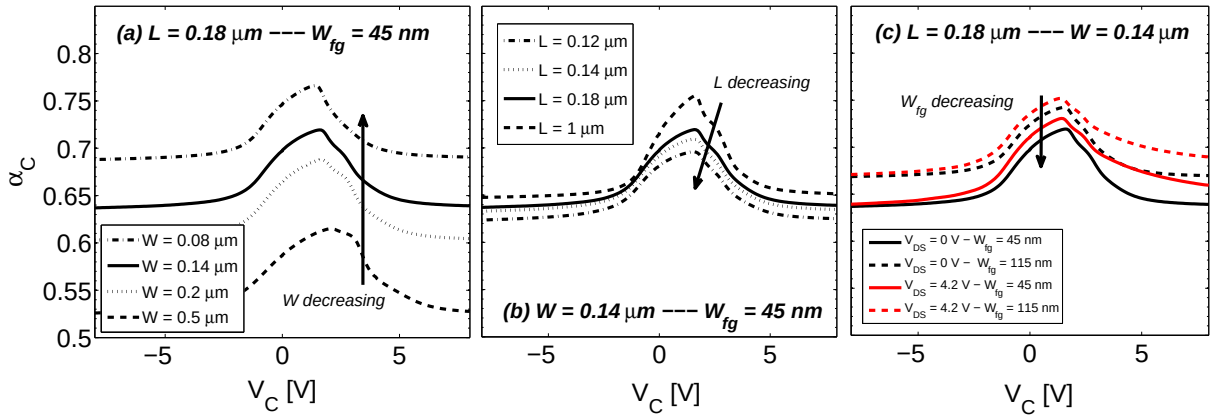


Figure 2-11 – Dependence of the control gate coupling coefficient α_C calculated with 3D TCAD simulations as a function of the control gate bias voltage V_{CS}, with the geometrical dimensions of the cell (width W (a), length L (b) and floating gate wing W_{fg} (c)).

(v) Overlap charge and capacitance models

The source and drain diffusions under the transistor gate create overlap voltage-dependent capacitances, which depend on both the gate-drain/source and gate-bulk voltages $V_{F\{D,S\}}$ and V_{FB}. A quantitative description of the overlap regions is obtained treating the low-doped-drain (LDD) regions as n^+-gate/oxide/n^+-bulk MOS capacitances, where the drain / source act as bulk terminals. In the n^+ overlap region, a non-uniform doping profile $N_{ov}(y)$ is associated to a body factor $\gamma_{ov}(y)$ and a flat band voltage $V_{fbov}(y)$. An analytical approach has been introduced to compute the overlap capacitances in all the regions of operations. A three-terminal charge sheet model based on Klein's approach [121] has been applied having a double-gaussian doping profile in the LDD junction region. The profile has been modeled using a 1D model along the length of the overlap region, modeling both

the high-level and the low-level doping profiles:

$$N_{ov}(y) = N_{LDD}\left[\exp\left(\frac{(y-y_{LDD})^2}{2\sigma_{LDD}^2}\right) + A\exp\left(\frac{(y-L_{ov})^2}{2\sigma_D^2}\right)\right], \tag{2.15}$$

with amplitudes N_{LDD} and AN_{LDD}, mean position y_{LDD} and L_{ov}, and variances σ_{LDD} and σ_D. A numerical integration over a limited amount of spatial points in the LDD is considered for the calculation of the total overlap charges. After integration along y, the bulk Q_{Bov} and inversion Q_{Iov} charge densities are added to the drain/source and bulk terminal charges of the intrinsic device, respectively.

The surface potential in the LDD region is approximated with:

$$\psi_{Sov} = \psi_{Sov}^{sub}\left(1 - \frac{1}{f_s\psi_{Sov}^{inv}\ln(1+\exp\left(f_s(\psi_{Sov}^{inv}-\psi_{Sov}^{sub})\right))}\right)^{-1}. \tag{2.16}$$

The smoothing parameter f_s is introduced for matching subthreshold and inversion regimes, where the surface potentials are determined using $\psi_{Sov}^{sub} = -\left(\frac{-\gamma_{ov}}{2} + \sqrt{\frac{\gamma_{ov}^2}{4} - (V_{FD}-V_{FB})}\right)^2$ and $\psi_{Sov}^{inv} = 2\psi_{Fov} + V_{BD} - 6\psi_T$, respectively. Being the overlap region n-doped, accumulation occurs when $V_{FD} > V_{fbov}$ and the surface potential is negligible due to charge-sheet layer of e^- at the Si interface. The majority carrier charge is then computed with:

$$\begin{aligned} Q_{Gov} &= Q_{Bov} = -C_{OX}(V_{FD}-V_{fbov}), \\ Q_{Iov} &= 0, \end{aligned} \tag{2.17}$$

On the other hand, in depletion or inversion the charges are computed from the surface potential using:

$$\begin{aligned} Q_{Bov} &= -\gamma C_{OX}\sqrt{-\psi_{Sov}}, \\ Q_{Iov} &= -C_{OX}(V_{FD}-V_{fbov}-\psi_{Sov}+\gamma\sqrt{-\psi_{Sov}}), \\ Q_{Gov} &= -C_{OX}(V_{FD}-V_{fbov}-\psi_{Sov}), \end{aligned} \tag{2.18}$$

A general model valid in all the regions of operation can be obtained using a smoothing function around the flat-band voltage [121].

In more compact approaches, complete analytical expressions are adopted. In such a case, the surface potential ψ_{Sov} is computed only in one point assuming a constant doping in the LDD [28]. The surface potential in the overlap region ψ_{Sov} is determined neglecting the minority carrier contribution to the space charge. The total charge in the overlap region is thus expressed by $Q_{Dov} = C_{FD}V_{OV}$ where the C_{FD} is the oxide capacitance in the overlap region, that can be calculated from layout and geometrical considerations, but is often treated as a fitting parameter, while $V_{OV} = V_{FD} - V_{fbov} - \psi_{Sov}$ is the oxide voltage drop in the overlap region.

(vi) Fringing capacitance models

Fringing capacitances in the overlap region of the MOSFET active component have to be modeled to achieve an accurate electrostatics of the device. As reported in [38], the coupling coefficients in the flash device are strongly affected by the overlap capacitance models and in particular by the fringing components above the LDD region and in the spacer region. Figure 2-12 indicates the field lines in the LDD and spacer regions together with the geometrical quantities required for the calculation.

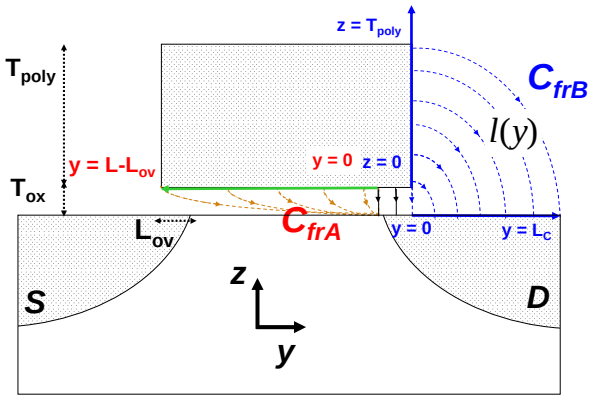


Figure 2-12 – Schematic of fringing capacitance models, approximating the electrical field lines from the floating gate to the drain/source in the spacer and in the LDD region with elliptical arcs or straight lines.

Previous approaches tends to underestimate the importance of $\alpha_{\{D,S\}}$ by considering only the parallel plate components in the overlap region. In his work, Larcher [38] introduces a geometrical technique for the determination of the fringing components considering geometrical parameters and linear field lines in the tunnel oxide and in the spacer region.

In this study, we propose an approach to take into account the fringing capacitance components with the distribution of the electric field lines from the floating gate terminal to the drain contact and LDD region and added to the overlap capacitance calculated in the previous subsection. Similarly to Eq. 2.4, arcs of ellipses could be considered for the calculation of the fringing capacitance in the overlap region. This leads to an expression for C_{fr} similar to Eq. 2.7, where the coefficients are expressed as in Table 2.2 for the two components.

The total capacitance is calculated summing the two fringe components with the parallel plate contribution. All extrinsic charges have been added to the charge-balance equation as well.

Overlap charges significantly impact the total amount of charges in the charge balance for short devices and thus must be accounted for in the calculation. Simulations have been performed progressively enabling the SCE and overlap capacitance models. In Figure 2-13, it can be seen that the intrinsic behavior of α_C is masked by SCE and overlap capacitance effects when the two models are enabled. When SCE are introduced, the peak is shifted to lower V_C (V_{th} roll–off) and its amplitude decreases. Including overlap and fringing capacitances in the calculation, the curve is further flattened and shifted to lower values of α_C.

Parameter	Value for C_{frA}	Value for C_{frB}
α_{fr}	3	$3\frac{T_{P0}^2}{D_{FS}^2} + 10\frac{T_{P0}}{D_{FS}} + 3$
β_{fr}	$10\ T_{OX}$	$10T_{OX} + 6T_{OX}\frac{T_{P0}}{D_{FS}}$
γ_{fr}	$3T_{OX}^2$	$3T_{OX}^2$
η_{fr}	$3\ T_{OX}$	$3\ T_{OX}$
λ_{fr}	3	$3\ (1 + \frac{T_{P0}}{D_{FS}})$
κ_{fr}	4W	4W

Table 2.2 – Model parameters used for the calculation of the fringe capacitance components in the drain/source contact region.

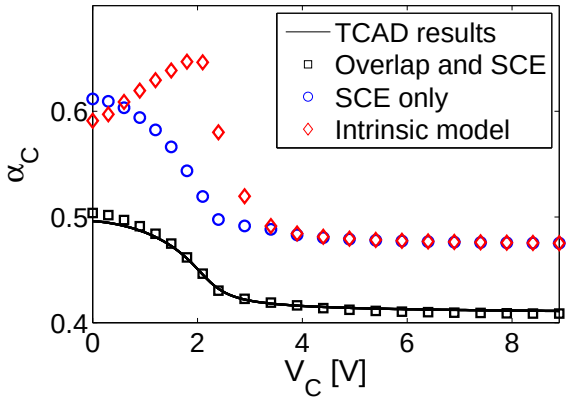


Figure 2-13 – Impact of extrinsic effects on the coupling coefficient α_C for different control gate voltages V_{CS} at $V_{DS} = V_{BS} = 0V$. The lines are referred to TCAD simulations, the symbols indicate model results.

(vii) Velocity saturation effects

Terminal charges and thus capacitive couplings and charge balance are affected by velocity saturation phenomena. The carriers' drift velocity v in the channel saturates at high lateral fields, reaching the value v_{sat}. To model this behavior, semi-empirical models [122] are introduced, determining v with:

$$v = \frac{\mu \cdot \frac{\partial \psi_S}{\partial y}}{\sqrt{1 + \left(\frac{\mu}{v_{sat}} \cdot \frac{\psi_S}{\partial y}\right)^2}}, \tag{2.19}$$

where μ is the carrier mobility in the channel (see Annex A.3 for details of the components affecting mobility degradation in ultrascaled devices). The lateral field is given by $F_{lat} = -\frac{\psi_S}{\partial y}$.

The approximate linear relationship between the inversion change and ψ_S is used to define a virtual inversion charge density Q_V [123], that differs from the usual expression

by a constant term, obtaining:

$$Q_V = Q_I - nC_{OX}\psi_T + \frac{I_{DS}}{Wv_{sat}}, \tag{2.20}$$

where:

$$n = 1 + \frac{\gamma}{2\sqrt{|\psi_m + \psi_T|}}, \tag{2.21}$$

and with ψ_m the surface potential at midpoint in the channel.

Given the dependence of Q_V on the current I_{DS}, it is evident that an expression of I_{DS} based on the virtual charge Q_V is not useful for compact modeling as it requires a non-linear solution. Eventually, in compact approaches where velocity saturation effects on the terminal charges are considered, the total inversion charge can be expressed as:

$$Q_I = \int_0^{L-\Delta L} q_i dy + \frac{\Delta L}{L} q_{isat}. \tag{2.22}$$

In other words, the channel can be divided in two separate regions (saturated and non-saturated). In the saturated region $[L - \Delta LL]$, the inversion charge density is assumed to be constant at q_{isat}, while in the non-saturated one $[0L - \Delta L]$ the usual expression holds. Expressing Q_I as a function of the virtual charges in the drain and source (Q_{VL} and Q_{V0}, respectively) [123], one has:

$$Q_I = \frac{L - \Delta L}{L}\left(\frac{2}{3}\frac{Q_{VL}^2 + Q_{VL}Q_{V0} + Q_{V0}^2}{Q_{VL} + Q_{V0}} + nC_{OX}\psi_T\right) - \frac{I_{DS}}{Wv_{sat}}, \tag{2.23}$$

Corrected expressions for the bulk and total charges (Section A.2 of Annex A) hold:

$$\begin{aligned} Q_B &= -\frac{n-1}{n}Q_I - sgn(\psi_m)\gamma C_{OX}\sqrt{\psi_m + \psi_T(\exp(-\psi_m/\psi_T) - 1)}, \\ Q_G &= -Q_B - Q_I, \end{aligned} \tag{2.24}$$

while after partitioning [109, 123, 124], one has:

$$\begin{aligned} Q_D &= \frac{L - \Delta L}{L}\left(\frac{6Q_{V0}^3 + 12Q_{VL}Q_{V0}^2 + 8Q_{VL}^2Q_{V0} + 4Q_{VL}^3}{15(Q_{VL} + Q_{V0})^2} + \frac{n}{2}C_{OX}\psi_T\right) - \frac{I_{DS}}{2Wv_{sat}}, \\ Q_S &= Q_I - Q_D. \end{aligned} \tag{2.25}$$

Velocity saturation occurs in short devices at high V_{DS} and V_{CS}, when the cell is usually programmed and an accurate model of α_{C} can offer a better estimation of the injection current. As depicted in Figure 2-14, we demonstrated that also this parameter can play a role on α_{C} bias dependency at the sub-micron scale.

In summary, the analytical simulations on sub-micron devices confirm that extrinsic

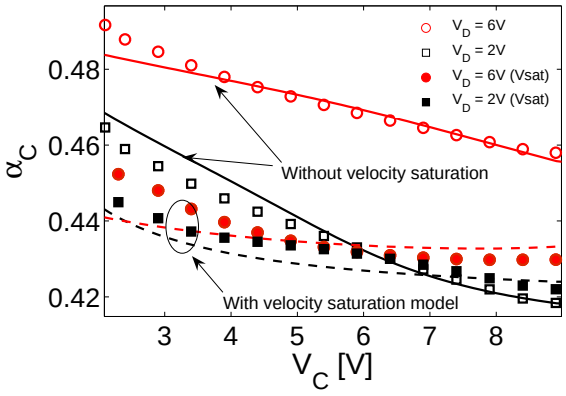


Figure 2-14 – Impact of velocity saturation on α_C for different values of V_{DS}. The lines are referred to TCAD simulations, the symbols indicate model results.

effects (charge sharing, overlaps and velocity saturation) dominate over the intrinsic behaviour of the cell, each playing an independent role on the coupling coefficient α_C.

2.2.2 Transient mechanisms

Transient phenomena have been considered by modeling the different current contributions in program, erase and disturb modes and including them in the charge balance.

(i) The general Tsu-Esaki tunneling model

The general Tsu-Esaki model for quantum tunneling in dielectric layers is usually adopted to derive and validate compact expression of the P/E transient currents for tunneling phenomena through a generic energy barrier [50]. In the flash structure, electrons tunneling through the two potential barriers of the tunnel oxide and of the ONO layer should be considered. The profile of the potential barrier strongly depends on quantum confinement and material band structure. Consequently an accurate calculation of the charge and voltage distributions in the device is often required. To this purpose, a 1D self-consistent Poisson-Schrödinger (PS) modeling tool has been developed to act as a reference and accurately calculate the barrier profile and the current through dielectric layers. Annex C describes more in detail the features, applications and implementation details of the aforementioned model. Figure 2-15 depicts the band diagram of a flash device calculated with the PS solver, where the charge balance equation has been integrated to take into account coupling effects and determine the floating gate potential of the device. In these results, the cell has been supposed erased, with no charge on the FG. Both the ONO and tunnel oxide barrier regions are shown in the insets.

The energy barrier representing a dielectric material separates two semiconductor or metallic regions (left electrode L and right electrode R). The total electron net current flowing through a barrier from the left electrode L to the right electrode R reads:

$$J = J_{L\to R} - J_{R\to L}, \tag{2.26}$$

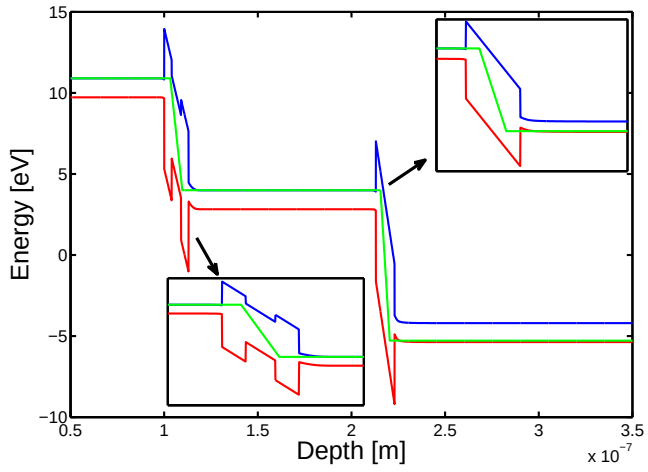


Figure 2-15 – Band diagram of a floating gate device calculated with a self-consistent Poisson-Schrödinger 1D solver at $V_{CS} = 1.9V$. The insets present the two barrier regions through which the tunneling current between two semiconductor regions needs to be evaluated. At high positive voltages, carrier confinement occurs in the channel where charges are distributed on well separated energy levels. In such a case, carriers tunnel from a quasi-bound to a free state in the gate [125]. For smaller bias voltages, a continuum of energetic levels is present in the substrate.

where $J_{L\to R}$ and $J_{R\to L}$ are the left to right and right to left contributions.

Each component of the tunneling current can be related to the barrier transparency T and the distribution functions in the left and right reservoirs (f_L and f_R, respectively):

$$\begin{aligned} J_{L\to R} &= q_0 \int_{\mathbf{k}} T(\mathbf{k}) v(\mathbf{k}) g_L(\mathbf{k}) f_L(\mathbf{k})(1 - f_R(\mathbf{k})) d\mathbf{k}, \\ J_{R\to L} &= q_0 \int_{\mathbf{k}} T(\mathbf{k}) v(\mathbf{k}) g_R(\mathbf{k}) f_R(\mathbf{k})(1 - f_L(\mathbf{k})) d\mathbf{k}, \end{aligned} \tag{2.27}$$

where the integration is performed on the wavevector $\mathbf{k} = k_x i + k_y j + k_z k$, $g(\mathbf{k}) = \frac{1}{4\pi^3}$ denotes the density of states per cubic volume in the momentum space, and $v(\mathbf{k})$ expresses the electron velocity. Assuming that the parallel momentum is conserved and that only transitions in the z directions are considered, the parallel wavevector $\mathbf{k}_\rho = k_x i + k_y j$ is not altered by the process. Moreover the barrier transparency T is assumed to be only dependent on the perpendicular momentum ($T(\mathbf{k}) = T(k_x = 0, k_y = 0, k_z) = T(k_z)$) Under these hypotheses one can write:

$$\begin{aligned} J_{L\to R} &= q_0 \int_{\mathbf{k}} T(k_z) v(k_z) g_L(k_z) f_L(\mathbf{k})(1 - f_R(\mathbf{k})) d\mathbf{k}, \\ J_{R\to L} &= q_0 \int_{\mathbf{k}} T(k_z) v(k_z) g_R(k_z) f_R(\mathbf{k})(1 - f_L(\mathbf{k})) d\mathbf{k}, \end{aligned} \tag{2.28}$$

Considering a parabolic single-band structure model under the effective mass approxima-

tion, the parabolic dispersion relation:

$$\begin{aligned} \mathcal{E} &= \frac{\hbar^2 \mathbf{k}}{2m_{\text{eff}}}, \\ v(k_z) &= \frac{\hbar k_z}{m_{\text{eff}}} \Rightarrow v(k_z)dk_z = \frac{d\mathcal{E}_z}{\hbar}, \end{aligned} \tag{2.29}$$

is valid. The carrier effective mass in the carrier reservoirs is m_{eff}. After separating the transversal energy levels $\mathcal{E}_z$ from the longitudinal one $\mathcal{E}_p$, one finds:

$$\begin{aligned} J_{L\to R} &= \frac{4\pi q_0 m_{\text{eff}}}{\hbar^3} \int_{\mathcal{E}_z} T(\mathcal{E}_z) \left(\int_{\mathcal{E}_\rho} f_L(\mathcal{E}_\rho, \mathcal{E}_z)(1 - f_R(\mathcal{E}_\rho, \mathcal{E}_z)) d\mathcal{E}_\rho \right) d\mathcal{E}_z, \\ J_{R\to L} &= \frac{4\pi q_0 m_{\text{eff}}}{\hbar^3} \int_{\mathcal{E}_z} T(\mathcal{E}_z) \left(\int_{\mathcal{E}_\rho} f_R(\mathcal{E}_\rho, \mathcal{E}_z)(1 - f_L(\mathcal{E}_\rho, \mathcal{E}_z)) d\mathcal{E}_\rho \right) d\mathcal{E}_z, \end{aligned} \tag{2.30}$$

$$J = \frac{4\pi q_0 m_{\text{eff}}}{\hbar^3} \int_{\mathcal{E}_z} T(\mathcal{E}_z) N(\mathcal{E}_z) d\mathcal{E}_z, \tag{2.31}$$

where $N(\mathcal{E}_z) = \int_{\mathcal{E}_\rho} \left(f_L(\mathcal{E}_\rho, \mathcal{E}_z) - f_R(\mathcal{E}_\rho, \mathcal{E}_z) \right) d\mathcal{E}_\rho$ is usually called supply function and only depends on the properties of the left and right reservoirs.

The calculation of $N(\mathcal{E}_z)$ requires supposing a carrier distribution $f(\mathcal{E}_z)$ in the reservoirs. In equilibrium conditions Maxwellian distributions can be applied. In particular with a Fermi-Dirac distribution in both the reservoirs one obtains:

$$N(\mathcal{E}_z) = k_B T \ln \left(\frac{1 + \exp\left(-\frac{\mathcal{E}_z - E_{FR}}{k_B T} \right)}{1 + \exp\left(-\frac{\mathcal{E}_z - E_{FL}}{k_B T} \right)} \right), \tag{2.32}$$

where E_{FL} and E_{FR} are the quasi Fermi levels in the left and right contacts.

In the determination of the barrier transmission, a rigorous approach would require the calculation of the potential profile in the structure such as in Figure 2-15, taking into account quantum effects and confinement. In most of the analytical models the barrier profiles are determined from material and geometrical parameters:

$$\begin{aligned} \mathcal{E}_C(z) &= \mathcal{E}_{c0}(z) - V_{ox}(z), \\ \mathcal{E}_V(z) &= \mathcal{E}_{v0}(z) - V_{ox}(z), \end{aligned} \tag{2.33}$$

being $\mathcal{E}_{c0}(z)$ and $\mathcal{E}_{v0}(z)$ the conduction and valence bands at flat band, respectively. Considering a constant electric field in the dielectric, in a flash device, at a given position y in the channel length, the potential $V_{ox}(z)$ across the oxide is linearly varying from the gate to the substrate with a drop $V_{FB} - \psi_S(y)$.

The transmission coefficient $T(\mathcal{E}_z)$ is defined as the ratio between the incident and

transmitted quantum-mechanical current densities across a potential barrier, from left to right. In particular, considering two plane waves in the two carrier reservoirs:

$$\begin{aligned} \psi_L(z) &= A_L \exp(ik_L z), \\ \psi_R(z) &= A_R \exp(ik_R z), \end{aligned} \tag{2.34}$$

the definition of barrier transmission holds:

$$T(E) = \frac{m_R k_R |A_R|^2}{m_L k_L |A_L|^2}. \tag{2.35}$$

The determination of the wave function amplitudes would normally require the solution of the stationary Schrödinger equation in the structure. This calculation can be performed numerically or analytically.

Numerical solutions include the Wentzel-Kramers-Brillouin (WKB) approximation [126], the transfer-matrix method [127] or the non-equilibrium Green's function (NEGF) approach [128]. Under the WKB approximation the transmission coefficient is calculated as:

$$T_{WKB}(\mathcal{E}_z) = \exp\left(-\frac{2}{\hbar}\int_{z_0}^{z_N} \sqrt{2m_{\text{ox}}(\mathcal{E}_C(z) - \mathcal{E}_z)}dz\right), \tag{2.36}$$

where m_{ox} is the carrier effective mass in the tunnel oxide layer. The domain of integration is included between the two classical turning points z_0 and z_N, representing the boundaries of the energy barrier where $\mathcal{E}_C(z) \geq \mathcal{E}_z$. However one has to remind that the WKB approximation is only valid when the variation of energy barrier is small. When considering abrupt barriers, this model cannot be applied in proximity of the classical turning points [56, 129]. When the integration is performed numerically, any non-abrupt arbitrary barrier profile can be used. The main limitations of this model rely on the semiclassical WKB approximation which does not consider wave-function interferences and thus it is not able to reproduce quantum mechanical oscillations in the transmission coefficient.

The NEGF approach takes into account quantum resonances, carrier confinement and interferences in multi-stacked layers [128]. NEGF finds extensive application for the design of resonant tunneling diode structures and permits the rigorous determination of the quantum current in elastic tunnelling transitions and of the local density of states (LDOS - Figure 2-16) in the device. However, although this numerical solution is suitable for any arbitrary potential barrier, it usually relies on computationally demanding numerical solutions, and thus it is not applicable to compact modelling. A detailed case study investigating quantum tunneling with the NEGF approach could be found in Annex C.

A comparison between the tunnel oxide barrier transparency calculated with the WKB and the NEGF approaches is provided in Figure 2-17 for different control gate bias voltages and for both the carriers. The WKB approach slightly underestimates the transparency at low carrier energies.

The tunneling current through the tunnel oxide layer has been calculated using both

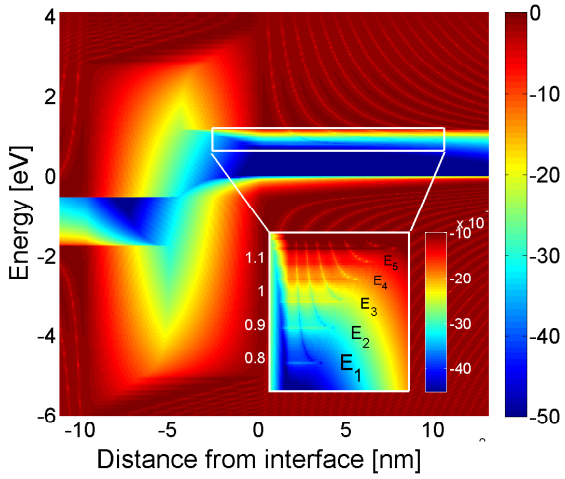


Figure 2-16 – Local density of states calculated with the NEGF solver in proximity of the tunnel oxide of a flash memory device at V_{CS} =1.9V. The inset indicates carrier confinement in discrete energy levels in proximity of the oxide/substrate interface. The energy reference is the Si valence band. The Si/SiO_2 interface is at 0nm.

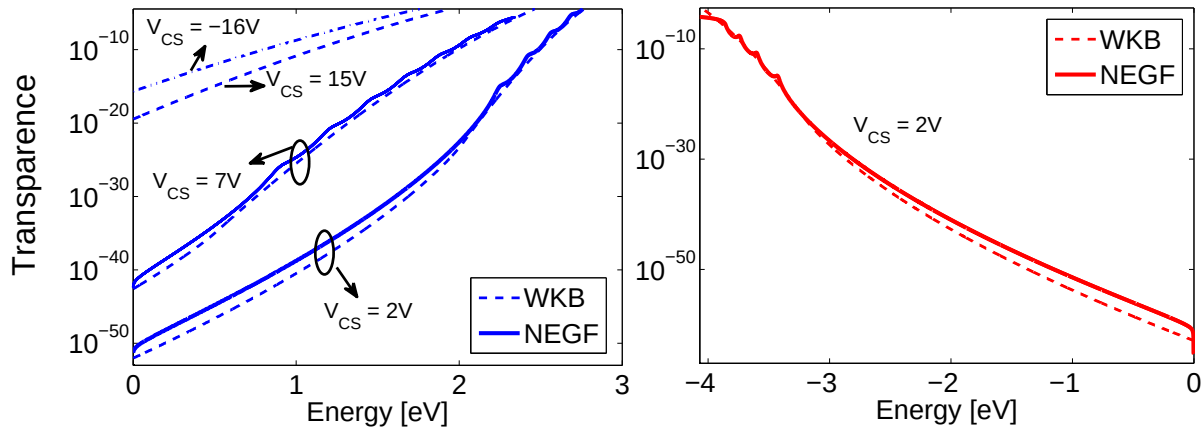


Figure 2-17 – Transparence versus carrier energy calculated with the WKB and NEGF approach for different voltage conditions from -16V to 15V. Left: electrons; right: holes.

the WKB and NEGF approaches and is indicated in Figure 2-18. The higher transmission in NEGF results in a higher tunneling current in low bias conditions. The WKB approach however remains quite suitable for an accurate estimation of the current through such a kind of thick mono-layers.

In case of complex dielectric stacks, such as for the determination of the leakage through the ONO layer, the WKB approach cannot be applied due to the presence of abrupt potential transitions in the stack. Moreover, quantum resonances appear to be dominant as indicated by NEGF simulations shown in Figure 2-19.

Compact expressions have been derived for direct tunneling conditions (rectangular or trapezoidal barriers) and Fowler-Norhdeim tunneling (triangular barriers). The WKB model offers analytical solutions when considering regular energy barriers, in particular for trapezoidal and triangular profiles. In the former case, the energy barrier can be expressed

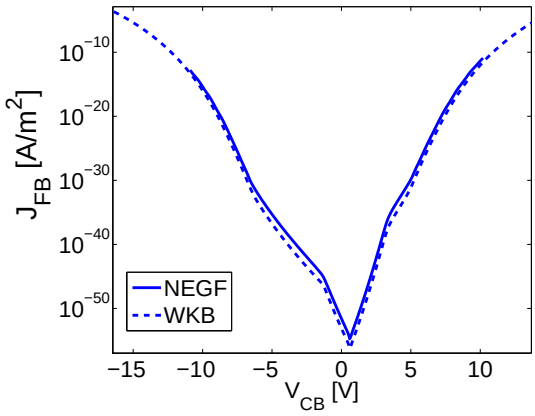


Figure 2-18 – Tunneling current density thought the tunnel oxide barrier of a floating gate device, calculated with both WKB and NEGF transparencies.

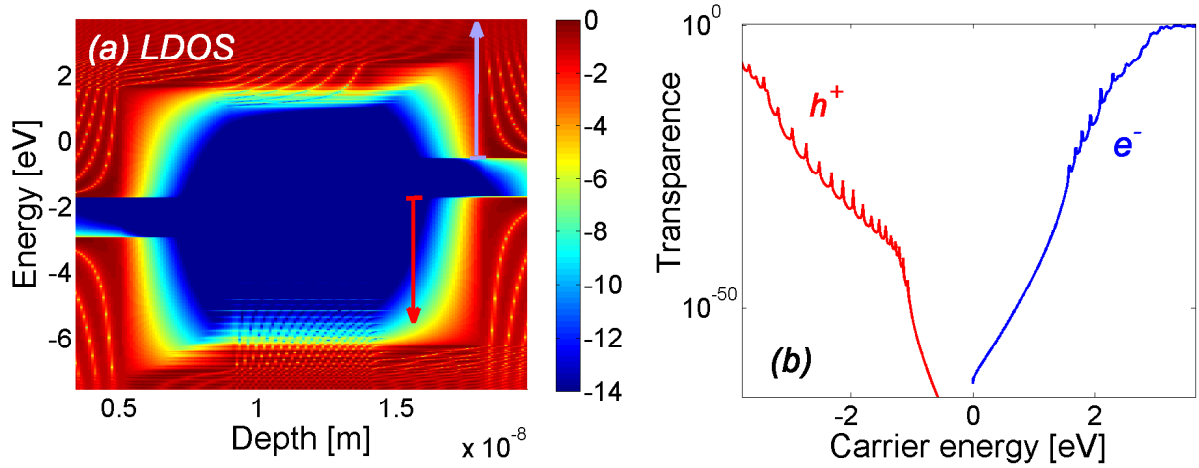


Figure 2-19 – (a) Local density of states and (b) barrier transparency calculated with a NEGF solver. Oscillations due to confined states in the nitride layer forming a quantum well are visible in the barrier transmission.

as $\mathcal{E}_C(z) = \psi_B - Fz$, with ψ_B indicating the barrier height and F_{eff} the uniform field in the oxide. Thus:

$$T_{WKB}(\mathcal{E}_z) = \exp\left(\frac{-4\sqrt{2T_{OX}}}{3\hbar q_0 F_{eff}}\left((q_0\psi_B - \mathcal{E}_z)^{3/2} - (q_0\psi_0 - \mathcal{E}_z)^{3/2}\right)\right), \tag{2.37}$$

where $\psi_0 = \psi_B - F_{eff}T_{OX}$. In the case of a linear barrier (Fowler-Nordheim regime), the integration is performed only when $\mathcal{E}_z > q_0\psi_0$ yielding

$$T_{WKB}(\mathcal{E}_z) = \exp\left(-\frac{4\sqrt{2m_{\text{ox}}}}{3\hbar q_0 F_{eff}}(q_0\psi_B - \mathcal{E}_C(z))^{3/2}\right). \tag{2.38}$$

This expression, together with a Fermi-Dirac distribution of carriers in the carrier

reservoirs L and R, leads to the well known Fowler-Nordheim current formula:

$$J = AF_{eff}^2 \exp\left(-\frac{B}{F_{eff}}\right), \tag{2.39}$$

where the two parameters have theoretical values of

$$\begin{aligned} A &= \frac{q_0^3 m_{\text{eff}}}{8\pi m_{ox} h q_0 \psi_B}, \\ B &= \frac{-4\sqrt{2m_{ox}(q_0\psi_B)^3}}{3\pi m_{ox} h q_0 \psi_B}. \end{aligned} \tag{2.40}$$

This method as been preferred for modeling the erase mechanisms by Fowler-Nordheim in the compact SPICE model. Indeed, it guarantees simplicity and preserves flexibility for parameter extraction. Quantization effects could also be added on this approximation applying corrections as in [56]. This approach has also been preferred over simplified versions of the Tsu-Esaki formula commonly included into surface potential based compact models [130], where the current expression is derived directly from the Taylor expansion of the transmission at midpoint and it is partitioned between the source/drain overlap regions.

(ii) Non-local and compact injection models

In CHEI program conditions, the high lateral electric field in the substrate causes carriers to gain energy and experience scattering events. At high lateral and vertical fields, Non-Maxellian out-of-equilibrium carrier distributions need to be introduced to model the tail of highly-energetic carriers in the channel. Analytical expressions of heated Maxwellians, non-Maxwellians [51,131], and compositions of Maxwellian distributions [53,61] have been extensively proposed to model cold and hot carrier populations in the substrate in hot-carrier injection regime.

A non-local semi-analytical model has been introduced in [132,133] to accurately calculate the distribution function inside the channel in non-equilibrium conditions. A 1D-real space probabilistic approach, inspired by the Lucky Electron Model formulation [134], has been applied including a non-local full-band description of transport. Considering a 2D (total kinetic carrier energy, space) system, the model determines the carrier concentration in each energy bin for each channel position [132]. The meshing in space and energy is calculated by a constant discretization of the channel potential profile. Two mechanisms are considered responsible for transport: the lateral variation of the potential is the main contributor of the carriers acceleration, while inelastic optical phonons are assumed to be responsible for energy loss in the channel. Similar assumptions have been adopted also in other approaches [52, 62, 135].

Due to the non-local aspect of the model, the carrier history, i.e. the number of transitions the carrier undergoes when traveling along the channel, affects its energy. Carriers flowing from S to D can either be ballistic or suffer an inelastic collision by emitting or

absorbing an energy quantum. After each collision, the carrier can travel either towards the D or backscatter towards the S. Therefore eight tunnelling fluxes have to be considered for each site in the energy-position 2D system [132]. Each tunneling event from one site to an adjacent one is characterized by a probability that can be expressed as a function of the channel position and the energy. Finally, the tunnelling fluxes are obtained for each site and used to calculate the carrier distribution in each energy bin.

In order to assess the validity of the approach, the distribution function given by the model has been compared to full-band Monte Carlo (MC) simulations [133] at different cross-sections along the channel (Figure 2-20).

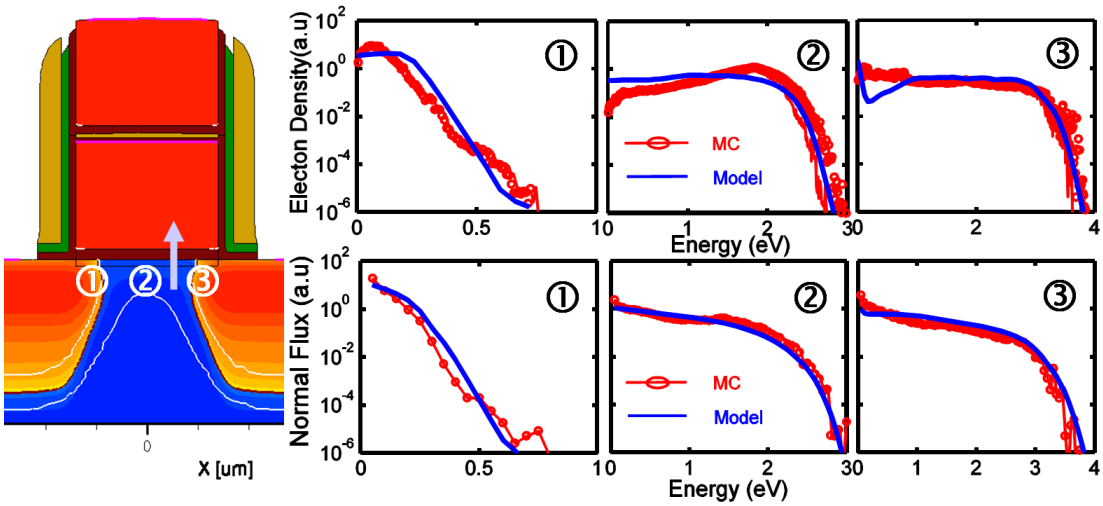


Figure 2-20 – Carrier energy distribution (top) and vertical current flux toward the floating gate (bottom) as a function of the electron energy, for three different positions along the channel (1 - in proximity of the source; 2 - in the middle of the channel; 3 - in the LDD. The non-local semi-numerical model has been validated on Monte Carlo simulations and the energy tail of the distributions in the proximity of the drain junction has been reproduced. Courtesy of A. Zaka [132].

In models more suitable to SPICE simulation, an approach similar to [55] and subsequently adopted also by Maure in [49] can be considered. Indeed, the injection current can be related to three probabilities of event: (1) electrons have to gain enough kinetic energy in the channel, (2) high energetic electrons should undergo elastic scattering events to be redirected towards the gate, (3) the same electrons should have enough energy to cross the energetic barrier at the Si/SiO_2 interface and tunnel without suffering any inelastic collision.

The probability P_1 for event (1) is expressed as in [136, 137] by:

$$P_1 = \frac{F_{lat}\lambda_{Si}}{4\psi_B} \exp\left(-\frac{\psi_B}{\lambda_{Si}F_{lat}}\right), \tag{2.41}$$

where F_{lat} is the lateral electric field, and λ_{Si} is the mean free path of the electron in silicon.

Integrating the carrier density $n(z)$ over the channel depth z, one is able to calculate the probability of electrons to reach the Si/SiO_2 interface without inelastic collisions:

$$P_2 = \frac{\int_0^\infty n(z) \exp\left(\frac{-z}{\lambda_{Si}}\right) dz}{\int_0^\infty n(z) dz}. \tag{2.42}$$

Finally, the probability of crossing the oxide energy barrier without collisions can be expressed in a similar way as a function of the equivalent oxide thickness T'_{OX} and the mean free path in the oxide λ_{OX}:

$$P_3 = \exp\left(-\frac{T'_{OX}}{\lambda_{OX}}\right). \tag{2.43}$$

The equivalent thickness T'_{OX} corresponds to the distance the electrons have to travel in the barrier and depends on the oxide electrical field F_{effov} in the overlap region:

$$T'_{OX} = \sqrt{\frac{q_0}{16\pi F_{effov} \epsilon_0 \epsilon_{ox}}}. \tag{2.44}$$

In such a compact approach the injection current is related to the drain-source current I_{DS} with:

$$I_{CHE} = I_{DS} \int_0^{L'} \frac{P_1 P_2 P_3}{\lambda_R} dy, \tag{2.45}$$

where L' is the distance travelled by the electrons from the source to the injection point in the LDD and λ_R an empirical parameter.

In ultrascaled technologies where high biases are applied, hot electrons can generate a couple of minority and majority carriers in the pinch-off LDD region. Given the complexity of carrier generation and energy exchange that require 2D/3D TCAD or MC numerical analysis, empirical approaches are adopted. The expression of the total injection current has been modified following [49, 55] and including a dependence on the weak-avalanche current I_{av} determined from the drain current as in [138]:

$$I_{CHE} = A_{CHEI} I_{av} \exp\left(-\frac{B_{CHEI}}{F_{effov}}\right). \tag{2.46}$$

2.3 Characterization and model validation

2.3.1 Test structure description

Several test structure configurations have been designed and integrated in a 65nm NVM CMOS technology. These include:

(*i*) individual cells/equivalent transistor devices where a single transistor is integrated without its surrounding matrix environment; these devices are required to extract the global model of the MOS transistor and they normally include long and/or large devices; since the cell structure, the ONO capacitance structure and surrounding environment are different from real flash devices, these structures are only used to extract initial values of the model parameters;

(*ii*) isolated single equivalent transistor/flash devices in a matrix environment; a single device surrounded by 100 WLs and BLs; these devices are the most suitable for DC and transient characterization as they better match the layout and environment of the device;

(*iii*) arrays of 80x80 equivalent transistor devices connected in parallel; in such a case 80 WLs and 80 BLs are short-circuited; this permits to obtain a large gate-channel area to measure gate, bulk and junction diode currents, which are usually too small to be measured on single devices; these devices are in the same environment of memory cells and they are surrounded as well by dummy devices to avoid edge effects;

(*iv*) arrays of NxM equivalent transistor cell devices connected in parallel, with drain and source short-circuited and where N and M vary from 10 to 2000 depending on the dimension of the individual cell; such structures are characterized to extract the AC model of the device, together with the overlap capacitances in the structure; de-embedding devices where the active part is removed, are also integrated and characterized to determine parasitic contributions of the interconnections and of the pads;

(*v*) arrays of fully addressable flash cells where 8 WLs and 10 BLs can be controlled individually and where the spacing between the floating gates and the distance between the cells has been varied; these structures are used to investigate cell-to-cell cross-couplings and V_{th} distributions (see Chapter 4).

All the matrices are in NOR configuration, with source and isolated substrates in common. Single cells are characterized in direct memory access (DMA) configuration directly connecting the pads to the terminals of the device.

2.3.2 Validation

A global parameter extraction has been preferred to verify scalability limits of the DC model. Therefore, all extractions are initially performed on the long and wide integrated devices, subsequently characterizing short and narrow devices to extract length and width dependencies. A detailed summary of the extraction methodology is reported in Annex B.

(i) Equivalent-transistor model extraction

AC characterization is performed on the NxM equivalent cell devices to extract the equivalent oxide thickness, the substrate doping and the flat band voltage using an Accretech

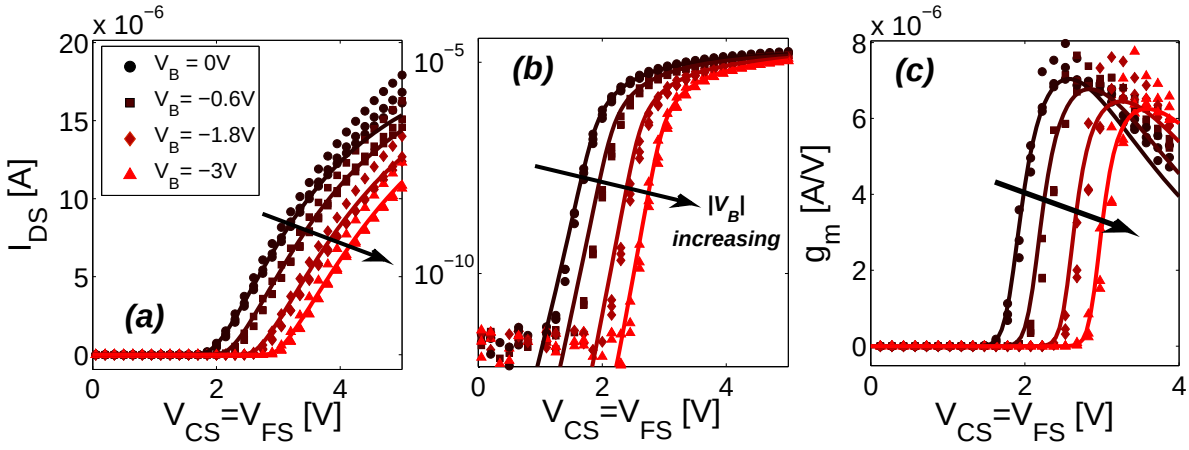


Figure 2-21 – Measured (symbols) and simulated (lines) $I_{DS}(V_{CS} = V_{FS}, V_{BS})$ characteristics performed on the equivalent transistor device in linear regime with $V_{DS} = 0.1V$. The quantity is shown both in linear and logarithmic scale in (a) and (b), respectively. In (c), the gate transconductance is illustrated.

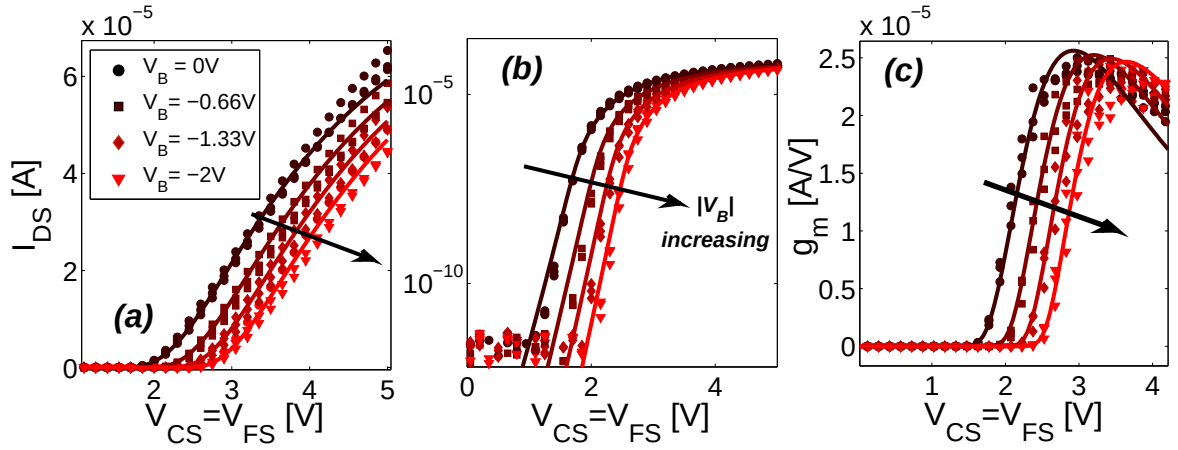


Figure 2-22 – Same measurement and simulation setup as in Figure 2-21 but in the read conditions of the flash device at $V_{DS} = 0.5V$.

UF3000 Full-automatic prober equipped with a 4076B Agilent Tester and an Agilent 4284A LCR meter. Overlap capacitances are extracted separating the drain/source and bulk capacitance and removing the parasitic components of the interconnections with specific de-embedding test structures. DC measurements on the individual and isolated equivalent transistor devices are performed with a HP 3458A multimeter. DC measurements are performed in a particular sequence, with degrading HV stress conditions applied at the end of the characterization session. The V_{th} variation is periodically monitored after each test to detect the presence of degradation effects that could compromise extraction. As a first screening, critical parameters such as the V_{th}, the maximum transconductance g_{mMax} and

the subthreshold slope are measured on all the 72 dies on the 300mm wafer. Subsequently, statistical data permit to extract reference values around which three golden dies are chosen to perform the complete characterization. Figure 2-21 presents measured and simulated DC characteristics of the channel current I_{DS} versus gate voltage $V_{FS} = V_{CS}$ on the equivalent transistor device in linear regime ($V_{DS} = 0.1V$). The current in (a-b) and gain transconductance g_m variations in (c) are reproduced as a function of isolated substrate bias V_{BS}. A similar extraction is performed in saturation regime, e.g. at the read voltage $V_{DS} = 0.5V$ of the cell (Figure 2-22). Extraction of the parameters controlling channel length modulation, saturation voltage and series output resistances is performed on $I_{DS}(V_{DS}, V_{CS})$ curves as reported in Figure 2-23(a). The output conductance g_{DS} in (b-c) is also aligned in high voltage conditions to achieve an accurate weak-avalanche current. Particular attention is taken to avoid device degradation and thus short integration time measurements are adopted. Flash measurement conditions cannot be achieved without reaching impact ionization conditions causing device degradation. Subsequently, $I_{DS}(V_{CS}, V_{DS})$ and $I_B(V_{CS}, V_{DS})$ measurements are performed. The impact ionization contribution is extracted from I_B curves at high positive voltage gate and drain biases (Figure 2-24). GIDL contribution and DIBL effects on V_{th} are also extracted in accumulation and sub-threshold conditions, respectively on arrays of 80x80 parallel equivalent transistors.

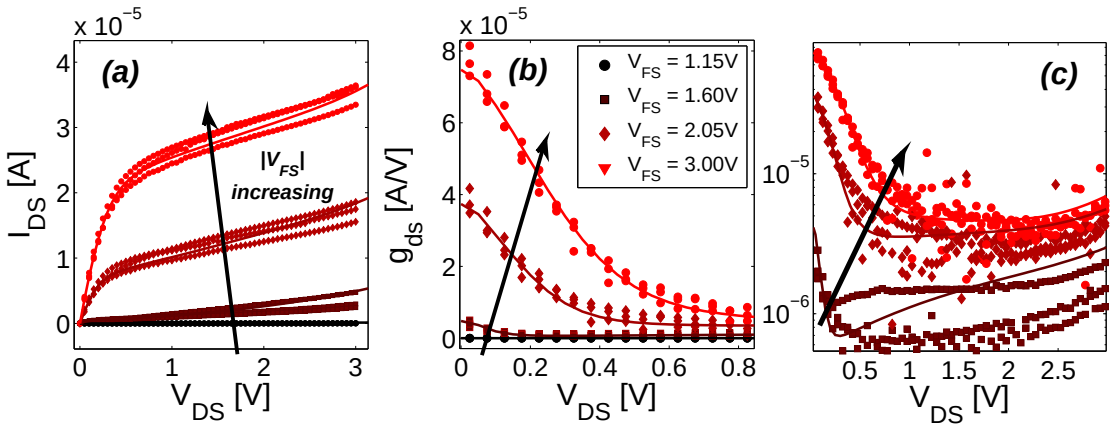


Figure 2-23 – In (a), measured (symbols) and simulated (lines) $I_{DS}(V_{DS}, V_{CS} = V_{FS})$ transfer characteristics performed on the equivalent transistor device. Channel modulation effects and impact ionization currents are well reproduced. The output conductance g_{DS} vs V_{DS} and V_{FS} in linear and logarithmic scale are also shown in (b) and (c), respectively.

Figure 2-25 shows the gate leakage current versus gate voltage measured on the structure. The current value and slope (in the inset) are calibrated in both inversion and accumulation conditions to have a reference starting point value for aligning erase transient measurements of the flash cell. Indeed, due to statistical variability, structural differences and couplings between the equivalent-transistor device and the flash cell, model parameters affecting I_G have to be slightly varied to match flash characteristics.

The equivalent-transistor model extraction is concluded analyzing the temperature de-

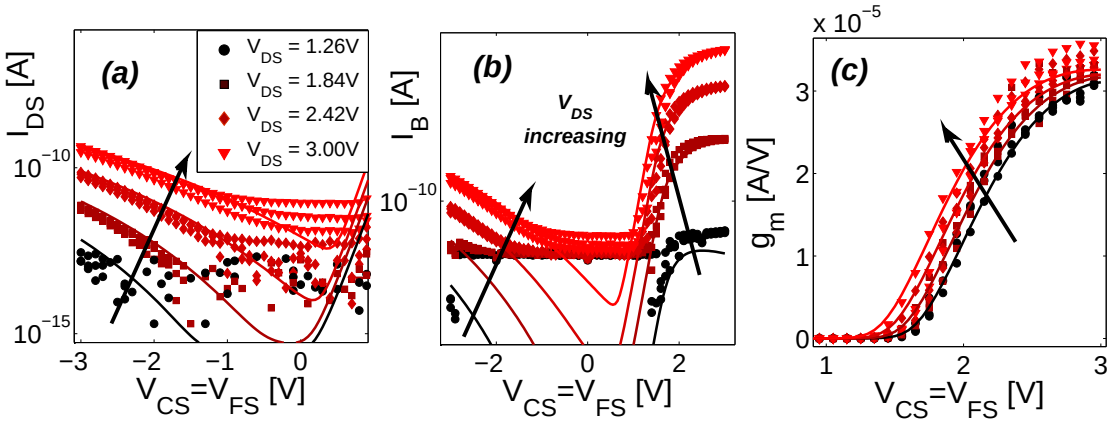


Figure 2-24 – Measured (symbols) and simulated (lines) $I_{DS}(V_{CS} = V_{FS}, V_{DS})$ characteristics performed on the equivalent transistor device. The quantity is shown both in accumulation to highlight in (a) the GIDL current to the drain terminal. A similar current is monitored on the bulk and can be seen in (b). At positive voltages, impact ionization current is monitored. In (c), the gate transconductance is also illustrated.

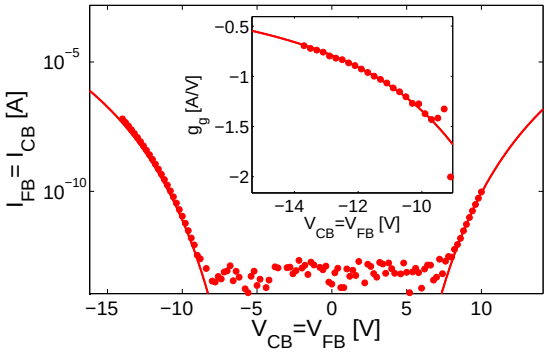


Figure 2-25 – Gate leakage current measured (symbols) on matrices of equivalent transistor devices where source, drain and bulk terminals are short-circuited. The Fowler-Nordheim model of Eq. 2.39 in lines well reproduces the leakage at high voltage conditions.

pendence. A set of parameters to reproduce the temperature effects on transfer characteristics are fitted over a temperature range from $-40°C$ to $85°C$. In particular the measured and simulated $I_{DS}(V_{FS} = V_{CS})$ characteristics in linear regime ($V_{DS} = 0.1V$) are presented in Figure 2-26.

(ii) DC flash model extraction

The flash model extraction is straightforward once the equivalent transistor characteristics are aligned. Enabling the charge balance equation solver and the ONO capacitance model, the ultraviolet state, i.e. having $Q_{F0} = 0$, is obtained.

An inherent mobility and subthreshold slope mismatch is present between the equivalent transistor and the memory cell. It is caused by the necessary layout differences between the two devices which can result in different levels of plasma-induced damage in the structures [139] (i.e. different concentration of interface defects in the flash device and different

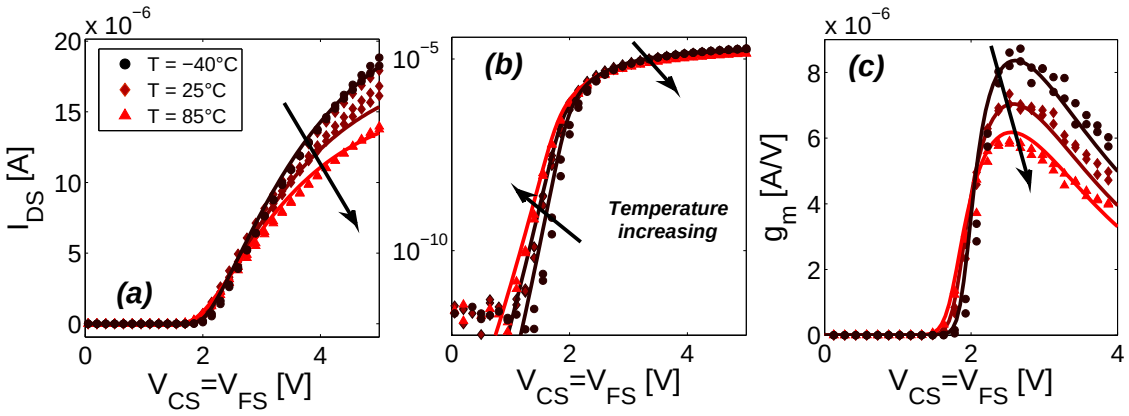


Figure 2-26 – Measured (symbols) and simulated (lines) $I_{DS}(V_{CS} = V_{FS})$ transfer linear characteristics performed on the equivalent transistor device at $V_{DS} = 0.1V$ for temperatures ranging from $-40°$C to $85°$C. The quantity is shown both in linear and logarithmic scale in (a) and (b), respectively. In (c), the gate transconductance g_m is illustrated.

parasitic couplings in the structure (gate - source/drain in particular)). Consequently, a marginal correction on the sub-threshold slope and on the mobility is required to align the transfer characteristics to the DC measurements on the cell.

The $I_{DS}(V_{CS})$ characteristic is then shifted along the V_{CS} axis, according to the charge stored on the floating gate in the program and erased states. Nominal program/erase conditions are applied to switch the state of the cell. The device is erased with a single step-like pulse 100ms long at -16V on the control gate, while it is programmed applying a 8.5V step-like pulse on the control gate and a 4.2V step-like pulse on the drain for 1μs. The two extracted charge concentrations for the simulated programmed and erased states are $Q^p_{FG} = -0.0518\frac{\mathrm{C}}{\mathrm{m}^2}$ and $Q^e_{FG} = 0.0147\frac{\mathrm{C}}{\mathrm{m}^2}$. Measurements performed on 20 different devices and several dies on the wafer show that the device and in particular the stored charge on the FG is strongly affected by process variations, which generate a wide V_{th} distribution. Consequently, the average erase and program threshold voltages (V^E_{th} and V^P_{th}, respectively measured with the constant current criterium at $I_{DScrit} = 8\mu$A) are considered, while the effects of statistical variations and the definition of process corners from statistical measurements and MC simulations will be discussed in Chapter 4. Similar variations are reported on the gate transconductance and sub-threshold slope.

(iii) Transient flash model extraction

P/E algorithms Transient measurements have been performed on single isolated flash cells using an Agilent 81110A pulse generator. The following programming and erase algorithms illustrated in Figure 2-28 have been adopted:

(a) *step-like program*: single pulses on the drain (bit line) and control gate (word line) terminals are applied. The total duration of the drain pulse plateau, which also corre-

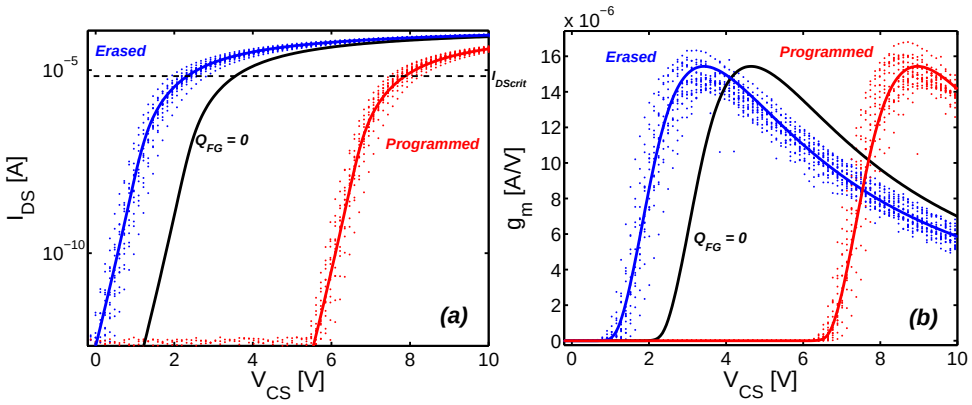


Figure 2-27 – $I_{DS}(V_{CS})$ in (a) and $g_m(V_{CS})$ (b) DC characteristics of the flash device measured (symbols) and simulated (lines) in programmed (red) and erased (blue) states at $V_{DS} = 0.5V$. Several devices have been characterized to show the spreading of the measured curves which generates a distribution of threshold voltages in both the states. The simulated state of the cell in which the stored charge in the floating gate is null (ultraviolet state) is reported in black. The current level $I_{DScrit} = 8\mu$A at which the threshold voltage is extracted is also indicated with a dashed line.

sponds to the duration of the programming phase is t_{prog}. The cell threshold voltage is measured before and after the program. This pulse is usually adopted in the technology development phase. Gate disturb is considered negligible when V_{DS} is in the low state;

(b) *progressive program*: the programming duration is divided in N steps having duration t_{prog}/N, realizing a partial programming and measuring the V_{th} after each pulse. This pulse well applied to the characterization of transient dynamics and for measuring V_{th} versus programming time. The threshold voltage is measured after each discrete pulse. The resolution N of the V_{th} versus time curve has to be high enough to validate the model and low enough to avoid perturbations by discretization or overshoot effects. Indeed, considering a typical V_{DS}=4.2V, a $t_{prog} = 1\mu$s and N=10, each step has a duration of 100ns, which is comparable to the 10ns rise/fall times of the drain pulse. Shorter rise/fall times down to 1ns can be applied, but the RC network of interconnections between the active device and the pulse generator acts as a pass-bass filter and prevents the application of abrupt transitions; it has been verified that at the end of the pulse, the V_{th} shift after a progressive program algorithm in the aforementioned conditions is identical to a single step-like program pulse;

(c) *incremental program*: the programming duration is divided in N steps whose duration progressively increases from t_{prog}/N to t_{prog}. After each step, the V_{th} is measured and the cell is cycled and erased; this method provides the most accurate V_{th} versus program time extraction as the rise/fall transition times are not cumulating as in the progressive approach; however, cycling the cell in nominal P/E conditions to restore

its erase threshold voltage is required after each discrete pulse;

(d) *step-like erase*: a single pulse of duration t_{erase} is applied on the control gate of the cell. The voltage can also be partially distributed on the bulk and/or on the source of the device. In typical product-level applications, a pulse of 8V amplitude is applied on the isolated p-well of the matrix, while -8V are forced on the word lines for a duration t_{erase} of several ms. The source is usually shorted with the bulk of the device while the drain is left floating;

(e) *progressive erase*: similarly to (b), the erase pulse can also be discretized, measuring V_{th} at each step; however, given the long duration of the erase pulse, the trade-off between resolution and overhead of rise/fall times is eliminated; given the exponential variation of the erase current versus time, the division in erase steps can also be performed in logarithmic scale;

(f) *progressive constant-current erase*: in this approach the FN erase current is maintained constant applying a progressive ramp on the control gate voltage of the cell. The number of steps N has to be high enough to avoid discretization effects on the voltage ramp. This method is commonly applied in product-level algorithms where narrow V_{th} distributions need to be achieved and an improved control on the final state of the cell is required. Constant-current step-like erase algorithms are also adopted.

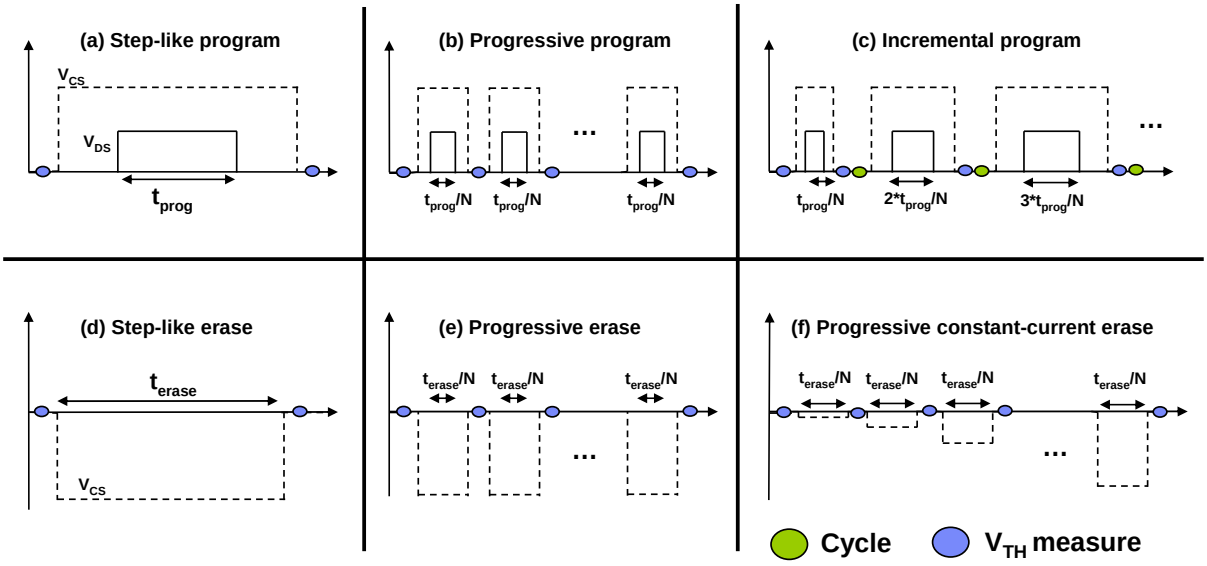


Figure 2-28 – Overview of programming and erase algorithms adopted in this study: (a) step-like program, (b) progressive program, (c) incremental program, (d) step-like erase, (f) progressive constant-current erase.

Program dynamics In static models only the two DC states are defined, while the programming and erasing transition cannot be simulated. In this approach, the compact CHEI and FN models provide a way to investigate the transient regimes with the evolution of the currents and power dissipation in the device versus P/E duration. The injection model has been calibrated over a broad range of programming durations and drain voltages, achieving a good agreement in most of the conditions. Measurements have been performed over several dies and using the progressive program algorithm of (b), by dividing the pulse in 10 steps. In Figure 2-29, simulated and measured programming dynamics are compared in linear and logarithm scale in (a) and (b), respectively. The total pulse duration varies from 1μs to 6μs while the drain bias ranges from 4.2V to 3.4V. Simulations have been performed using a single step-like pulse and the V_{th} variation is calculated from the injected charge in the FG and from the initial $V_{th}^{E} \approx 2.5V$. The variation on the initial V_{th} plays a minor role on the programming dynamics as it has been verified that a difference of less than 1V on the initial V_{th} is not altering the final programmed state. At V_{DS} =4.2V, an over exponential behavior is found and the first measured points are altered by discretization and by the influence of the 10ns-long rise/fall times of the pulse.

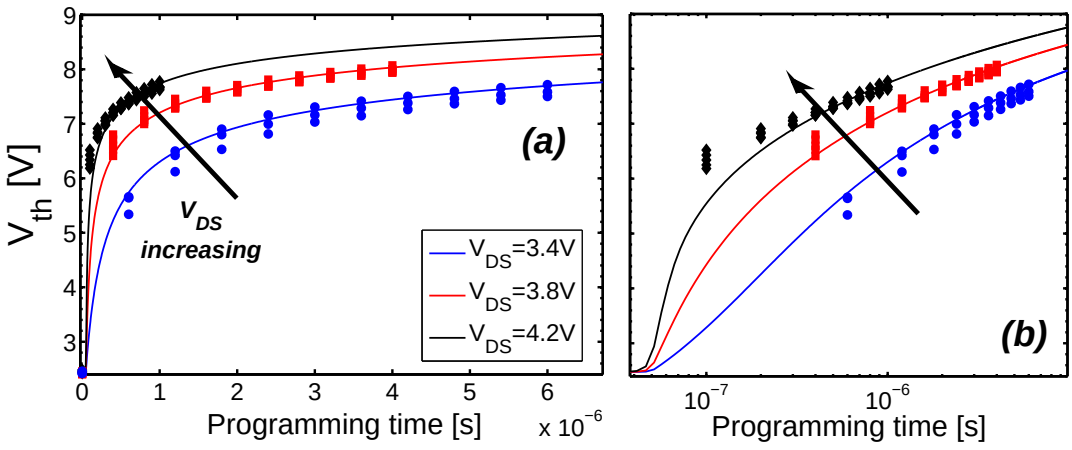


Figure 2-29 – Measured (symbols) and simulated (lines) threshold voltage dynamics versus programming time during a square box pulse on the drain of the cell in linear (a) and logarithmic (b) scales. Simulations have been performed maintaining the control gate voltage at 8.5V and varying the drain voltage from V_{DS} =3.4V to 4.2V. V_{th} measurements at V_{DS}=0.5V on several dies have been extracted, dividing the duration of the pulse in 10 parts. The same initial erase V_{th} is restored for the three voltage conditions. Good agreement is obtained with the compact model of Eq. 2.46.

The transient model permits to evaluate relevant design quantities that could facilitate worst-case analysis and IP block design. In particular, Figure 2-30 groups a collection of plots showing the evolution of these variables during programming. A step-like program simulation is performed. In (a), the pulses applied on the gate (dashed lines) and on the drain (solid lines) are shown for different drain voltages. Being the cell initially in the erased conductive state, with the increase of V_{DS} the drain current in (b) increases

and weak avalanche phenomena occur in the channel. This initially further increases the drain current, whose overshoot becomes more pronounced for high V_{DS}. Subsequently, the negative charge injected to the FG reduces the floating gate voltage of the cell and the drain current exponentially decreases. The reduction also decreases the injected current and charge in (c) and (d), respectively. CHEI programming is a self-limiting process, as when the device is turned off by the V_{FB} decrease shown in (e), injection stops. Drain coupling effects are also visible on the increase of the floating gate voltage with V_{DS} in the initial part of (e). Instantaneous power consumption on the drain and on the FG can be estimated with $P_D = I_{DS} \cdot V_{DS}$ and $P_F = I_{FD} \cdot V_{FB}$, respectively. As indicated in (f-g) subplots, CHEI is a very inefficient process with only a minimal portion of the drain current injected through the tunnel oxide.

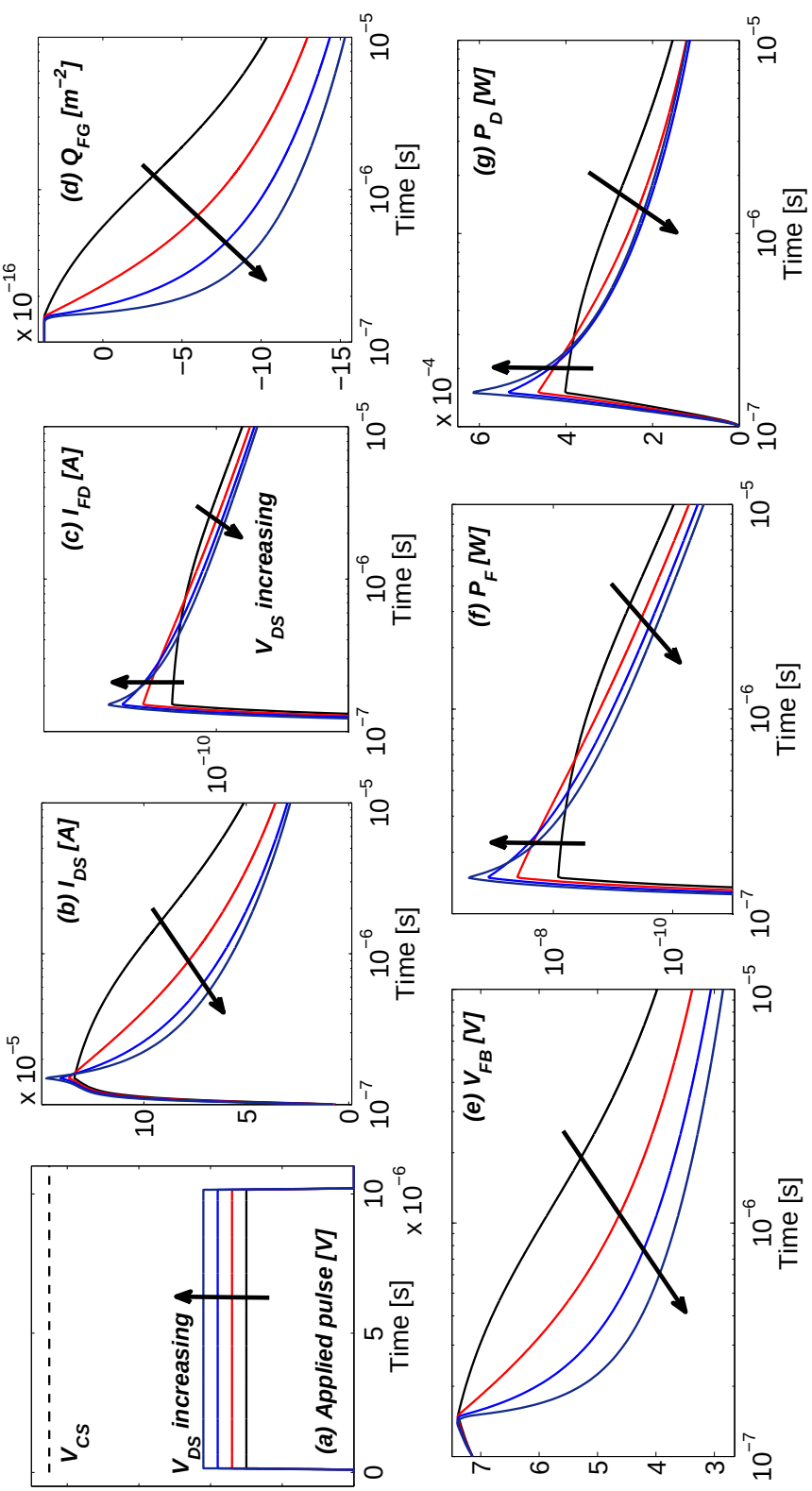


Figure 2-30 – Simulations results of the programming dynamics in Figure 2-29, showing; (a) the applied pulse on the CG and Drain of the device; (b) the drain current I_{DS}, whose overshoot and decay strongly depend on the drain voltage during programming, affecting cell endurance and program efficiency; (c) the injection current I_{FD} from the drain to the floating gate; for low V_{DS} voltages the overshoot is drastically reduced, generating the gently decrease of the floating gate charge in (d); a similar tendency is also shown in the variation of the floating gate voltage V_{FB} in (e) which increases with V_{DS} due to coupling effects and then starts to decrease as electrons are injected. In (f) and (g), the instantaneous dissipated power on the floating gate and drain terminals, respectively, are calculated.

Bulk polarization is applied to evidence its effects on the characteristics. Measurements have been performed forcing a constant voltage on the isolated p-well of the matrix array from 0V to -1V during the entire programming sequence. The drain voltage is decreased to avoid junction breakdown. Figures 2-31 and 2-32 show the good agreement obtained between measurements and simulations for V_{DS}=3.8V and 3.4V, respectively. Results are illustrated both in linear and logarithmic time scales in (a) and (b), respectively. The black dashed line corresponds to the (V_{DS}=4.2V,V_{BS}=0V) condition and has to be compared with the dark cyan curve at (V_{DS}=3.8V,V_{BS}=-0.5V). High substrate voltages increase the slope of the curve in (b) and consequently lower drain voltages can be applied to achieve the same V_{th} shift after a given programming time.

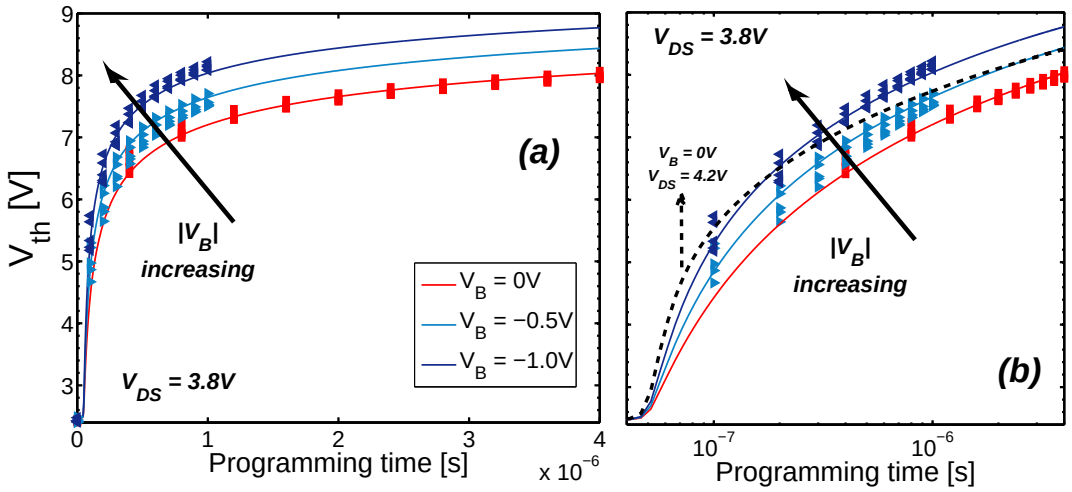


Figure 2-31 – Same dynamics shown in Figure 2-29, but for a drain pulse to 3.8V and varying the V_{BS} from 0V to -1V. At high V_{BS} programming by CHISEL occurs boosting programming dynamics. The model shows good agreement with measurements over a good range of voltage and programming duration.

Simulation results in Figure 2-33 indicate the variation of the drain current, floating gate charge and voltage, and power consumption in the device. A representative metric for injection efficiency is provided by the ratio between the injected and the drain current. The dependence of this quantity on the threshold voltage of the cell is illustrated in Figure 2-34 as a function of both V_{DS} (a) and of V_{BS} (b). A strict trade-off is established between several critical design specifications including: programming time, bias voltages, program efficiency, power consumption, endurance, drain disturb, spread of V_{th} distributions and complexity of CP circuitry. Requirements of short programming times (equivalently, high V_{th} shifts) imply high bias conditions, which provide better efficiency as shown in Figure 2-34(a). On the other hand, an increase of V_{DS} exponentially increases drain disturb on the cells connected to the same BL of the target device (cfr. Section 4.3) and current overshoots during voltage ramp-up. This phenomenon also boosts CHEI-induced degrada-

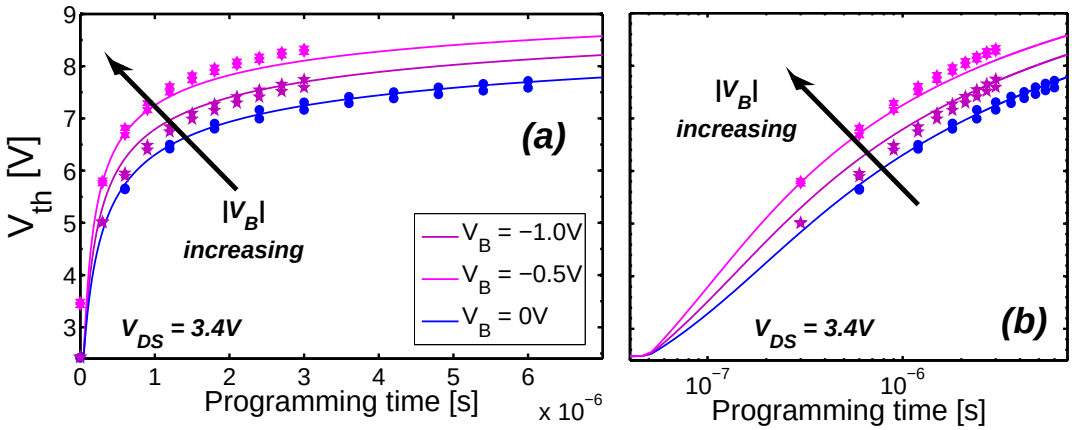


Figure 2-32 – Same dynamics as in Figure 2-31 but for a drain pulse to 3.4V.

tion, with the creation of interface and oxide defects in the tunnel oxide, that compromise memory endurance and retention performances (cfr. Section 4.2). Circuit design complexity is also negatively affected by aggressive device performances, in particular high demands of current and voltage conditions for charge pump circuits. Also, bulk bias voltages are increasingly adopted to overcome high voltage and current requirements in embedded applications. Figure 2-33(a) indicates the drain current peak reduction, when the drain voltage is decreased and the bulk voltage is raised of the same quantity. For this reason, CHISEL programming offers a good trade-off between program efficiency and high voltage requirements. The reduced control of the V_{th}^{P} distribution in CHEI conditions requires the application of Program/Verify algorithms, increasing the complexity of HV circuitry and reducing programming time [140, 141]. The role of the different tradeoffs in programming conditions will be further discussed in Chapter 5.

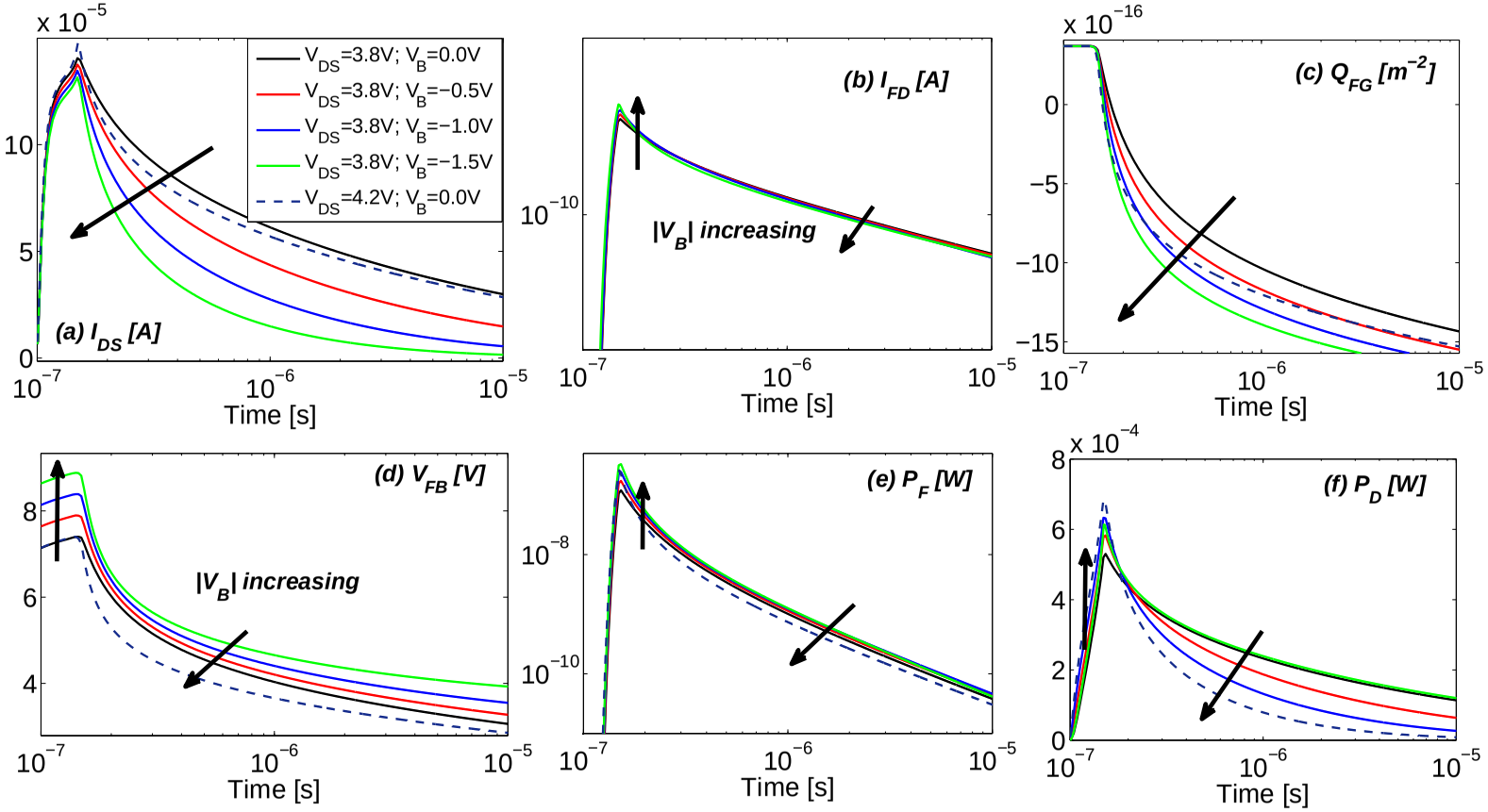


Figure 2-33 – Same quantities as in Figure 2-30 from (b) to (g) but varying the bulk voltage V_{BS} and maintaining the drain voltage to $V_{DS} = 3.8V$. With $V_B = -0.5V$, a similar final charge is obtained as the case at $V_{DS} = 4.2V$ shown in Figure 2-30 and reported in these plots in dashed.

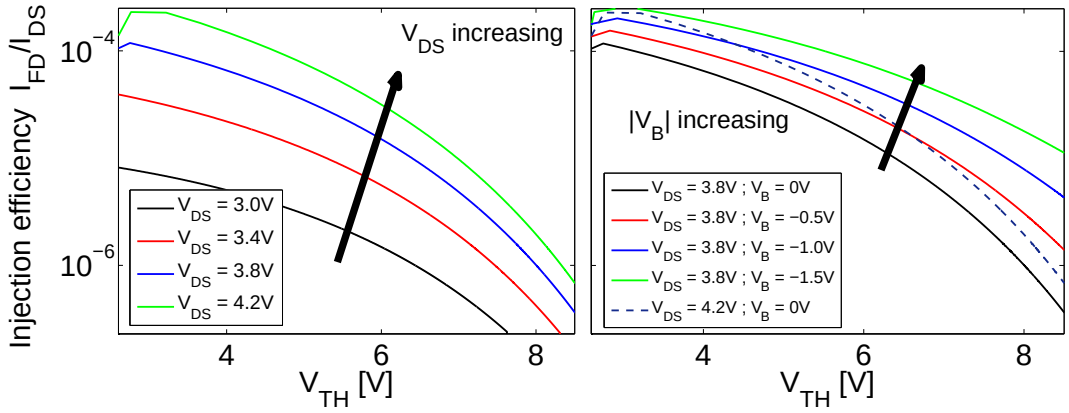


Figure 2-34 – Program efficiency I_{FD}/I_{DS} versus threshold voltage V_{th} simulated for different drain voltage conditions V_{DS} at $V_{BS} = 0V$ (a), and for different V_{BS} maintaining V_{DS}=3.8V. The dashed curve indicates a simulation performed with V_{DS}=4.2V and V_{BS}=0V and illustrates how efficiency is flattened and optimized on the entire programming phase with back–bias voltages. This permits to decrease the required maximum drain current and relax CP design specifications.

The compact model of Eq. 2.46 guarantees good scalability around the operating conditions, but its predictivity is reduced for lower drain voltages below 3V. In such a case, non-local or Monte Carlo numerical solutions should be adopted to reproduce the dependence of the carrier distribution on V_{DS}. Furthermore, measurement results performed for drain voltages ranging from 1.5V to 4.2V with a progressive algorithm indicate an increase of variability and dependence on the initial state of the cell for very low bias conditions (Figure 2-35).

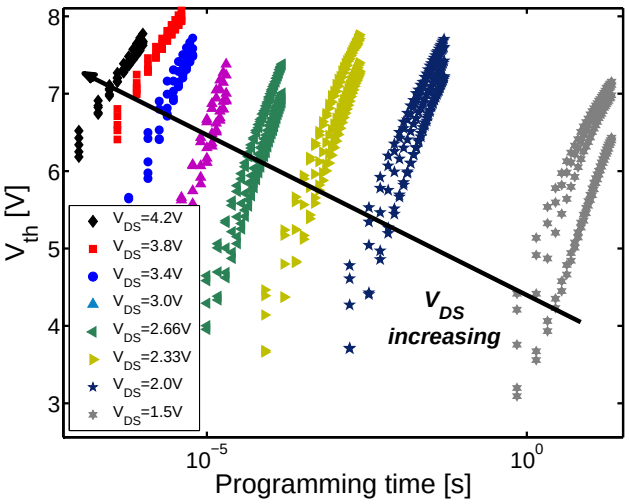


Figure 2-35 – Measured threshold voltage variation as a function of programming time, upon the application of a step-like pulse with maximum $V_{CB} = 8.5$V and drain voltage from 1.5V to 4.2V.

Erase dynamics Erase transient characteristics have been measured characterizing the variation of V_{th} versus erase time using the progressive erase algorithm. The 100ms pulse applied to the word line of the device is discretized in 10 steps having 10ms duration and 10μs rise/fall time. For the nominal condition at V_{CB}=-16V the discretization is performed in 20 pulses in logarithmic spacing to characterize the V_{th} variation for very short erase times. Three dies are characterized and a smaller spreading than for program characteristics is found. Three voltage conditions are investigated and reproduced with the FN current model of Eq. 2.39 (Figure 2-36).

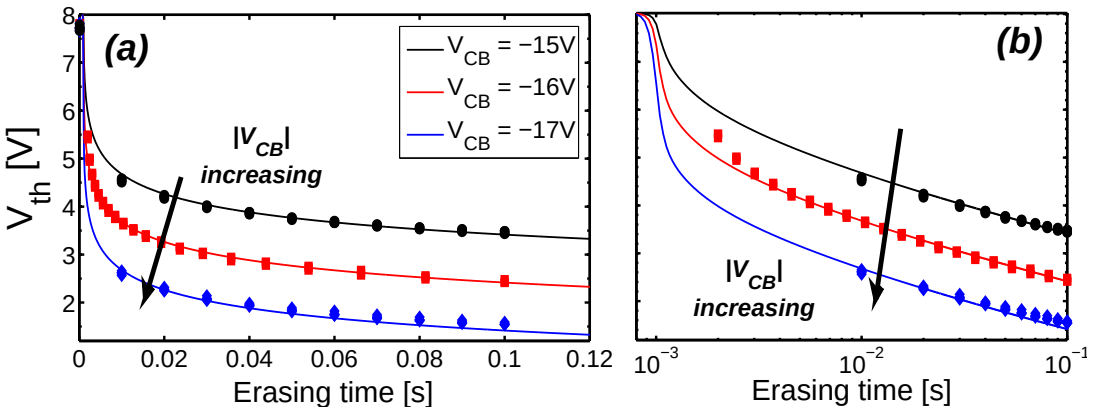


Figure 2-36 – Simulated (lines) and measured (symbols) variation of the cell threshold voltage as a function of the erase time upon the application of a square pulse on the control gate terminal. Measurements are extracted dividing the erase pulse in 20 smaller steps and measuring the threshold voltage of the cell after each erase sequence; good agreement with simulations has been found.

Measurements in Figure 2-37 illustrate that the voltage partitioning on the control gate or on the bulk can be chosen without affecting the erase dynamics. Therefore, only the ramp on the control gate voltage pulse is controlled, maintaining the bulk of the matrix array grounded. Constant-current erase algorithms are increasingly applied in modern embedded memories to improve controllability of the erase V_{th} distribution and relax requirements of CP regulators [142]. In these approaches, a voltage ramp, whose slew rate can be accurately controlled by the feedback regulators in switched-capacitor circuits, is applied to the bulk of the device while the WL is biased at high voltages. Since the current exponentially increases with the electrical oxide field, it is maintained constant on the voltage ramp. Lower ramp rates result in less efficient and slower erase. Figure 2-38 shows the good agreement found with simulations, varying the control gate voltage from 0V to -16V with a variable ramp rate from 80V/s to 53V/s. In such an approach depletion conditions with threshold voltages at 0V can be easily achieved.

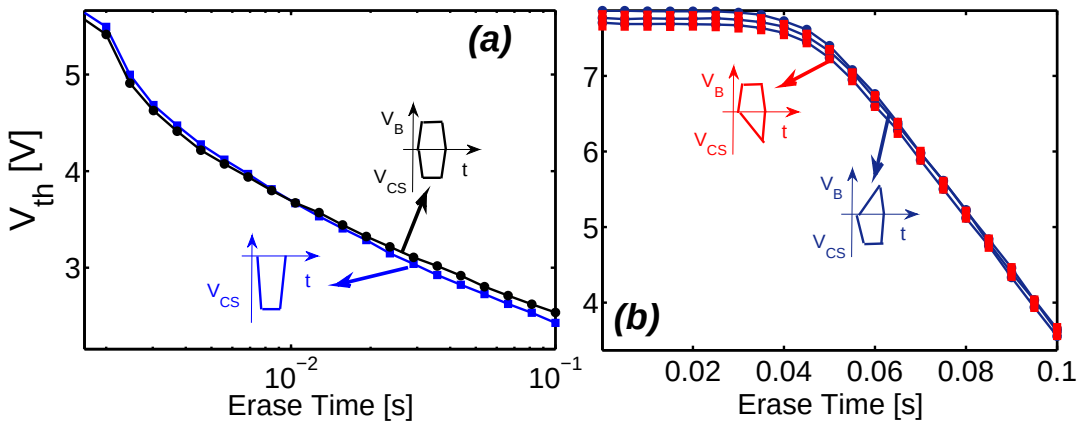


Figure 2-37 – Measured V_{th} versus erasing time for different erase methods. (a) Comparison between progressive constant gate voltage erase at -16V and progressive constant gate and bulk voltage erase at -8V and +8V, respectively. Negligible effects are found. (b) Comparison between progressive constant current erase ramping V_{CS} from 0 to -8V and progressive constant current erase ramping V_B from 0 to +8V, both in 100ms. Dynamics are equivalent.

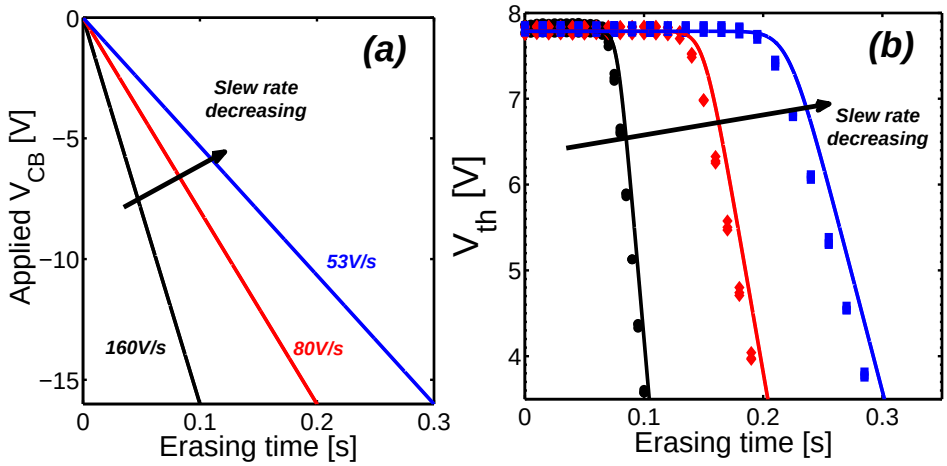


Figure 2-38 – (a) Pulse applied during constant-current erase operation where the bias voltage on the gate is increased from 0V to -16V with a specific slew rate. (b) Variation of the cell threshold voltage as a function of the total erasing time and for three different slew rates. Simulated results (lines) indicate a linear decrease of the V_{th} with time and are in good agreement with measurements (symbols).

2.4 Conclusion

In this chapter, a new SPICE model for the flash device, developed on the paradigm of the charge balance on the FG, has been described in detail. The model has been validated by means of 2D and 3D TCAD AC simulations obtaining good agreement and scalability of

the coupling coefficient of the structure, in various bias conditions. A complete extraction methodology, relying on specifically designed test structures and convergence algorithms, has been developed. Characterization and model extraction has been performed on devices integrated in a 65nm derivative technology, for DC, AC and transient regimes. The results indicate that the model offers good scalability with bias voltages and well reproduces the behaviour of the cell upon the application of various program/erase algorithms.

Additional effects include disturb effects, cross–couplings between the cells in the array, device aging and endurance/retention reduction and statistical process variation. These contributions will be described in detail in Chapter 4. The assessment of device endurance and the modeling of charge trapping phenomena requires a complete numerical quantum study that has been detailed in the following Chapter.

Chapter 3

Charge trapping effects in CMOS technologies

3.1 Introduction

The influence of oxide defects in advanced CMOS technologies is becoming more prominent as aggressive low power requirements and high endurance performances are targeted. The degradation of both oxide layers and Si/SiO_2 interfaces in MOS devices is responsible of uncontrolled threshold voltage V_{th} shifts [143, 144], mobility and gate trasconductance g_m reduction [145, 146], gate leakage [68], Flicker noise [147] and general device lifetime reduction by oxide breakdown [148]. The effects of defects on device performances are not limited to MOS devices, but they also extend to Flash, SONOS, nanocrystal and other charge trapping memories [149, 150]. Indeed, it is commonly believed that highly-energetic carriers in programming by Channel Hot Electron Injection release hydrogen, which then diffuses at the Si/SiO_2 interface and generates interfacial states [151, 152]. Both defect states and fixed oxide charges affect DC and AC performances, program/erase dynamics and memory endurance [68,149,153]. Moreover, a low-field leakage current also called stress-induced leakage current caused by oxide traps appears and drastically reduces memory retention [68] [69]. The understanding of the effects of traps requires advanced physical models to evaluate the mechanisms underlying the V_{th} window degradation in flash devices and the impact of traps on memory retention.

In this chapter, we will review and investigate the effects of traps using a non-radiative multiphonon-assisted (MPA) model suitable for DC, AC and transient analysis. The nature, the mechanisms of C/E of carriers and the general effects of point defects on the characteristics of electrically stressed CMOS dielectrics are discussed. Three of the most common defect characterization techniques, the capacitance/conductance-voltage (CGV) method, the trap-assisted-tunneling (TAT) characterization and multifrequency charge pumping (MFCP) are also introduced. Section 3.2 deals with the description of the MPA model. A novel method for the calculation of the MOSFET gate impedance in presence of traps has been applied to the analysis of AC MOS characteristics and a new CGV extraction method-

ology by reverse modeling is proposed. This approach ensures the physical modeling of trap filling, frequency response and device electrostatics. In Section 3.4 trap-assisted-tunneling in steady state DC conditions is analysed, investigating the role of the oxide regions contributing to leakage. Transient effects linked to charge trapping and the role of the C/E time constants are detailed in Section 3.5. Charge pumping techniques have been applied to the determination of the probed region in the oxide depth and in trap energy using the considered MPA model.

Using the MPA model as a reference, effective trap capture cross-sections and their oxide depth and energy dependencies could be extracted. These parameters are critical for the formulation of compact and analytical approaches.

3.1.1 Interface and bulk point defects

Typical point defects at the interfaces are represented by threefold-coordinated unoxidizing Si atoms, which contain unsaturated valency (dangling bonds). This kind of defect is characterized by an unpaired electron localized in a sp^3 like silicon dangling bond at the interface [154] and can thus be detected using Electron Paramagnetic Resonance (EPR) spectroscopy [155–157]. In these experiments, the source of the dominating paramagnetic signal (P_b) has been identified as a dangling Si bond at the Si/SiO_2 interface. Its center is located at the substrate part of the interface and its bond is pointing towards the oxide layer [158].

Concerning the charge attributed to the defects, Lenahan et al. demonstrated that in the lower part of the Si bandgap (around 1/3 eV below midgap), P_b centers behave as diamagnetic donor-like interface defects, positively charged when occupied. As one moves towards the bandgap, the defects become paramagnetic and neutral. The states in the upper part of the bandgap peaking at 1/3 eV above midgap behave like diamagnetic acceptor-like centers, negatively charged when occupied by an electron. The amphoteric nature of P_b centers drastically alters the electrical characteristics of CMOS devices. Brower [159] and Cartier [160] [161] studied the role of hydrogen in the passivation of Si dangling bonds by annealing [162] [163]. Electron Spin Resonance (ESR) spectra revealed both the atomic structure, the density and the electric level of P_b centers [157]. Since it is known that the total interface state density extracted from various techniques (AC measurements, Charge Pumping characterization, DCIV measurements, ...) is in close relationship with the P_b-center density [161,164], the determination of both their properties and nature has become of considerable importance in those modern technologies where novel oxide stacks are investigated.

In parallel to experimental investigations, atomic scale modeling (ASM) studies have been performed to analyze the mechanisms of generation and passivation of defects, the atomic structural rearrangement and electrical characteristics [165] [166]. Ab-initio studies performed by [158, 167] on clusters of Si atoms permitted to investigate the mechanisms of structural relaxation with a fairy good agreement with ESR spectra. Bulk oxide defects generated by oxide vacancy V_O have also been detected in amorphous SiO_2 [168]. These point impurities have been labelled as E' centers after EPR analysis on neutron irradi-

ated α-quartz [169], revealing the presence of a single unpaired electron generating the observed doublet in the hyperfine spectrum. Four different E' centers have been identified in amorphous Si [168] differing in the local atomic arrangements and charge states. E'_1 on α-quartz and E'_γ centers on γ-ray irradiated silica glass have the same EPR spectra, nature and origin from the vacancy of a O atom in SiO_2 [170]. The presence of negative g-shifts in ESR spectra revealed that the charging state of these defects is originated by electron trapping [169, 171] but their charge state is still under debate as both ab-initio simulations [172] and CV measurements evidence the creation of a $Si\uparrow - Si^+$ pair around a missing O center with positive charge when the defect is occupied and spin localized around a single Si atom [173].

Among the effects of point defects on CMOS technologies, one of the strongest reliability concerns is represented by Negative Bias Temperature Instability (NBTI) [174] which manifests as V_{th} and g_m variations in PMOS devices. Defect extraction techniques [175, 176] on electrically-stressed devices by high negative applied voltage revealed an increase of states at the Si/SiO_2 interface and trapping of bulk charges in the oxide stack. Similar effects can be seen in NMOS devices upon the application of electrical stress due to a high positive applied voltage (Positive Bias Temperature Instability - PBTI [177]).

The physical nature of the defects originating NBTI is still under debate: even if these effects are created after the device undergoes electrical stress, recent results confirmed the presence of neutral point defects, which behave as amphoteric charges and whose recovery phenomena are consistent with the dynamic recovery of NBTI. Indeed, Campbell et al. illustrated the role of (100) P_{b0} and the nearly identical (111) P_{b1} centers on conventional EPR and spin dependent recombination (SDR) detected EPR and correlate these results to DCIV measurements on electrically stressed devices [162]. From the modeling point of view, the correlation between microscopic physical effects and macroscopic NBTI empirical effects is also under discussion. Indeed, one of the peculiar aspects of NBTI degradation is the presence of both a quickly recoverable component and a slowly recovering (or permanent) component [174,178,179]. Recent studies seemed to reinforce the hypothesis that the recoverable part is due to permanent oxide charge trapping, while the presence of interface states originated from P_b centers is causing the permanent part [180]. This theory has been confirmed by showing the close relationship between Random Telegraph Noise (RTN) and NBTI recovery, and considering interface and oxide defects having a wide spread distribution of capture/emission (C/E) charge trapping rates. The stochastic nature of the defects is generally taken into account in such approaches [181]. Since RTN is considered the main responsible for the fluctuation of individual trap occupancies, its physical explanation has to be found in the trapping dynamics of single defects [182]. Moreover, by showing the strong decorrelation between C/E time constants, clear experimental evidence has been provided that recovery is due to independent defects and that the characteristic switching behavior is not due to a diffusive process. This hypothesis is assumed in the popular and historical reaction-diffusion (RD) theory [183, 184]. Evidence is found that the C/E transitions are controlled by multiphonon-assisted transitions [185–187]. Increasing effort has been performed to determine the relationship between the macroscopic time constants and the microscopic variables predicted by multiphonon theory and associated to the tunneling

interaction between a carrier reservoir and the point defect position.

3.1.2 Trap characterization

The electrical characterization of the amount of trapping states at the oxide-semiconductor interface, can be performed using several techniques. Various researchers in the field use well-known CGV characterization methods [72, 73, 188] for extracting trap concentrations at Si/SiO_2 interfaces. The energy distribution of interface defects in the band-gap can be extracted comparing Quasi-Static (QS) and High Frequency (HF) CV curves and reveals two peaks across the Si bandgap corresponding to interfacial P_b centers defects [160, 161, 188]. However, due to the wide distribution of C/E rates, only small portions of the trap distribution can be characterized with common CGV measurement equipment [189]. Indeed, while traps near the substrate band-edges have characteristic frequencies which are too high to be measured, deeper traps in energy and depth are usually way too slow [190].

Charge pumping (CP) methods are currently one of the most widely used Si/SiO_2 interface trap electrical characterization techniques [191–195]. Similar characterization approaches have been introduced for the first time in 1969 [196], applying periodic pulses on the gate of a MOSFET, while keeping the other terminals grounded. During each pulse, at the high level voltage plateau, channel inversion occurs and the oxide and Si/SiO_2 interface defects capture electrons from the Si substrate conduction edge band. On the other hand, when the voltage decreases to the low level region, trapped electrons recombine with holes captured from the valence band [194, 197, 198]. The total net current, measured on the bulk contact and integrated on an entire pulse, represents the recombination current in the trap sites, and is maximum when the defects capture both electrons in inversion and holes in accumulation. The value of the current peak gives an indication of the distribution of defects at the Si/SiO_2 interface. Multifrequency CP techniques are currently used to probe specific parts of the spatial and energetic distribution of defects [198, 199]. Additionally, lateral profiling techniques relying on CP measurements are largely adopted to characterize the asymmetrical defect distribution along the channel of MOSFET devices after stress by hot carrier injection.

While the characterization of traps having long time constants cannot be performed using CGV or charge pumping techniques, other techniques can be adopted to analyze these defects; one of the oldest has been found in the Photoluminescence Intensity Characterization [200] (PIC), applied for interface states present in III-V bulk materials. However this approach permits to extract only the total amount of all interface states and its relative insensitivity of the technique represents a strong limitation.

For this reason, other solutions have been recently found to characterize single defects and capture the microscopic properties of impurities having long time constants. These methods, being performed on small-area devices, are not suffering from the averaging out of the individual defect properties and benefit from the unambiguous identification of their properties [201]. The Time Dependent Defect Spectroscopy (TDDS) [202, 203] analyses a high number of stress/recovery experiments where single C/E events can be identified by step-like traces on macroscopic parameters, e.g. V_{th} and g_m degradation. Typical defect

time windows that can be analyzed range from few μs up to minutes or even hours [201], while the gate bias dependence can be monitored on a wider range than in RTN [204]. Single defect characterization has been performed also by Haider in [205] using Scanning Tunneling Spectroscopy (STS) to image Si dangling bonds in the lattice and study their reciprocal couplings. Other approaches are adopted to accurately determine microscopic properties of multiphonon-assisted transitions [206].

3.1.3 Charge trapping models - state of the art

Even though in the past 30 years CGV methods have been considered reliable approaches [72, 74,189,207], they had to face also critics by a part of the scientific community due to the limitations of their empirical extraction models [73, 208]. Since the formulation of pioneering models investigating the impact of defect states that reside on the semiconductor surface in MOS capacitors [72], a considerable amount of effort has been devoted to obtaining a wider understanding of the behavior of deeper traps in dielectric layers [209, 210]. Most of these latter approaches rely on the extension of the Shockley-Read-Hall (SRH) recombination theory [211] considering elastic tunneling through the oxide barrier. Multi-Frequency CV (MFCV - [74, 189, 210]), Deep-Level Transient Spectroscopy (DLTS - [212, 213]), Trap Assisted Tunneling (TAT- [68,214]) and Multi Frequency Charge Pumping (MFCP - [215–217] experiences have been analysed to extract the spatial and energetic distribution of defects. However, the models adopted in the aforementioned extractions can lead to major approximations in the estimation of the total trap density and incorrect temperature dependence [218]. Indeed, as recently pointed out, the standard extension of the SRH recombination theory essentially predicts a correlation between the time constants and depth-wise position, which is not verified experimentally [219].

Carrier capture/emission by means of non-radiative MultiPhonon-Assisted processes [220, 221] have been studied in its application to TAT in SiO_2 [70, 71, 83, 214, 222] and HfO_2 [86] dielectrics. The applications of the studies performed by Palma et al. confirmed that the C/E time constants observed in RTN experiments, as long as their bias and temperature dependencies, can be accurately reproduced with this approach [214]. The exponential dependence of the trapping rates with the distance from the interface is in line with the large measured spread of the C/E rates. However, an approach consisting of coupling an accurate MPA model to physical MOS impedance models to reproduce trapping effects has not been developed yet.

3.2 Multiphonon-assisted trapping model

A multiphonon-assisted charge trapping model has been implemented into a self consistent 1D Poisson-Schrödinger solver [153] to investigate electrically-stressed MOSFET oxide stacks. The model building blocks are detailed in Figure 3-1.

This section discusses the hypotheses and the physical principles of the considered trapping model. In Subsection 3.2.1, the rate equation controlling the exchange fluxes

with the substrate and gate reservoirs is recalled. The main hypotheses on trap modeling are explained. The solution of the equation requires determining the C/E fluxes of carriers Φ_c/Φ_e from one reservoir into the trap position in the oxide using multiphonon theory. Once the fluxes are obtained, the rate equation is solved for the quasi Fermi level in the dielectric and the trapped charge distribution ρ_T is found and applied to the Poisson solver (Subsections 3.2.4).

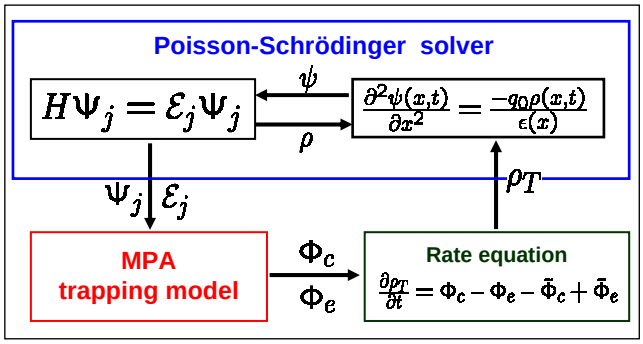


Figure 3-1 – Scheme indicating the model building blocks: a multi-band Poisson/Schrödinger solver based on a k.p closed boundary method, a multiphonon-assisted trapping model for the calculation of the C/E fluxes Φ_c/Φ_e from the carrier wavefunctions Ψ_j and the energetic levels $\mathcal{E}_j$, and the rate equation solver for the determination of the trapped charge ρ_T.

A novel approach for small signal AC analysis is introduced using a limited development of the rate equation and the total impedance of the system is found (Subsection 3.2.5).

3.2.1 Trap modeling and rate equation

Ab-initio density functional theory (DFT) simulations indicate that the Si/SiO_2 interface should be considered as a gradual transition from bulk Si to bulk SiO_2 more likely formed of a $Si_xO_{2(1-x)}$ material [223]. The extension of this region is expected to be ranging from 2Å to 10Å [223,224] and to present a large number of unpassivated dangling bonds after electrical stress [225,226]. Consequently, a clear distinction between P_b and E' defects is difficult to be established in such a region. A general approach can be adopted to model a wide variety of P_b, E' and other types of defects in the dielectric. However, in the analysis of impedance effects, interfacial amphoteric defects generally represented by Si dangling bonds, have been found to dominate the DC stretch-out and AC response in the electrical characteristics. The signature of these defects on ESR spectra revealed that E' defects tend to rapidly recover after the removal of the stress condition [227]. Furthermore, bulk traps are unlikely to respond to the AC small-signal component due to their C/E frequencies smaller than the small-signal frequency [228].

Defects are distributed in the oxide layer in all spatial directions and, as highlighted in [229], they can have different energy levels. At a given cut across the oxide stack at position x, the spatial distribution of traps in the transversal plane can be related to a wide energy distribution. In this work, a time-independent spatial and energy distribution of the defects $N_T(x, E_T)$ in the oxide layer stack has been adopted. We did not consider the possible change of the distribution vs. time, as in multi-state models linking the

carrier C/E mechanisms with the generation/annealing dynamics within the NBTI and RTN contexts [179, 181, 182, 201, 230, 231]. For impedance modeling, it has been verified that no further degradation is added when performing admittance measurements once the device has been stressed.

The dynamics of trap occupation are controlled by the rate equation [70]:

$$\frac{\partial \rho_T(x, E_T, t)}{\partial t} = \Phi_c(x, t, E_T) - \Phi_e(x, t, E_T) - \tilde{\Phi}_c(x, t, E_T) + \tilde{\Phi}_e(x, t, E_T), \tag{3.1}$$

relating the variation in time of $\rho_T(x, E_T, t)$ with the electrons e^- (holes h^+, respectively) capture and emission flows $\Phi_c(x, E_T, t)$ and $\Phi_e(x, E_T, t)$ ($\tilde{\Phi}_c(x, E_T, t)$ and $\tilde{\Phi}_e(x, E_T, t)$, respectively). The trap is assumed to be placed at position x in the oxide and at energy E_T.

As ESR and STS experiments [157, 167] extensively demonstrated, P_b centers exhibit an *amphoteric* charge nature, being neutral when unoccupied and having the possibility to capture a h^+ or an e^- ($+ \rightarrow 0 \rightarrow -$). In addition, depending on their nature (E'_γ, E'_δ centers, ... [229]), oxide/border defects can be *acceptor*-like traps, i.e. neutral defects becoming negatively charged ($0 \rightarrow -$) when an e^- is captured or a h^+ is emitted, or *donor*-like traps, which are positively charged defects becoming neutralized ($+ \rightarrow 0$) when an e^- is captured or a h^+ is emitted.

The total trapped charge $\rho_T(x, t)$ for acceptor and donor, respectively, can be calculated using:

$$\begin{aligned} \rho_T(x,t) &= -\int_{E_T} f_T(x, E_T, E_F, t) N_T dE_T, \\ \rho_T(x,t) &= \int_{E_T} \left(1 - f_T(x, E_T, E_F, t)\right) N_T dE_T, \end{aligned} \tag{3.2}$$

where $f_T(x, E_T, E_F(x,t), t)$ expresses the occupation function of the trap, i.e. the probability a trap placed at energy E_T and position x is occupied by an electron at time t. The quasi Fermi level at position x is indicated with $E_F(x, t, E_T)$.

Rigorous modeling of amphoteric states with doubled occupancy would require to account for the Coulombic energy shift between the (0/-) and the (+/0) energy transition. For Si dangling bonds such energy shift can be quite large (e.g. approximatively 0.75eV for defects on (111) Si surfaces [232]) and a rigorous calculation would require to determine two occupation functions f_T^0 and f_T^- for each occupancy state [233]. However, in order to extend the formalism developed in this paper to amphoteric-like defects we neglected the Coulombic energy and approximate the total charge with:

$$\rho_T(x,t) = \int_{E_T} \left(1 - 2 f_T(x, E_T, E_F, t)\right) N_T dE_T, \tag{3.3}$$

In steady-state, the defects occupancy functions are at equilibrium with the carrier reservoirs (channel/gate). Thus, the concept of a quasi Fermi level $E_F(x)$ valid in the

substrate can be extended to the trap site at position x. Therefore, summing left and right-hand sides of the set of equations expressed by Eq. 3.1 for each energy level, the rate equation reduces to:

$$\int_{E_T} gN_T(x,E_T)\frac{\partial f_T}{\partial t}dE_T = \int_{E_T} [\Phi_c(x,E_T,t) + \tilde{\Phi}_e(x,E_T,t) - \Phi_e(x,E_T,t) - \tilde{\Phi}_c(x,E_T,t)]dE_T, \quad (3.4)$$

The expression of ρ_T in the left-hand side of Eq. 3.1 has been replaced by one of the expressions in Eqs. 3.2 3.3, with a different defect charge factor $g = \{-1,-2\}$, resulting from the derivative of ρ_T (acceptor/donor and amphoteric, respectively). The validity of Eq. 3.4 for the DC and AC domains is discussed in Subsection 3.5.

3.2.2 Capture/emission flows

The C/E flows Φ_c/Φ_e in Eqs. 3.1 and 3.4 and have been obtained using the theory of non-radiative trapping of carriers by multiphonon-assisted processes [185, 187, 234].

The e^- flows are expressed using:

$$\begin{aligned} \Phi_c(x,E_T,t) &= N_T(x,E_T)\,(1 - f_T(x,E_T,E_F,t))\,\tau_c^{-1}(x,E_T,t), \\ \Phi_e(x,E_T,t) &= N_T(x,E_T) f_T(x,E_T,E_F,t)\tau_e^{-1}(x,E_T,t). \end{aligned} \quad (3.5)$$

For h^+ similar expressions are used:

$$\begin{aligned} \tilde{\Phi}_e(x,E_T,t) &= N_T(x,E_T)(1 - f_T(x,E_T,E_F,t))\tilde{\tau}_e^{-1}(x,E_T,t), \\ \tilde{\Phi}_c(x,E_T,t) &= N_T(x,E_T) f_T(x,E_T,E_F,t)\tilde{\tau}_c^{-1}(x,E_T,t), \end{aligned} \quad (3.6)$$

where $\tau_c^{-1}(x,E_T,t)$ ($\tilde{\tau}_c^{-1}(x,E_T,t)$, respectively) is the capture rate for e^- (h^+, respectively), and $\tau_e^{-1}(x,E_T,t)$ ($\tilde{\tau}_e^{-1}(x,E_T,t)$, respectively) is the emission rate for e^- (h^+, respectively). Their evaluation requires considering the multiphonon theory for the calculation of C/E carrier tunneling probabilities W_c/W_e.

3.2.3 Multiphonon transitions

Figure 3-2(a) illustrates the multiphonon-assisted tunneling mechanisms involved and the C/E trapping flows considered at one interface. In the considered model, an e^- (h^+, respectively) can be captured (emitted) from (to) the energy level $\mathcal{E}_j$ ($\tilde{\mathcal{E}}_j$, respectively) in a carrier reservoir (gate or channel) to an unoccupied trap site at energy E_T. This transition is assisted by phonon emission/absorption associated to a structural lattice energy relaxation ΔE. As will be shown in the following sections, the capture rate exponentially depends on ΔE. Once the carrier is captured, any further structural relaxation or defect annealing modifying the trap wavefunction, its potential or energy E_T, e.g. through meta-stable states as described in [179, 181, 182, 201, 229–231], is neglected in our model.

The theory of multiphonon-assisted interaction [185, 187, 234] predicts capture probabilities W_c given by:

$$W_c(x, \mathcal{E}_j, E_T) = \frac{2\pi}{\hbar} R(\Delta E)|V|^2 \exp\left(\frac{F^2}{F_C^2}\right) \times \times \left[rS\left(1 - \frac{\Delta E}{\hbar\omega S}\right)^2 + (1-r)\sqrt{\left(\frac{\Delta E}{\hbar\omega S}\right)^2 + 4\overline{n}(\overline{n}+1)} \right], \quad (3.7)$$

where $\Delta E = \mathcal{E}_j - E_T$ and $\overline{n} = (\exp(\hbar\omega/(k_B\mathrm{T})) - 1)^{-1}$ is the average number of phonons of pulsation ω given by Bose-Einstein statistic, r and the Huang-Rhys factor S are model parameters [220].

The term $R(\Delta E)$ is expressed as:

$$R(\Delta E) = \frac{1}{\hbar\omega} \exp\left[-(2\overline{n}+1)S + \frac{\Delta E}{2k_B\mathrm{T}}\right] \sum_m I_m(\xi)\delta(m\hbar\omega - \Delta E), \quad (3.8)$$

being $I_m(\xi)$ the reduced Bessel function of order m [235], T the device temperature and $\xi = 2S\sqrt{\overline{n}(\overline{n}+1)}$. The factor $\exp(F^2/F_C^2)$ has been included to take into account the exponential dependence of the capture probability with respect to the electric field F in the oxide. The critical field F_C is a model parameter that can be extracted from TDDS and STS measurements [182, 184].

The calculation of the emission probability W_e assumes that the trap is in thermal equilibrium with the gate/channel. From this hypothesis and applying the detailed balance principle [211], one has:

$$W_e(x, \mathcal{E}_j, E_T) = e^{\frac{E_T - \mathcal{E}_j}{k_B\mathrm{T}}} W_c(x, \mathcal{E}_j, E_T). \quad (3.9)$$

In a complete MOSFET system, the fluxes at the two oxide interfaces (left - L: gate/oxide interface and right - R: oxide/substrate interface) are indicated in Figures 3-2(b) as Φ^L and Φ^R. Both the e^- and h^+ contributions have been included in Eq. 3.1. In addition to TAT tunneling mechanisms, direct quantum tunneling has been considered using a conventional Tsu-Esaki model and the WKB approximation for calculating the barrier transmission (Chapter 2).

(i) Defect wave function model

In order to calculate the wave function overlap $|V|^2$ and the microscopic model parameters S and r, a model for the defect wave function is needed. The billiard-ball (BB) model at the trapping center proposed by Ridley [236] is used:

$$|\Psi_b(x)\rangle = \begin{cases} 1/\sqrt{x_L^3}, & \text{for } |x - x_0| \leq \frac{x_0}{2}, \\ 0, & \text{for } |x - x_0| > \frac{x_0}{2}, \end{cases} \quad (3.10)$$

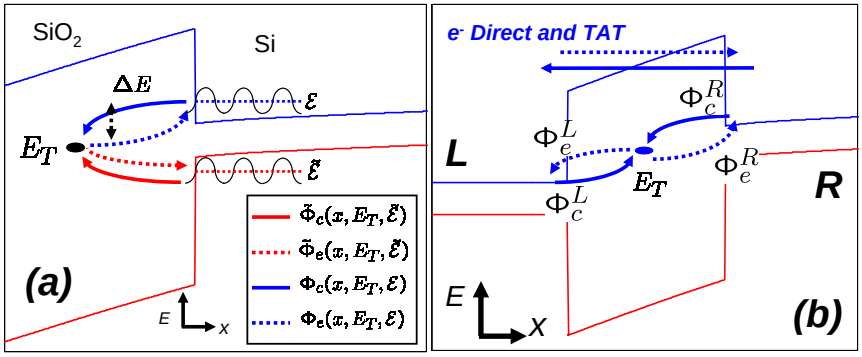


Figure 3-2 – Multiphonon-assisted trapping mechanisms considered in the proposed model. (a) An e^- (h^+, respectively) from the channel at energy $\mathcal{E}$ ($\tilde{\mathcal{E}}$, respectively) is captured by an inelastic tunneling event through the energy barrier at the Si/SiO_2 interface losing energy ΔE (four C/E fluxes for both e^- Φ_c / Φ_e and h^+ $\tilde{\Phi}_c$ / $\tilde{\Phi}_e$ are considered). In (b), the electron fluxes cross the oxide with direct and trap assisted tunneling. Similar considerations apply to h^+.

where x_0 is the extension of the trap center, x_L is the edge of the cube which is isovolumetric to the sphere of the defect [220]. Thus, $x_L = a_T(\frac{4\pi}{3})^{1/3}$ and $a_T = \frac{\hbar}{\sqrt{2m^*(E_c(x)-E_T)}}$, considering $E_c(x)$ the conduction band level at position x and m^* carrier mass in the oxide. The influence of the trapped charge and state on the trap potential and wavefunction has been neglected.

(ii) Microscopic parameters calculation

The transition matrix element between states b and j [70] has been expressed as:

$$|V|^2 = |\langle\Psi_b|U|\Psi_j\rangle|^2 = 5\pi S(\hbar\omega^2 a_T)^2 \int_{x-x_L/2}^{x+x_L/2} |\zeta(z,\mathcal{E}_j)|^2 dz, \tag{3.11}$$

where $\zeta(z,\mathcal{E}_j)$ is the wavefunction calculated using the Schrödinger solver.

The Huang-Rhys factor S in the BB model can be determined using:

$$S = \frac{27}{4(\hbar\omega)^2(q_D x_L)^3} \cdot \frac{D_{ph}^2}{M_r\omega/\hbar}, \tag{3.12}$$

where $M_r\omega/\hbar = 1/(\rho_m\omega)$ and the phonon-deformation potential $D_{ph} = 6 \cdot 10^8$ eV/cm is a fitting parameter which can be extracted from the measured values of the Huang-Rhys factor. The phonon frequency ω and the material density ρ_m are intrinsic properties of the material. A single phonon frequency has been assumed in this work considering an average value determined in the studies of Berthe [237]. The Debye cutoff wavevector q_D is defined as $q_D = (6\pi^2/a_0)^{1/3}$, with a_0 equal to $(0.25^{1/3})$ of the lattice constant in Zinc-Blend lattice.

Scanning tunneling spectroscopy studies in [237, 238] performed on Si substrates estab-

lish a value of S = 9.6 and $\hbar\omega$ = 32.5meV for P_b center defects. These constant values have been used in this study, but the dependence of the AC characteristics on the Huang-Rhys factor S has also been investigated varying its value from 0.1 to 50; the calculation and measurement of this parameter and of its field dependency is still under debate; literature studies performed on bulk materials reported values ranging from 50 down to 1 [185,187,234]. The influence of S and F_C on the admittance and transient characteristics has been investigated in [239] and in Annex D.

(iii) Capture/emission rates

The C/E rates of e^- are defined as:

$$\begin{aligned} \tau_c^{-1}(x, E_T) &= \sum_j W_c(x, \mathcal{E}_j, E_T, t) g_{2D} D^-(\mathcal{E}_j), \\ \tau_e^{-1}(x, E_T) &= \sum_j W_e(x, \mathcal{E}_j, E_T, t) g_{2D} D^0(\mathcal{E}_j), \end{aligned} \tag{3.13}$$

where $g_{2D} = \frac{m^*_{Si}}{\pi\hbar^2}$ represents the 2D density of states in the semiconductor carrier reservoir and m^*_{Si} is the effective mass of the carriers in Si ($0.19m_0$ and $0.514m_0$ for e^- and h^+, respectively, compared to the free e^- mass m_0). The quantized energetic levels in the reservoir $\mathcal{E}_j$ are determined using the PS solver [153]. The coefficients D are expressed as:

$$\begin{aligned} D^-(\mathcal{E}_j) &= k_B T \ \ln\left[1 + \exp\left(\frac{-(\mathcal{E}_j - E_F)}{k_B T}\right)\right], \\ D^0(\mathcal{E}_j) &= k_B T \ \ln\left[1 + \exp\left(\frac{-(\mathcal{E}_j - E_F)}{k_B T}\right)\right] \exp\left(\frac{\mathcal{E}_j - E_F}{k_B T}\right), \end{aligned} \tag{3.14}$$

Similar considerations are applied for the calculation of the C/E rates of holes $\tilde{\tau}_c^{-1}/\tilde{\tau}_e^{-1}$. Considering the system in thermal equilibrium, a Fermi-Dirac distribution can be used to describe the carrier distribution function $f(x, \mathcal{E}, E_F, t)$ in the semiconductor.

The model predictions have been compared to those obtained using a multiphonon capture model applied to the study of TAT current [70]. Same structure and model parameters as in [70] have been used for the comparison. A good qualitative alignment on multiphonon C/E frequencies calculated with Eqs. 3.13 is observed (Figure 3-3(a)). Quantitative discrepancies have been attributed to some differences between the model in [70] and this work. Indeed, we have considered both gate and substrate contributions in the rate equation, the impact of the electric field on h^+ and e^- fluxes, inelastic emission towards both the gate and channel reservoirs and self-consistent simulation of device electrostatics. The importance of computing a self-consistent device electrostatics is shown in the band diagram of Figure 3-3(b). The charge trapped in the oxide modifies the potential profile of the structure and thus needs to be taken into account in the Poisson equation. For an accurate analysis in all bias conditions, both gate and substrate fluxes should be taken into account considering also the h^+ trapping currents towards the valence bands. Inelastic and

direct tunneling towards the gate can also occur with important effects on the TAT current and trap occupation.

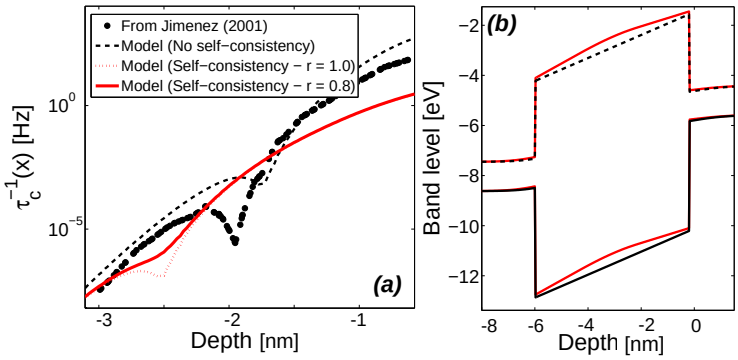


Figure 3-3 – (a) e^- capture frequencies τ_c^{-1} versus the position x in the oxide, calculated with the proposed model and with the model in [70]. Structure parameters are detailed in [70]. (b) Band diagram of the structure showing the conduction and valence bands bending induced by the negatively charged traps in the oxide layer.

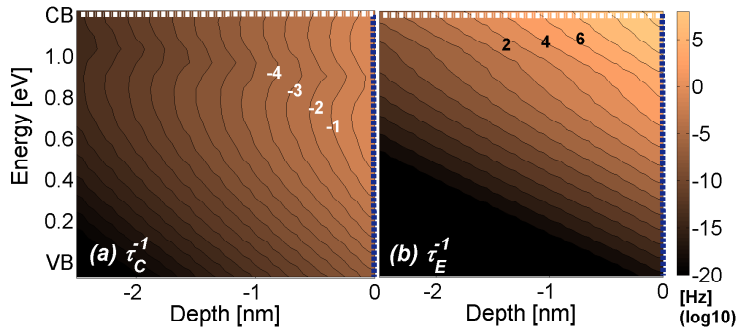


Figure 3-4 – Simulated capture (a) and emission (b) frequencies of electrons in weak inversion conditions (V=0.9V) as a function of the trap energy and depth in the oxide layer. The oxide/substrate interface is placed at position $x = 0$ with the substrate conduction and valence band edges at 0eV and 1.2eV. A zoom of the band diagram in Figure 3-3(b) in proximity of the Si/SiO_2 interface is shown. Cuts in energy and depth are shown in Figure 3. The predictions of the multiphonon-assisted model show a strong decrease of the C/E rates in oxide depth and trap energy.

A theoretical study on the C/E rates dependence in energy, position and bias has been conducted and is reported in [218]. Capture cross-sections have also been extracted to provide a references for compact degradation models [240]. Figure 3-4 shows a zoom of the oxide region in Figure 3-3(b) in weak inversion conditions (V=0.9V). In (a), the higher probability of e^- of being captured in the vicinity of the Si conduction band edge is shown. The energy dependence is explained considering that transition of carriers is less favourable when a larger amount of phonons is emitted/absorbed. The emission rates in (b) indicate a stronger energy dependence given by the detailed balance in Eq. 3.9. For the considered voltages, the oxide field remains smaller or comparable to the critical field F_C and thus this empirical parameter is assumed to have a reduced effect on the characteristics; in strong

inversion, the capture frequency exponentially increases with the applied bias voltage. A similar exponential decrease in energy and position has been observed for holes C/E rates.

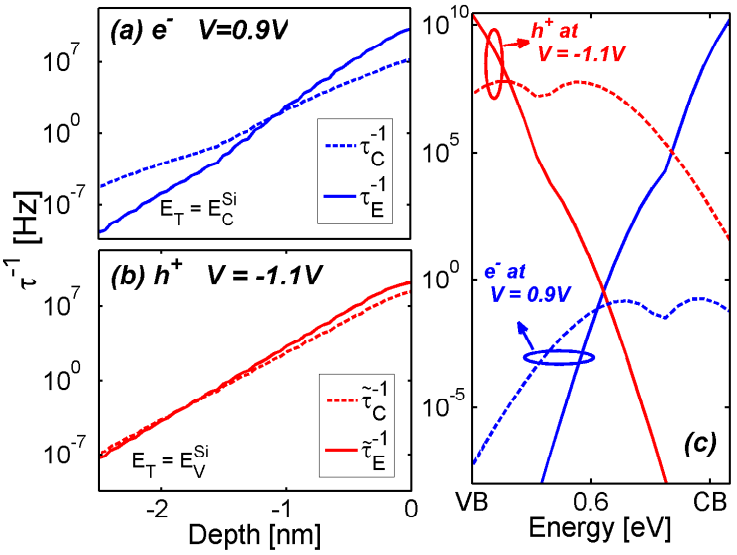


Figure 3-5 – Capture and emission rates as a function of the oxide depth (e^- in (a) and h^+ in (b)) and the trap energy (c) related to the cuts of Figure 3-4 along the dotted lines. The capture/emission frequencies are higher near the Si/SiO_2 interface. The hole capture/emission rates $\tilde{\tau}_c/\tilde{\tau}_e$ are higher in proximity of the Si valence band $E_T = E_V^{Si}$, while τ_c/τ_e peak near the conduction band $E_T = E_C^{Si} = 1.2V$. The quantities referred to h^+ are illustrated for V = -1.1V, while those related to e^- for weak inversion conditions at V=0.9V.

To further quantitatively illustrate the phenomena, the energy and depth dependencies of τ_C^{-1}/τ_E^{-1} have been illustrated at the Si/SiO_2 interface (x=0nm in Figure 3-4) and at the Si conduction/valence band edges E_C^{Si}/E_V^{Si} (Figure 3-5). The exponential dependence due to the wavefunction penetration in the oxide is stronger than the dependence in energy. The e^- capture rate in weak inversion is of several orders of magnitude smaller than both the emission rate and the capture rate of holes in accumulation, due to the reduced inversion charge in the channel.

3.2.4 Steady-state DC regime

In steady-state conditions, i.e. for $\frac{\partial \rho_t(x,t)}{\partial t} = 0$, Eq. 3.4 simplifies according to

$$\int_{E_T} \left[\Phi_c(x, E_T, t) + \tilde{\Phi}_e(x, E_T, t) - \Phi_e(x, E_T, t) - \tilde{\Phi}_c(x, E_T, t) \right] dE_T = 0. \tag{3.15}$$

The solution of this non-linear equation provides the value of the quasi Fermi level at the trap position. The quasi Fermi level calculated from Eq. 3.15 in equilibrium conditions remains equal to the quasi Fermi level in the gate/channel, justifying the commonly adopted detailed balance hypothesis. Subsequently, the trap occupation f_T and the distribution of trapped charges ρ_T can be calculated supposing a Fermi-Dirac distribution at the trap position.

Partitioning is required to attribute the trapped charge to its appropriate energy band and interface in the calculation of TAT and AC response. The criteria adopted in this work

consider the net tunneling fluxes $\Phi_n = \Phi_c - \Phi_e$ between the different bands and interfaces as weighting factors.

The four net fluxes Φ_n^L, Φ_n^R, $\tilde{\Phi}_n^L$, $\tilde{\Phi}_n^R$ with respect to the e^- and h^+ flows at the left and right oxide interfaces are determined. The charge partitioning is obtained using:

$$\begin{aligned} \rho_T^k &= -\frac{\Phi_n^k}{\Phi_n^L - \tilde{\Phi}_n^L - (\Phi_n^R - \tilde{\Phi}_n^R)}\rho_T, \\ \tilde{\rho}_T^k &= \frac{\tilde{\Phi}_n^k}{\Phi_n^L - \tilde{\Phi}_n^L - (\Phi_n^R - \tilde{\Phi}_n^R)}\rho_T, \end{aligned} \tag{3.16}$$

where the index $k = \{R, L\}$ has been added and refers to the left or right oxide interfaces (L and R in Figure 3-2(b)). A similar approach is applied for weighting the TAT currents.

Steady-state Trap-Assisted-Tunneling current calculation requires the fluxes determination to be performed for the two interfaces and for both the gate and the channel from which tunneling can occur.

The net current density flowing at one interface is given by:

$$J_{TAT}^k(x) = \int_{E_T} [\Phi_c^k(x, E_T) + \tilde{\Phi}_e^k(x, E_T) - \tilde{\Phi}_c^k(x, E_T) - \Phi_e^k(x, E_T)] dE_T, \tag{3.17}$$

where index $k = \{R, L\}$ has been added. The total net steady-state current density is expressed as:

$$J_{TAT}^{SS} = \int_x J_{TAT}^R(x)dx = -\int_x J_{TAT}^L(x)dx. \tag{3.18}$$

For thinner oxides, the direct tunneling current component J_{WKB} has been considered as well using a Tsu-Esaki model and adopting a WKB approximation for the determination of the barrier transparency [50, 241].

3.2.5 Transient and AC analysis

The transient TAT current at time t is represented by the left hand side of Eq. 3.4 and defined as:

$$J_{TAT}^T(t) = \int_x \int_{E_T} N_T(x, E_T) g \frac{\delta f_T}{\delta t} dE_T dx. \tag{3.19}$$

The transient rate equation is solved using a finite difference approach for the trap occupation.

From a practical point of view it is possible to obtain the gate impedance of the device from a transient simulation, applying a small sinusoidal signal on the gate. However, such

an approach requires a large computational effort and thus, an alternate small-signal model is preferred. A novel small-signal approach has been developed and validated, and could be found in [75, 153]. Equations for such a model are described and reported in the following sections.

The total system admittances Y are calculated with:

$$Y(V,\omega) = Y^V(V,\omega) + Y^T(\omega), \tag{3.20}$$

where ω is the angular frequency of the applied sinusoidal signal $V(t) = V_0 + v(t)$. The first contribution $Y^T(\omega)$ represents the intrinsic response of the defects at a given ω due to local quasi Fermi level variation. The second contribution $Y^V(V,\omega)$ corresponding to the influence of the defects on the MOS charges should also be taken into account. In the calculation that follows all references to position x have been intentionally omitted.

(i) Intrinsic trap AC response

A general frequency-dependent model suitable for multiphonon C/E is formulated, performing a small-signal analysis of the rate equation in the Fourier domain. Applying a gate voltage signal $V(t)$, a Taylor expansion is adopted around the DC operating V_0, yielding:

$$\begin{aligned} f_T(x,E_T,V(t)) &\approx f_{T0} + \delta f_T, \\ E_F(x,V(t)) &\approx E_{F0} + \delta E_F, \\ \tau^{-1}(x,E_T,V(t)) &\approx \tau_0^{-1} + \delta\tau^{-1}, \end{aligned} \tag{3.21}$$

where f_{T0}, E_{F0} and τ_0^{-1} are the steady state values of the trap distribution, quasi Fermi level and capture/emission times.

Considering Eq. 3.4, we replace $\frac{\partial}{\partial t}$ with $j\omega = j2\pi\nu$ and the expressions of the capture/emission fluxes in Eq. 3.5 for both holes and electrons. Adding the flux contributions and removing the steady-state flux contributions, it can be seen that after mathematical derivation Eq. 3.4 reduces to:

$$\begin{aligned} j\omega \int_{E_T} g\delta f_T N_T(x,E_T)dE_T &= \int_{E_T} N_T(x,E_T)[(1-f_{T0})(\delta\tau_c^{-1} - \delta\tau_e^{-1}) + \\ &\quad + f_{T0}(\delta\tilde{\tau}_c^{-1} - \delta\tilde{\tau}_e^{-1}) - \delta f_T \tau_t^{-1}]dE_T, \end{aligned} \tag{3.22}$$

In such an expression, we expressed the sum of all the trapping as a characteristic trap frequency τ_t^{-1}:

$$\tau_t^{-1} = \tau_{c0}^{-1} + \delta\tau_c^{-1} + \tau_{e0}^{-1} + \delta\tau_e^{-1} - (\tilde{\tau}_{c0}^{-1} + \delta\tilde{\tau}_c^{-1} + \tilde{\tau}_{e0}^{-1} + \delta\tilde{\tau}_e^{-1}). \tag{3.23}$$

The importance of this quantity relies on its direct relation with the cut-off frequency of the MOS admittances as clearly shown in Section 3.3.1 and [218].

Expressing $f_T(x,E_T,V(t))$ using Taylor series expansion around the DC operating

point, yields to:

$$\delta f_T = -\delta E_F \frac{(1-f_{T0})f_{T0}}{k_B T}, \tag{3.24}$$

and thus:

$$\delta E_F = \frac{k_B T \int_{E_T} N_T((1-f_{T0})(\delta\tau_c^{-1} - \delta\tau_e^{-1}) + f_{T0}(\delta\tilde{\tau}_c^{-1} - \delta\tilde{\tau}_e^{-1}))dE_T}{\int_{E_T} N_T(gj\omega + \tau_t^{-1})(1-f_{T0})f_{T0}dE_T}. \tag{3.25}$$

The admittance component is calculated using the trapped charge expression in the left hand side of Eq. 3.4:

$$\begin{aligned} y^T(x,\omega) &= j\omega\frac{\partial\rho_T}{\partial V} = j\omega g\int_{E_T}\frac{N_T(x,E_T)\delta f_T}{v_{ac}}dE_T, \\ Y^T(\omega) &= \int_x y^T(x,\omega)dx, \end{aligned} \tag{3.26}$$

where v_{ac} is the amplitude of the small signal pulse.

Eqs. 3.25 and 3.26 are key equations derived to obtain the total intrinsic trap conductance $G^T = \mathrm{Re}(Y^T(\omega))$ and capacitance $C^T = \mathrm{Im}(Y^T(\omega))/\omega$. In the derivation of Eq. 3.25 and 3.26, we have only considered the charge trapping fluxes exchanged with the substrate R. Extending the method including the capture/emission with the gate reservoir L and separating each contribution between holes and electron contributions, leads to consider 4 admittance components $y_T^R(x,\omega)$, $y_T^L(x,\omega)$, $\tilde{y}_T^R(x,\omega)$, $\tilde{y}_T^L(x,\omega)$.

Upon the application of a low-frequency sinusoidal signal to the gate, all traps are expected to respond to the bias voltage [73]. In such a case, the transient fluxes are considered null, with the traps reaching the equilibrium with the carrier reservoirs whenever the bias voltage changes. Consequently, G^T is null and the intrinsic quasi-static trap capacitance C^T can be defined as:

$$C_{DC}^T \equiv C^T(\omega\to 0) = \frac{\partial Q_T}{\partial V} = \frac{\partial\int\int\rho_{T0}dE_T dx}{\partial V}. \tag{3.27}$$

As ω increases, the trap C/E rates become comparable to the frequency of the small-signal component and filling of defects and trap response is reduced [73]. When driving the device at high frequency, traps do not follow the small-signal voltage applied to the gate and can be out of equilibrium with the reservoirs. Consequently, for a given DC bias, the charge density due to filled traps is a distribution of fixed charges and only their influence on device electrostatics is considered.

(ii) Complete MOS admittance model

Trap filling induces a modification of the electrostatics of the system, affecting the MOSFET bulk and inversion charges Q_B and Q_I. This corresponds to two admittance com-

ponents Y^V_{GB} and Y^V_{GC}, respectively. These quantities do not only depend on the applied voltage V but also on the total trapped charge $Q_T = Q_{T0} + \delta Q_T(t)$ and consequently on the frequency. The charges can be linearised using Taylor expansion around the DC point $O\ (V_0, Q_{T0})$, where the device and the defects are in equilibrium conditions. For simplicity the calculation is shown for Q_B, the same considerations applying to Q_I:

$$Q_B(V, Q_T, t) = Q_B(V_0, Q_{T0}) + \frac{\partial Q_B(V, Q_T)}{\partial Q_T}\Big|_O \delta Q_T(t) + \frac{\partial Q_B(V, Q_T)}{\partial V}\Big|_O v(t), \quad (3.28)$$

and the related admittances as:

$$Y^V_{GB}(V, Q_T) = j\omega \frac{\partial Q_B(V, Q_T)}{\partial V} = \frac{\partial Q_B(V, Q_T)}{\partial Q_T}\Big|_O Y^T(\omega) + j\omega \frac{\partial Q_B(V, Q_T)}{\partial V}\Big|_O, \quad (3.29)$$

Two contributions enter in Eq. 3.29. Recalling Eq. 3.26, we can notice that the first contribution depends on $j\omega\frac{\partial Q_T}{\partial V} = Y^T(\omega)$ with the corrective term $\frac{\partial Q_B(V,Q_T)}{\partial Q_T}\Big|_O$. This can be determined expressing the charge variation as:

$$\frac{\partial Q_B(V, Q_T)}{\partial Q_T}\Big|_O = \frac{Q_B^{HF} - Q_B^{LF}}{Q_T^{HF} - Q_T^{LF}} = \frac{C_{GB}(\omega \to \infty) - C_{GB}(\omega \to 0)}{C^T(\omega \to \infty) - C^T(\omega \to 0)}, \quad (3.30)$$

with $Q_B^{HF} = Q_B(V, \omega \to \infty)$ and $Q_B^{LF} = Q_B(V, \omega \to 0)$.

The gate-to-bulk $C_{GB}(\omega \to \infty) = \frac{Q_B^{HF} - Q_{B0}}{v_{ac}}$ is calculated from the device electrostatics simulated at $V = V_0 + v_{ac}$ with the self-consistent PS solver, when all the traps are supposed not to respond to the frequency and are considered as fixed charges. In such a case they do not follow the small signal bias voltage and their charge distribution remains the same as the one calculated in DC conditions (modulated through f_{T0}), as at infinite frequency the defects have not the time to respond to a small signal voltage variation applied to the gate. The capacitance $C_{GB}(\omega \to 0) = \frac{Q_B^{LF} - Q_{B0}}{v_{ac}}$ on the other hand is obtained from a PS simulation applying the same DC voltage $V = V_0 + v_{ac}$, with f_T calculated from the new bias conditions. Similar considerations apply for the gate-to-channel capacitances $C_{GC}(\omega \to \infty)$ and $C_{GC}(\omega \to 0)$ using Q_I.

Knowing that the trap response is null at high frequency ($\omega \to \infty \Rightarrow \delta Q_T \to 0$), the second contribution in Eq. 3.29 is:

$$Y^V_{GB}(\omega \to \infty) = j\omega \frac{\partial Q_B(V, Q_T)}{\partial V}\Big|_O = G_{GB}(\omega \to \infty) - j\omega C_{GB}(\omega \to \infty). \quad (3.31)$$

The conductance $G_{GB} = \Delta(J_{WKB} + J^{SS}_{TAT})/v_{ac}$ is determined from the steady-state trap assisted J^{SS}_{TAT} and the direct tunneling J_{WKB} current components.

We replace the two contributions with the expressions obtained in Eqs. 3.30 and 3.31.

Therefore Eq. 3.29 can be rewritten as:

$$Y^V_{GB}(V, Q_T) = Y^V_{GB}(\omega \to \infty) - Y^T(\omega)\frac{C_{GB}(\omega \to \infty) - C_{GB}(\omega \to 0)}{C^T(\omega \to 0)}, \tag{3.32}$$

As detailed in Eq. 3.32, the total admittances can be expressed from the high frequency term and a correction term proportional to the intrinsic total trap admittance.

(iii) AC model validation

The proposed AC signal model has been compared to the transient approach simulating the response upon the application of a sinusoidal voltage pulse and extracting all the displacement currents in the system. The structure taken into account is a 50Å-thick SiO_2 NMOS device where a Gaussian trap distribution in energy and position is placed at 2Å from the Si/SiO_2 interface. Figure 3-6 compares the trap response obtained with the two approaches for an *amphoteric*-like trap distribution. The displacement currents determined from transient simulations are used for extracting the admittance components of the system. In particular, Y^V_{GB}, Y^T_{GB}, Y^V_{GC} and Y^T_{GC} have been extracted from the amplitudes of the current densities $J^V_{GB} = \frac{\partial Q_B}{\partial t}$, $J^T_{GB} = \frac{\partial \tilde{Q}_T}{\partial t}$, $J^V_{GC} = \frac{\partial Q_I}{\partial t}$ and $J^T_{GC} = \frac{\partial Q_T}{\partial t}$ using $Y = |J|/v_{ac}\exp(j\Delta\Phi)$; $\Delta\Phi$ is the phase shift between J and $v(t)$. In Figure 3-7, the applied AC sinusoidal pulse and the transient current flowing towards the conduction band upon the application of a small-signal voltage is plotted as a function of time. The admittances are then extracted from the phase shift Φ and amplitude v_{ac} of the response signal with respect to the applied sinusoidal pulse. The real and imaginary parts of the admittances enable the calculation of the conductances and capacitances, respectively.

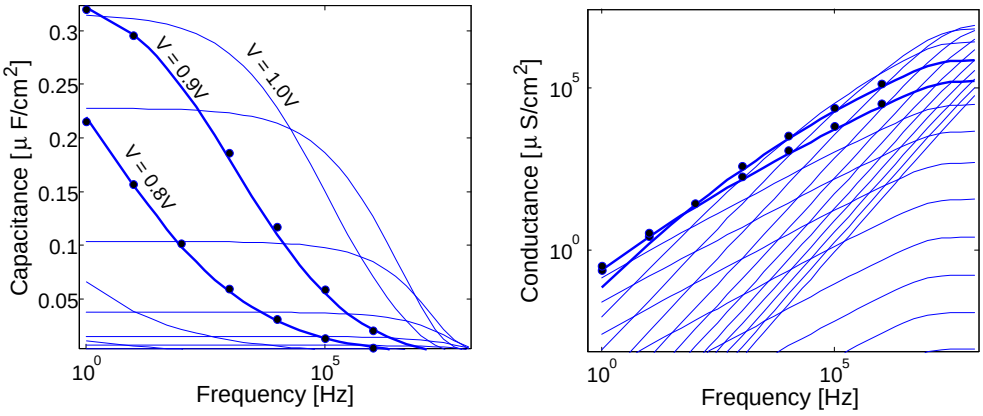


Figure 3-6 – Trap capacitance C^T_{CB} (a) and conductance G^T_{CB} (b) calculated with the AC small-signal model (lines) and the transient extraction (symbols) as a function of the AC frequency. The applied DC voltages range from 0V to 2.0V with steps of 0.1V. The cut frequency is related to the characteristic trap frequency τ_t^{-1} determined with Eq. 3.23.

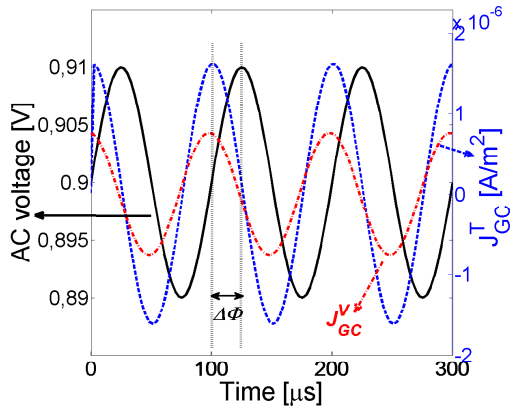


Figure 3-7 – Applied AC pulse (black solid line), channel displacement current J^V_{GC} (red dot-dashed line) and trap displacement current J^T_{GC} (blue dashed line) as a function of time simulated in transient regime. The calculation of the admittance component is performed determining a phase shift $\Delta\Phi$ and the ratio of the amplitudes between the curves and the applied pulse.

3.3 AC analysis

AC characteristics (conductance and capacitance) of multi–fingered NMOS devices integrated in a 65nm derivative technology have been measured using a HP4284A LCR-meter. The bulk and drain/source components have been separated and the parasitic components of the contact pads and metal lines appropriately removed from the contribution of the active device. The transistors have been subject to electrical stress by applying a positive constant gate voltage stress (PCVS) through the oxide stack, keeping the other nodes grounded and varying the duration of the pulse t_{str} or the applied stress voltage V_{str}.

3.3.1 CV characteristics

Upon the application of electrical stress, P_b-center defects are usually created in the interfacial layer between Si and SiO_2 at a distance of a few Å from the interface with the substrate. To study this effect with the proposed model, an amphoteric distribution of traps has been placed from the Si conduction band to the Si valance band as in Figure 3-8. In the extraction of the spatial/energetic defect profile, we considered a simple form for the energetic and spatial distribution of defects, i.e. two Gaussian distributions in energy and trap depth. The spatial mean value is placed at 2Å from the Si/SiO_2, the energy mean is in correspondence of the Si midgap, the spatial variance σ_x is 2Å and the energetic variance σ_E is 0.6eV (almost uniform energetic profile in the midgap). The validity of this trap distribution is further discussed in Subsection 3.3.3.

Figures 3-9 and 3-10 compare the split CV characteristics C_{GC} and C_{GB} of fresh and degraded devices, after different stress conditions. In (a), the characteristics of an unstressed device having an oxide thickness of $T_{OX} = 65$ Å have been fitted without traps in the oxide stack. Different PCVS stress conditions have been applied: (b) $t_{str} = 1000s$ at $V_{str} = 6V$, (c) $t_{str} = 100s$ at $V_{str} = 6.7V$, (d) $t_{str} = 1000s$ at $V_{str} = 6.7V$. The same profile has been adopted for all the cases. Only the total defect concentration is varied: (b) $N_T = 0.4 \cdot 10^{12}$ cm^{-2}, (c) $N_T = 0.78 \cdot 10^{12}$ cm^{-2}, (d) $N_T = 1.82 \cdot 10^{12}$ cm^{-2}. In addition, fixed charges are expected to be stuck in the oxide during degradation by Fowler-

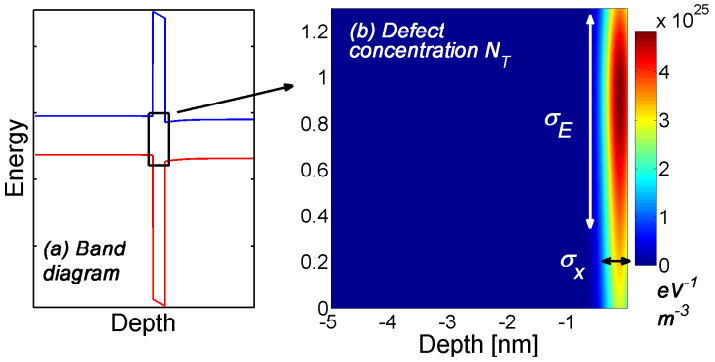


Figure 3-8 – (a) Band diagram of a FET device where traps placed into the oxide stack model the distribution of amphoteric defects for stressed devices. The defects are placed at the Si/SiO_2 interface at 2Å from the interface and with variance 2Å, while in energy the profile is centered at 0.8eV from the Si valence band edge and has a variance σ_E = 0.6eV.

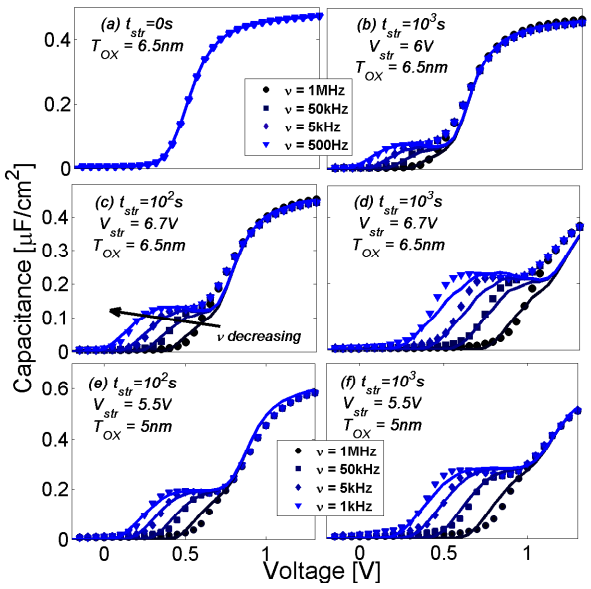


Figure 3-9 – Gate-channel capacitance C_{GC} as a function of the bias voltage for different small-signal frequencies and stress durations. In (a), the characteristic of a fresh MOS device is shown. From (b-d) the device has been stressed for different conditions using positive CVS. Measured results (symbols) well match capacitance values calculated with the multiphonon-assisted AC model (lines).

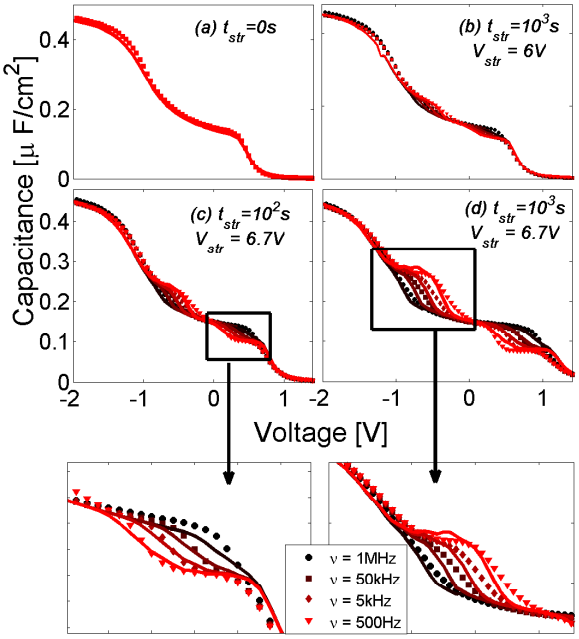


Figure 3-10 – Identical measurement conditions as in Figure 3-9(a) through (d), but showing the gate-bulk capacitance. Also in this case the effects are reproduced, including the frequency-dependence of bulk charges in weak inversion. Zooms are provided in the insets at the bottom of the figure.

Nordheim tunneling [242] causing a rigid voltage shift of the characteristics. The positive voltage shift varies with the stress condition and remains relatively small compared to the the 600-700mV stretch-out V_{th} shift due to trap filling: (a) ΔV = 50mV, correspond-

ing to a surface concentration of $1.56 \cdot 10^{11}$ cm^{-2}, (b) $\Delta V = 100$mV, corresponding to $3.12 \cdot 10^{11}$ cm^{-2} and (c) $\Delta V = 250$mV, with $7.8 \cdot 10^{11}$ cm^{-2}. Both the threshold voltage shift and the parasitic capacitive component of the traps are reproduced for a frequency range from 500Hz to 1MHz. A similar extraction has been performed only on C_{GC} in Figure 3-9(e) and (f) on a device with $T_{OX} = 50$Å, stressed for 100s and 1000s of PCVS at $V_{str} = 5.5V$. Minimal variations in the defect concentration energy profile have been applied; the extracted total concentration is $1.3 \cdot 10^{12}$ cm^{-2} and 2.05 10^{12} cm^{-2}, respectively.

As evidenced by all the CV curves, several effects attributed to charge trapping mechanisms emerge in stressed devices:

(i) inversion/accumulation are delayed to higher voltages: C_{GC}/C_{GB} are subject to a shift and a stretch-out [73]);

(ii) the C_{GC} capacitance presents a frequency-dependent increase in proximity of the weak inversion region whose amplitude depends on t_{str} and V_{str} ;

(iii) the C_{GB} capacitance presents a frequency-dependent increase in depletion region and near the flat-band voltage V_{fb};

(iv) the C_{GB} capacitance presents a frequency-dependent reduction in the weak inversion region in correspondence of effect.

The stretch-out of the capacitance in both inversion and accumulation is attributed to defect charging [143] and is correlated with the amplitude of the peaks in weak inversion and depletion. Since the CV stretch-out is only caused by a change in the electrostatics of the system in *equilibrium* conditions, effect (i) is not dependent on the frequency of the applied pulse. In equilibrium DC conditions the quasi Fermi level is pinned to the substrate and gate reservoirs, and the trap occupation follows this level. However, it should be pointed out that this is not the case in transient simulations where traps respond with a wide spread of C/E rates. In a similar way and due to the amphoteric nature of the defects, as the applied voltage decreases and the accumulation layer of h^+ is formed, capture events from the Si valence band become more probable and traps become globally positively charged.

To explain the frequency dependence of both C_{GC} and C_{GB} (effects (ii) and (iii)), one should consider that at a given frequency, only traps with C/E rates higher than the measurement frequency are able to follow the AC small-signal voltage. The strong dependence of the rates with respect to trap position results from the exponential decrease of the evanescent carrier wavefunction in the oxide depth [240]. Additionally, the dependence with trap energy is related to the lower probability of absorbing/emitting phonons and to the decrease of the capture trap radius for deeper traps in energy.

Given the wide distribution of C/E rates in energy and position, only traps in proximity of the Si conduction band capture/emit electrons and respond to the AC small-signal. This is highlighted in Figure 3-11, where the characteristic trap frequency τ_t^{-1}, calculated with Eq. 3.23 is illustrated as a function of trap energy and position. Similar results can be obtained for h^+ in corresponding bias conditions. The accessible regions in energy

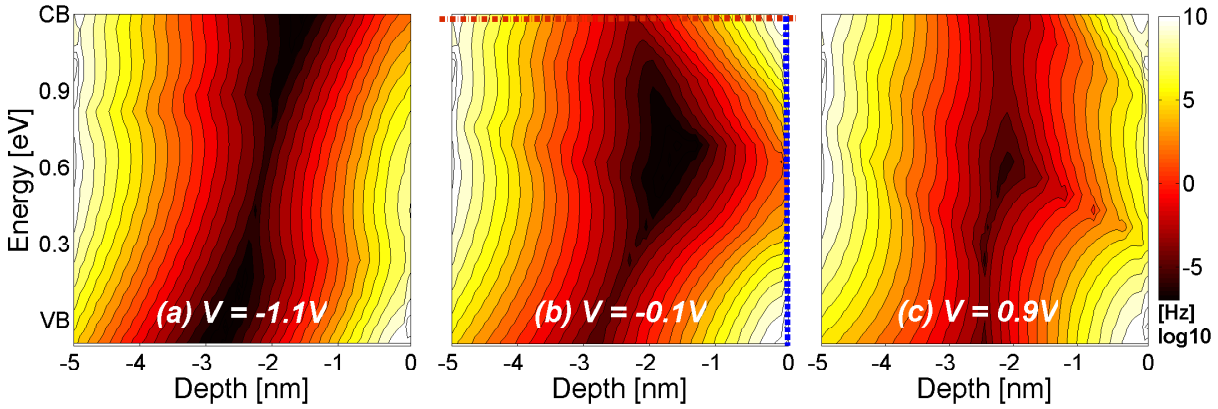


Figure 3-11 – Characteristic response τ_t^{-1} of the defects in the oxide layer for flat band (a), depletion (b) and weak inversion regions (c) indicating the frequency at which one trap can communicate with the gate and channel reservoir at different bias voltage. The interface with the substrate is at x=0. The energy reference is at the Si valence band. Capture/emission are favorable in proximity of the conduction and valence bands of the Si substrate. In proximity of the gate, traps can easily communicate with the gate reservoir, but the AC response is negligible due to their distance from the Si surface. Cuts in energy and depth are shown in Figure 3-12.

and position can be more quantitatively identified in the cuts in energy and depth shown in Figure 3-12(a-b). A similar analysis has been performed in depletion ($V = -0.1V$) and for different device temperatures (from -40°C to 100°C). In Figure 3-12(c) the dependence on oxide depth is shown in proximity of the midgap for h^+. In Figure 3-12(d), the cut-off frequency exponential variation in trap energy at the Si/SiO_2 interface further illustrates the reduced response of midgap defects in depletion regime. The exponential dependence in temperature is given by the multiphonon assisted C/E rates. The wide exponential distribution of C/E rates in energy could be at the origin of an apparent lack of correlation between the capture rate and the defect depth, as reported in [219, 243]. The large exponential decrease of τ_t^{-1} in energy and position explains the strong frequency dependence of the intrinsic trap capacitance: in weak/strong inversion, traps near the Si conduction band easily exchange carriers with the channel and their capacitance response is maximum and saturating. In depletion, both the carrier concentrations and the cut-off frequency are reduced, with lower frequencies required to scan the traps; consequently, for a given frequency, the capacitive response decreases. At flat-band and in accumulation, the h^+ trapping probability increases and defects near the Si valence band having a cut-off frequency higher than the AC signal frequency can be characterized [240].

(i) Temperature dependence

Multiphonon assisted transitions generally exhibit a strong temperature dependence. Figure 3-13(a) indicates the intrinsic parasitic trap capacitance for different bias conditions and frequencies, after decomposing the trap response from the system capacitances. For two voltage conditions, the trap capacitance has been calculated varying the device tem-

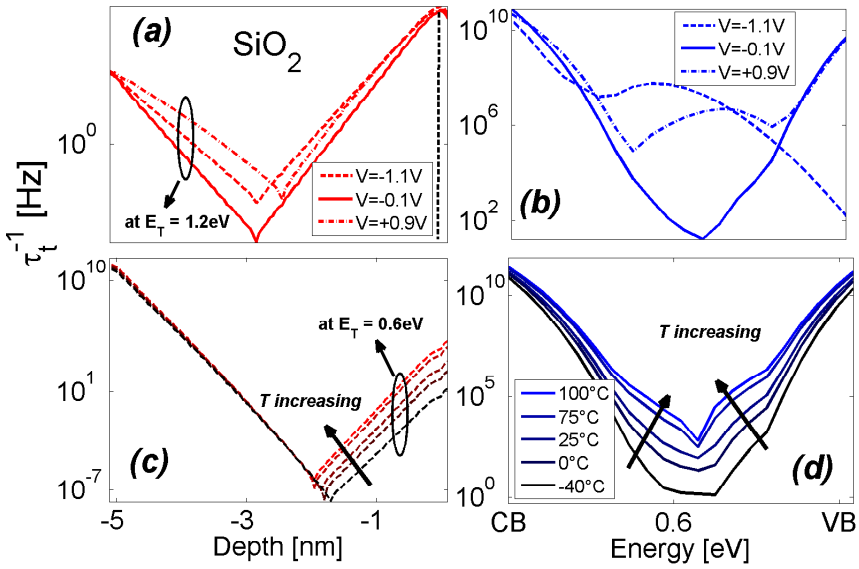


Figure 3-12 – Depth (a) and energy (b) dependencies of the trap response frequency τ_t^{-1}. The cuts correspond to the lines indicated in Figure 3-11 for the three considered bias voltages. In (a), τ_t^{-1} is taken at the conduction band level, while in (b) the results are shown at the Si/SiO_2 interface. In depletion (solid line), the exponential decrease of the trap response prevents the characterization of midgap traps, which respond with very low frequencies. (c) Trap cut-off frequency vs. trap depth in the oxide layer for different temperatures from $-40°C$ to $100°C$ and at gate voltage $V = 0.9V$. Results in proximity of the midgap are shown. In (d), τ_t^{-1} is shown as a function of defect energy and temperature. The response of traps deeper in the oxide layer and in energy exponentially decreases, while it increases with temperature.

perature over a wide range of values (Figure 3-13(b)) and indicate a large variation of the cut-off frequency. CV curves of the stressed device in Figure 3-9(f) have also been measured at lower temperatures (Figure 3-14(a-b)), where multiphonon transitions are slower due to the decrease of the phonon occupation factor ($\overline{n}$ in Eq. 3.7). Consequently, the trap cut-off frequency decreases and thus the trap capacitance measured at a given frequency is reduced (Figure 3-12(d)). It should be pointed out that the magnitude of the parasitic capacitance contribution and the quantity of stretch-out remain constant, which implies that the total defect concentration does not vary for low temperatures.

The CV measurement has also been performed at higher temperatures. In such a case the parasitic trap capacitance is enhanced at intermediate frequencies (i.e. larger spreading of the curves in frequency). In addition, evidence of trap recovery is found in the reduction of both trap capacitance peaks and stretch-out effects with temperature increase. Recovery has been taken into account in the model by decreasing the total trap concentration to $1.82 \cdot 10^{12}$ cm^{-2}. Good agreement with measurements has been found both in terms of stretch-out and magnitude of parasitic peaks. In addition, the spreading in frequency is well aligned with the increase of C/E rates and of the capacitive response

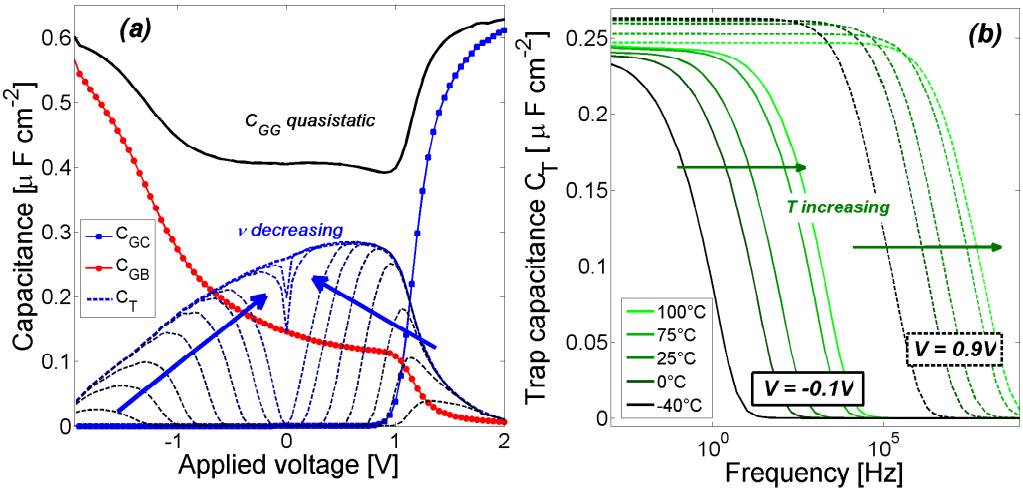


Figure 3-13 – (a) Simulation results showing the intrinsic parasitic trap capacitance of device in Figure 3-9(f) for different bias conditions and frequencies. The parasitic trap contribution C_T strongly depends on the frequency of the applied AC signal (dashed lines). At low frequency, all the traps are able to follow the AC component and the response saturates. The parasitic trap component is added to the capacitances of the MOS system (symbols). In quasi-static conditions ($\nu \to 0$) the solid black line of C_{GG} is obtained. (b) Simulated trap capacitance as a function of small-signal frequency and temperature. Calculations for two voltage conditions are shown. The extracted cut-off frequency corresponds to the τ_t^{-1} frequencies in Figure 3-11, determined for the distribution of defects. The exponential variation of 3 orders of magnitude with temperature varying from $-40°$C to $100°$C correlates with the variations of τ_t^{-1} shown in Figure 3-12(d).

at lower frequencies.

(ii) Oxide thickness and device configuration

Similar CV analysis has been performed on different oxide stacks, including the tunnel oxides of flash devices. The characterization results comparing fresh device characteristics with those after 1000s of positive constant current stress (PCCS), are shown in Figure 3-15 and evidence the presence of similar degradation phenomena.

PMOS devices have also been investigated, noticing qualitatively similar effects found on NMOS devices. Figures 3-16(b-d) show C_{GB} and C_{GC} capacitances of stressed PMOS 50Å devices, measured after applying a negative constant-current stress, as a function of bias voltage and small-signal frequency. The capacitances present the same effects previously illustrated and similar cut-off frequencies can be noticed.

3.3.2 GV characteristics

Figure 3-17 shows the simulated intrinsic response of traps on the conductance G_{GG}^T as a function of the small-signal frequency and for different DC voltages. The capacitance

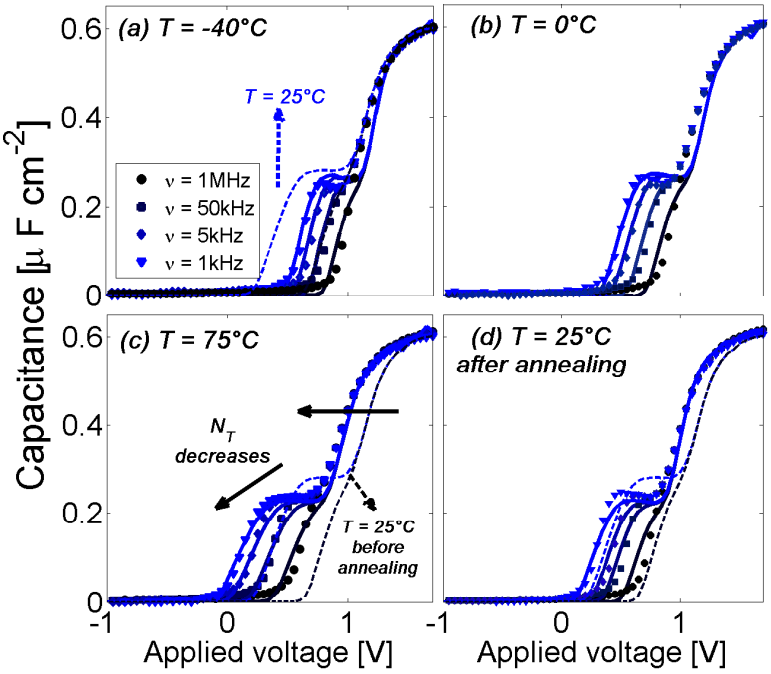


Figure 3-14 – Measured (symbols) and simulated (lines) C_{GC} capacitance in weak inversion as a function of applied bias, for different frequencies and temperatures. The measurements are performed in a specific order. Low temperature measurements at -40°C and 0°C (a-b) show a decrease of the trap frequency response with a narrowing of the parasitic response. At 75°C, a part of the defects relaxes due to the long bench stabilization time (10 minutes) before the measurement (c). Finally, in (d) the capacitance has been measured again after the characterization at 75°C to highlight the recovery.

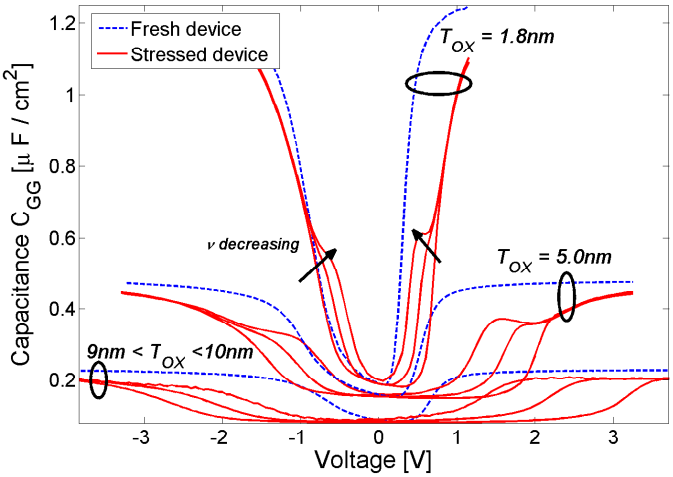


Figure 3-15 – Measured gate capacitances as a function of applied voltage for different devices with oxide thickness ranging from 18Å to 100Å. Both the characteristics of fresh and stressed devices are shown. For the stressed devices, three frequencies have been measured to highlight the parasitic trap response.

decrease at high frequency (inset) corresponds to the exponential increase of the trap conductance.

Figure 3-18 illustrates the measured MOSFET conductances under the same stress conditions as in Figure 3-9(d). Parasitic peaks can be seen on both the G_{GC} (in (a) - blue solid shading) and G_{GB} (in (b) - red shading) conductances after 1000s of electrical stress. In addition to the positive parasitic peak near the flat band voltage (solid lines), the bulk conductance shows a negative parasitic contribution in correspondence of the peak on G_{GC} in weak inversion (dashed lines). This indicates the response of the bulk charges

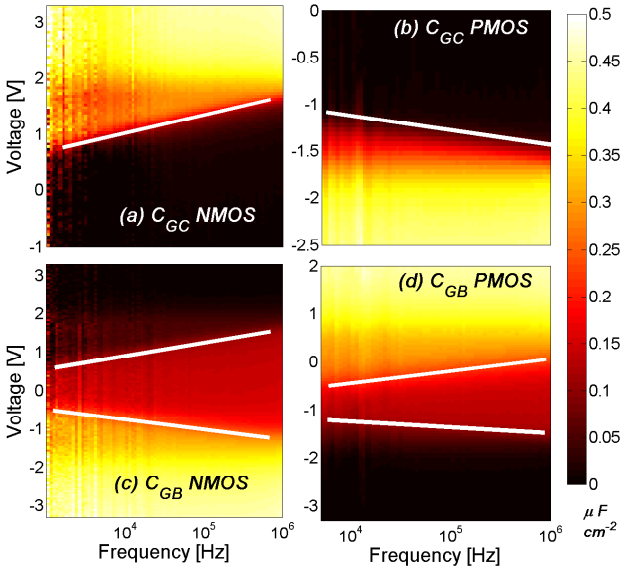


Figure 3-16 – Measured C_{GC} (a-b) and C_{GC} (c-d) capacitances for NFET and PFET with $T_{OX} = 50$Å, respectively, as a function of applied voltage and small signal frequency. The characteristic cut-off poles are indicated in white for both the capacitances. Positive CCS stress has been applied in both the cases. The degradation the PMOS device undergoes is smaller than the NMOS device.

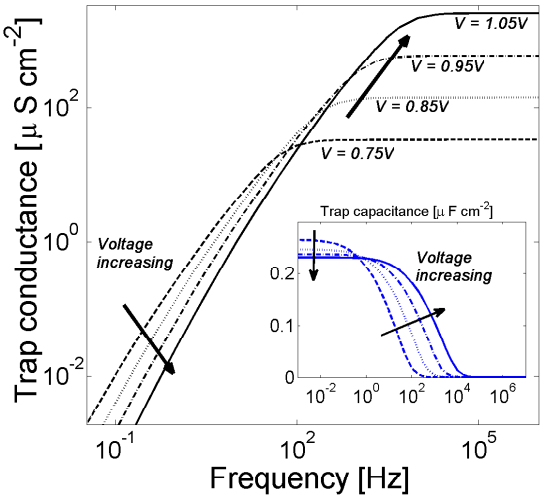


Figure 3-17 – Simulated parasitic trap conductance for device in Figure 3-9(f) as a function of the small-signal frequency at four different applied voltages. The total trap concentration is $N_T = 2.05 \cdot 10^{12}\text{cm}^{-2}$. The simulated trap capacitance is also shown in the inset. The pole frequencies can be identified in both curves.

to the filling of traps by electrons and corresponds to the frequency dependence of C_{GB} in inversion. From a physical point of view, the presence of a conductance peak can be interpreted as a current component that is added to the steady state TAT. The magnitude of all the peaks increases at high frequency. It should be pointed out that the negative peak on G_{GB} is masked by the rise of G_{GC} in inversion and thus it cannot be identified when only measuring the total conductance G_{GG}. The negative conductance implies that the trap site behaves like an active element. Indeed, in the same way as current through positive resistance implies that energy is being dissipated, the current through a negative resistance implies a source of energy.

Figure 3-19 presents conductance measurements as a function of applied voltage per-

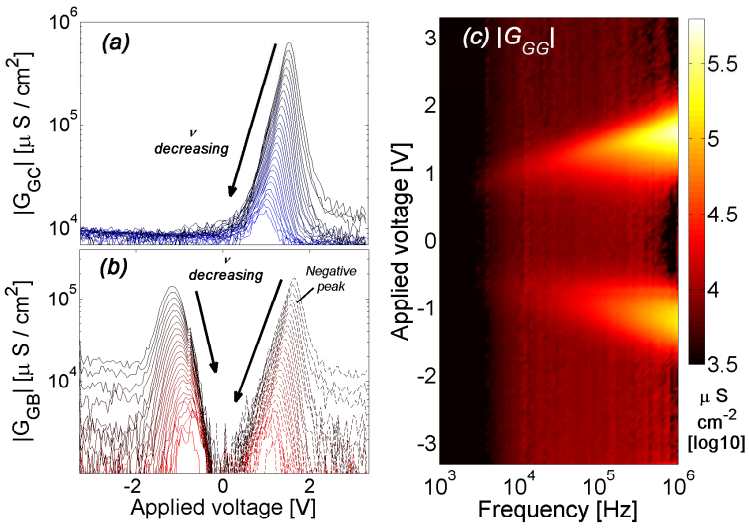


Figure 3-18 – Measured conductance as a function of the applied bias for after 1000s of positive CVS for a device having a 50 Å SiO_2 oxide stack. In (a), a frequency dependent peak is present on the G_{GC} conductance in weak inversion conditions. This peak corresponds to the parasitic trap response on the capacitance in weak inversion due to the communication of e^- in the substrate conduction band. A similar effect is present at flat band conditions on G_{GB} conductance. In weak inversion, a negative conductance peak on G_{GB} is measured as well (dashed lines in (b)). The total MOS conductance is shown in (c).

formed after 1000s of PCCS at $I_{str} = 10^{-8}A$ using a device with $T_{ox} = 18$Å. Simulations obtained using $N_T = 2.8 \cdot 10^{12}\text{cm}^{-2}$ are also shown in (b), with a distribution similar to the one adopted for Figure 3-9. Both the channel $|G_{GC}|$ in (a) and bulk $|G_{GB}|$ conductances in (b) are matching well with measurements and show the exponential decrease of the peaks in frequency.

Additionally, both the steady-state TAT and the direct tunneling currents calculated with Eq. 3.18 have been taken into account. This is particularly evident at higher bias voltages, where the conductance increase due to oxide leakage current is present on both simulation and measurement results.

3.3.3 Discussion

(i) Probed region

In AC analysis, the probed region, i.e. the area of the dielectric in energy and depth contributing to the trap capacitance response, can be determined by the defect AC capacitance response normalized with respect to the applied trap concentration $N_T(x, E_T)$:

$$p_{AC}(x, E_T, \nu) = \int_V \frac{C^T(x, E_T, V, \nu)}{N_T(x, E_T)} dV, \tag{3.33}$$

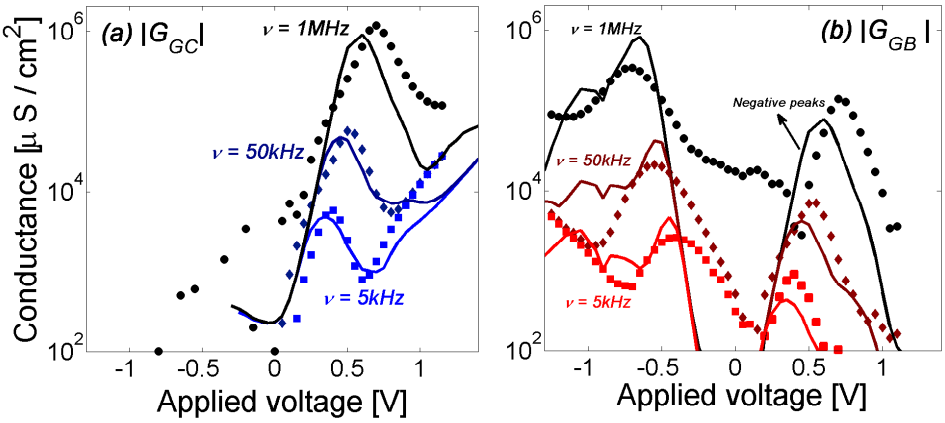


Figure 3-19 – Measured (symbols) and simulated (lines) conductances as a function of applied voltage and frequency (from 5kHz to 1MHz) for a device with $T_{ox} = 18$Å and after 1000s of constant current stress at $I_{str} = 10^{-8}A$. Good agreement is found for both $|G_{GC}|$ in (a) and $|G_{GB}|$ in (b). In particular, the negative peak on G_{GB} in weak inversion is also well reproduced.

Since we are interested in the scanned region, independently from the trap concentration, normalization over $N_T(x, E_T)$ is needed. For each gate voltage V, a specific region of the energy band-gap is probed. The regions contributing to the CV response are indicated in Figure 3-20 as a function of the oxide depth and energy for two different frequencies and DC voltages. Figure 3-20 also shows the increase of the depth of the probed region when the small-signal frequency is lowered. This is in good agreement with previously shown results on C/E rates.

The region probed during a complete CV measurement can be calculated integrating over all the DC voltages. Due to band bending effects, the shape of p_{AC} is deformed to scan higher energetic and deeper defects.

Figures 3-21(a-b) show the region probed with an entire voltage sweep, simulated for a frequency of 20kHz in the dielectric of a CMOS device having T_{OX}=28Å. The accessible region is localized at the Si/SiO_2 interface, in proximity of the Si conduction E_C^{Si} and valence bands E_V^{Si}, and presents a strong dependence on ν. Figures 3-21(b-c) illustrate the region calculated for frequencies of $\nu = 20$kHz and $\nu = 1$MHz. In Figure 3-21(d), the probed region has been calculated at $\nu = 1$MHz for a thicker device having $T_{OX} = 50$Å. Lower frequencies are able to scan deeper traps in the oxide depth and in energy. As frequency decreases, the region extends from the Si conduction and valence band edges towards midgap energies and deeper traps. Mid-gap defects require low frequencies to be scanned, due to the low density of carriers when the device operates in depletion. During the extraction, their response is always present and it cannot be discriminated from the contribution of the scanned defects. A frequency of 1MHz is able to probe 1nm depth defects and in an energy range of more than 2/3 of the Si band-gap. Comparing (c) and (d), it can be noticed that AC characterization is able to scan deeper defects in thinner oxides. On the other hand, traps near the substrate band-edges have characteristic frequencies too

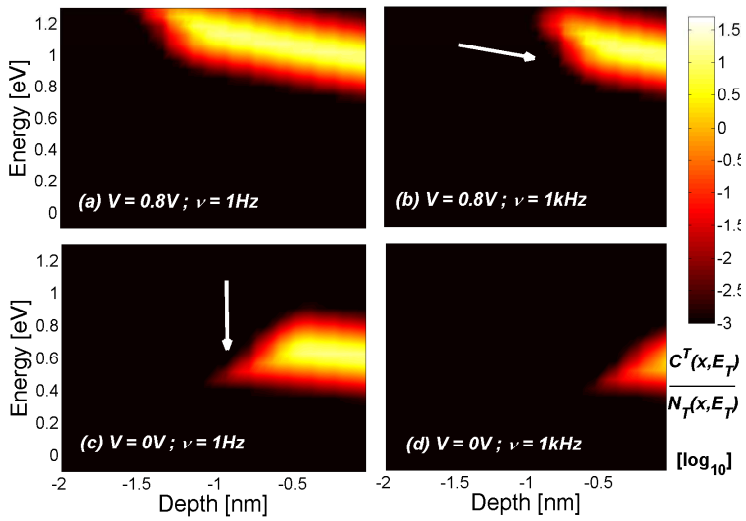


Figure 3-20 – Simulated and normalized capacitive response of traps in oxide depth and trap energy, for two frequencies (1Hz and 1kHz) and two DC voltages in depletion and accumulation on device in Figure 3-9(f). Lower frequencies scan deeper in the oxide. Lower capacitance response is obtained in depletion regime.

high and thus their contribution cannot be reduced for the frequencies available in common LCR meters [218]. This confirms that CV measurements are able to scan only shallow traps in energy and at the Si/SiO_2 interface. Inversely, this also means that the decrease of trap concentrations at Si midgap revealed by empirical models can be attributed to the constant cross-sections approximation which is adopted in most of these extraction techniques.

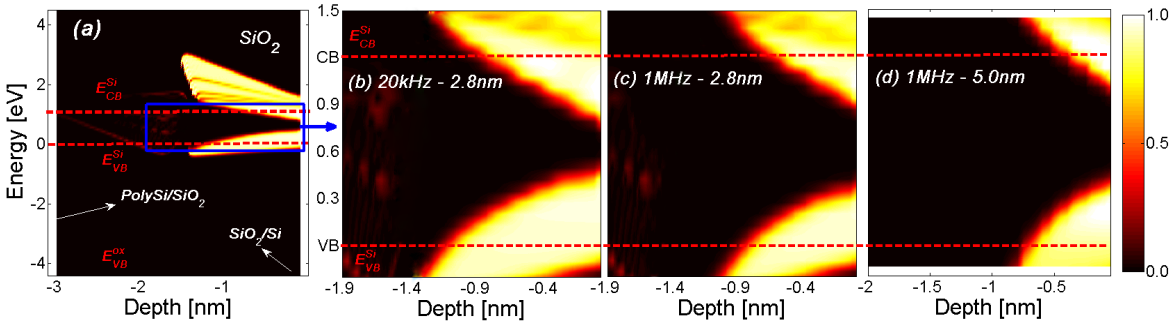


Figure 3-21 – (a) Probed region in the entire oxide layer vs. trap energy and depth from the interface, for the device having T_{OX}=28Å. Band bending occurs depending on the applied voltage and a trapezoidal shape is found both near the conduction and valence bands of Si. (b-c) Zoom from (a) of the region at the Si/SiO_2 interface. Defects at 1nm from the oxide/channel interface are accessible with AC characterization analysis. (d) Zoom of the probed region at the Si/SiO_2 interface for a 5nm-thick device. The same qualitative shape is observed but a shallower region is probed. The reference energy is the Si valence band edge placed at 0eV. The Si/SiO_2 interface is at depth x=0nm.

It is worth noting that the extraction of the trap concentrations with AC characterization depends on the microscopic multiphonon trapping parameters S and F_C in Eqs. 3.7 and 3.8. A complete modeling study of the dependency of the probed region

on model parameters is in Annex D.

(ii) Defect distribution extraction

A relevant question relates to whether trap parameters, such as the Huang-Rhys factor and the spatial/energetic trap distribution, can be determined experimentally. Inversely, could the presented model be used for the interpretation of measurements, such as CV vs. frequency and capture cross sections results associated with deep oxide defects? Some authors claim it is possible and interesting to do so [189, 234, 244], although as pointed out by Stoneham, fitting the defect parameters with a simplified model would certainly cause a false illusion of accuracy [245]. This results from the fact that one tries to describe a complex system using models including coarse approximations (single phonon frequency, detailed balance, field dependence capture probability, etc). The parameters should be regarded as effective parameters rather than microscopic parameters, and the fit may not be unique.

Nevertheless, some interesting qualitative features can be investigated concerning the trap charging model, as well as their energetic and spatial distribution. In the following study, the device analysed in Figure 3-9(f) has been used as a reference, maintaining the total trap concentration constant to achieve a similar impact on the device electrostatics.

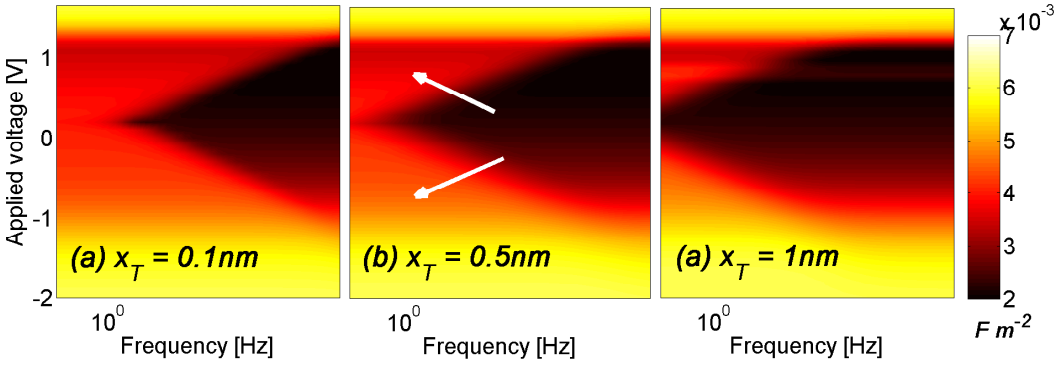


Figure 3-22 – Colour plots showing the simulated gate capacitance vs. the applied voltage and the small-signal frequency for three different mean positions of the spatial trap distribution, from 1 Åto 10 Å. Deeper traps have longer time constants causing the communication with the Si conduction and valence bands of the gate and channel to decrease.

In Figure 3-22, the depth of defects in the oxide layer is progressively increased from 1Å to 10Å. Traps located in the middle of the oxide have a lower frequency response and thus very low frequencies are needed to characterize them. As it will be illustrated in the following section, deep defects can be characterized with SILC measurements, as they communicate with both the substrate (high carrier capture probability) and gate reservoirs (high carrier emission probability) [70]. Figure 3-23 shows the impact of the position in energy of the trap distribution. As the distribution approaches the Si conduction band edge, the trap frequency response in inversion and on C_{GC} increases (a). With a distribution

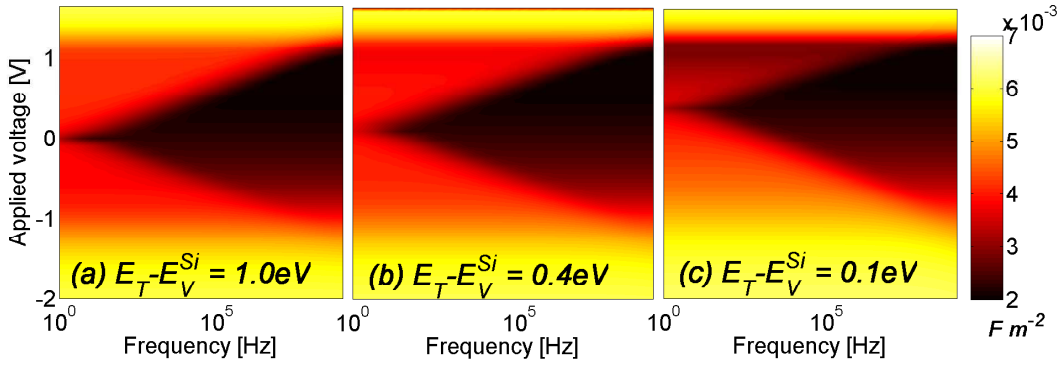


Figure 3-23 – Gate capacitance as a function of the applied gate voltage and the small signal frequency, varying the energy position of the defect concentration peak of the spatial Gaussian distribution. In this case the variance in energy has been reduced to 0.2eV to evidence the effect.

peaking at the Si valence band edge, the response is larger on C_{GB}, while it decreases on C_{GC} (c). Other parameter variations can be found on [239] and in Annex D.

3.4 Trap-Assisted-Tunneling analysis

As indicated in the previous Section, CGV methods fail to scan defects placed deep into the oxide. Trap-assisted-tunneling methods permit to have a better estimation of these traps. Steady-state TAT simulations and measurements have been performed on thin-oxide devices (T_{OX} = 28Å) with the purpose of investigating the defect concentration in the dielectric and assessing the extraction efficacy of this technique for defect characterization.

3.4.1 Response of deep defects

The metric used for evaluating the portion of dielectric region contributing to TAT is given by:

$$p_{TAT}(x, E_T) = \int_V \frac{J_{TAT}^{SS}(x, E_T, V)}{N_T(x, E_T)} dV, \tag{3.34}$$

where the net current density $J_{TAT}^{SS}(x, E_T, V)$ is calculated as in Eq. 3.18, taking into account both e^- and h^+ tunneling and all the C/E fluxes from the two interfaces. The region probed with TAT analysis results from considering all the bias voltages in the sweep and normalizing with respect to the trap distribution.

In the analysis of $p_{TAT}(x, E_T)$, two quantities appear relevant: the spatial and energetic position of the probed region and the magnitude of the current. Both are strongly varying with the bias condition. Figure 3-24(a) shows the normalized probed region for the 28Å structure for a complete gate voltage sweep from -2V to 4.5V, while in (b) the maximum current peak is shown as a function of the bias voltage. In (a), the bright regions

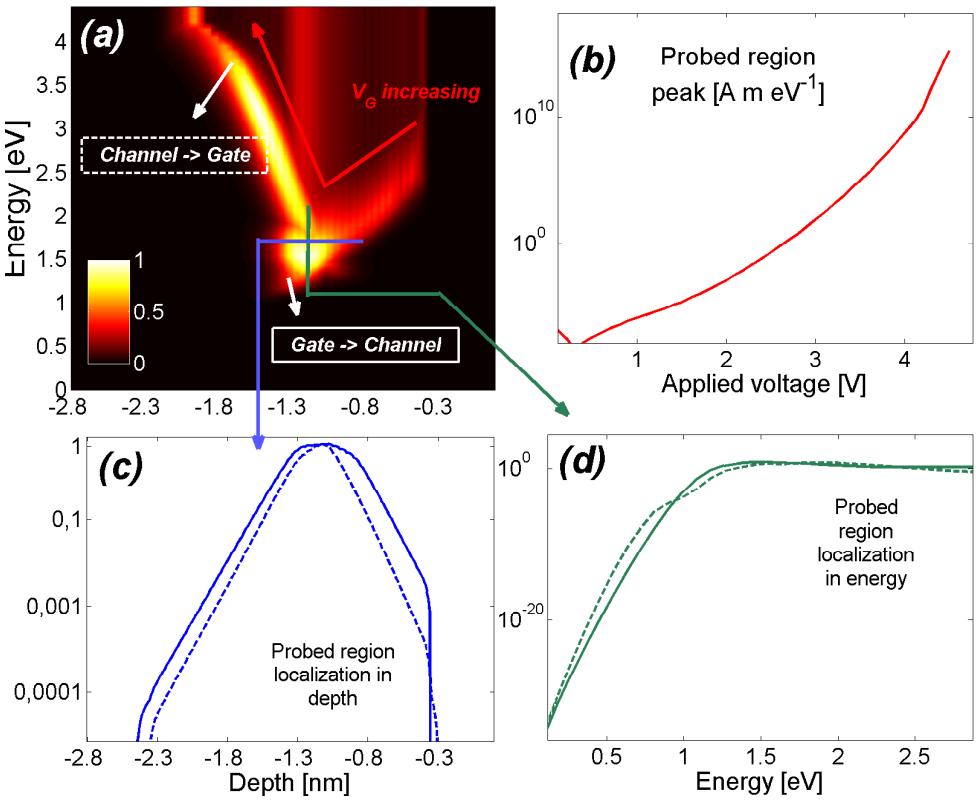


Figure 3-24 – (a) Band diagram of the oxide layer for a device having $T_{OX} = 28$Å, where the regions contributing to TAT during a complete gate voltage sweep from -2V to 4.5V are illustrated in color and as a function of energy and trap position. The Si/SiO_2 interface is located at depth x=0nm and the valence band edge of Si is at 0eV. Bright regions correspond to the parts of the oxide mostly contributing to TAT. The probed region corresponding to the current from the substrate to the gate is shown and, for increasing V, it moves to higher energy levels and towards the gate due to band bending. The region contributing to the current from the gate to the substrate is also illustrated for negative bias voltages. More energetic traps placed closer to the substrate interface are scanned in accumulation and depletion. (b) Normalization factor of the J^{SS}_{TAT} current vs. bias voltage showing a pronounced increase of the weight of highly energetic defects. (c) Cut along the trap depth at $E_T = 1.5eV$ showing the exponential reduction of the contribution far away from the center of the oxide due to the exponential wave function overlapping decrease. (d) Cut along the energy axis of (a) in proximity of the middle of the oxide; dashed lines indicate the region probed for positive V, solid lines the one for negative biases, indicating the dominant influence of highly energetic traps on the probe current.

where defects are participating to TAT are localized at a significant distance from the dielectric/substrate interface, almost in the middle of the oxide. Indeed, for positive voltage biases, defects have both large e^- capture rates from the substrate, and e^- emission rates towards the gate. When considering the individual contributions for each V, it is clear that

the region moves toward high energy levels, as a result of band bending and of the potential distribution in the dielectric. Two cuts in the dielectric depth (Figure 3-24(c)) and energy (Figure 3-24(d)) permit to clearly identify the exponential wavefunction penetration in the oxide layer and the exponential dependence in the oxide gap, respectively. Additionally the magnitude/weight of the contribution also increases due to the higher probability of C/E at higher voltages. This quantity is shown in Figure 3-24(b) and can be used to evaluate the weight of the current in a specific voltage condition. Similar considerations apply to the probed region calculated for a voltage sweep towards accumulation. In such a case, tunneling from the gate to the substrate dominates and a symmetrical region can be observed near the substrate. Although high V conditions permit to probe traps placed at higher energies in the dielectric layer, in such regimes the TAT contribution is masked by the direct tunneling component and thus trap contribution cannot be assessed from measurements.

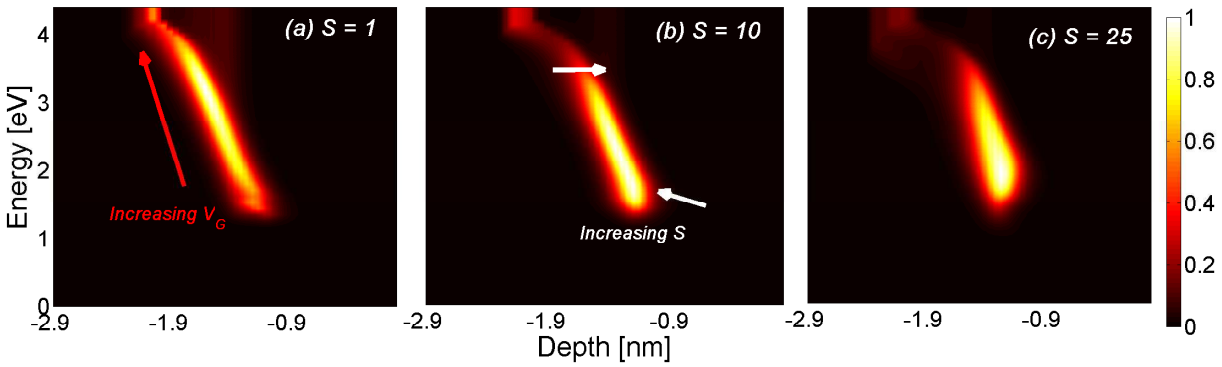


Figure 3-25 – Dependence of the TAT probed region with the Huang-Rhys factor, plotted as a function of trap depth and energy and for a voltage range from 0.1V to 4.6V for three values of S. The Si/SiO_2 interface is located at depth x=0nm and the valence band edge of Si is at 0eV.

Since TAT events are strongly dependent on the value of the transition rates, the trap model parameters play a major role in the current estimation and thus on the extension of the probed region. Figure 3-25 illustrates the dependence of the region on the Huang-Rhys factor. For increasing values of S, the extension of the probed region is reduced and its maximum tends to remain localized in the middle of the oxide layer. The extension in depth is also correlated with the transition of the trap quasi Fermi level [75] from the equilibrium value in the gate to the one in the substrate (Figure 3-26(b)).

Similar considerations apply to the dependence of the probed region on the critical field F_C, as presented in Figure 3-27. Small values of F_C determine an exponential increase in high field conditions. For higher values, the exponential dependence of TAT on V is strongly reduced.

Finally, the probed region that is accessible using TAT has been calculated for different device temperatures (Figure 3-28). At high temperatures the peak is localized close to the substrate since the phonon-assisted transition is favourable. For decreasing temperatures,

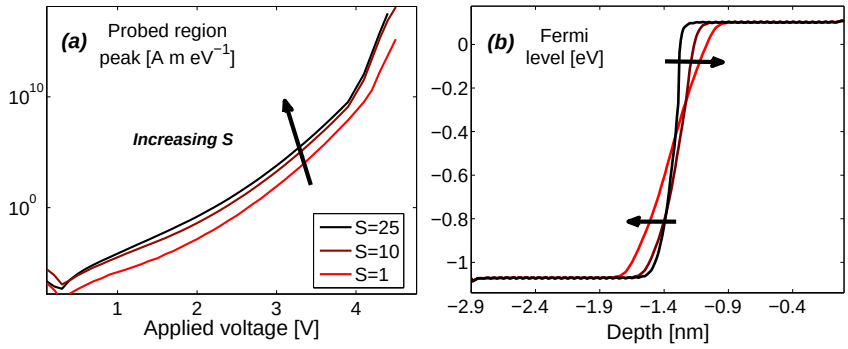


Figure 3-26 – (a) Increasing maximum peak of the probed region vs. applied voltage for increasing S. Higher values of S result in a more localized probed region. (b) Quasi Fermi level calculated in the oxide depth for different values of S and at $V = 1.4V$. The localization of the probed region in depth corresponds to the areas of the oxide where the carrier is not in equilibrium either with the substrate or the gate ($E_F \neq E_{F_{Si}}$ and $E_F \neq E_{F_{Poly}}$). The increase of the width of the transition in E_F for lower values of S, corresponds to an increase of width of the probed region.

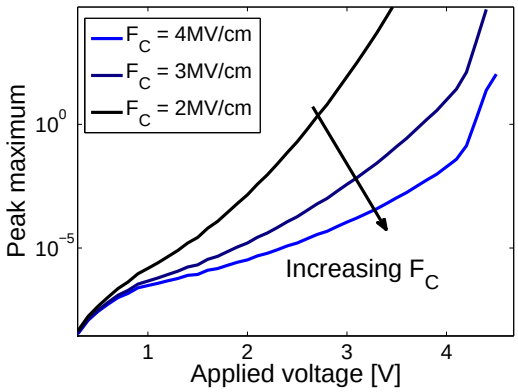


Figure 3-27 – Dependence of the peak of the TAT probed region with V and F_C. The weight of the TAT contribution exponentially increases for increasing F_C. The probed region position in space and energy is not considerably modified by this parameter.

the C/E rates exponentially decrease and the maximum region contributing to the current moves to higher voltages.

3.4.2 Extraction of oxide defect concentrations

In Section 3.3 a distribution of interface and border defects created after electrical stress has been extracted from CV characteristics and attributed to P_b center defects, given the amphoteric nature of the stretch-out. In this Section, the extraction methodology is applied to TAT, relying on the depth and energy behavior of the probed regions. This permits to estimate the distribution of E' defects placed at higher energy levels and deeper in the oxide.

In a first step, the direct tunneling current of the fresh device is extracted having ob-

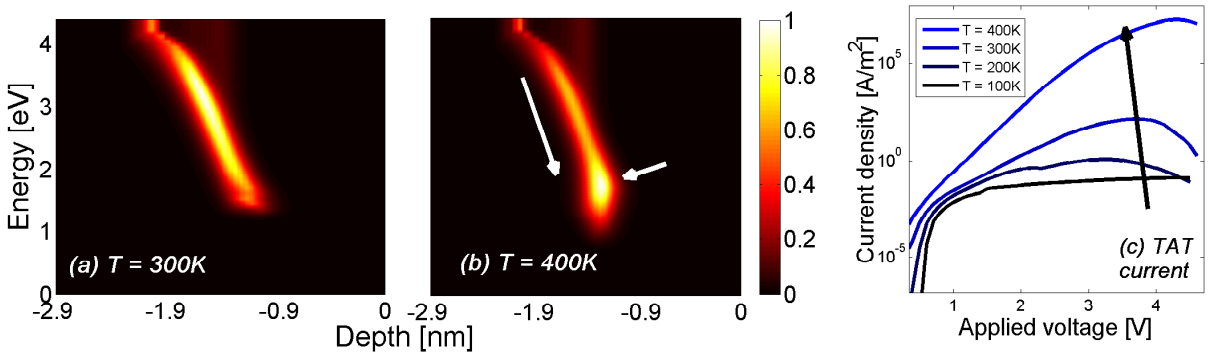


Figure 3-28 – (a-b) Dependence of the TAT probed region with device temperature T, plotted as a function of trap depth and energy and for a voltage range from 0.1V to 4.6V. The Si/SiO_2 interface is located at depth x=0nm and the valence band edge of Si is at 0eV. (c) Increase of the TAT current vs. temperature from 100K to 400K.

tained the correct device electrostatics from fresh AC characteristics. The parameters of the tunneling model, e.g. carrier masses in the oxide, extension of the interface layers [223,228] and band-offsets, can be extracted. Subsequently, a Gaussian trap distribution placed in proximity of the Si conduction band and following Figure 3-8 is defined, extending deeper in the oxide layer and presenting a larger spatial variance than amphoteric defects. This distribution can be associated to E' defects localized at 7 Å from the interface and exhibiting a donor nature. Good alignment has been found on the spatial profiles extracted from CP measurements and showing two Gaussian distributions for P_b and E' defect distributions at the border interface [246]. For these defects, the model parameters of E' defects are extracted from literature ($S = 1.5$ and F_C=3MV/cm [184,247]).

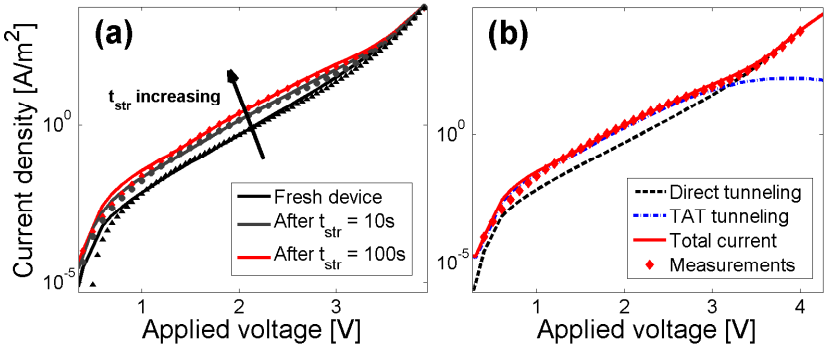


Figure 3-29 – (a) Gate current density as a function of applied gate bias on device with $T_{OX} = 28$Å, after fabrication and after two stress conditions at constant voltage bias of 4.2V for 10s and 100s. Measurement results are indicated with symbols. (b) In weak inversion, the direct tunneling component and the TAT current could be clearly distinguished while for $V > 3.1$ direct tunneling dominates.

Figure 3-29(a) presents the measured and simulated gate current for the fresh device and

after two different stress durations (10s - $N_T = 0.1 \cdot 10^{12}\text{cm}^{-2}$; 100s - $N_T = 0.32 \cdot 10^{12}\text{cm}^{-2}$). The total trap concentration and the spatial and energetic profile are extracted from the most stressed condition. The same extracted profile is also applied for shorter stresses reducing the total concentration. The components of the current are shown in (b), where the direct tunneling current coincides with the leakage in the fresh device in which TAT is negligible. After degradation, a TAT component appears in low voltage conditions. For higher voltages the direct component masks the TAT contribution, and thus its value cannot be determined for $V > 3.1V$. Consequently, the trap concentration in the probed region corresponding to this range of voltages remains uncertain, only using TAT characterization. However, eventual defects in this region would contribute to TAT with the weight reported in Figure 3-29(b), generating an exponential increase of the TAT component overcoming the direct tunneling component. Measurement results seem to disqualify this hypothesis, since the TAT component is observed to decrease at intermediate voltages for higher stress conditions. This tends to indicate that the high energy region is not populated by traps. Second-order effects related to the influence of defects on the direct tunneling current have also been noticed but appear negligible with respect to the exponential increase of the TAT current for increasing V. It has also been determined that for longer stresses, soft breakdown occurs and percolation strongly increases the TAT current. In such a case, a different approach is required to model TAT [248, 249].

3.5 Transient analysis

Having discussed charge trapping phenomena in DC and AC conditions with TAT and AC analysis, a third trap characterization method based on MFCP transient measurements has been analysed. Transient simulations require analyzing the validity of Eq. 3.4 for trap equilibrium approximation (Subsection 3.5.1). Out-of-equilibrium conditions and transient effects in MFCP characteristics have been investigated in Subsection 3.5.2. This section also deals with the dependencies of the CP current and of the extension of the probed region on CP pulse parameters. Subsection 3.5.4 covers other transient effects related to the C/E dynamics of oxide traps and details their role in hysteresis loops on AC characteristics. Finally, recovery dynamics of stressed devices after various degradation conditions are also compared in Subsection 3.5.5.

3.5.1 Equilibrium conditions and trapping dynamics

In the previous sections, the DC operating point has been calculated in *equilibrium* quasi-static conditions, neglecting all the large-signal transient effects. Consequently, for each simulated voltage, all the border traps have been assumed in equilibrium with the gate or the channel, while in measurements this is not always the case. Indeed, it has been shown that deeper oxide traps have C/E time constants much higher than the measurement time between two voltages [218]. Consequently, it is important to discuss the validity of this approximation in detail.

Let us consider a cross-section plane in the dielectric at a given oxide depth x. During a transient pulse, we can assume that each individual trap at depth x exchanges carriers only with the channel and the gate. In the present 1D model, transverse directions (y, z) are not explicitly accounted for. The large number and variety of defects that can be present in the cross section plane at x have been associated to an energy distribution. Moreover carrier exchange by trap to trap tunneling have been neglected. The rate equation needs to be solved for each energetic level, calculating the quasi Fermi level E_F in the oxide stack for all the possible trap energies E_T. The general approach of Eq. 3.1 is required to accurately take into account the out-of-equilibrium conditions of the system.

It should be pointed out that the attribution of a quasi-Fermi level to a single defect has been only performed in the view of better illustrating the *equilibrium* steady state condition. Indeed, in the previous section, the DC operating point for the AC analysis has been calculated in quasi-static conditions, neglecting all the large-signal transient effects. After a sufficiently long time, the defects establish a common equilibrium condition with the substrate/gate and their quasi Fermi levels equal the one in the carrier reservoirs. Under this equilibrium condition, an uniform quasi-Fermi level is eventually established for all the defects located on an entire device cross-section x. In practice, since $E_F(x)$ is the same for the totality of defects at depth x, the rate equations describing the dynamics of the different traps can be summed term-to-term and a single rate equation controlling the trapping dynamics is obtained.

Eq. 3.4 differs from Eq. 3.1 as it includes the integration over trap energy E_T. In the following, this approach will be referred to as "quasi-equilibrium" (QE) approximation. In such an approach, non-equilibrium effects on f_T are neglected. As pointed out in this section, the QE approximation can lead to unphysical interpretations when large signals are applied. Moreover, since the approximation still remain applicable to small-signal AC analysis and this solution is sensibly more efficient compared to the general approach, it is important to discuss the range of validity of this solution and its consequences on the misleading interpretations of transient results.

3.5.2 Multi frequency charge pumping

(i) Calibration methodology

MFCP and AC results have been analysed to investigate the importance of transient out-of-equilibrium effects on CP characteristics and the relevance of the methodology applied for the extraction of trap concentration profiles.

In a first step, AC characteristics, measured on electrically stressed multi–fingered NMOS devices integrated in a 65nm technology, have been compared to simulations results, by solving Eq. 3.4 in DC equilibrium conditions. The extracted spatial/energetic defect profile follows a bivariate Gaussian distribution, whose validity has been discussed in [75] and in the previous section.

In a second step, for the same stress conditions, charge pumping measurements have been performed applying periodic pulses on the gate of the CMOS device. Figure 3-30(a)

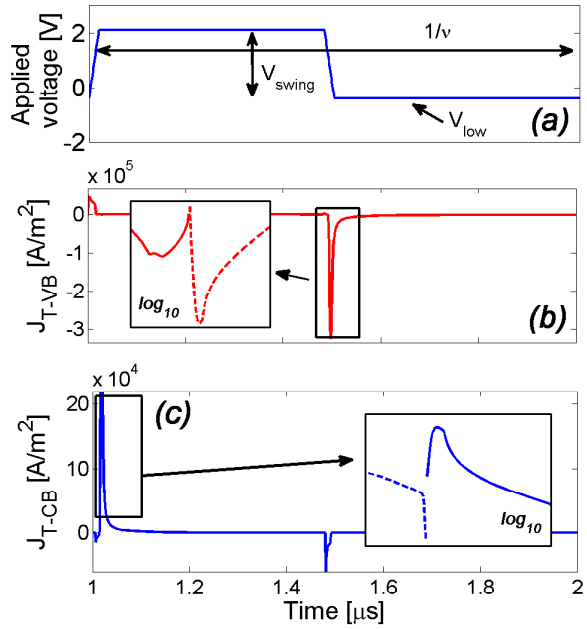


Figure 3-30 – (a) Pulse applied for charge pumping analysis. Both the amplitudes V_{low}/V_{high} of the pulse, the voltage swing V_{sw} and the frequency have been varied. (b) Simulated current density generated by the capture emission of holes from the Si valence band of the substrate. The same calculation can be performed for the C/E flows of electrons on the conduction band of the substrate in (c). Insets indicate the absolute value of the currents in logarithmic scale during the transition phases; solid red: h^+ emission, solid blue: e^+ capture, dashed red: h^+ capture, dashed blue: e^+ emission.

shows a single simulated CP pulse applied to the device. The analysis has been performed varying the low base voltage bias V_{low} from -2V to 0.5V, while the amplitude of the applied pulse, i.e. the swing voltage $V_{sw} = V_{high} - V_{low}$, ranges from 0.8V to 2.5V. The signal frequency ν is varied from 1kHz to 1MHz. Three square pulses have been simulated to avoid the presence of undesired transient effects originated from the initial equilibrium condition.

The transient trapping currents are shown in Figure 3-30(b-c) and can be expressed by the net e^-/h^+ fluxes:

$$\begin{aligned} J_{T_{VB}}(t) &= \int_x \int_{E_T} [\tilde{\Phi}_C(x, E_T, t) - \tilde{\Phi}_E(x, E_T, t)] dE_T dx \\ J_{T_{CB}}(t) &= \int_x \int_{E_T} [\Phi_C(x, E_T, t) - \Phi_E(x, E_T, t)] dE_T dx, \end{aligned} \tag{3.35}$$

Figure 3-30(b) shows that for the CP pulses where the MOS operates both in accumulation and inversion (V_{sw}=2.5V, $V_{low} = -0.4V$), the $V_{high} \rightarrow V_{low}$ transition generates a hole capture current (dashed red curve), which exponentially decreases when the voltage is maintained at V_{low}. Electron emission can also be noticed in this phase. On the other hand, during the $V_{low} \rightarrow V_{high}$ transition shown in the inset of Figure 3-30(c), electron capture (solid blue) dominates in weak/strong inversion, while hole emission occurs when the device is in depletion. The current contributions during the transitions must be carefully taken into account due to their important role in the estimation of the CP current.

The total charge pumping current J_{cp}, depicted in Figure 3-31(b) as a function of V_{low}, has been calculated integrating the net transient currents $J_{T_{VB}}(t)$ and $J_{T_{CB}}(t)$ over an

entire square pulse:

$$J_{cp} = \int_{T'} J_{T_{VB}}(t)dt = -\int_{T} J_{T_{CB}}(t)dt. \tag{3.36}$$

where T' corresponds to a full period of the CP pulse $T' = 1/\nu$.

Figure 3-31(a) indicates the AC characteristics of the MOS device for various small-signal frequencies from 1kHz to 1Mhz. In (b), MFCP measurements and simulations performed using the same distribution extracted from AC characteristics following the methodology adopted in [75] are reported. For each V_{low} in Figure 3-31(b), a full transient simulation has been performed with the general transient model of Eq. 3.1 and the CP current has been calculated. Good correlation with transient results for different frequencies, and literature has been found [250].

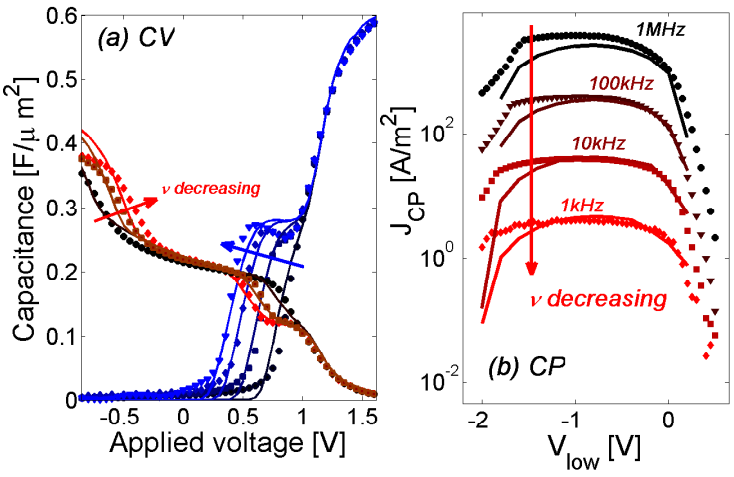


Figure 3-31 – (a) Gate-channel (in blue) and bulk (in red) capacitances as a function of the bias voltage for different small-signal frequencies and after a positive constant voltage stress of 1000s at 5.5V. Symbols indicate measurement results, while simulations are shown in a solid line. Both the threshold voltage shift and the parasitic capacitive component of the traps are accurately reproduced for a frequency range from 1kHz to 1MHz [75]. (b) Charge pumping current vs. V_{low} voltage for different CP pulse frequencies. The same spatial/energetic trap distribution profile has been adopted to reproduce all the results.

3.5.3 On the validity of QE approximation

Figure 3-32 shows the calculated E_F and f_T as a function of energy and position in the dielectric for two timesteps of a CP simulation. At the beginning of the pulse (plots (a-b)), the trap distribution is in equilibrium with one of the reservoirs (DC conditions; $V = V_{low}$ =-1.2V), following the quasi Fermi level in proximity of the Si/SiO_2 interface regions. In the middle of the oxide, when taking into account the gate capture/emission contributions, different equilibrium conditions dependent on the energetic positions of the defects are created, and as a consequence both E_F and f_T vary with energy. If one considers the QE approximation (Figure 3-32(c)), a similar equilibrium condition on f_T is found in proximity of the two interfaces, e.g. within 1nm–depth, but differences are found in the transition region, i.e. for a depth between 1nm and 3nm.

On the rising edge of the voltage pulse, defects in proximity of E_C^{Si} are progressively filled by e^-, while traps near E_V^{Si} are emptied of holes. The equilibrium front propagates from the two Si band edges as charge exchange driven by the capture/emission time constants continues [218]. When decreasing the applied voltage, e^- are emitted towards the Si conduction band and traps at the midgap capture h^+ from the valence band.

After an entire CP pulse, deep defects do not reach equilibrium (Figure 3-32(d)) and remain filled of e^- following a spatial and energetic f_T distribution (Figure 3-32(e)). If one chooses to apply the QE approximation to transient simulations, the capture/emission fronts only propagate along the direction of oxide depth while the propagation along the direction of the energy level is not present. As a consequence, the charge distribution in out-of-equilibrium conditions after a CP pulse is uniform in energy and forms a peak at a given depth.

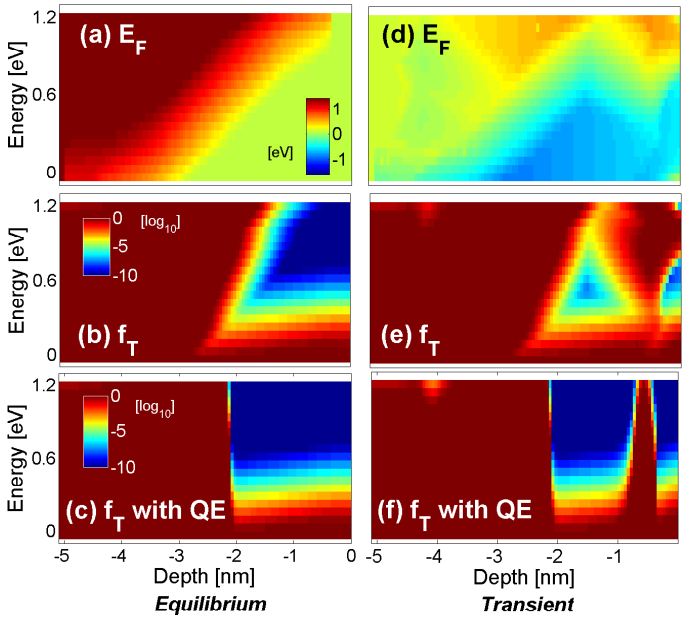


Figure 3-32 – Calculated quasi-Fermi level (a,d) and trap occupation (b,e) distributions in energy and oxide depth for different bias conditions during a CP pulse using the transient model of Eq. 3.1. The Si/SiO_2 interface is placed at position $x = 0$ with the substrate conduction and valence band edges at 0eV and 1.2eV. Initial DC equilibrium conditions at $V_{low} = -1.2V$ are shown in (a) and (b). Transient conditions after two CP cycles are represented in (d) and (e). Drastical differences are noticed when the QE approximation is considered. In transient conditions, only the dependence of C/E rates in depth is considered (plots (c,f)), causing border/oxide defects to remain filled almost independently from their energetic level.

Figure 3-33 shows the differences on the maximum CP current obtained with the QE approximation, illustrating the effects of neglecting the energy dependence of the trapping time constants during the extraction and confirming its importance in a general transient

analysis. In the following, the general transient approach is used without considering the QE approximation.

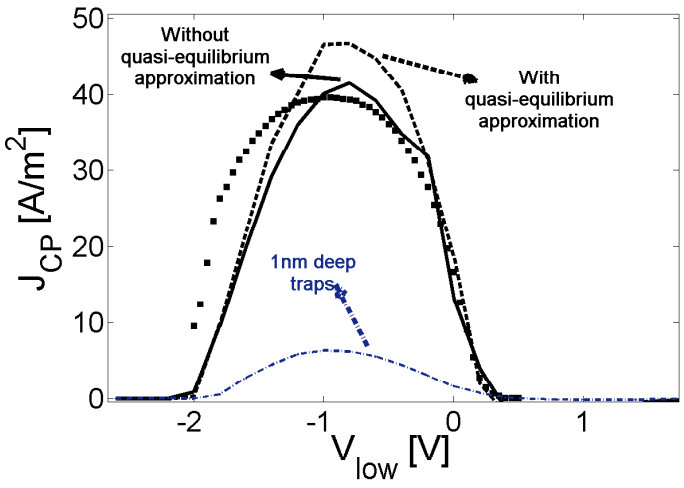


Figure 3-33 – Measured (symbols) and simulated (lines) charge pumping current as a function of V_{low}. Simulation results for the same trap distribution but shifted 1nm away from the Si/SiO_2 are also shown in dot-dashed line.

(i) Pumped and probed CP regions

Two representative metrics are used to assess the limitations of the MFCP technique in the extraction of the defects distribution, namely the pumped and effectively probed oxide regions. The pumped region represents the spatial and energetic portion of the oxide where traps are exchanging carriers with the substrate during a single CP pulse. This quantity has been calculated using an analytical approach similar to the one adopted in [198, 199] from the maximum electron occupations and maximum hole occupations, using:

$$g_{CP}(x, E_T) = (1 - f_T(x, E_T, t_A)) f_T(x, E_T, t_B), \tag{3.37}$$

where t_A and t_B represent the timestep which occur after the low level voltage plateau (corresponding to the maximum capture of h^+), and after the high level voltage plateau (corresponding to the maximum capture of e^-), respectively.

Figure 3-34 shows the pumped zone for different V_{low} voltages and for a 1kHz CP signal, and indicates the regions exchanging carriers with the substrate during the voltage pulse. Charge exchange is higher (bright color) in proximity of both the substrate conduction and valence band edges located at 0eV and 1.2eV at position x=0. Additionally, midgap traps are able to capture electrons and emit holes contributing as well to the pumped region.

The probed region offers a better estimation of the degraded oxide regions contributing to carrier recombination and to the charge pumping current calculated with Eq. 3.36. This quantity can be interpreted as the impulse response of the CP current due to a single trap placed at (x, E_T) and is determined as:

$$p_{CP}(x, E_T) = \frac{\int_{T'} (\tilde{\Phi}_C(x, E_T, t) - \tilde{\Phi}_E(x, E_T, t)) dt}{N_T(x, E_T)}, \tag{3.38}$$

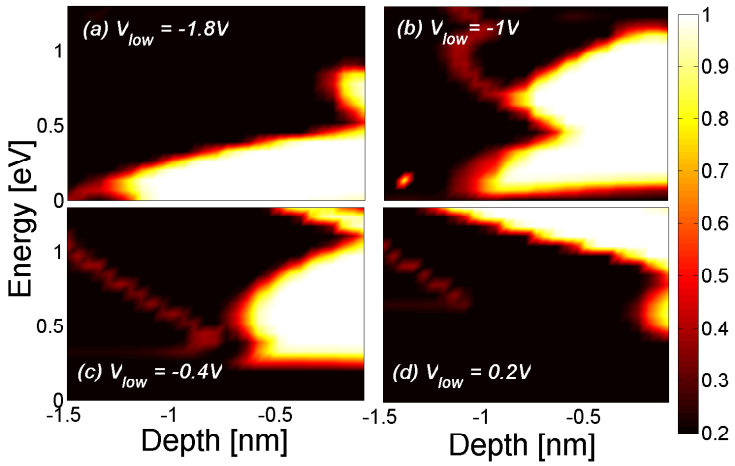


Figure 3-34 – Pumped region calculated using Eq. 3.37 for various V_{low} voltages and using a 1kHz pulse. The Si/SiO_2 interface is placed at position x=0 while the Si conduction and valence bands are respectively at 0eV and 1.2eV. The bright regions at the interface and near E_C^{Si} and E_V^{Si} are able to communicate with the substrate during the CP pulse. The value has been normalized between 0, corresponding to no carrier exchange, and 1, indicating maximum capture/emission.

Figure 3-35 shows the calculated $p_{CP}(x, E_T)$ for a signal frequency of 1kHz and for different V_{low} conditions. Only midgap defects placed at the Si/SiO_2 interface are probed, as carrier recombination occurs in these energetic positions at the Si/SiO_2 interface with the substrate. This is observed on Figure 3-35 by the presence of a bright area at midgap representing the probed region vs. energy and oxide depth. Defects placed near the conduction and valence bands have too high capture/emission rates to contribute to the charge pumping current: electrons in the conduction band captured into the defects during the rising edge are immediately emitted towards the same band during the falling edge, and thus do not contribute to the CP current. Deeper defects have large time constants and are not probed, as they are not able to exchange carriers with the channel at the considered pulse frequency. Different V_{low} voltage conditions are shown and indicate that only specific midgap regions and at the Si/SiO_2 interface can be probed. Figure 3-33 shows CP simulations for the same defect distribution profile but shifted 1.2nm away from the interface. Different V_{low} voltage conditions are shown and indicate that only specific midgap regions and at the Si/SiO_2 interface can be probed.

Comparing the pumped region g_{CP} and the probed region p_{CP} in Figures 3-34 and 3-35 respectively, one evidences that the pumped zone includes the probed region. Indeed, g_{CP} includes two parasitic *wings* near E_C^{Si} and E_V^{Si}, corresponding to the regions where both capture and emission of electrons or holes occur, which should not be taken into account in the determination of the probed region and J_{cp} current calculation.

Major discrepancies are observed on the pumped region when the QE approximation is introduced (Figure 3-36): the energy dependence of the pumped region on the trap energy is reduced and thus a large portion of the energy gap is pumped. However, in this case,

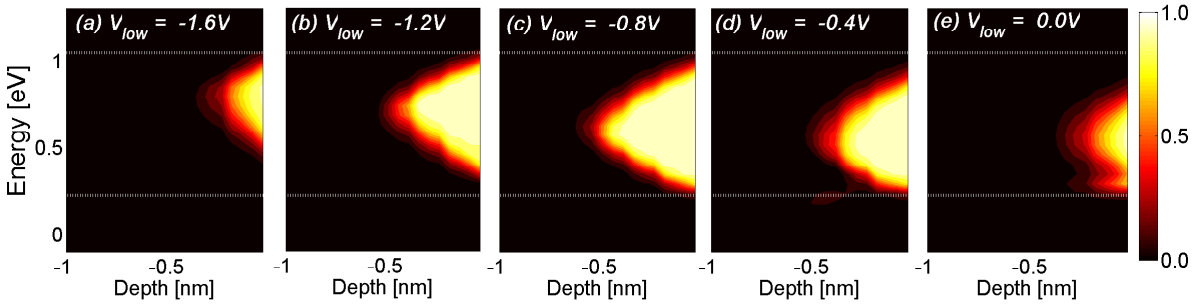


Figure 3-35 – Probed CP region vs. trap energy and position in the oxide layer, calculated with Eq. 3.38 at $\nu = 1kHz$, $V_{sw} = 2.5V$ and $T = 300K$. The current peak in Figure 3-33 corresponds to the condition where the probed region is maximum in extension and the operating regime varies from accumulation to inversion as in (b). The Si/SiO_2 interface is placed at position $x = 0$ while the Si conduction and valence bands are at 0eV and 1.2eV, respectively.

the probed regions cannot be derived from the fluxes contributions using Eq. 3-34, since this would conduct to an incorrect interpretation, concluding that fluxes near E_C^{Si} and E_V^{Si} dominate over the fluxes at midgap [240].

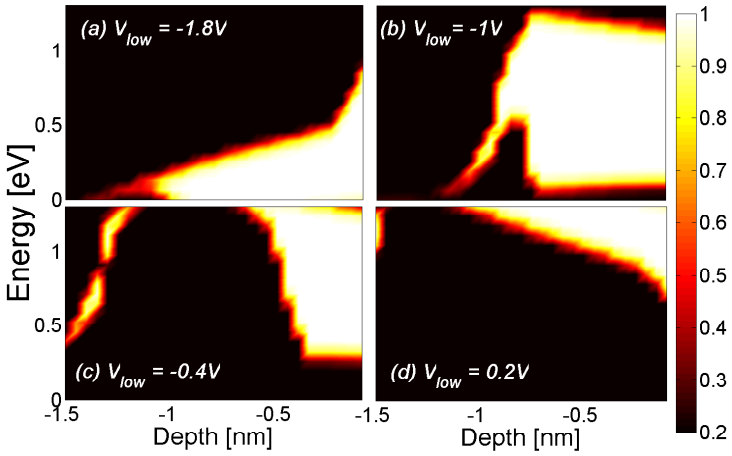


Figure 3-36 – Pumped CP region vs. trap depth and energy, calculated using Eq. 3.37 for different V_{low} conditions and using the QE approximation ($\nu = 1$kHz, $V_{sw} = 2.5V$ and $T = 300K$). The reference energy is the Si valence band, while the Si/SiO_2 is at depth x=0nm. A trapezoidal region is found in (b), which relates to the V_{low} condition where the peak in J_{cp} is maximum. This result is in accordance with calculations performed in [198]. The extension of the region in energy is larger due to the fact that traps are in equilibrium with the substrate. The extension in depth remains comparable to the results shown in Figure 3-35.

The extension of p_{CP} in oxide depth and energy strongly varies with the pulse characteristics (swing, frequency, low-base bias, ...). For this reason, MFCP techniques are adopted

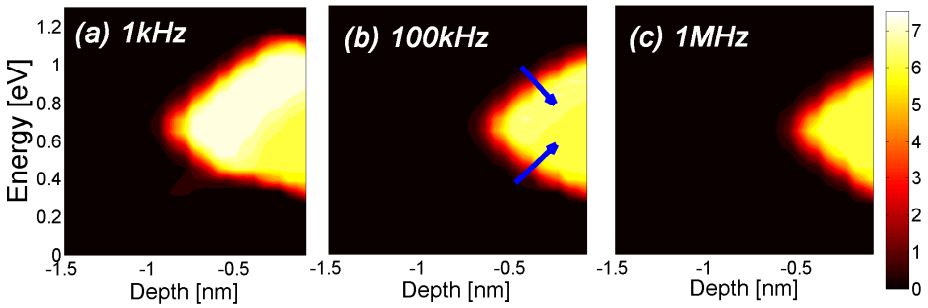


Figure 3-37 – Probed CP region vs. trap depth and energy, calculated using CP pulses with $V_{low} = -1.4V$ and different frequencies. Deeper defects having reduced time constants are scanned using low-frequency pulses.

to scan particular regions of the oxide layer by comparing the measured CP currents in response to diverse pulse conditions. Figure 3-37 illustrates the simulated p_{CP} for three frequencies (1kHz, 10kHz, 1MHz) at $T = 300K$ and $V_{sw} = 2.5V$. The maximum probed regions achievable are illustrated at V_{low} corresponding to the maximum CP current. The trap concentration in specific areas of the oxide (in energy and depth) can be extracted by comparing the current contribution of the additional probed regions scanned with lower frequencies signals. Similar considerations can be applied considering the variation of J_{CP} with V_{sw}, shown in Figure 3-38 from $V_{sw} = 1$V to 2.5V. 1nm-deep defects at the midgap can be probed at approximately ν =1kHz and using $V_{sw} = 2.5V$, while the defects closer to the Si band edges require much lower frequencies.

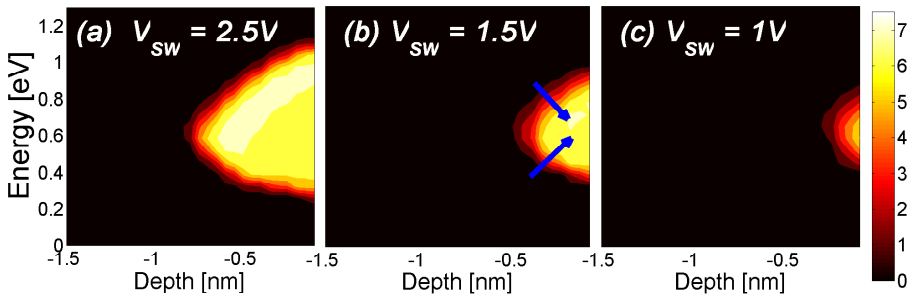


Figure 3-38 – V_{sw} dependence of the probed region at $\nu = 10$kHz, $V_{low} = -1.2V$ and $T = 300K$. The reference energy is the Si valence band, while the Si/SiO_2 is at depth x=0nm. Higher V_{sw} permits to scan deeper in the oxide depth and at energies closer to the Si band edges.

Temperature also plays an important role in multiphonon-assisted transitions. Figure 3-39 shows the variation of the probed region extension at (a) T=200K and (b) T=400K, indicating the increase of the probed region extension when temperature is raised. The rise/fall times and the duty cycle of the pulse have been found to play a minor role on the CP current and on the simulated probed region.

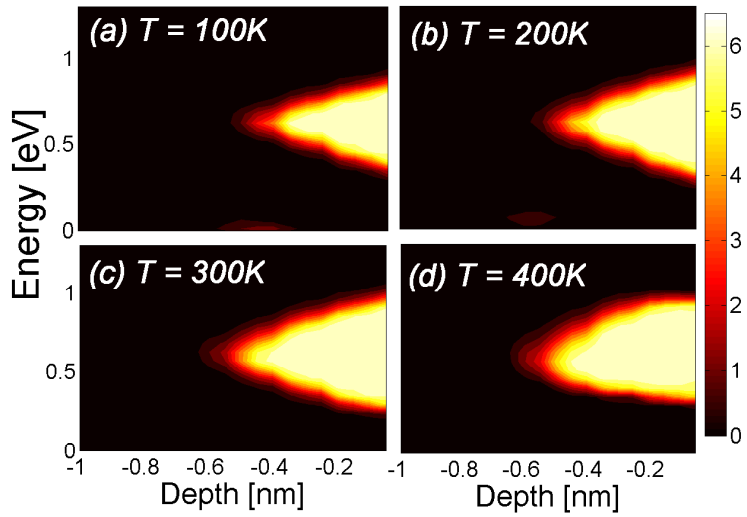


Figure 3-39 – Temperature dependence of the probed CP region at $\nu = 1MHz$, $V_{low} = -1.2V$ and $V_{sw} = 2.5V$. The reference energy is the Si valence band, while the Si/SiO_2 is at depth x=0nm. Higher temperatures scan deeper in the oxide depth.

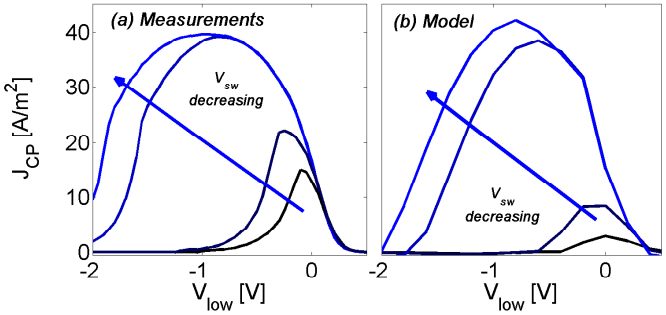


Figure 3-40 – Charge pumping current as a function of the low level V_{low} voltage and for different pulse swings V_{sw}. Measurement results at ν=10kHz in (a) are in qualitative agreement with simulations in (b) and with literature [194, 250–252].

Figure 3-40 shows the dependence of the charge pumping current with V_{sw} for a 10kHz CP signal frequency. The charge pumping current is plotted as a function of V_{low} and for different voltage swings V_{sw}: increasing the magnitude of the swing allows to scan deeper traps and the peak increases since carrier capture increases. A reasonable agreement has been obtained using the same defect distribution than in Figure 3-31.

The rise/fall times and the duty cycle of the pulse have been found to play a role on the maximum CP current and on the simulated probed region.

Finally, Figure 3-41 presents charge pumping current results of amplitude-sweep simulations which have been performed varying the swing amplitude by keeping the same V_{low} voltage (charging mode in (a)) and or by keeping the same V_{high} voltage (discharging model (b)). These extraction techniques are usually adopted to measure the magnitude of the trapped charge and to extract the energy dependence of trap distribution profiles. In (a), for low swing voltages the current remains low as the device always operate in accumulation/depletion and does not enter in inversion. When this latter regime is reached, electrons can recombine with holes at the trap sites and the current increases. In (b), the V_{high} voltage level of the pulse is higher than V_{th} and only for very low V_{low} voltages a

recombination current emerges.

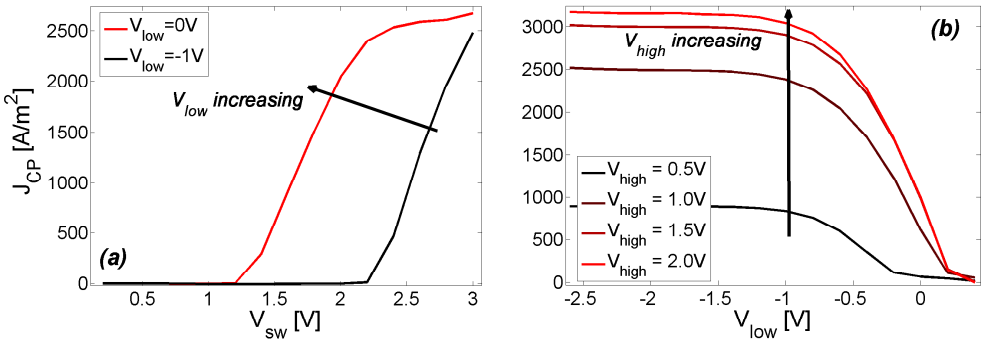


Figure 3-41 – (a) Simulated charge pumping current vs. V_{sw} voltage during a charging amplitude-sweep CP technique for different V_{low} voltages. (b) Simulated charge pumping current vs. V_{low} for a discharging amplitude-sweep CP for V_{high} voltages.

3.5.4 Hysteresis effects

Hysteresis loops can be noticed on measured AC characteristics in both weak inversion and accumulation; the interpretation offered by the model indicates that they are due to the long time constants of border and oxide traps which get filled during the DC voltage sweep. Figure 3-42(a) shows the measured C_{GG} capacitance after 1000s of constant voltage stress, where AC measurements have been performed ramping up and down the applied voltage. As can be seen, an hysteresis loop of magnitude Δh, is formed in weak inversion and accumulation. The full pulse applied by the AC analyzer and simulated with the model is shown in Figure 3-42(b). During the voltage ramp-up, deep traps having time constants of the order of tens of seconds/minutes (the time elapsed during the ramp up) progressively get filled and the electrostatics of the system changes with a positive increasing stretch-out. During ramp-down, the system is in a different condition with respect to the ramp-up as deep traps are filled and require a given amount of time to emit carriers. Consequently, the dynamic of charge trapping is slightly altered, causing the hysteresis loop. The magnitude of the latter is found to be slightly dependent on the frequency of the small signal pulse, as can be noticed on Figure 3-42(c). It should be pointed out that the hysteresis loops in Figure 3-42(a) cross each other in depletion, revealing electron filling in inversion (V_{th} increase) and hole capture in accumulation (V_{fb} decrease).

The effect can be seen with a full transient simulation using the multiphonon model and considering the pulse in Figure 3-42(b). A low frequency large signal sawtooth pulse is applied and when reaching a given voltage, a small signal sinusoidal pulse permits the determination of the transient currents and admittances of the system calculated as in Section 3.2. This operation is performed on both the ramping up and ramping down of the large signal pulse, for a limited set of DC voltages and frequencies. The simulated capaci-

tances shown in Figure 3-42(d) indicate that the hysteresis loops on the C_{GC} capacitance is in qualitative agreement with measurements in Figure 3-42(c).

This phenomenon further confirms the presence of a wide spread distribution of C/E rates for deeper traps in energy and depth, in accordance with multiphonon model results and recent TDDS measurements of the trap time constants [184, 201]. Additionally, the stretch-out shown for all the previously shown CGV curves is affected by the total measurement time and thus the history of the system complicates the accurate determination of the threshold voltage shift.

The magnitude of the hysteresis loops strongly depends on the total rise time of the DC pulse (not shown here): with a more rapid increase of DC voltage, deeper traps have less time available to stabilize and hysteresis decreases.

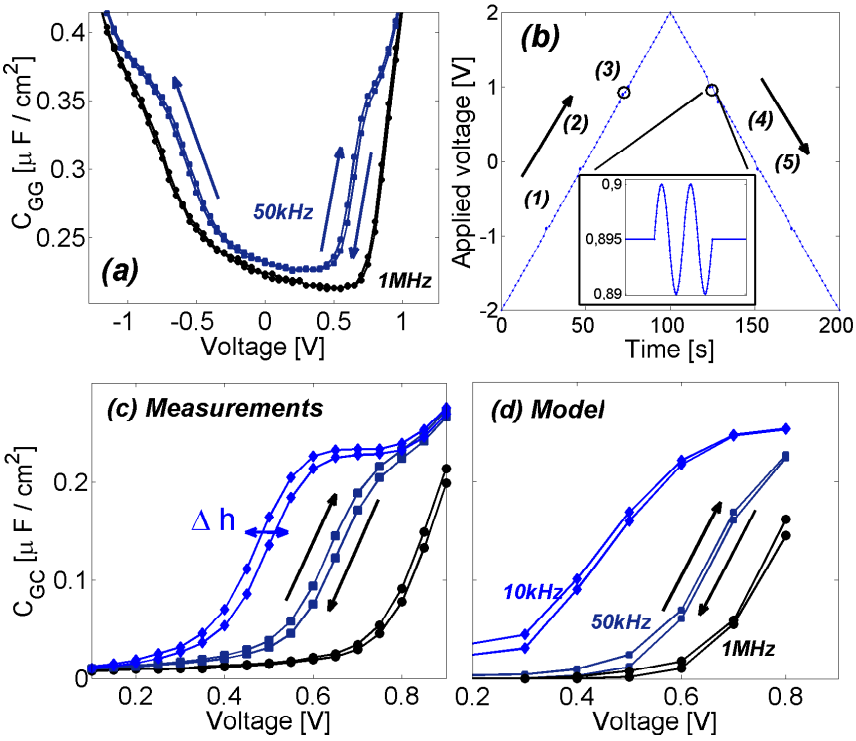


Figure 3-42 – (a) Gate capacitance measurements performed after 1000s PCVS, ramping-up and subsequently ramping down the large signal voltage for two frequencies. (b) Applied voltage vs. time for the simulation and the analysis of hysteresis effects. A slowly varying DC voltages is applied to the device, while an AC small signal pulse (inset) is used to calculate the AC response of the device from the trap displacement currents. The AC small signal pulse is considered on both the rise and the fall of the DC voltage and is performed only for a limited number of DC voltages. (c) Detail of the hysteresis loops on C_{GC} capacitance measured in the frequency range 10kHz-1MHz. The width of the hysteresis loops Δh increases due to the fact that at lower frequencies the total measurement time increases and more time is let to deep traps to capture/emit carriers, increasing the magnitude of the phenomenon. (d) Simulated hysteresis loops in the channel capacitance C_{GC} for different frequencies. The arrows indicate the versus of the applied DC voltage.

3.5.5 Trap recovery

Trap recovery dynamics for NFET and PFET devices have been characterized using two approaches. In the static approach, the NMOS device with $T_{OX} = 50$Å is stressed for 1000s forcing a constant current $I = 10^{-8}A$ across the oxide stack; the AC characteristics reported in Figure 3-43 are measured immediately after stress, after 6 hours and after 18 days to monitor the annealing time constants of defects. After short time periods, only the recovery of E' defects occurs causing partial stretch-out recovery in both accumulation and inversion; the width of the hysteresis loops decreases but the parasitic trap capacitance remains unchanged. On the other hand, only a partial recovery of P_b centers occurs after days/months, causing a decrease of both the AC trap response and of the stretch-out. The curves have been measured again after 59 days and no further P_b center recovery was present. In other words, two typologies of defects presenting strongly different recovery time constants are present in the oxide layer: P_b center defects with long recovery times, E' traps recover in shorter periods. This further indicates the bivalent permanent and recoverable nature of trap recovery of NBTI and PBTI [184]. As evidenced in literature [75, 253], P_b centers can be healed annealing the device at high temperatures. Similar characteristics after negative stress at $I = -10^{-8}A$, indicate a stronger stretch-out recovery in accumulation.

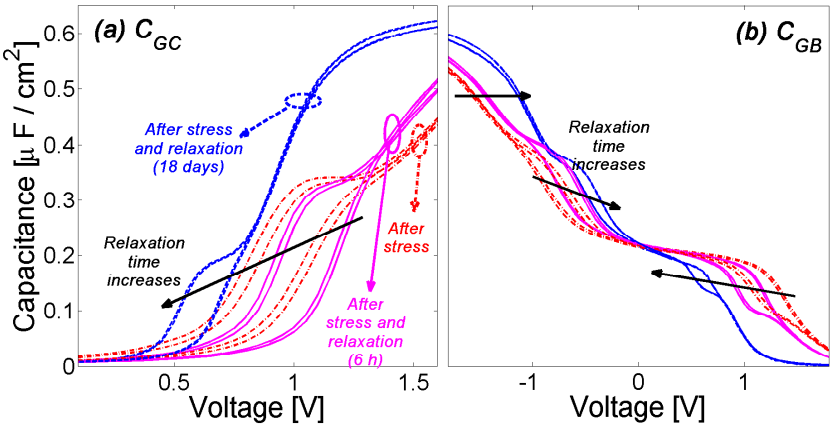


Figure 3-43 – Recovery of traps visible on the C_{GC} and C_{GB} capacitances vs. applied voltage for the NMOS device. The capacitances are measured for frequencies ranging from 1MHz and 50kHz using a sawtooth pulse for the DC voltage to show the recovery of hysteresis effects after defect concentration reduction. Dashed red curves are measured immediately after 1000s of *positive* CCS at $I = 10^{-9}$A. The device is then measured again after a relaxation time of 6 hours (magenta solid curves). The stretch-out of the curve is reduced and the magnitude of the hysteresis loop slightly decreases, while the frequency peaks remain unchanged. Results after a relaxation time of 18 days is also shown. In this case, in addition to the previously mentioned effects, the magnitude of the frequency bump also decreases.

To further highlight the relatively rapid reduction of the effects of E' defects with respect to P_b centers, a continuous measurement of the AC characteristics at 50kHz has

been performed on the same devices after stressing them with positive constant current stress. Each measurement includes a full DC voltage ramp-up and ramp-down from -3V to 3V and lasts approximately 1 minute. In the NFET characteristics showed in Figure 3-44 after stress, and after 1, 61 and 121 measurement cycles, trap recovery is indicated by a reduction of the stretch-out. The amplitude of the hysteresis loops is also progressively reduced. The magnitude of both the conductance and capacitance peaks corresponding to the frequency response are not decreasing, indicating that P_b centers originating them are not relaxing in this time frame. A similar behavior has been found after negative stress.

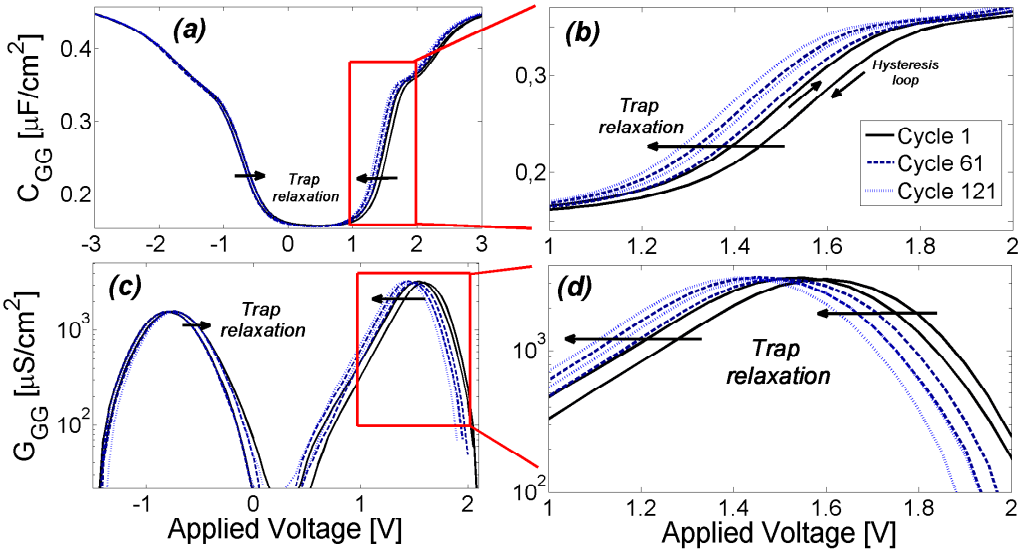


Figure 3-44 – Total gate capacitance C_{GG} (a) and conductance G_{GG} (c) characteristics measured in a continuous way after the application of 1000s of *positive* CCS at $I = 10^{-8}$A on device having $T_{OX} = 50$Å oxide thickness. The inset of subplot (a) shows the DC voltage vs. time applied during the measurement. The characteristics after 1, 61 and 121 cycles illustrate the effects of trap recovery on both the AC characteristics: a zoom in weak inversion is provided in (b) and (d) and clearly indicate the effect of E' recovery on the threshold voltage of the device. Hysteresis loops can be also identified.

The variation of critical electrical parameters has been monitored for both NFET and PFET devices under different conditions of stress. In particular Figure 3-45 presents the defect reduction after positive constant current stress on an NMOS device. Both the threshold voltage V_{th} in (a), the flatband voltage V_{fb} in (b) and the magnitude of the hysteresis loop Δh (c) are shown. For positive stress, the biggest variation is found on V_{th}, while V_{fb} is subject to higher recovery after negative stress. Similar tendencies are evidenced for PFET devices.

Figure 3-46 illustrates how similar recovery phenomena are visible on the electrical parameters of PFET devices after negative constant current stress conditions. Both the flat band voltage V_{fb} (a), the threshold voltage V_{th} (b) and the magnitude of hysteresis (c) are recovered.

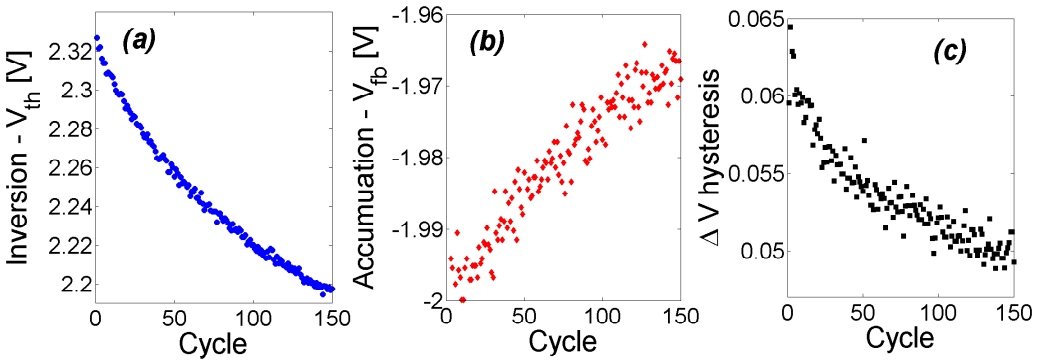


Figure 3-45 – Quantitative variation of electrical parameters as a function of the number of AC measurement cycles (PBTI stress conditions). Each cycle is performed in about 1 minute. As time progresses, the threshold voltage V_{th} in (a) and the flatband voltage in (b), taken in Figure 3-44 when $C_{GG} = 0.4V\,\mu\mathrm{F/cm^2}$, are respectively decreasing and increasing, globally indicating a reduction of the stretch-out effect. The threshold voltage V_{th}, which undergoes a stronger variation during stress, is relaxing more rapidly. In (c) the decrease of the hysteresis loop amplitude taken at $C_{GG} = 0.25V\,\mu\mathrm{F/cm^2}$ is also shown.

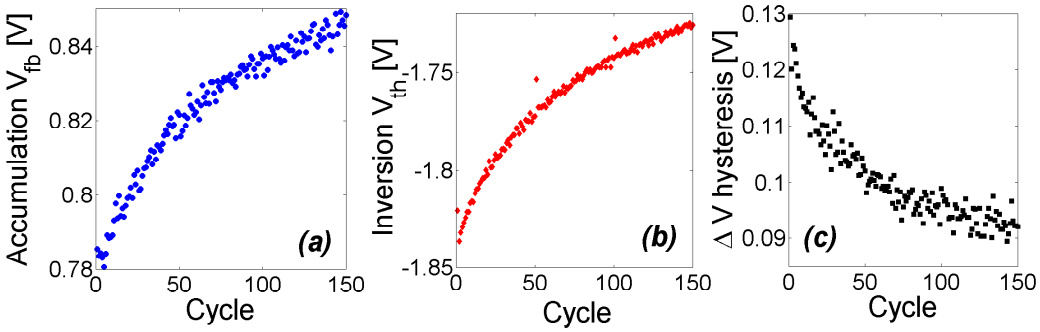


Figure 3-46 – Same extraction as in Figure 3-45 but performed on the PFET device after *negative* CCS at $I = 10^{-9}$A (NBTI stress conditions). In this case the variation seen on Vth, which strongly decreases during stress, is larger than on V_{fb}.

3.6 Comparison of characterization techniques

The probed regions calculated with multi-frequency AC, multi-frequency CP and TAT current analysis, and their dependence on model parameters have been presented in the previous Sections. Figure 3-47 indicates the complete energetic and spatial regions of the dielectric where traps can be probed using the three techniques. The accessible region includes the Si/SiO_2 interface region and the highly energetic traps in the bulk oxide. This information and the adopted modeling methodology permit a more meaningful comparison of the three characterization techniques evaluating their drawbacks and advantages. In [239], we review and compare in detail the three aforementioned characterization meth-

ods.

Multi-frequency CV characterization is adopted to scan interface regions in proximity of the Si conduction and valence bands. The main limitation of this technique only relates to the reduced extension in depth and energy of the scanned region. For decreasing frequencies, deeper defects in energy and space having lower C/E rates are scanned and thus specific regions of the interface can be probed comparing low frequency with high frequency measurements. This comparison is only possible within the range of frequencies provided by the AC analyzer. Indeed, the measurement frequency should be low enough to scan deep defects, and sufficiently high to discriminate the contributions of defects near the Si conduction and valence band edges. For high gate biases (strong inversion), the defects in the close proximity of the Si conduction and valence bands respond very rapidly and thus have cut-off frequencies too high to be distinguished from the contribution of deeper defects. Moreover the analysis of CV results relies in the transient trapping mechanisms that could affect the AC characteristics. These have been investigated in detail in [254] and can be accurately represented by hysteresis loops caused by the slow communication and out-of-equilibrium conditions of deeper defects. Midgap defects cannot be probed due to the reduced mobile carrier density in depletion. From the previous analyses, it is also clear that deeper defects require very reduced measurement frequencies to be characterized. The main advantages of multi frequency CV include the simplicity of the measurement technique, the rapid extraction of the defect profile and the good accuracy of the approach when relying on multiphonon-assisted capture models [75]. Additionally, the probed region and the trap response on the capacitances present a good invariance with respect to model parameters and a strong temperature dependence which further confirms the role that multiphonon C/E transitions have in charge trapping.

TAT current analysis can be adopted to probe the concentration of oxide defects in the middle of the dielectric layer and at higher energetic levels close to the Si conduction band edge. Depending on the DC bias voltage, reverse extraction can be applied in the domain where the TAT current component can be effectively isolated from direct tunneling. For increasing voltage conditions, different regions of the oxide bandgap are scanned, extending from the Si conduction band edge to the SiO_2 conduction band edge. In such a case, band bending plays an important role on the characterized region as the probed zone is strongly affected by the dielectric band structure and stack configuration. Higher voltage conditions also permit to probe deeper defects and reverse extraction at high voltage can be used to evaluate the presence of highly energetic traps. However, the major drawback of reverse extraction based on TAT relates to the strong sensitivity to model parameters (mainly F_C and S in the multiphonon-assisted model). For different categories of defects, (P_b, E'), empirical values have been determined from ab-initio simulations or measured from single defect studies, and can be adopted as a reference. Finally, this technique is very sensitive to oxide breakdown effects and dielectric percolation. Indeed, higher oxide stresses are responsible of the creation of percolation paths where the traps act as stepping stones in the carrier tunneling through the dielectric.

MFCP analysis permits to extract defects placed at recombination sites in the gate-substrate interface and in the middle of the energy bandgap. The depth and the energy

extension strongly depends on the CP pulse, and 1nm-deep defects are probed for realistic pulse parameters. The comparison between the probed CP regions indicates that a similar depth range can be achieved with CV characterization. However, using this method, a complete transient simulation is needed for which large computational resources are required. Considering the importance of transient effects and charge trapping currents, the extracted region is strongly dependent on S and on the CP pulse configuration. For these reasons, the extraction of the trap concentration in the midgap should be initially determined by extrapolation of CV results or quasi-static analysis, and the results verified with CP techniques. Simplified models could also be applied to model the probed region in compact approaches [198], permitting a more rapid estimation of the total trap concentration at the midgap.

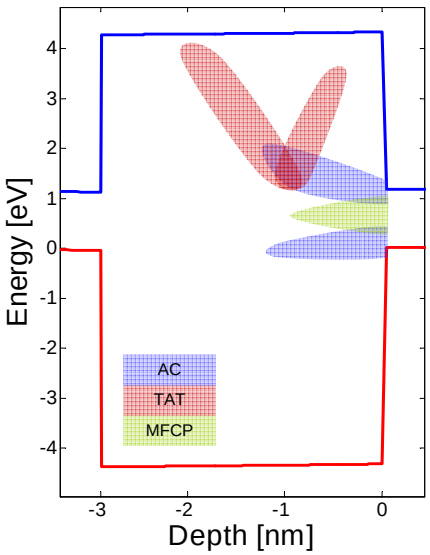


Figure 3-47 – Band diagram of the oxide stack showing the regions that can be probed using the three investigated techniques.

The overlapping of the probed regions is of major concern for the efficacy of extraction. Defects placed in multiple probed regions alter the corresponding electrical characteristics and thus several extraction iterations could be required. Moreover, an extraction based on several characterization techniques permits to improve the level of confidence on the defect profile. The dependence of the probed region with trap filling has not been investigated in detail in this study but, considering the reduced trap concentrations used, second-order effects on the electrostatics calculated by the self-consistent Poisson-Schrödinger solver are expected.

The extraction methodology could similarly be performed on advanced multilayered dielectrics. Figure 3-48 illustrates the calculated probed region by the TAT method for a HighK Metal Gate stack. The substrate/IL interface is located at depth x=0nm and the valence band edge of Si is at 0eV. In aggressively scaled stacks the penetration of the wave function near the interface causes a reduction of the electric field and an increase of the trapping rates. Consequently, the dependence of the probed region on the critical field is not negligible as in thicker SiO_2 dielectrics.

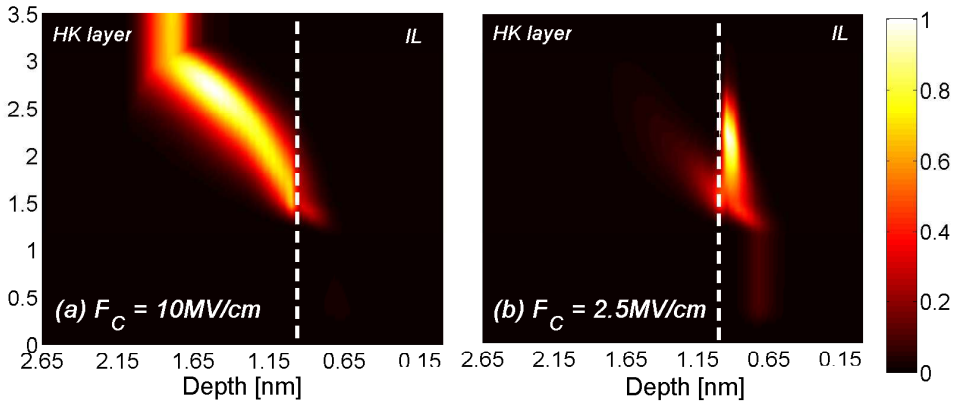


Figure 3-48 – TAT Probed region versus energy and trap depth in a HKMG oxide stack for two values of the critical field. Since the electric field in the structure is high, the critical field parameter affects the oxide probed region. Higher values than the electric field in the interface layer (IL) indicate that the scanned region includes both the layers, while lower critical fields show that the TAT contribution is mainly due to IL traps.

3.7 Conclusion

Multiphonon models are promising approaches for the investigation of charge trapping effects in oxide layers which have so far not been applied to CGV methods for trap extraction. A general model for the calculation of MOS impedances and supporting multiphonon C/E theory has been derived. It includes TAT and direct tunneling effects across the oxide. This model intrinsically accounts for capacitance stretch-out, trap frequency response and temperature dependence. Additionally, the frequency-dependent peaks and the effects of direct tunneling on conductance curves have been reproduced.

The wide spread distribution of C/E times is responsible of the strong frequency-dependence of the capacitances and conductances of stressed devices. Correlation between the C/E constants of the system with the AC response in frequency and the stretch-out of the curves has been found. This result confirms the measured trap C/E constants recently reported in literature [182], where the capture rates range from nanoseconds to months or years.

Multi-frequency AC, multi-frequency CP and TAT current characterization techniques have been investigated by means of multiphonon-assisted charge trapping simulations. We compared: (a) the extension of the accessible regions in energy and position in CMOS dielectric layers, (b) the methodology of defect concentration extraction by reverse modeling and (c) the impact of model parameters on the characteristics (Annex D). Correlating the three extracted profiles provides an improvement in the confidence level of both the extraction methodology and the multiphonon-assisted charge trapping models.

The proposed model can be used for the analysis of the physics underlying the charge trapping mechanism, the extraction of trap concentrations from electrical characteristics

and to evaluate the quality of oxide layers in modern nanoscale technologies. This model has been used as a reference for the understanding of charge trapping mechanisms in electrically-stressed oxides, and in particular for the interpretation of CHEI-induced degradation in flash tunnel oxides [255, 256]. Additionally, the semi-analytical model presented in Chapter 4 will be based on a similar multiphonon approach for the analysis of charge trapping effects on the DC and transient characteristics of flash devices.

Chapter 4

Complementary effects in compact flash modeling

4.1 Introduction

In Chapter 2 we analysed the effects of intra–cell couplings and geometrical dimensions on flash electrostatics in DC and transient regimes. In this Chapter we investigate long-term parasitic effects in a compact design-oriented approach. In particular in Section 4.2, we apply the considerations on charge trapping in MOS devices detailed in Chapter 3 to the analysis of tunnel oxide degradation and flash cell endurance performances. Furthermore, in Section 4.3 a compact model has been developed to reproduce drain disturbs present in NOR configurations during program operation. The model has been validated on measurements for a broad range of bias voltages. Cross-coupling effects are detailed in Section 4.4, which also deals with the extraction of the capacitances between the floating gates of the cells in the matrix environment by means of 3D TCAD simulations. Finally, worst case analysis adopted by circuit designers requires the model to be compatible with process corners and statistical simulations. Therefore, statistical measurements and Monte Carlo simulations have been performed to reproduce the V_{th} distribution of the states and their effects on cell performances.

4.2 Endurance modeling

In this Section, a methodology to model cell aging effects on flash endurance is described. Relying on the consideration and on the rigorous 1D approach developed in Chapter 3, a semi-analytical version of the multiphonon-assisted charge trapping model is presented. Subsequently, the effects of degradation on flash endurance are described focusing on DC and transient performances altered by charge trapping at the Si/T_{OX} interface. A compact semi-empirical solution is finally adopted and integrated in the NVM-SPICE model of Chapter 2 to demonstrate its applicability to memory design. The methodology for endurance investigation is detailed also in [255].

4.2.1 Multiphonon charge trapping in a 2D approach

Using an analytical version of the multiphonon charge trapping approach similar to the one illustrated in Chapter 3, the effects of charge trapping on flash performance have been studied in detail. However, since a Poisson-Schrödinger solution in 2D would be too much computationally expensive, a semi-analytical approach preferred. In particular, the energy levels $\mathcal{E}_j$ of the confined carrier in the substrate are approximated with [257]:

$$\mathcal{E}_j(y) = \frac{1}{q_0} \cdot \left(\frac{\hbar^2}{2m_{eff}}\right)^{\frac{1}{3}} \cdot \left(\frac{3q_0\pi F_{eff}(y)(j-1+\frac{3}{4})}{2}\right)^{\frac{2}{3}}, \tag{4.1}$$

where $j = [1, \infty)$ indicates the index of the level and $F_{eff}(y)$ the vertical field at a given position y in the channel length calculated from the surface potential $\psi_S(y)$ as in Chapter 2. The evanescent wave $\Psi_{ox}(y, z)$ in the oxide layer has a wavevector k_j:

$$k_j(y, z) = \frac{1}{\hbar}\sqrt{2q_0 m_{ox}(\mathcal{E}_c(y, z) - \mathcal{E}_j(y))}, \tag{4.2}$$

where $\mathcal{E}_c(y, z)$ is the potential barrier of the oxide at a point in the dielectric with distance z from the substrate. The barrier is modelled assuming a constant uniform field in the dielectric. The evanescent wave $\Psi_{ox}(y, z)$ is thus expressed by:

$$\Psi_{ox}(y, z) = \Psi_{ox}(y, z, 0)\exp(k_j(y, z)z). \tag{4.3}$$

In the semiconductor, Schrödinger's equation evolves into the Airy differential equation [257–260], whose solutions are the Airy functions [261]. Consequently, the wavefunctions in the semiconductor can be expressed by a linear superposition of Airy functions Ai:

$$\Psi_{Si}(y, z) = CAi(u_{Ai}(y, z)), \tag{4.4}$$

where:

$$u_{Ai}(y, z) = \left(\frac{2m_{eff}}{\hbar^2 F_{eff}(z)^2}\right)^{\frac{1}{3}} q_0(F_{eff}(z)z - \mathcal{E}_j), \tag{4.5}$$

and C is a factor obtained matching the wave function ($\Psi_{Si}(y, 0) = \Psi_{ox}(y, 0)$) at the Si/SiO_2 interface. Having analytically determined the wavefunction and the energy level of the carrier, a similar multiphonon-assisted charge trapping approach of Chapter 3 can be adopted solving the rate equation at each position in the channel. The trap occupancy $f_T(y, z)$, calculated from the capture/emission fluxes, and the total trapped charge $\rho_T(y, z)$ are used to determine macromodel parameters applied to the MOSFET model. In particular, since compact approaches rely on the symmetrization of the charge in the channel or linearization around a point in the channel, it has been chosen to reduce all the trapped charge in the channel to the mid point. Consequently the total trapped charge in the oxide

is calculated integrating:

$$Q_T = \int_{T_{OX}} \int_L \rho_T(y,z) dz dy, \tag{4.6}$$

Recall Eq. 2.1 from Chapter 2. In this semi-analytical approach, the surface potential equation is modified taking into account the total trapped charge at each iteration:

$$\begin{aligned}(V_{FB} - V_{fb} - \psi_S(z))^2 &= \left(\frac{Q_T}{C_{OX}}\right)^2 + \gamma^2 \psi_T \left(e^{-\frac{\psi_S(z)}{\psi_T}} + \right. \\ &\left. + \frac{\psi_S(z)}{\psi_T} - 1 + e^{-\frac{2\psi_F}{\psi_T}} \left(e^{\frac{\psi_S(z) - V_C(z)}{\psi_T}} - \frac{\psi_S(z)}{\psi_T} - 1 \right) \right). \end{aligned} \tag{4.7}$$

A mobility reduction term is also added in Matthiessen's rule to take into account mobility degradation due to surface scattering at the interface:

$$\begin{aligned}\frac{1}{\mu_{eff}} &= \frac{1}{\mu_{tr}} + \frac{1}{\mu_{cs}} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}} \\ \mu_{tr} &= k_T \left(\frac{Q_T^2}{(Q_B + Q_I)^2} \right) \end{aligned} \tag{4.8}$$

where k_T is an empirical parameter. The progressive filling of traps as the gate voltage increases causes a degradation of the electrostatic through ψ_S and of the carrier mobility (i.e. gate transconductance) in the channel. The model has been used for the interpretation of endurance characteristics in flash devices.

4.2.2 Degradation effects in flash cells

We can now analyse the effects of degradation on the flash cells and the phenomena determining endurance performances. When cycling the device, the tunnel oxide is subject to FN electrical stress during erase and CHEI stress during program. Figure 4-1(a) shows the influence of electrical stress during cycling, inducing a modification of the V_{th} window, $W = V_{th}^P - V_{th}^E$. Two cycling sequences have been adopted for programming/erasing the cell: FN/FN and CHEI/FN operation. In the former case, the cell is both programmed and erased by FN operation (program: pulse width 10ms; $V_{CB} = 18.9V$ - erase: pulse width 10ms - $V_{CB} = -17.65V$). The V_{th} is measured after a given amount of cycles. Two effects can be identified in this configuration: (a) both the V_{th}^E and V_{th}^P increase, (b) the increase of V_{th}^P is less pronounced than V_{th}^E (thus W decreases). In the latter case, the endurance characterization has been performed by cycling the cell with CHEI for programming and FN for erasing (program: pulse width 4μs; $V_{CB} = 8.5V$, $V_{CB} = 4.2V$ - erase: pulse width 1ms - $V_{CB} = -17.65V$). Similarly to the previous case, two phenomena are identified: (a) V_{th}^E increases due to the progressive filling of interface traps delaying inversion; (b) V_{th}^P initially decreases, inducing the closure of the window after moderate cycling, and then increases. The threshold voltage window is reduced after a considerable number of cycles in both the cases, compromising device endurance.

Hereafter, it will be shown that all the previously mentioned effects are attributed to the influence of defects on flash characteristics. A physical interpretation of these effects is of critical importance to improve endurance performances during technology development. This permits to reconstruct the window evolution from DC and transient characteristics measured after a given amount of cycles, where the effects of the defects are correctly separated and taken into account (Figure 4-1(b)). In this methodology, trapping effects result in different threshold voltage variation dynamics ΔV_{th} versus the number of cycles N_{cycles}, for the two states. The variation ΔV_{th} can be separated in three parts:

- ΔV_{th}^{R}: threshold voltage variation directly related to charge trapping effects on the active MOSFET part of the device (subthreshold slope degradation, fixed charges trapped in the T_{OX}, g_m reduction);
- $\Delta V_{th}^{Eeff}(N_{cycles})$: threshold voltage variation component after N_{cycles}, attributed to the reduction of erase efficiency induced by cell ageing;
- $\Delta V_{th}^{Peff}(N_{cycles})$: threshold voltage variation component after N_{cycles} due to the reduction of program efficiency after cycling.

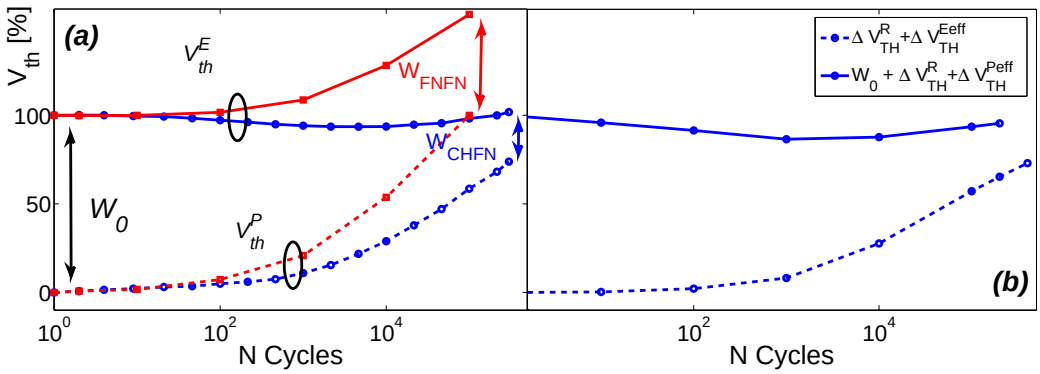


Figure 4-1 – Threshold voltage window $W = V_{th}^{P} - V_{th}^{E}$ versus number of P/E cycles. In (a), direct V_{th} measurements after cycling are shown, using FN/FN and CHEI/FN operations for programming and erasing the cell, respectively. In (b), window extraction for CHEI/FN regime using the proposed extraction methodology to decouple transient and DC degradation effects.

(i) DC effects on the active device

Border and interface traps are responsible of the variation of the device electrostatics when the cell is electrically stressed during cycling [151]. Figure 4-2(a) shows measured $I_{DS}(V_{CS}, N_{cycles})$ characteristics performed on flash devices during cycling. As the number of P/E cycles increases, the subthreshold slope is degraded and the gate transconductance is reduced. Similar effects are visible in (b) for equivalent transistor devices, after having undergone electrical stress. For these devices, the floating gate voltage during stress

has been calculated from considerations on the cell coupling coefficients. It should be pointed out that, while the V_{th} of the equivalent transistor (taken at an arbitrary current I_{DScrit}) increases, the V_{th} of the two states of the cell behave differently (V_{th}^E increases, V_{th}^P decreases).

The charge trapping model in Section 4.2.1 has been adopted to identify the reason of this discrepancy. In Figure 4-2(c), the read characteristics of the flash device have been simulated. Dashed curves indicate fresh devices, while solid curves indicate the results when a Gaussian distribution of traps is added in proximity of the T_{OX}/substrate interface. Both sub-threshold slope, transconductance and V_{th} are affected. In the simulations the programming/erase efficiencies are not altered as the charge on the floating gate Q_{F0} is directly varied to restore the DC state of the cell.

Consequently one can estimate the effects of defects in DC read conditions with the V_{th} voltage variation ΔV_{th}^R, in any of the two states. This quantity can be extracted also from erase transient characteristics as shown in Figure 4-3(a) and with:

$$\Delta V_{th}^R(N_{cycles}, t_R) = V_{th}(N_{cycles}, t_R) - V_{th}(0, t_R) \tag{4.9}$$

where $t_R = 0.2$ms is a time duration chosen long enough such that the initial state of the device does not affect the dynamics. The apparent discrepancy between (a) and (b) of Figure 4-2 is explained considering the efficiency of the injection mechanism.

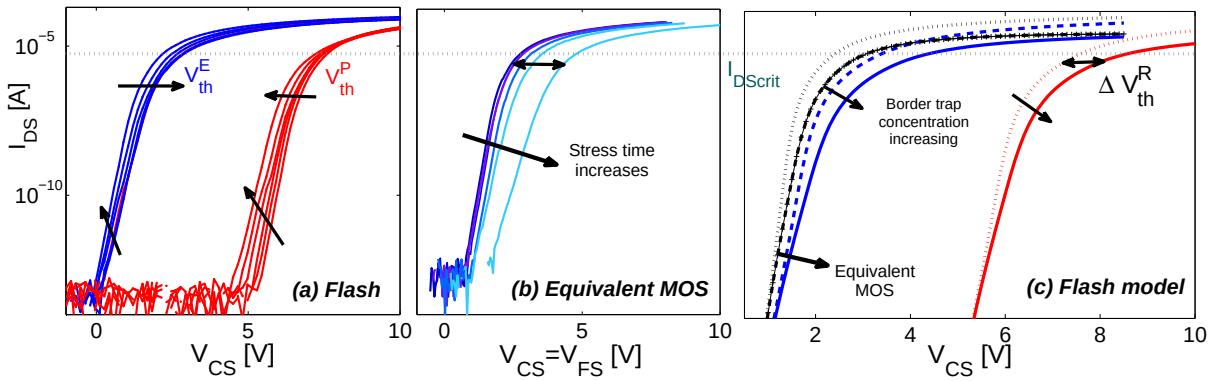


Figure 4-2 – (a) Measured I_{DS} versus V_{CS} characteristics of a flash cell in erased (blue) and programmed (red) states during 1000 P/E cycling by CHEI/FN. As the cell is cycled and stressed, the subthreshold slope increases and the g_m decreases. However, the two V_{th} behave differently: the erased V_{th} increases of ΔV_{th}^R, while the programmed one is slightly reduced. (b) On an equivalent-transistor cell that experiences electrical stress conditions similar to the device in (a), qualitatively similar effects are visible. In (c), simulation results indicate the impact of defects on the electrostatics without considering program efficiency reduction (dashed curves: fresh devices, solid curves: with a Gaussian distribution of traps in proximity of the T_{OX}/substrate interface).

(ii) Program/Erase efficiencies

The influence of degradation on erase mechanisms is analyzed in Figure 4-3(a). Due to trap filling in inversion, ΔV_{th}^{R} affects the V_{th} versus erase time measurements. This contribution can be removed by vertically shifting all the curves (inset of Figure 4-3(b)). The erase efficiency degradation, corresponding to the threshold voltage variation ΔV_{th}^{Eeff}, can be identified as the variation of the slope of the curve in the latter plot. A quantifiable estimation of this degradation is given by:

$$\Delta V_{th}^{Eeff}(N_{cycles}) = V_{th}(t_E, N_{cycles}) - \Delta V_{th}^{R}(N_{cycles}) - V_{th}(t_E, 0). \tag{4.10}$$

This represents the portion of V_{th} shift that cannot be restored when erasing the aged cell for a given erase time t_E =1ms. From this result, one can thus conclude that the erase efficiency degradation is marginal with respect to the effect of trap filling ΔV_{th}^{R} on the electrostatics. Indeed, the apparent erase performance degradation, i.e. the V_{th}^{E} increase in Figure 4-1(a), mostly corresponds to a change of the device electrostatics due to traps filling, i.e. ΔV_{th}^{R} on Figure 4-3(b).

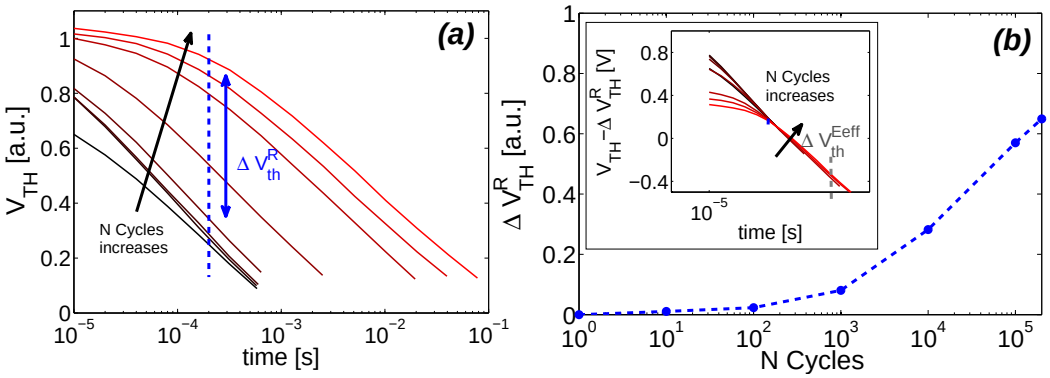


Figure 4-3 – Extraction methodology to separate ΔV_{th}^{R} and ΔV_{th}^{Eeff} with DC and transient erase measurements. (a) Raw transient measurements performed after a different number of cycles. Due to the contribution of traps filling, the initial state of the cell is not the same and thus measurements need to be aligned by subtracting $\Delta V_{th}^{R}(N_{cycles})$ (inset of (b)). $\Delta V_{th}^{R}(N_{cycles})$ can also be extracted from DC measurements in Figure 4-2(b). The inset shows how erase efficiency is minimally affected by degradation (identical slope of the curves). (b) ΔV_{th}^{R} versus N_{cycles} following the trend of the window baseline.

A similar extraction procedure has been performed in program conditions (Figure 4-4). In this case, the cell has been previously over-erased to exhibit the full transient dynamics. Figure 4-4(a) shows raw transient measurements before data processing; in the inset of (b), the ΔV_{th}^{R} contribution is removed so that the initial charge on the floating gate be the same for all the stress/cycling conditions. In the inset of (a), $V_{th} - \Delta V_{th}^{R}$ is plotted as a function of the effective program time, measured starting from the instant where the threshold voltage is V_{th}^{E}. The program efficiency degradation can be identified in the

latter plot as the threshold voltage variation ΔV_{th}^{Peff} that cannot be restored after a given arbitrary program time $t_P = 4\mu$s:

$$\Delta V_{th}^{Peff}(N_{cycles}) = V_{th}(t_P, N_{cycles}) - \Delta V_{th}^{R}(N_{cycles}) - V_{th}(t_P, 0) \tag{4.11}$$

The program efficiency is thus sensibly affected by device degradation after stress and plays an important role on the window dynamics.

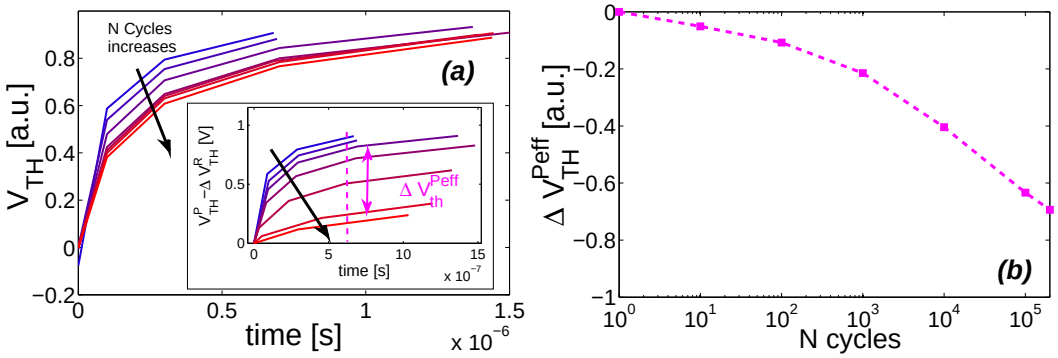


Figure 4-4 – Extraction methodology to determine ΔV_{th}^{Peff} from program transient measurements of V_{th} versus programming time. (a) Raw transient program measurements performed after a different number of cycles. In this case normalization has to be performed removing the contribution ΔV_{th}^{R} of filled traps (inset). (b) ΔV_{th}^{Peff} versus N_{cycles} representing the degradation of program efficiency.

The presence of trapped charges delays inversion and reduces the tunneling current, causing program efficiency degradation in both forward FN and CHEI regimes. As the CHEI current is also dependent on the channel current I_{DS}, the CHEI efficiency is more sensitive to degradation. From MPA simulations of gate leakage, it can be seen that erase efficiency is marginally reduced after electrical stress. This has been attributed in [262] to stacked oxide charges, that globally shift the leakage characteristics to higher positive voltages [255].

The semi-analytical model has been used to confirm experimental results of P/E performance degradation. Figure 4-5 shows the effects of the total trap concentration in the oxide for FN and CHEI program transient operations, respectively. Program efficiency is strongly reduced as interface trap concentration increases and g_m decreases.

(iii) Process optimization and correlation with equivalent–transistor devices

The proposed extraction method can be applied to the physical understanding of endurance characteristics. In industrial technology development, several process splits are processed to align technological parameters on final specifications. The main issue for the understanding of endurance relies on the large number of effects involved that challenges the separation of the different contributions and compromises the identification of the regions to be optimized. Three-dimensional TCAD simulations have been performed to study

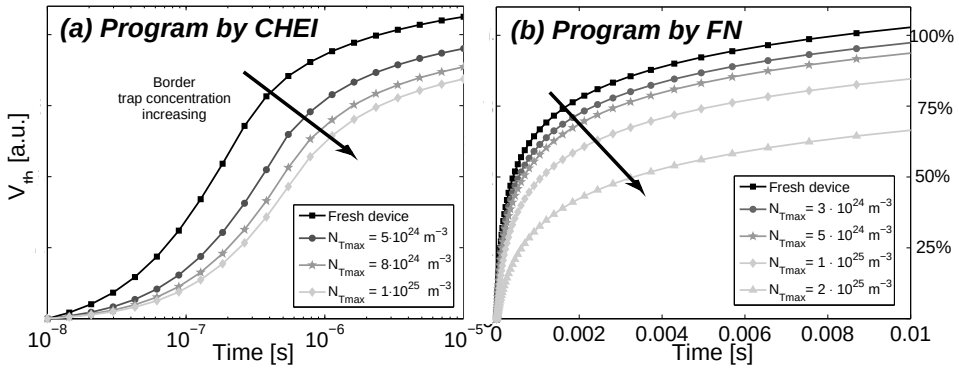


Figure 4-5 – Transient simulations showing the effects of device degradation on CHEI (a) and FN (b) program efficiencies for increasing trap concentrations. In both cases, uniform Gaussian trap distributions along the channel are placed at the Si/SiO_2 interface and the peak N_{Tmax} of the concentration has been increased. In (a), the gain is reduced due to the presence of interface traps decreasing the channel current, the carrier distribution and consequently the injection current. In (b), trap filling causes inversion to be delayed, reducing the tunneling current.

degradation phenomena and identify the defect localization. e^- tunneling in FN regime is concentrated in the floating gate *divot* region (Figure 4-6(a)). On the other hand, CHE injection mainly occurs close to the drain/channel junction (Figure 4-6(b)).

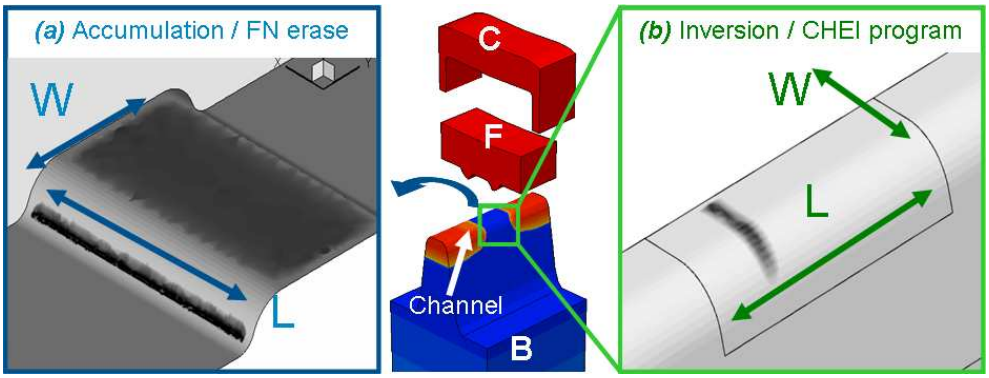


Figure 4-6 – 3D TCAD simulations to identify the points subject to high electrical stress during in accumulation/erase (a) and inversion/program (b). In the former case the current flows through the floating gate divot region. In the latter case, the current is concentrated in the LDD region where electrical field and injection are higher. Darker areas identify regions with higher current densities.

Process development requires identifying which critical fabrication parameter can play a role on the areas to be optimized [263]. Figure 4-7 shows the correlation between the V_{th} of the equivalent transistor device (V_{th}^{TREQ}) after 100s stress in both FN and CHEI and the cell window after 200k cycles. When altering the LDD profile, the characteristics of the injection region change and thus the V_{th}^{TREQ} (blue curve) variation under CHEI stress

conditions is affected as well. The variation of the tendency is in good alignment with the variation measured on the window W. On the other hand, modifying the process deposition of the tunnel oxide layer has a higher influence on the V_{th}^{TREQ} measured after FN stress. This modification is reflected on the flash window as well.

Figures 4-8(a-b) show the large spread in W, obtained grouping the results of process variations where the LDD implant characteristics are varied (P1,P2,P3), while Figures 4-8(c-d) present similar results for devices where the oxide formation process is varied (P4,P5,P6). The window W variations and the V_{th}^{E} baseline trends for both the subsets are presented in Figure 4-8(b-d). Both show the direct correlation existing between the baseline and the window.

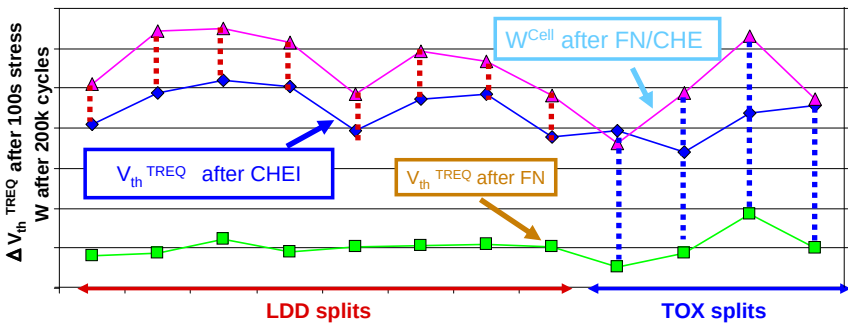


Figure 4-7 – Correlation between the threshold voltage V_{th} of stressed equivalent transistor devices (V_{th}^{TREQ}) and the flash cell window W after 200k cycles. Equivalent transistors have been stressed using both FN and CHEI stress. These results show how the oxide process variations influence V_{th}^{TREQ} after FN stress, while LDD process variations affect V_{th}^{TREQ} after CHEI stress. In both cases, the variation on V_{th}^{TREQ} is reflected on the window of the stressed flash device.

4.2.3 Compact modeling of flash endurance

Compact approaches for modeling flash endurance require two aspects to be taken into account: (a) the physical modeling of the effects of the defects on the device characteristics, (b) the physical degradation model which generates the defects. Numerical approaches have been proposed for point (a), although compact models need to be efficient and fast enough to permit designers to have optimized results in short simulation times. Consequently, a compact empirical method has been preferred for this case, to preserve speed and convergence of the charge balance algorithm, with the defects having three effects: a rigid variation of the flat band voltage $\Delta V_{fb} = -\frac{Q_T}{C_{OX}}$, where Q_T and trap filling are modulated by the quasi Fermi level in the substrate, an empirical variation of the sub-threshold slope parameters of the MOSFET model, and a reduction of the mobility using a factor as in Eq. 4.9. A single parameter N_{IT} has been used to model the trap concentration at the oxide interface.

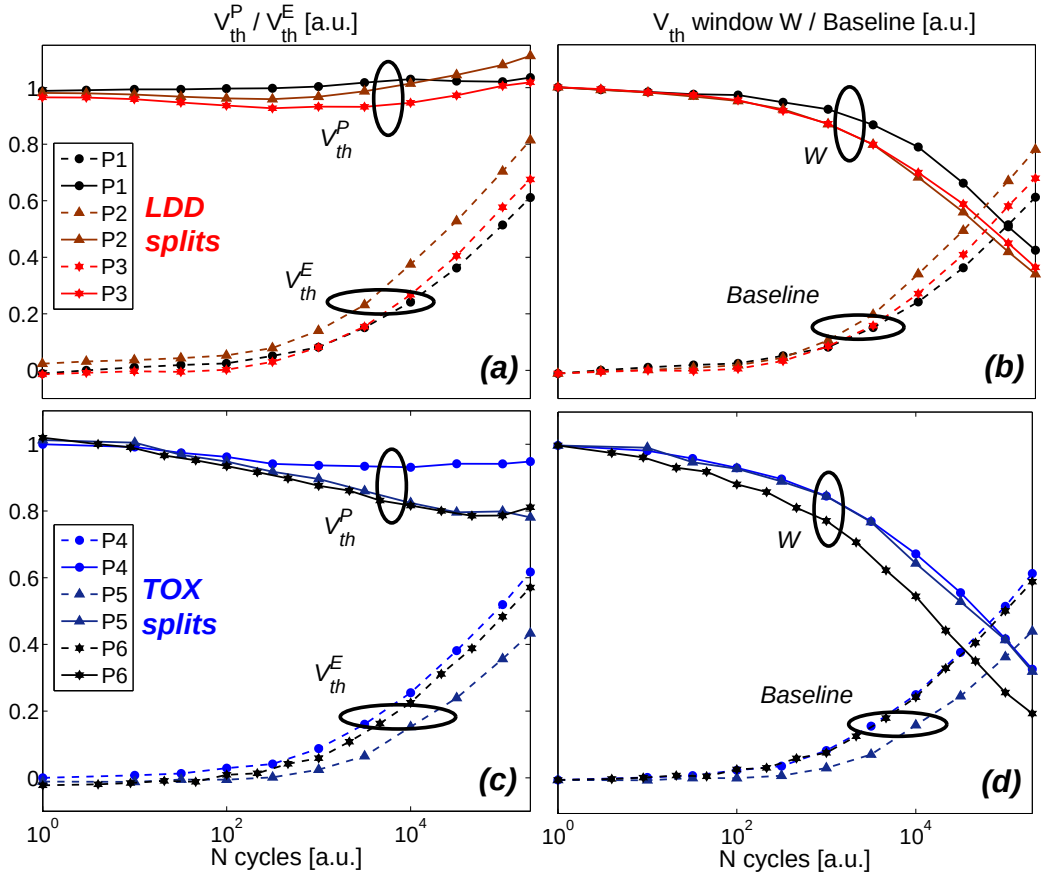


Figure 4-8 – Endurance characteristics for different process splits on LDD doping (a-b) and oxide growth (c-d). In (b-d) the window reduction is perfectly correlated with the baseline of the V_{th}^{E} in case of process variations on the T_{OX} growth.

A model for the variation of the defect density as a function of the applied bias or currents in the device is required. In this study, it has been chosen to model the degradation using:

$$N_{IT}(t) = \int_0^t \left(A_{dg} |I_{FD}(t)|^{C_{dg}} + B_{dg} |I_{FB}(t)|^{D_{dg}} \right) dt', \tag{4.12}$$

where the empirical parameters A_{dg}, B_{dg}, C_{dg} and D_{dg} are extracted from DC measurements after degradation. These parameters could be extracted from empirical analysis using an approach similar to [264] based on subthreshold slope and transconductance variations vs. stress voltages and durations. The parameters A_{dg} and B_{dg} could be calibrated evaluating the dependence of the equivalent transistor parameters with CHE stress, while C_{dg} and D_{dg} are extracted from Fowler-Nordheim stress.

More accurate approaches could be applied to take into account different energy-driven degradation modes as in [264–266] where aging coefficients can be extracted for various

drain and floating gate bias coefficients. This solution requires a large set of degradation measurements on both flash devices and equivalent transistor cells to calibrate the bias and stress time dependence of degradation. With the purpose of illustrating the applicability of the compact model to endurance analysis, the model of Eq. 4.12 has been preferred for this study. Furthermore, an empirical exponential factor dependent on the total trapped charge has been introduced in Eq. 2.46 to model the reduction of program CHEI efficiency. Conversely, given the considerations in the previous Section, it has been assumed that the erase efficiency is not affected by the increase of border defects at equivalent electric field.

Figure 4-9 illustrates the methodology to simulate endurance degradation in flash devices using the NVM-SPICE compact model. Each cycle is represented by four consecutive SPICE simulations, where the calculated FG charge Q_{F0} is applied as instance parameter. After erase and program simulations, two read DC simulations are required to assess the V_{th} of the cell with sub–threshold slope effects. After a complete cycle, the tunnelling and injection currents are used in the ageing model of Eq. 4.12 that computes a new value of effective trap concentration for the following cycle.

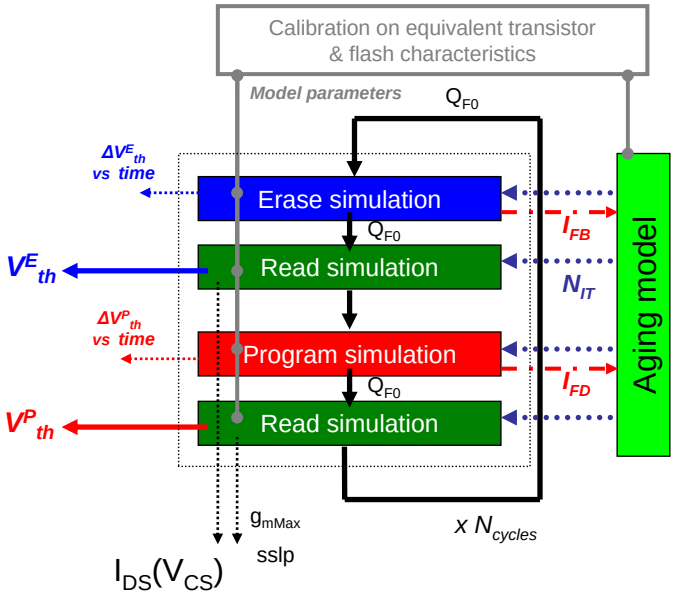


Figure 4-9 – Modeling methodology adopted to take into account ageing effects on flash cell characteristics.

Endurance measurements have been performed cycling the device up to 10^5 times with nominal P/E conditions. After a given amount of cycles, the $I_{DS}(V_{CS})$ characteristic is measured in read conditions and the V_{th}, maximum gate transconductance g_{mMax} and subthreshold slope are extracted for both the states. Since the simulation of the entire cycling sequence requires a long simulation time, it has been chosen to apply a conversion of the number of cycles N_{cycles} to an effective number of iterations $N_{effcyc} = N_{cycles}^{2/5}$. In Figure 4-10(a), the normalized V_{th} window is shown as a function of the number of simulated effective cycles N_{effcyc}. Measurement results are in symbols. The parameter extraction methodology relies on decoupling the effects as indicated in the previous section. In particular, the extraction should consider both DC and transient effects, where

the degradation of the gate transconductance and of the sub–threshold slope shown in (b) and (c), respectively, should be taken into account. The simulated relative variation of the P/E efficiencies is shown in the inset of Figure 4-11 and indicates the significant reduction of CHE injection.

This extraction methodology can be applied to different types of pulses, and static models of N_{IT} vs N_{cycles} can be elaborated for several operating conditions (Figure 4-11). This permits the designer to directly set the worst condition state of the aged cell and analyse the implications on the product specifications without cycling the device 10^5 times. The main limitations of this approach rely on the fact that a set of N_{IT} vs N_{cycles} models need to be extracted, corresponding to each pulse condition under investigation. Therefore this solution cannot offer the flexibility of the model in Eq. 4.12.

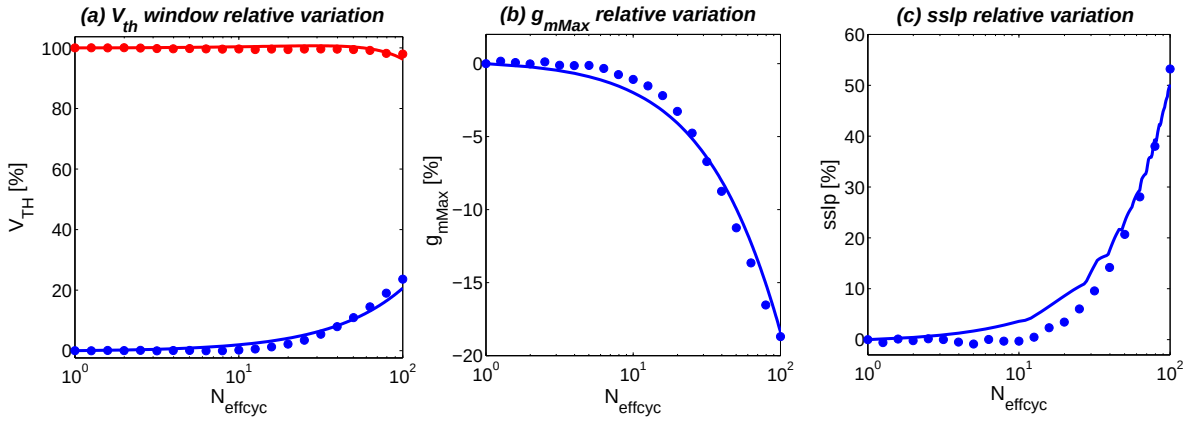


Figure 4-10 – Relative variation of the threshold voltage window (a), maximum gate transconductance g_{mMax} (b), and sub–threshold slope (c) as a function of the effective number of cycles. Measurements are indicated in symbols while simulations are presented in lines. The effective number of cycles is obtained using $N_{effcyc} = (N_{cycles})^{2/5}$ to avoid excessively long SPICE simulations.

4.3 Disturb mechanisms

In aggressive technologies, disturb dynamics monitor the V_{th} variation in a given state, during program or read operation of adjacent cells. In these conditions, high voltage biases are applied on the terminals of the flash device for long time durations. These effects should be carefully considered in the tradeoff balance.

Disturb effects can be divided in bit line (or drain) disturbs and word line (or gate) disturbs, depending on which node the high voltage condition in applied. Dynamics also strongly depend on the state of the disturbed cell.

Let us consider a single bit line of flash devices organized in NOR configuration. All the devices have been programmed and share the same bit line. When a given cell is cycled, high negative voltages are applied on the drain of all the cells during the programming phase

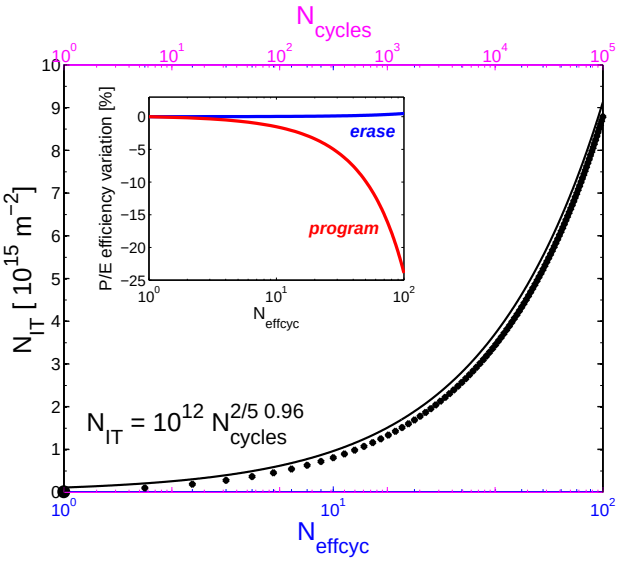


Figure 4-11 – Increase of the simulated concentration of interface defects N_{IT} at the Si/SiO_2 surface as a function of the number of effective cycles N_{effcyc}. The conversion to the measured number of cycles is also provided in magenta. (Inset) Simulated variation of the program and erase efficiencies versus N_{effcyc}.

and, after a high number of cycles, the threshold voltage of the programmed aggressed cells tends to decrease. This stress condition can be reproduced applying a high voltage DC bias on the BL of a programmed device and assessing the V_{th} decrease.

Even though the injection efficacy is considerably smaller than for CHEI, the total cumulative disturb stress time can induce large V_{th} variations. Disturb dynamics have been characterized using progressive algorithms (see Figure 2-28 in Chapter 2), dividing the disturb stress time in small pulses, after which the V_{th} of the device is measured. All the measurements have been performed on three dies to illustrate statistical variability. In a first case, the cell is initially programmed at V_{th} ≈7.7V and subsequently stressed by only applying a high bias voltage on the drain terminal of the device. The WL of the stressed cell remains grounded. Figure 4-12 reports these BL disturb measurements performed on programmed devices, showing the V_{th}^{P} decrease as a function of the cumulative disturb time for several V_{DS} and V_{BS} stress conditions. For voltages lower than V_{DS}=3.4V disturbs have not been characterized as they would have required too long measurements and programming at such small V_{DS} voltages is rarely applied. In Figure 4-12(b), the dependence on the bulk voltage is also illustrated.

Since a negative voltage is present on the FG of the cell when the device is in the programmed state, disturb effects are suspected to be caused by band-to-band tunneling and electron-hole generation in the space charge region of the drain-bulk junction [11]. Due to the high drain-bulk voltage, holes can be accelerated towards the isolated substrate of the device (hot-holes). If the energy of the carriers is high enough, ionization can occur generating highly energetic tertiary e^-/h^+ couples. While e^- are collected at the drain, the energy of h^+ can be high enough to permit the carriers to surmount the energy barrier of the tunnel oxide [11]. The ionization process can also occur multiple times, generating additional carrier pairs. It has been shown that tertiary holes injection is greatly respon-

sible of threshold voltage decrease [11]. These approaches require high computationally demanding numerical simulations and usually rely on full band MC analysis to analyse scattering and carrier distribution in the LDD.

BL disturb has been modeled adding a current contribution from the drain to the floating gate of the device [105]. An empirical expression has been proposed to take into account the dependence on the stress time duration and on the field across the tunnel oxide, using:

$$I_{dist} = A_{dist} I_{GIDL} \exp\left(\frac{-F_{FD}}{B_{dist}} + C_{dist} V_{DB}\right), \tag{4.13}$$

where the dependence on the floating gate voltage is taken into account in both the gate induced drain leakage (GIDL) current I_{GIDL} and the exponential factor on the field F_{FD} across the T_{OX} in proximity of the LDD junction. The current contribution I_{GIDL} has been calibrated on measurements performed on equivalent transistor test structures and is modeled with the usual band-to-band current approximation in [267]. The dependence on the floating gate voltage is obtained evaluating the CBE at each timestep. The model parameters A_{dist}, B_{dist} and C_{dist} have been adjusted to reproduce BL disturb measurements at the different drain and bulk bias voltages.

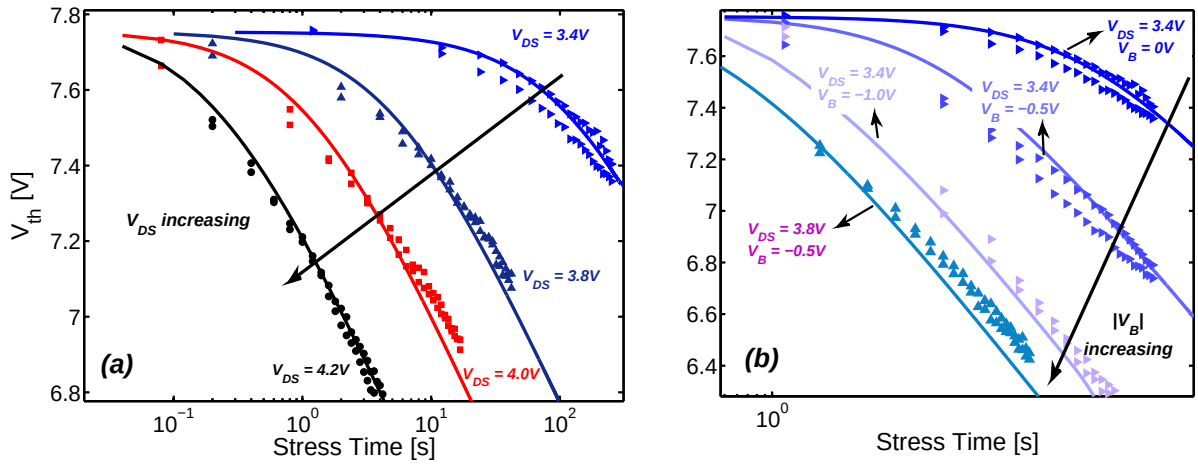


Figure 4-12 – (a) Simulated (lines) and measured (symbols) threshold voltage variation versus disturb stress time for a programmed cell at different drain bias voltages at V_B=0V. The bulk bias dependence is also shown in (b) for two V_{DS}=3.4V and 3.8V and for V_B ranging from -1V to 0V.

The application of a V_{BS} bias allows to greatly improve program efficacy, but it has a negative impact on BL disturb as it enhances the hot-hole injection phenomenon.

Similarly, when the cell is in an erased state, the floating gate potential of the device is high and consequently the hole population is not present in the LDD region. Thus, HHI cannot occur. On the other hand, although CHEI is very reduced due to the low V_{CS} bias,

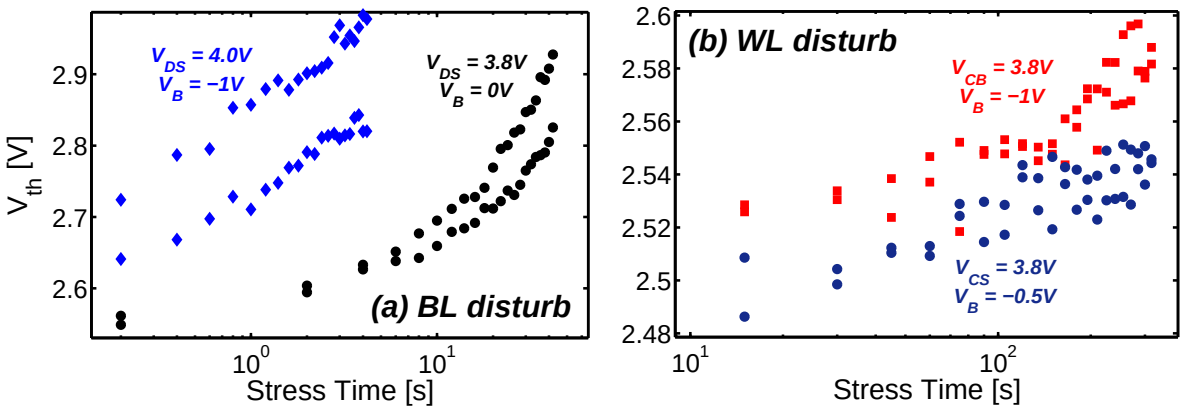


Figure 4-13 – (a) Threshold voltage variation versus disturb time for an erased cell measured for different drain and bulk bias voltages under BL disturb conditions. A variation of less than 0.5V after 4 seconds is reported, much smaller than the V_{th} shift on programmed devices. (b) Threshold voltage variation versus disturb time for an erased cell measured for different WL voltage biases under WL disturb conditions. Carrier tunneling is very reduced resulting in a V_{th} variation of less than 100mV after 300s.

a portion of e^- can be injected from the channel and LDD to the floating gate with an undesirable soft programming effect that raises the V_{th} of the device. The V_{th} variation reported in Figure 4-13(a) appears very reduced with respect to the one previously shown for the programmed device.

WL disturb measurements have been performed as well on the erased cell and are illustrated in Figure 4-13(b). Under the programming phase, cells that share with the target device the same WL are subject to a high voltage condition on the control gate terminal. Consequently, WL disturb can occur due to undesired e^- tunneling from the channel to the floating gate terminal. Reduced variations are observed due to the very low leakage of the tunnel oxide. WL disturbs are not present when the device is programmed as the stored charge decreases the floating gate voltage, reduces the field in the oxide layer and prevents e^- tunneling.

4.4 Cross-couplings

As the dimension of the memory matrix decreases, the influence of neighbouring cells on the electrostatics of a single device raises. To this purpose, cross-coupling effects have been analyzed by means of 3D TCAD simulations and measurements on miniarrays in NOR configuration.

Starting from the 3D structure presented in Chapter 2, the back-end stack of interconnections has been added using the physical dimensions in the layout of the matrix. In a NOR configuration, due to the asymmetry of the back-end stack, the two schemes illustrated in Figure 4-14 should be considered to evaluate all the capacitive contributions

between the floating gates in the array. The simulated 3D structures include 4 flash memory devices and are reported in Figure 4-15, with 2D cross-sections along the channel length and width. The values of the capacitive components are reported in Table 4.1 for both the configurations. A large coupling, higher than the capacitance C_{FD} between the floating gate and the drain of a cell, is observed between the bit lines. The capacitances between the floating gates of the cells remain considerably smaller due to the shielding of the source metal lines. The two configurations present similar values of the capacitance which implies that the back end stack asymmetry is only marginally influencing the capacitance values. Furthermore, the analysis of the complete back-end stack permits to extract the contributions of the parasitic components and estimate the load of BLs and WLs in circuit simulations for IC design (Section 5.3.1(i)).

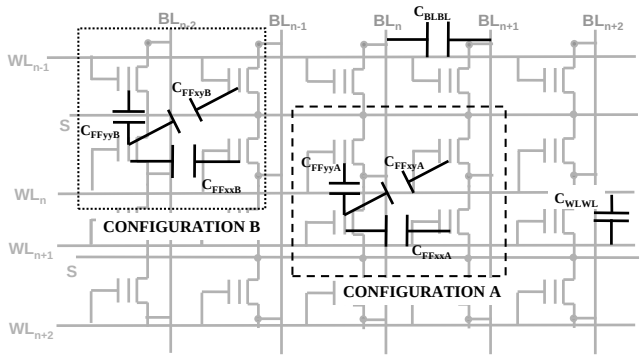
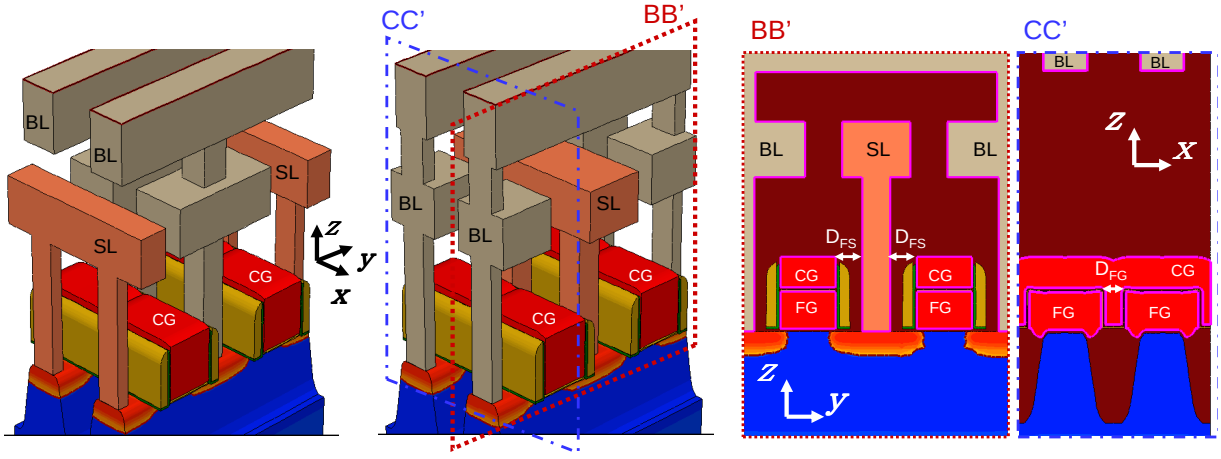


Figure 4-14 – Schematic of the matrix in NOR configuration illustrating the cross-coupling capacitances. In configuration A, a single source line is simulated while in B, the interconnections are different due to the presence of a common drain between two cells. The couplings between WLs or BLs pairs are extracted as well.



(a) Configuration A (b) Configuration B (c) 2D cross-sections

Figure 4-15 – (a-b) Three dimensional TCAD simulations performed on the configurations indicated in Figure 4-14 memory. The geometry and architecture of the back-end structure is aligned with the layout of the matrix. In (c), two cross-sections of Configuration B are shown.

Component	Configuration A [aF/cell]	Configuration B [aF/cell]
C_{FFxx}	1.752	1.772
C_{FFxy}	0.049	0.049
C_{FFyy}	0.177	0.173
C_{WLWL}	1.010	1.008
C_{BLBL}	46.67	46.67
C_{FD}	37.8	37.5

Table 4.1 – Values of the extracted cross-coupling capacitances for the two configurations indicated Figure 4-15 normalized on the cell dimension.

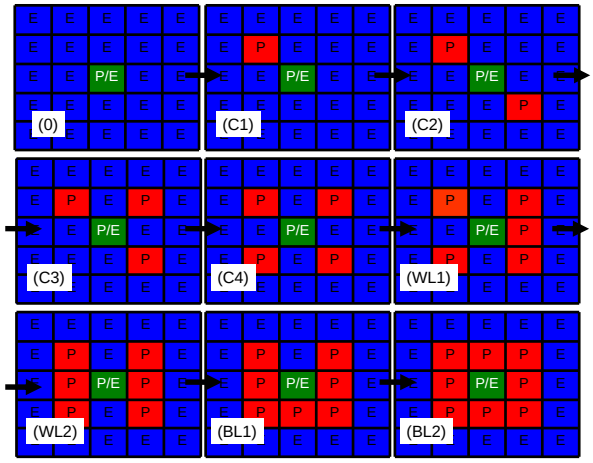


Figure 4-16 – Scheme for the characterization of cross-coupling effects in NOR configuration. The target cell in green is placed in the middle of a mini-array and set to a specific programmed/erased state. Surrounding devices are progressively programmed.

Indirect measurements of cross-coupling effects in a memory matrix have been performed on the fully addressable mini-arrays where the effects of the 8 aggressor devices surrounding a target cell can be studied. The influence of farther cells on the target device has been neglected. The entire matrix is initially erased. Subsequently, the 8 cells surrounding a target device are progressively programmed starting from the ones in the corners (C1-4), followed by the ones sharing the same WL (WL1-2) and the same BL (BL1-2) (Figure 4-16). After each programming step, the threshold voltage of the target cell is measured. Two types of effects have been studied depending on the state of the target device.

The target cell is initially erased and its floating gate voltage remains high during the entire characterization. When programming an adjacent cell (i, j), its floating gate voltage $V_{FS_{i,j}}$ decreases and the couplings in the device are negatively affected. The decrease of the gate coupling α_C increases the threshold voltage of the cell and reduces P/E performances. Threshold voltage variations of the order of 140mV have been found. Three dies have been characterized: on each of them, three configurations of mini-arrays are investigated, increasing the distance DX between two floating gates in the bit-line direction from 80nm

to 130nm. As the distance separating the BLs and WLs gets smaller, the cross-coupling capacitances $C_{FF_{i,j}}$ linearly increase. The distance between the drain/source contacts and the floating gate of the cells decreases as well, resulting in higher parasitic drain/source coupling. The increase of the D/S coupling has effects on both the aggressor and the target devices. On the target device, V_{th} slightly raises as a consequence of the higher $C_{FF_{i,j}}$ coupling, but the programming dynamics are exponentially affected by the α_D coupling degradation. Consequently, in the charge balance equation the reduction of $(V_{FS} - V_{FS_{i,j}})$ appears more pronounced than the increase of $C_{FF_{i,j}}$ couplings. This implies that cross-coupling effects are slightly lower in arrays where the distance between two floating gates is higher.

In the programmed state, minimal effects are found. The measured V_{th} variation of the target cell is reported in Figure 4-16(a). A variation of V_{th} lower than 40mV is extracted when surrounding cells are programmed. This is attributed to the reduced $(V_{FS} - V_{FS_{i,j}})$ factor when the aggressor adjacent cells are programmed.

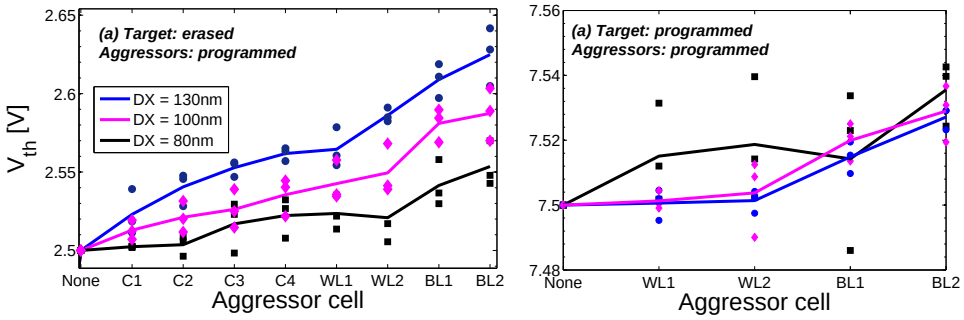


Figure 4-17 – Measurements of cross-coupling effects of programmed aggressor cells on a target erased (a) and programmed (b) device. An increase of V_{th} is found in both cases when adjacent cells are progressively programmed. The influence of corner cells is smaller than the one of BL and WL aggressor devices. Variations with the distance DX between the floating gates in in the BL direction are also reported. Lines indicate the average of measured results in symbols.

Cross-coupling effects in the array can be modeled and included in the compact approach considering the capacitive components of the cell FG of the target with the surrounding devices in the matrix. In such a case, additional terms should be considered in the CBE that in DC conditions reads:

$$Q_G(V_{FS}, V_{DS}, V_{BS}) + C_{CF}(V_{FS} - V_{CS}) + \sum_{j=-1,0,1} \sum_{i=-1,0,1} C_{FF_{i,j}}(V_{FS} - V_{FS_{i,j}}) = Q_{F0} = \text{const.} \tag{4.14}$$

The nodes $V_{FS_{i,j}}$ could be forced to fixed voltage values or simulated connecting them to the FGs of the neighbouring aggressor cells.

4.5 Statistical variations

Worst-case analysis is usually performed by IC designers to assess circuit performances in harsh environment conditions. This requires the definition of statistical process corners in device models. Industrial design kits define process corners by a set of statistical variations of particular process parameters around nominal values. For example, incertitude in the resolution of lithographic patterns corresponds to the definition of statistical distributions for the physical dimensions of the device.

In flash devices, adopting a similar approach to the one on MOSFET models, process corners for both erased and programmed states could be defined. Figure 4-18 illustrates the variation of the drain current and gate transconductance with V_{CS} in read conditions ($V_{DS} = 0.5$V) for the different process corners. Red and blue solid curves indicate the spread of DC characteristics of the device in programmed and erased states simulated with Monte Carlo analysis. Gaussian distributions are adopted to reproduce variability on the physical dimensions, tunnel oxide thickness, substrate doping, carrier mobility and flat band voltage. In DC regime, the initial stored charge is also assumed to be distributed with an uniform distribution around the nominal conditions of both the states. Statistical parameter variations determine oscillations on the device threshold voltage, subthreshold slope and maximum gain transconductance.

Other specific static process corners could be defined, to simulate the flash cell in particular conditions. For example, embedded devices usually adopt P/E-verify algorithms to check the efficacy of the writing process and reduce the spread of V_{th} distributions [141]. Consequently, it is useful to define P/E-verify corners from final product specifications (dashed and dashed-dotted lines in Figure 4-18), which can be used to assess device performances under soft-program/erase operation [268]. These static corners are defined varying the stored charge in the floating gate of the device from the nominal P/E state values. Additionally, in applications where sector-erase functionality is exploited, a wide set of devices can undergo several consecutive erase sequences, inducing parasitic tails in V_{th} distributions due to leakage (see Chapter 5). A depletion-verify corner is also defined and indicated with a dotted line. It is used to define a state of over-erased depleted cells whose V_{th} needs to be raised to reach the nominal erased state.

The statistical variation of electrical parameters is extracted from the program and erased state characteristics measured on all the fully addressable mini-arrays integrated on the wafer. During the measurement, for each word line, all the devices on the different bit lines are characterized. The $I_{DS}(V_{CB})$ characteristic of each cell is measured three times: at the initial state, after a program sequence and after applying an erase pulse. The program step-like pulse has $V_{CS} = 8.5V, V_{DS} = 4.2V, V_{BS} = 0V$ for a duration of 1μs, while the step-like erase pulse reaches $V_{CB} = -16$V for 100ms. The variances of V_{th} and g_{mMax} distributions for both the P/E states have been extracted from results presented in Figure 4-19(a-b). Assuming a Gaussian distribution of V_{th}, a variance of $\sigma_{V_{th}^E} = 0.15$V and $\sigma_{V_{th}^P} = 0.13$V has been extracted for erased and program states, respectively. An

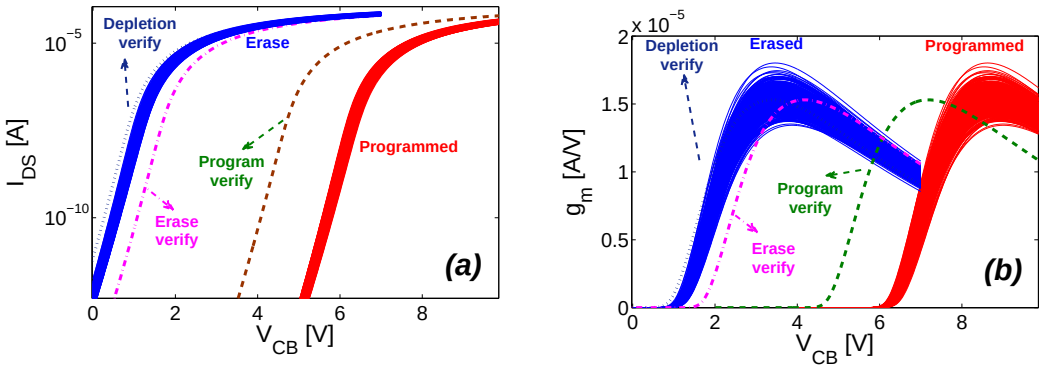


Figure 4-18 – (a) I_{DS} and (b) gate transconductance g_m characteristics simulated with Monte Carlo statistical analysis for the erased (blue) and programmed (red) states at V_{DS}=0.5V. A large spreading of V_{th} can be noticed. Program, erase and depletion-verify static corners are also included.

indication of the V_{th} distributions for virgin (after fabrication) and over-erased[1] cells is also provided (black histogram). The virgin V_{th} has been found to be around 5.3V and can be measured only for devices connected to the first BL. Indeed, when moving along a bit line, the page erase pulse erases all the cells sharing the same WL, while the program sequence increases the V_{th} of the target cell only. Consequently, the cells on the same WL become more and more over-erased while moving towards the BLs, generating a skewed distribution of low V_{th}^E. In Figure 4-19(b) the distribution of the gate transconductance peak is shown. A Gaussan distribution with $\sigma_{g_{mMax}} = 0.11$S can be assumed. A slight discrepancy between the programmed and erased states is noticed and has been attributed to different control gate-drain couplings under DC measurement conditions. Figure 4-19(c-e) illustrates the spread of the V_{th}, g_{mmax} and subthreshold slope distributions obtained from MC simulations on 400 samples. The subthreshold slope has been extracted from $I_{DS}(V_{CS})$ curves in linear regime between 1nA and 10nA.

4.6 Conclusion

In this Chapter, parasitic and long term effects have been illustrated and modeled. The analysis on endurance performances in flash devices permitted to decouple the effects of the defects on the cell electrostatics and P/E efficiencies. Transient and endurance characteristics of embedded flash memory devices have been characterized and modeled using a physical extraction methodology, a semi-analytical approach and a compact design-oriented model. This approach intends to present the basic methodology for the modeling of degradation for circuit simulations. However, more complete and physical models for defect creation can be adopted to improve the predictions of the compact model. This per-

[1]Because of the particular NOR architecture, a cell can be subject to multiple erase sequences, having its V_{th} strongly reduced with respect to the nominal V_{th}^E after a single erase.

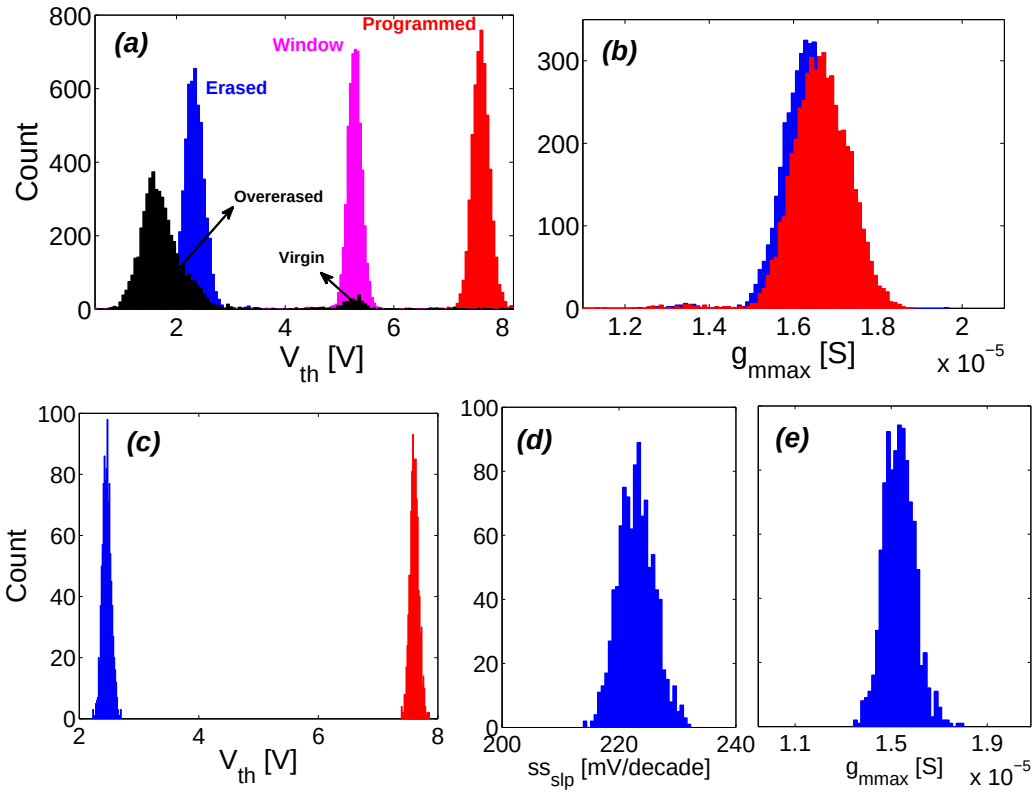


Figure 4-19 – (a) V_{th} distributions measured on mini-arrays of 80 addressable devices for 72 dies. Program (red), erase (blue) and initial (black) distributions are shown. The magenta histogram indicates the threshold voltage window W distribution. (b) Measured distribution of maximum transconductance gain for programmed and erased states. (c-e) Simulated V_{th}, gain and sub–threshold slope distributions simulated with Monte Carlo analysis applied to the extracted flash SPICE model.

spective analysis would permit to model endurance characteristics and cell performance reduction for a wide range of stress bias voltages and P/E algorithms. The phenomenon of BL disturb has been characterized, modeled and included in the compact NVM-SPICE solver. The role of cross-talking effects has been characterized and parasitic capacitances extracted from 3D TCAD simulations. As a result, the values of these components have also been applied in Chapter 5 into BL and WL models. Statistical process corners permit to analyse worst-case cell conditions and define process targets.

Chapter 5

From compact modeling to IC design

5.1 Introduction

In this chapter, the previously described compact model has been applied in a process development and design perspective. On one hand, technology development requires a profound understanding of trade-offs in flash devices, which affect DC, transient and long term performances. In Section 5.2 the main metrics adopted by process integration engineers and NVM designers for the evaluation of the cell figure of merits are described [105]. Parametric analysis is applied in transient regimes, studying the effects of program/erase pulses on relevant quantities involved during the switching (threshold voltage V_{th} shift, maximum drain current peak during programming, power consumption, cell disturb, degradation, etc.). The shape of the voltage pulses applied to both word lines (WLs) and bit lines (BLs) largely impact the final state, the short term and long term performances of the device. Consequently, one of the main tasks of IC designers consists in defining the pulses to be applied in order to optimize the cell and satisfy product specifications. Additionally, some cell specifications strongly affect the requirements of other building blocks surrounding the memory sector. In this view, an optimization methodology using the NVM-SPICE model is proposed. Section 5.3 discusses the design, integration and characterization of a 40KB memory sector for smart card applications. The integrated building blocks are highlighted, with particular focus on the direct use of the flash compact model for their optimization. This section also includes a description of BL and WL parametric cells conceived to simulate a complete line of devices in the memory sector. Layout considerations involving the design techniques for high-voltage flash technologies are explained. Characterization results and functionality testing of the memory test-chip conclude the chapter.

5.2 NVM-SPICE model applications

5.2.1 Performance metrics and parametric analysis

The discussion of trade-offs and pulse optimization requires a clear definition of the different metrics to assess device performance. In particular, in this study the following metrics can be considered:

- program/erase time
- power consumption
- maximum program/erase current
- WL/BL disturb
- injection efficiency
- interface and oxide degradation
- width of V_{th} distribution

It has been chosen to detail here only the analysis of the programming pulses applied to BLs and WLs are pulsed as more trade-offs are involved. Nevertheless, the methodology is also applicable to the erase operation. The evaluation can be performed varying a limited set of pulse parameters or discretizing the applied voltage pulse in a given number of timesteps. However, we simplified the problem investigating only a family of trapezoidal WL and BL pulses, characterized by two parameters for each pulse as indicated in Figure 5-1. This choice has been made also considering that linear regulation is usually adopted in switched capacitor regulation blocks and its importance will be justified in the next Section. Each WL/BL pulse is composed of a ramp-up step and a plateau part. The total duration of the ramping phase is t^C_{rise} (t^D_{rise} for the BL pulse, respectively), while its initial voltage is V^C_{min} (V^D_{min}, respectively). In this example, the maximum voltages on the plateaux $V^C_{max} = 8.5V$ ($V^D_{max} = 4.2V$, respectively) have been maintained for a constant total pulse duration of $1\mu s$. Furthermore, the impact of the isolated p-well bulk bias can also be controlled. In the following examples, V_B has been kept constant at 0V.

As a result, Figure 5-2 shows the simulated programming time t_{prg} variation with the pulse parameter. The programming time corresponds to the instant where the V_{th} of the cell reaches 7.5V. This quantity is of major importance as it discriminates the two set/reset states and determines the threshold voltage window. A linear variation of t_{prg} with t^C_{rise} and t^D_{rise} is found.

Low-power applications require optimization of the program/erase efficiency. The total power consumption during a single programming pulse is shown in Figure 5-3 peaking at intermediate V^C_{min} and V^D_{min} voltages, and increasing for slower ramps, as the program efficiency is reduced. The dependence of the programming time on V_{DS} pulse parameters is higher than the dependence on V_{CS} variations. As a consequence, it is more efficient

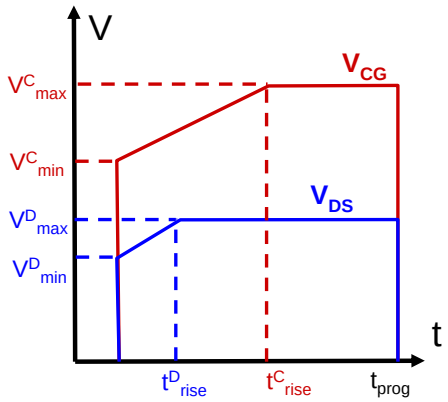


Figure 5-1 – Schematic defining the pulse configuration parameters for pulses applied on the control gate and drain of the selected cell. The investigated family of pulses is represented by a trapezoidal voltage pulse where the rise time and the low level bias are varied.

to vary the BL pulse in order to optimize the power consumption. Square-box pulses obtained for t_{rise}=0s or $V_{min} = V_{max}$ achieve the largest V_{th} variation with the lowest power consumption. The trend is also correlated with the average injection efficiency.

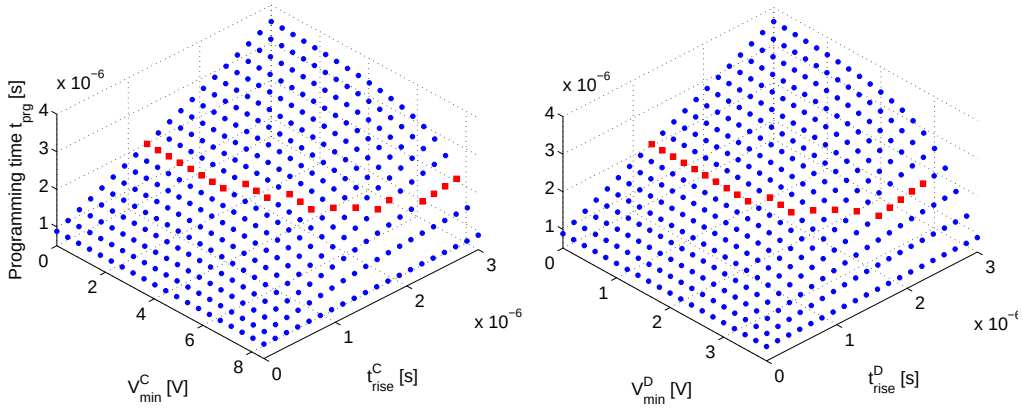


Figure 5-2 – Simulated program time needed to achieve a final V_{th} of 7.5V, as a function of the parameters indicated in Figure 5-1, i.e. rise times and starting voltage biases of both control gate (left) and drain (right) pulses. In the left plot, a square box voltage pulse with $V^D_{max} = 4.2$V has been applied to the drain of the cell, while in the right plot the control gate is biased with a square box pulse having $V^C_{max} = 8.5$V. Red symbols highlight the pulses programming the cell after $t_{prog} = 2\mu$s, as in Figure 5-2.

Another parameter to be considered is the maximum drain current I_{DS} which defines the performances of BL charge pump circuits corresponding to the maximum current deliverable by the switched capacitor network. As a result, Figure 5-4 shows the maximum current variation as a function of the pulse parameters. Lower currents are obtained with intermediate V^C_{min}/V^D_{min} voltages and for slow ramps, while the current peaks for square-box pulses. A tradeoff is established between maximum program current and power consumption. Since the control gate coupling coefficient is higher than the drain coupling and the

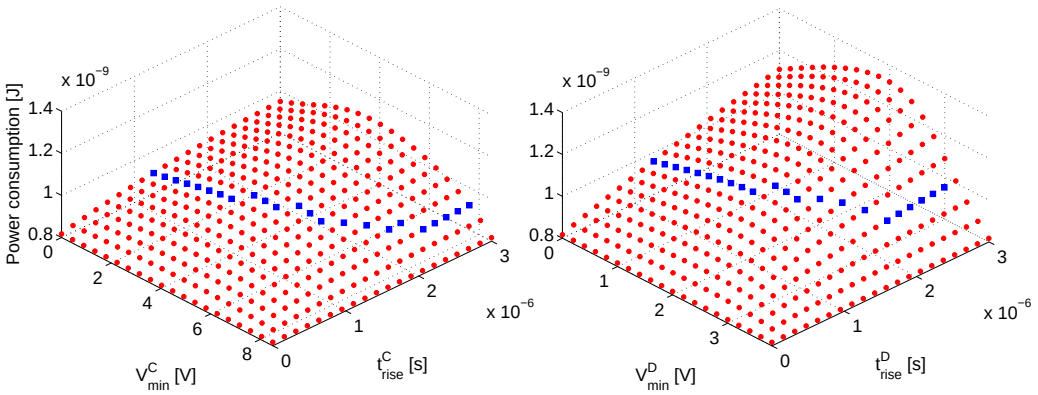


Figure 5-3 – Same programming conditions as in Figure 5-2 but showing the power consumption during a single program pulse. Square box pulses achieve minimum power consumption. Highlighted in blue symbols the isolines with $t_{prog} = 2\mu$s as in Figure 5-2.

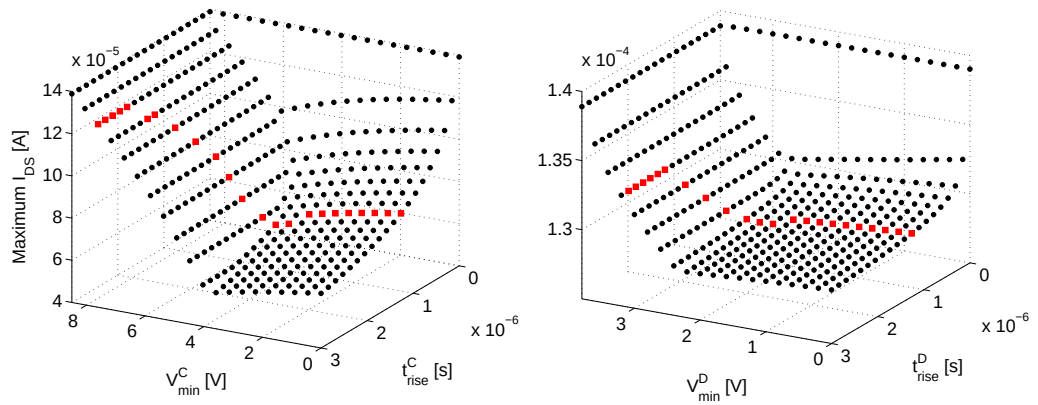


Figure 5-4 – Same programming conditions as in Figure 5-4 but showing the maximum current absorbed during program operation. Red symbols highlight the pulses programming the cell after $t_{prog} = 2\mu$s, as in Figure 5-2.

cell is programmed in saturation, the control of the channel current is very sensitive to the WL pulse parameters. Consequently, it is more convenient to adapt the WL pulse parameters in order to optimize the maximum I_{DS} current.

Drain current optimization in program is also correlated to degradation. Indeed, following [142] and the derivations in Chapter 4, a characteristic ageing coefficient can be computed as well with Eq. 4.12 from the drain and bulk currents and adopted as degradation metric or using the considerations in [264]. Square-box pulses are characterized by high degradation effects and consequently alternative solutions should be preferred in products where endurance is a priority.

Another metric is associated with drain disturb in program conditions, which represents a major issue in NOR flash technologies. Disturb effects can be evaluated applying a pulse on the BL and calculating the variation of the V_{th} of an aggressed programmed cell sharing the same BL. However, due to the large time constant, a more relevant approach considers the integral of the disturb injection current I_{FD} from drain to gate calculated with Eq. 4.13. Simulation results for this metric are shown in Figure 5-5 for a BL pulse with $V^D_{max} = 4.2V$ and a constant voltage $V^C_{max} = 0V$ on the control gate. As expected, an exponential increase is observed for higher values of V^D_{min}, while the dependence on the ramp time t^D_{rise} leads to a linear tendency on the disturb metric. Also in this case, higher disturb is achieved for square box pulses with larger V_{FD} voltage drop.

Other metrics or a composition of them could be defined by the designer who also has to evaluate the impact of pulse and cell optimization on the surrounding circuitry. Indeed, reliability concerns should be considered on the degradation of the HV decoding and regulation circuitry. Several bias conditions can be assessed, also considering the bulk substrate bias voltage V_B used to improve program efficiency.

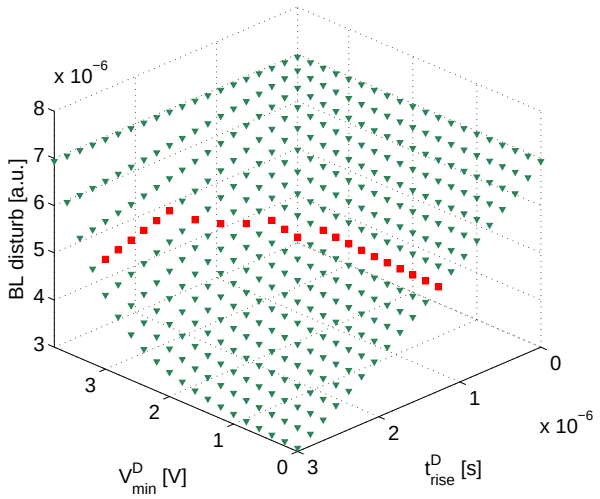


Figure 5-5 – Same programming conditions as in Figure 5-2, but showing the BL disturb metric as a function of the BL pulse configuration parameters. No pulse is applied on the control gate of the cell. The latter represents a device which shares the same BL with a programmed cell. Red symbols highlight the pulses programming the cell after $t_{prog} = 2\mu$s, as in Figure 5-2.

5.2.2 The quest for the best program pulse

In industry, ramped algorithms are applied to the cells in program operation to limit the program current and relax charge pump circuit complexity, and enable better controlling the V_{th} variation or cell aging effects. Various program algorithms have been recently proposed, one of the most diffused being a ramped pulse on the WL, while maintaining the BL at a given program voltage [142]. Indeed, as shown in the following, such an approach permits to avoid the drain current peak. The floating gate voltage is linearly increasing with the control gate voltage and the current can be maintained constant for the entire ramp up of the WL pulse. The maximum current is set by the designer depending on the desired charge pump specifications. Figure 5-6 shows simulation results after the

application of two pulse configurations achieving the same specifications on the threshold voltage variation and on the programming time. A given maximum program current has been imposed; with this constraint, the control gate voltage at each time step is computed. As shown in (b) and (c), the injection and drain currents remain constant for the entire ramp up preventing overshoots. This avoids inefficient programming in presence of high initial current peaks followed by lower currents. This method also permits to prevent the abrupt exponential variation of V_{th} that is observed using square-box pulses. Indeed, in presence of such a rapid variation, the final state of the cell becomes strongly dependent on variability, resulting in larger V_{th} distributions of the final programmed state. While in [142] the calculation of the equilibrium condition on V_{FS} needed to achieve best efficiency is determined using approximated models where empirical and not-measurable parameters are considered, NVM-SPICE permits to accurately perform this assessment. Two pulse configurations are shown for two different V_{max}^{D} constraints.

A third pulse can also be applied on the isolated p-well of the array, biasing the common substrate at a negative voltage. This has the purpose on one hand of reducing the drain current during writing and on the other hand of improving the injection current by increasing the voltage drop on the junction.

A general optimization methodology has been developed for the extraction of pulse configuration parameters depending on a given set of specifications and constraints (a patent disclosure has been submitted and is currently under evaluation on this subject).

Finally, Monte Carlo statistical and process corner simulations can be adopted for analysing the impact of pulse configurations on V_{th} distributions narrowing. The designer is able to apply a program–verify sequence [141] to investigate its effects on the distributions of states. Statistical models included in NVM-SPICE permit to elaborate a circuit structure and a methodology to remove statistical variability (another patent disclosure is under evaluation on this aspect).

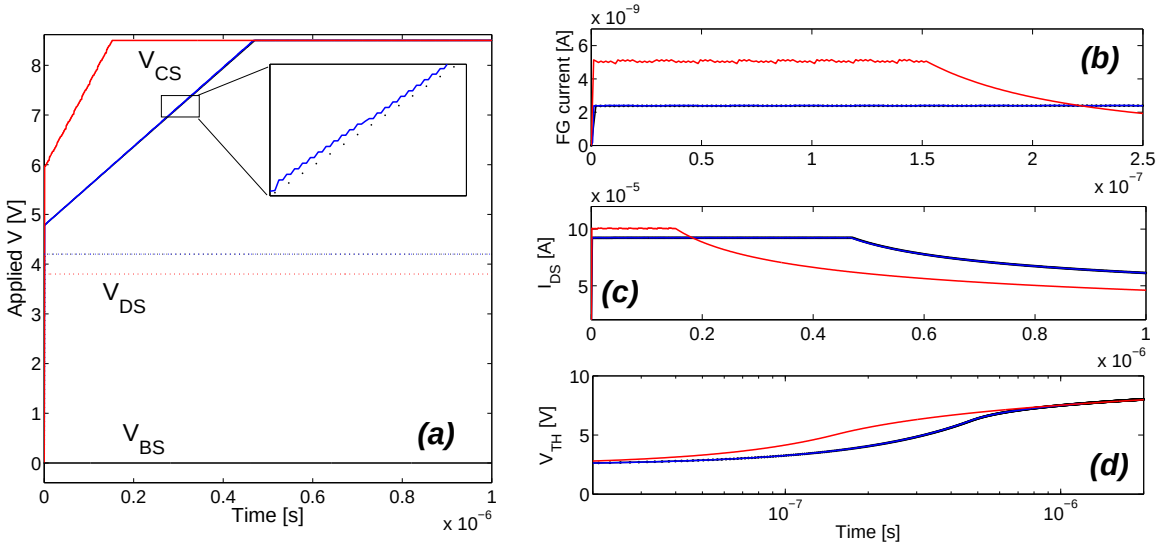


Figure 5-6 – (a) Program pulse configurations for V_{CS} and V_{DS} simulated with the NVM-SPICE model for two maximum drain biases. On the right, the floating gate and drain currents vs. program time are also shown in (b) and (c), respectively. Constant current program is achieved maintaining the injection current constant for the entire ramp up of the control gate voltage. On the voltage plateau, the current decreases with a reduction of V_{FS}. In (d), the threshold voltage variation is reported for the two pulses, achieving the same program time specification at $1\mu s$ with different program dynamics.

5.3 40KB eNVM sector design

A 40KB memory test-chip has been designed and integrated in a derivative eNVM 65nm technology. The following blocks have been designed and are presented in this section. They include:

- the memory sector, which surrounds the WL decoder circuit and has been divided in two blocks to enable reuse for future improvements;
- the WL decoder, which provides the conductive path from the HV switches to the local WLs of the sector;
- the BL decoder, which creates the connections between the local BLs and the external IO terminals;
- the HV management system, including a set of HV switches, to create the conductive path from the external HV pads to the decoders for the HV signals, and high/low level shifters to increase the dynamic range of digital LV control signals controlling the switches;
- the combinatory logic block that provides the control signals required by all the building blocks in the different operating modes.

A global system floorplan of the test-chip, showing the connections between the main building blocks is illustrated in Figure 5-7, while a photograph of the integrated testchip is provided in Figure 5-8, also showing the test structures integrated for reliability investigations. The main signals are summarized in Table 5.1, differentiating the signals between external biases and internal controls. The first category includes logic control signals, e.g. MODE_LV<0:1> which determine the operating mode or SCTERA_LV to enable sector erase functionality, and HV signals, which are used to supply WL and BL decoding circuits and the flash device under program/erase operation. The derivative eNVM technology supports HV devices with thick oxide layers and robust junctions, to sustain voltages up to 11V. Additionally, triple-well HV devices have been adopted to integrate low-level shifter circuits and provide negative WL bias voltages in erase and read modes. Charge pumps and sense amplifiers have not been integrated in the implemented testchip. As a consequence, HV pulses required to program/erase the devices are provided externally during testing (signals VX, VNEG, VY and VBULK). Signals VX and VY are also used to supply the HV management system and the decoders. Given the absence of sense amplifiers to measure the V_{th} of the device, a direct memory access (DMA) mode has been chosen. In such a way the BLs of the memory array could be directly contacted and drain voltages applied in read/program operation. Such an approach provides a more attentive and accurate characterization, since one can directly access the current flowing in the cell; however access time strongly increases. The IO bus provides access to 10 bits in parallel.

The internal signals connecting the different building blocks of the test-chip are indicated in the bottom part of Table 5.1. All the blocks require logic signals to enable different functionalities according to the operating mode. The role of pre-decoded and HV biases in decoders is illustrated in the following subsections.

5.3.1 Building blocks

(i) Memory matrix and WL/BL parametric models

Several factors should be considered in establishing the array organization. In EEPROMs, access time constraints impose limitations on row and column lengths implying the integration of several different WL and BL decoders in a single memory array. The impact of electrical stress and disturbs also needs to be considered.

A memory array is usually organized in such a way that it can be read and programmed using a parallelism for the output data either by 8 (byte) or by 16 (word). In our case, given the lack of internal blocks for testability and built-in self test functionalities, we extended the dimension of the IO data buffer to 10 signals, 1 byte + 2 bits to support error detection and correction code (ECC) functionality in future implementations. The array has been divided in 2 subsectors having 320 BLs and 512 WLs each, forming an array of 327680 cells (32KB + 8KB for ECC functionality). The IOs are divided in two sets and connect the BLs of the two sub-sectors. Each IO can address 64 BLs. The number of cells connected to one WL should be chosen depending on specifications on the access time, WL decoder area and WL disturb. On the other hand, increasing the number of WLs increases

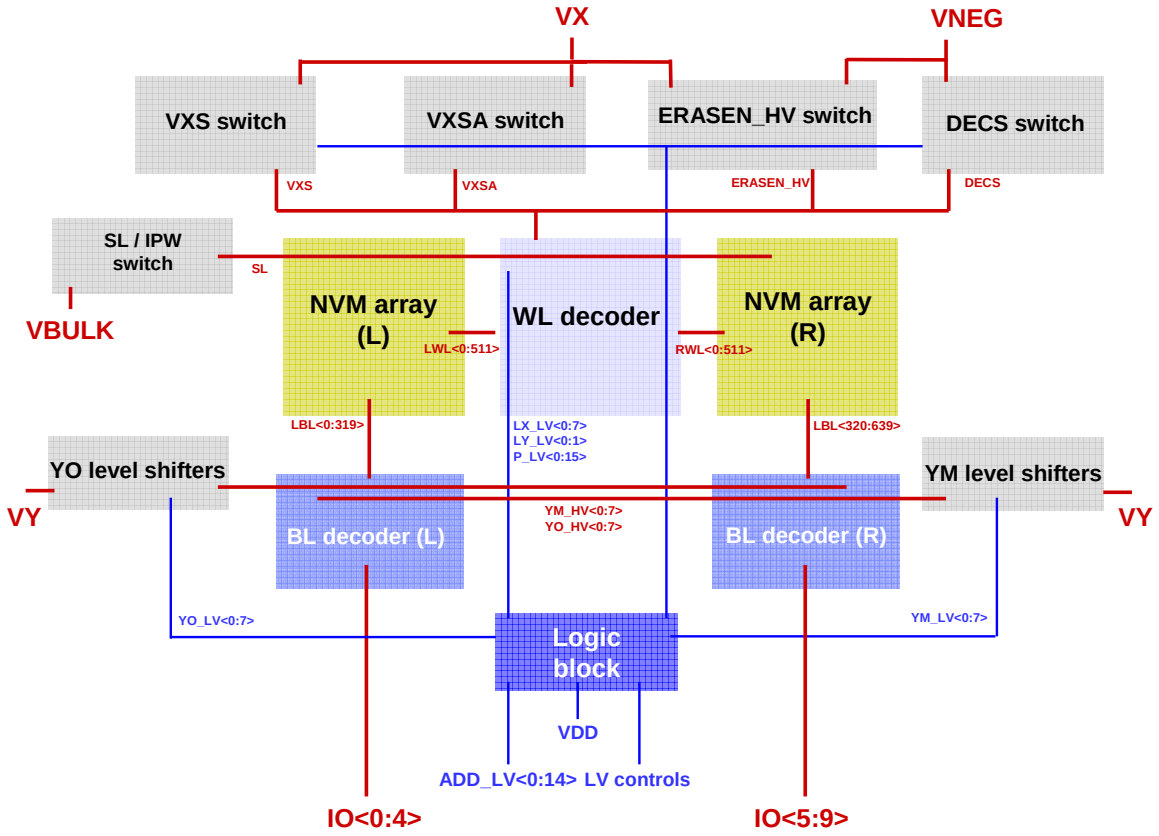


Figure 5-7 – Building blocks of the 40KB memory sector. Blue lines indicate LV signals; red lines represent HV biases. The BL and WL decoder are connected to the NVM subsectors and controlled by the predecoded LV buses from the logic block. The HV management system constituted by the 5 HV switches creates the path for the HV biases from the decoders to the external inputs.

the impact of BL disturb and can reduce sense amplifier efficacy due to the influence of potentially over-erased cells in read mode. These effects also involve limitations on power consumption. The WL decoder selects a single WL in read or program operation, while in erase single or multiple WLs can be selected. Indeed, the array organization is also driven by particular erase functionalities that in Flash memory permit to restore the logic '1' on the entire array (sector erase procedure). In NOR configuration, both page (or WL) erase and sector erase operations could be performed, since FN erase does not represent a major issue for power consumption as program by CHEI. The capability of erasing the entire block has been included in our sector, so that the WL decoder is able to select all the 512 WLs of both the subsectors.

An accurate transistor-level model of the flash device strongly improves design optimization. Indeed, common design techniques for establishing the WL/BL loads adopt coupling considerations and approximated distributed networks with passive elements only. Eventually, a more accurate solution requires to take into account the cell capacitances in different

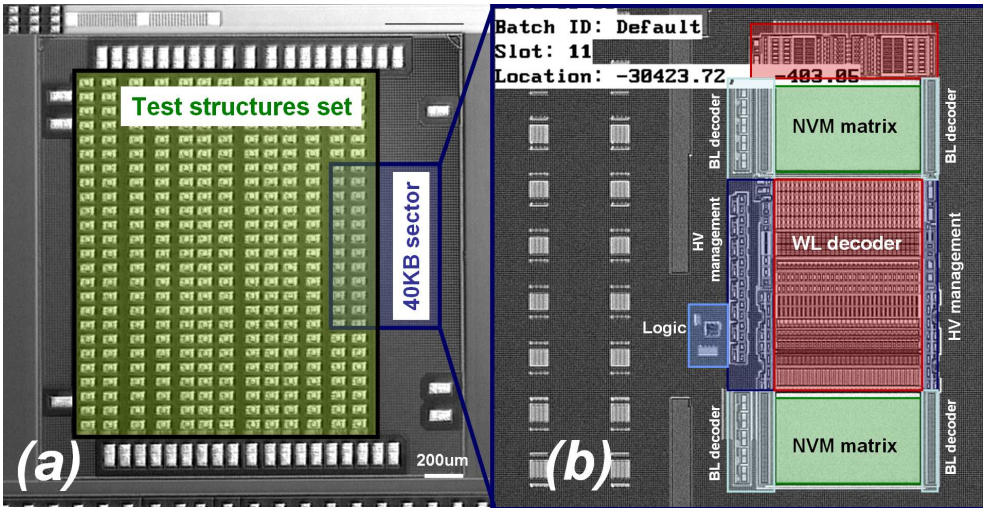


Figure 5-8 – Photograph of the designed testchip. (a) The 40KB sector is placed on the right and a large portion of the area is occupied by test structures to individually characterize the HV circuit blocks. (b) Zoom of the region including the 40KB testchip, where the different building blocks are indicated.

Pin name	Description	Voltage range
VDD	Logic supply	1.2V
GND	Ground reference	0V
ADD_LV<0:14>	Address bus	0/1.2 V
MODE_LV<0:1>	Operating mode selection	0/1.2 V
SCTERA_LV	Sector erase control signal	0/1.2 V
DECODENABLE_LV	Address enable control signal	0/1.2 V
GATENEG_LV	Negative WL control signal	0/1.2 V
VX	WL decoder supply WL voltage in P/R modes	0/10 V
VY	BL decoder supply	0/10 V
VNEG	WL decoder supply for E mode	0/10 V
VBULK	Isolated P-Well voltage in E mode	0/10 V
IO<0:9>	IO bus (DMA mode)	0/5 V
READ_LV, PROGR_LV, ERASE_LV, IDLE_LV	Operating mode signals	0/1.2 V
LS_LV	Sector-enable signal	0/1.2 V
LX_LV<0:7>, LY_LV<0:3>, P_LV<0:15>	Predecoded signals for WL decoder	0/1.2 V
YO_LV<0:7>,YM_LV<0:7>	Predecoded signals for local/global BL decoders	0/1.2 V
DISCHARGE_LV	Discharge control signal	0/1.2 V
YO_HV<0:7>,YM_HV<0:7>	HV signals for local/global BL decoders	0/10 V
WL<0:511>	WL signals	0/5 V
LBL<0:639>	Local BL signals	0/5 V
VXS	Selected WL bias in P/R modes	0/10 V
VXSA	Bias for WL decoder	0/10 V
DECS	Selected WL bias in E mode	-10/0 V
VSL	Source and P-Well bias	0/10 V

Table 5.1 – Description and voltage range of signals in the 40KB sector.

operating conditions and states. To this purpose, the model of Chapters 2 and 4 can be employed for circuit simulation and for the development of parametric distributed models representing entire WLs or BLs. Parasitic resistances and capacitances on the transmission line can be extracted from 3D TCAD simulations, from the layout of the memory sector using parasitic extraction tools or with geometrical approximation based on process characteristics. Transient simulations can be performed with the parametric cells proposed in Figures 5-9 and 5-10, representing distributed models for one entire WL or BL. A better trade-off between simulation time and accuracy is reached when dividing the set of addressable cells in three parts: the cells physically located ahead the addressed target device, the addressed selected cell and the set of devices physically located downstream the selected cell. An additional part can be added in presence of dummy unconnected devices at layout level. For example, in a WL that addresses 64 cells, only one device is selected in read operation. The propagation delay of the voltage pulse on the gate of the cell depends on the index of the selected device. In our testchip, considering 5 IOs, i.e. 320 total BLs, several WL models have been cascaded to simulate the voltage propagation delay on the last cell of the WL. At process level, cell access resistances can be reduced with silicidation or with metal line strapping. This permits to strongly relax the tradeoff between access time and row/column length.

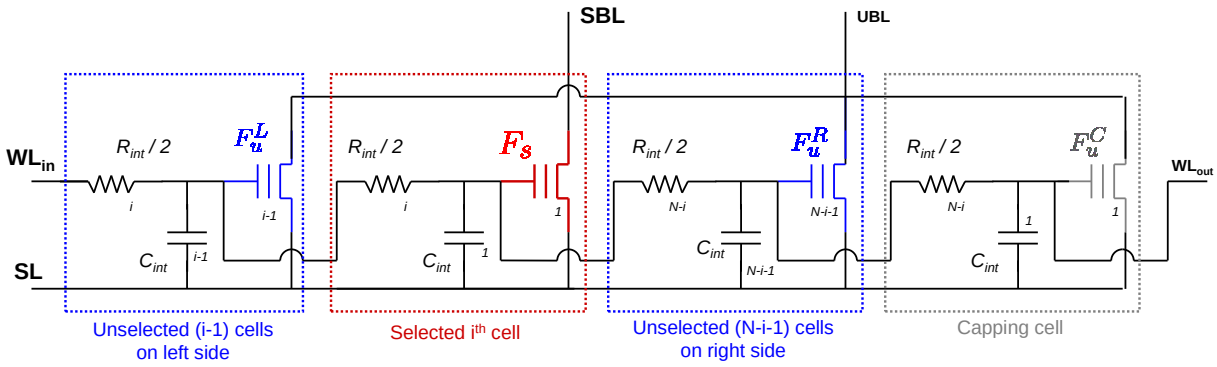


Figure 5-9 – Parametric model of a WL used for estimating the load of the WL drivers. Model parameters include the index i of the selected BL, the total number of addressable cells N on the WL, the number of dummy devices at the boundaries of the line and parasitic components, that can be extracted from layout and geometrical considerations.

Similar considerations are applicable to the sizing of the array along the BL direction, using parametric models as in Figure 5-10. In such a case, not only access resistances and parasitic capacitances limit the cell access time, but also BL disturb effects need to be considered. Indeed, longer BLs cause disturb effects to be present on large portions of the memory sector. Transient disturb simulations using model in Chapter 2 represent a precious tool to evaluate the effect on programmed devices and the maximum drain current.

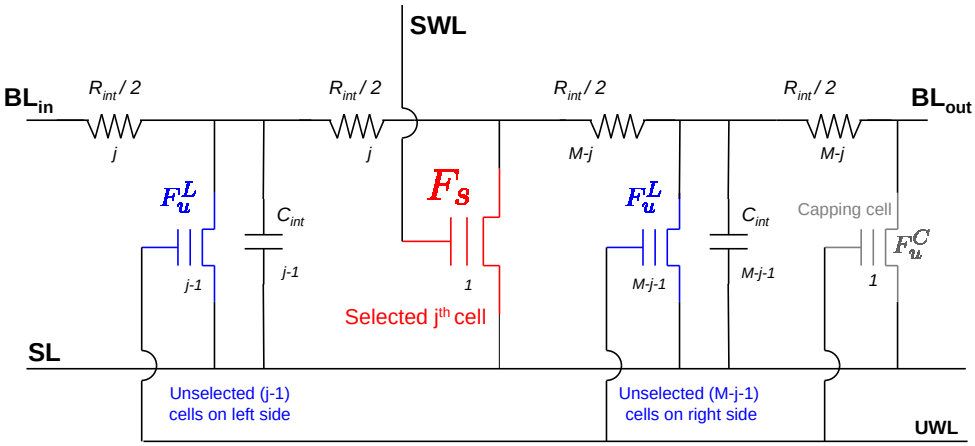


Figure 5-10 – Parametric model of a BL used for estimating the load of the pass-gates in the local BL decoder. Model parameters include the index j of the selected WL, the total number of addressable devices M sharing the same BL, the number of dummy devices at the boundaries of the line and parasitic components R_{int} and C_{int}.

(ii) WL decoder

One of the major challenges in memory design is represented by the integration of row and column decoders. Two factors challenge decoder design: the addition of new functionalities, including various operating voltages and modes of operation, and the continuous reduction of cell dimension leading to an exponential array density increase.

Hierarchical decoding The row decoder is the first block to be designed after the memory array. In order to enable random access to each cell, a path should be realized using a specific device address. Each line can be individually addressed through the WL decoder, in contrast to the operation of BL decoders where bytes or words are decoded in parallel. This approach has advantages in terms of access time, since one single row must be addressed to read the bits of the same byte, which are placed on different columns [90]. In our testchip, the WL decoder addresses $512 = 2^9$ rows; thus 9 address bits ADD_LV<0:8> are required. The address bus is biased at VDD=1.2V, while various HV biases need to be applied on the flash device. In particular, Table 5.2 summarizes the WL DC voltage specifications for the different operating modes. In read mode, we would like to bias the control gate of the selected device at a voltage equal to 5V, while keeping the unselected word lines grounded. When long BLs are integrated and when over-erased cells are affecting current/V_{th} measurement, it might also be convenient to apply negative biases in read mode on unselected devices to reduce parasitic leakage. Design complexity also increases due to the fact that the erase operation requires high negative biases to be applied on the WL. At layout level, issues emerge if one considers that the last driver stage of the decoder should fit into the row pitch (footprinting issue) [90]. Consequently, if one excludes the memory array, the WL decoder represents the most compact portion of the testchip. On the other hand,

device reliability and reusability concerns impose layout design with relaxed rules. Finally, a large mismatch between the minimum size of HV devices and flash cell dimensions also exists which further challenges integration.

WL voltage	R	P	E
Selected	0/10 V	0/10 V	-10/-5 V
Unselected	-2/0 V	0 V	0 V

Table 5.2 – Summary of the WL voltage range DC specifications during read (R), program (P) and erase (E) operations. Unselected WLs can be biased at negative voltages in read mode, to reduce contributions of over-erased cells sharing the BL with the selected device.

Hierarchical approaches are generally preferred to reduce area occupation and layout complexity. Figure 5-11 indicates the decoding scheme and the distribution of the signals from the addresses to the WLs. The 9 LV address bits are decoded independently in three groups. Three predecoded buses (LX_LV<0:7>, LY_LV<0:3>, P_LV<0:15>) are obtained with conventional n-to-2^n logic *pre-decoders*. The dynamics of the logic signals are raised to the desired HV bias in a second stage (*selector* - blocks NAND and P). The logic AND operation is performed between each LX_LV and LY_LV couple. Therefore, the 512WLs are divided in 32 groups of 16 rows each. The selection of each group is performed with the predecoded signals LX_LV<0:7> and LY_LV<0:3> in NAND-block circuits, while signals P_LV<0:15> select one of the WLs in the group in P-block circuits. The hierarchical solution simplifies layout complexity and signal routing as only the predecoded signal buses have to be routed across the memory array. The last block is represented by the final driver, which selects and transfers the analog voltage to one of the 512 WLs. Figure 5-12 shows the hierarchical view of the WL decoder divided in global (a) and local (b) blocks.

WL driver The single WL driver circuit that biases the line must be designed considering the voltage requirements depicted in Table 5.2. Read and program modes share the same lines, as the HV decoding path that needs to be created is the same for both, varying only the applied voltages and timings. The WL is only selected when both the NAND-block and P-block are selected, while deselection of the WL occurs when the predecoding signals identifying the group of 16 or the single WL in the group are lowered. The WL driver circuit is supplied by the HV signals VXS and DECS provided by the HV management system. The amplitude of both varies depending on the operating mode: in read/program VXS ranges from 0V to 10V, while it is grounded in erase. The DECS signal has a dynamic range from -2V to GND in read/program while it decreases to VNEG in erase mode. The logic of HV switches will be described in Section (iv). The WL driver forces a stable voltage on the WL disregarding the selection or operating mode. Indeed, it is not only important that a WL is selected but also that all the rows that are not addressed are forced to a defined voltage, to avoid floating nodes and unwanted leakages. Consequently, the selected

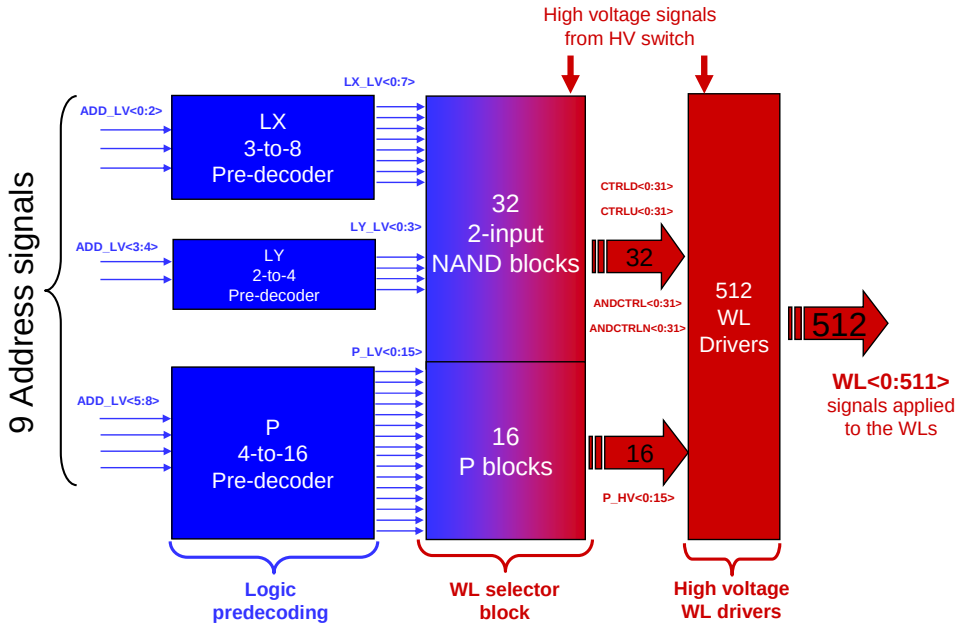


Figure 5-11 – Hierarchical decoding scheme for WL decoding. Red blocks indicate HV circuits [-8V, 9V], while blue blocks represent LV circuitry [0V, 1.2V].

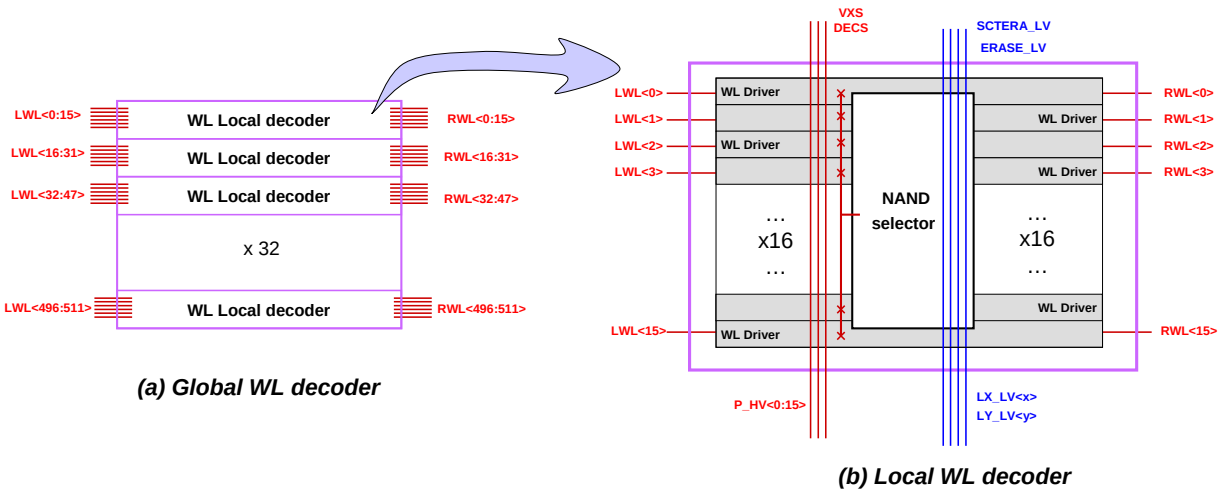


Figure 5-12 – Architecture of the global (a) and local (b) WL decoders. HV and LV signals are indicated in red and blue, respectively.

WL in read/program states and erase state is forced to VXS and DECS, respectively, while the unselected WLs are forced to DECS and to VXS, respectively.

The circuit in Figure 5-13 performs the biasing of the individual WL, using the complementary decoded signals ANDCTRL/ANDCTRLN and CTRLU/CTRLD provided by

each NAND block, and the signal P_HV provided by each P-block (see the next section for an explanation of the generation of these signals). The two outputs LWL and RWL correspond to the WL connections of the left and right subsectors and are driven by HV buffers. Sizing the two output buffers can be performed using the previously described WL models, evaluating the tradeoff between access time, reliability and occupied area. The node FN is forced to VXS or DECS in all the operating modes. When the NAND block is selected, both transistors M1n and M1p are disabled and the inverter with P_HV as input is enabled. A signal path is thus created for P_HV, which also determines the final WL voltage. When the NAND block is disabled, the WL is forced to a stable state by M1n and M1p, independently from the P_HV voltage, to avoid the presence of floating gate nodes and stability issues. For example, in erase mode M1n is enabled applying VXS on its gate and forcing FN to DECS. WL drivers should fit in the WL pitch of the array and thus full custom design of devices of HV minimal dimensions is usually considered. The blocks should also be highly modular to preserve symmetry and permit reusability. The truth table of this circuit is shown in Table 5.3.

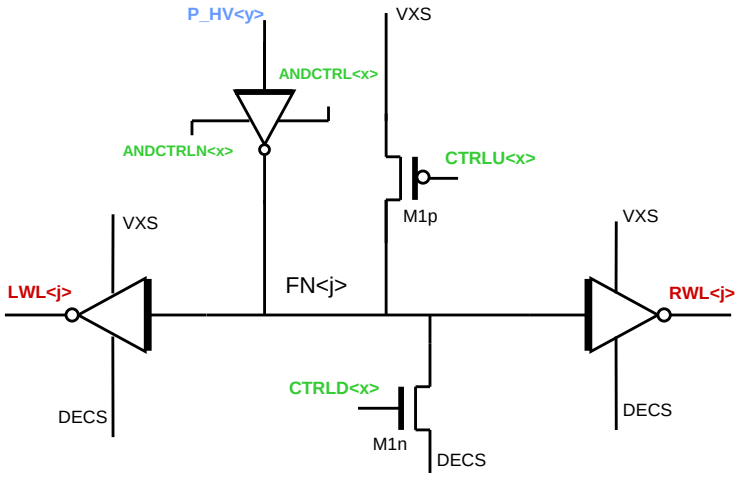


Figure 5-13 – Circuit schematic of the WL driver. Signals from the x^{th} NAND block are indicated in green, while the signal P_HV<y> is the HV output of the y^{th} P block. Ouput signals in red are connected to the j^{th} WL of the NVM matrix.

NAND and P blocks In the decoding flow, the dynamic of low voltage signals needs to be increased from [0 VDD] to the voltage range [DECS VXS] required to program/erase the cell. Level shifting circuits are adopted to this purpose in both the NAND blocks and the P blocks forming the selector. In NVM design, both high level and down level shifters need to be designed. The high-level shifter converts a LV logic level in the range of [0 VDD] to a high voltage signal [0 VXS]. On the other hand, low level shifters extend the dynamic range of the signal toward negative HV values [DECS VDD].

A common architecture is illustrated in Figure 5-14. These conventional levels shifters are not suitable for deep-submicron LV applications, where the supply voltage is reduced below 1.2V and the NMOS drivers are not able to switch the two branches of the circuit. Furthermore, reliability issues become prominent when HV biases are considered. Due to

Mode	The WL is:	ANDCTRL	ANDCTRLN	CTRLU	CTRLD	FN	P_HV	LWL=RWL
R/P	Selected	VXS	DECS	VXS	DECS	DECS	VXS	VXS
R/P	Unselected via P block	VXS	DECS	VXS	DECS	VXS	DECS	DECS
R/P	Unselected via NAND block	DECS	VXS	DECS	DECS	VXS	VXS	DECS
R/P	Unselected via both	DECS	VXS	DECS	DECS	VXS	DECS	DECS
E	Selected	VXS	DECS	VXS	DECS	VXS	DECS	DECS
E	Unselected via P block	VXS	DECS	VXS	DECS	DECS	VXS	VXS
E	Unselected via NAND block	DECS	VXS	VXS	VXS	DECS	DECS	VXS
E	Unselected via both	DECS	VXS	VXS	VXS	DECS	VXS	VXS

Table 5.3 – Summary of the voltage values for the signals in the final WL driver depending on the mode of operation or the addressing state of the device.

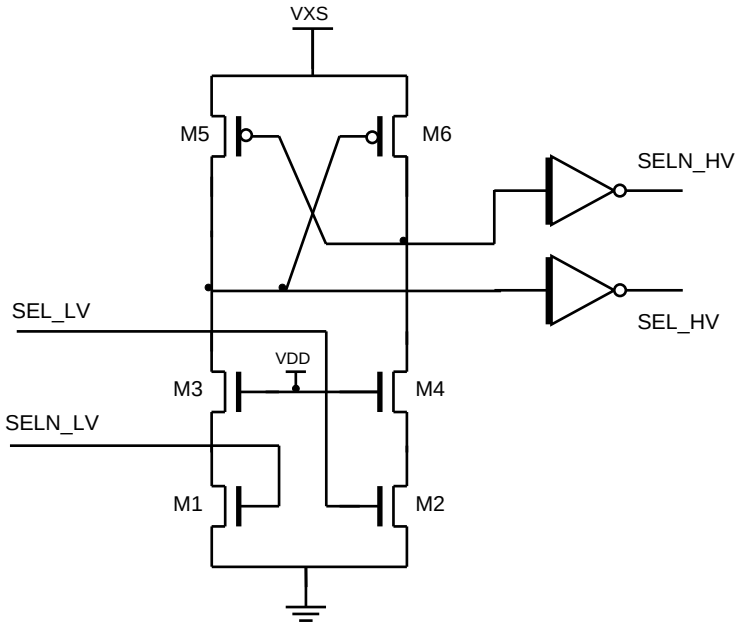


Figure 5-14 – Traditional high level shifter architecture. Complementary low voltage signals SEL_LV and SELN_LV are applied on the two branches on the gates of M1 and M2. The HV devices M3 and M4 are used to reduce device degradation of M1 and M2. M5 and M6 constitute the latch to maintain the signal stable at the HV bias VXS on the output of the level shifter. Finally, the two buffers are used to satisfy output load requirements.

the too rapid VDD scaling, tradeoffs in level shifters become more tight and new architectures need to be introduced. This is due to the fact that while VDD scales down, the V_{th} of HV and IO MOSFETs remains too high to allow these devices to be turned on [91, 269]. Consequently, new solutions have been introduced to reduce the gap between VDD and the V_{th} of HV devices. Some solutions use core NMOS devices to get a lower V_{th} but suffer from reliability and breakdown issues. Other approaches pump-up the supply signal to intermediate voltages, e.g. 2VDD [270], but they require extra circuits and power for the generation of the intermediate bias. High level shifter circuits have been designed using an approach similar to [96]. Figure 5-15 shows the circuit schematic of the adopted high level

shifting architecture. The complementary logic signals SEL_LV/SELN_LV drive both the LV toggling devices Mn1/Mn2 and Mp1/Mp2. The presence of the PMOS devices in the pull-up network improves the switching speed, reducing the contention between the two branches [96]. The zero-V_{th} Mn3 and Mn4 HV devices are used to protect the toggling NMOS devices which suffer from reliability issues, without altering the toggling capability. All the devices have been sized trading off switching speed with drive capabilities.

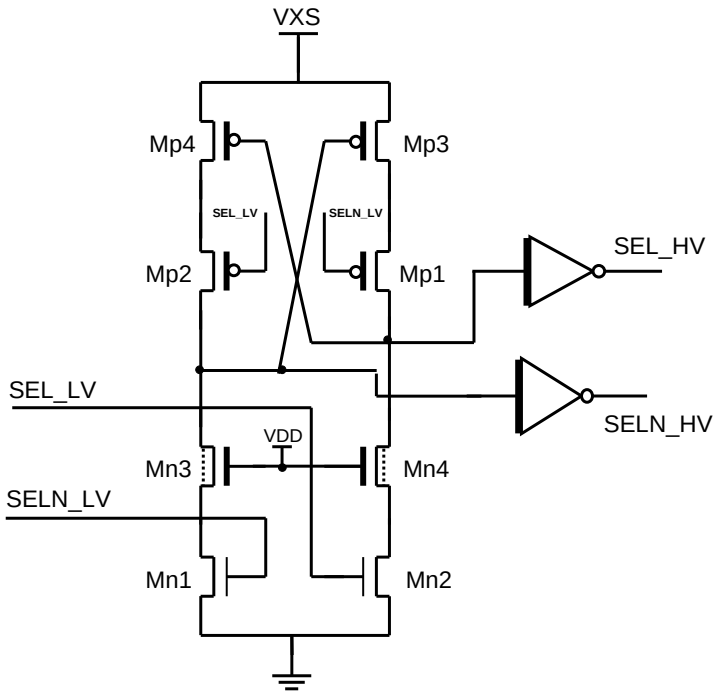


Figure 5-15 – Circuit schematic of a contention-reduced high-level shifter increasing the dynamic voltage range of SEL_LV/SELN_LV from [0 VDD] to [0 VXS]. Contention reduction is implemented with the integration of PFETs Mp1 and Mp2. Zero-V_{th} HV devices Mn3 and Mn4 protect transistors Mn1 and Mn2 from degradation and breakdown.

Low level shifters are integrated to increase the signal dynamics at the negative scale. A scheme that is symmetrical to the high level shifter of Figure 5-15 can be realized using triple–well NFETs with isolated substrate bias to build the latch, and PFETs to drive the toggling. The isolated well of the NMOS devices is biased at high negative voltages. An example of low level shifter will be discussed in Subsection (iv).

High-low level shifters, increasing the voltage dynamic range in both directions, are intensively applied in decoding circuits. Although various complex architectures exist [271] a widely adopted approach consists in cascading a high and a low level shifter circuit, as presented in Figure 5-16. In this way, the dynamic range of signals SEL_LV/SELN_LV is increased in the positive direction from [0 VDD] to [0 VXSA] by the high level shifter and subsequently further extended for negative biases down to [DECS VXS] (with VXS≤VXSA). An isolation scheme is commonly employed to connect the two blocks and prevent both the application of a voltage (VXSA+DECS) on the oxide of HV device that would induce breakdown, and junction forward biasing with potential latch-up effects.

The structure of NAND and P blocks is presented in Figure 5-17. Both blocks include the high-low level shifter of Figure 5-15 and a low-voltage logic circuit that determines

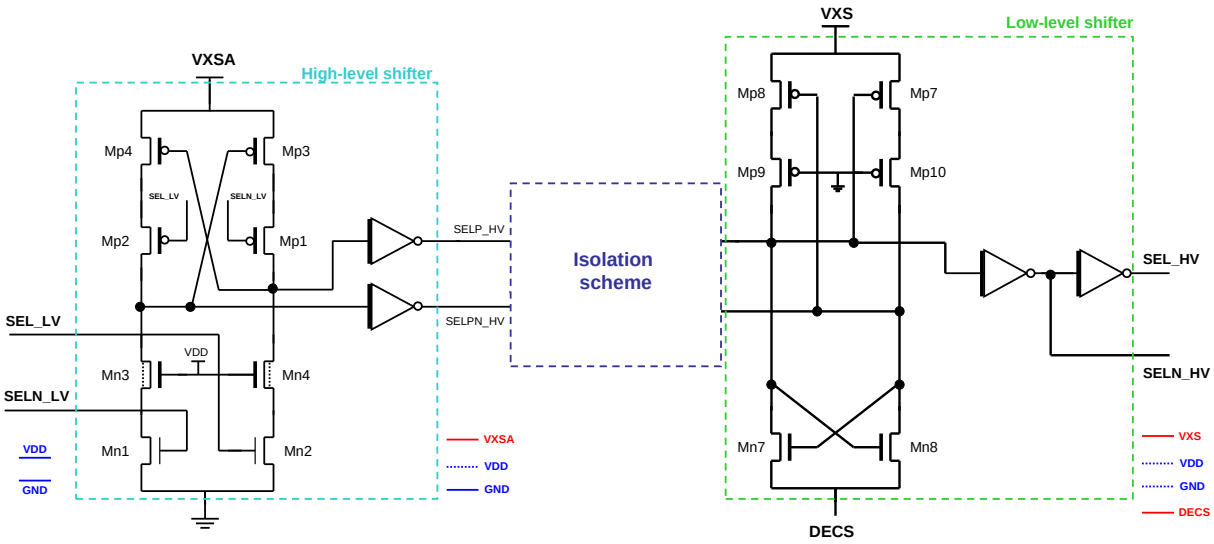


Figure 5-16 – Circuit schematic of a high-low level shifter where two levels shifter are cascaded to increase the dynamic voltage range of SEL_LV/SELN_LV from [0 VDD] to [DECS VXS]. The isolation scheme constituted by Mn5, Mn6, Mp5 and Mp6 devices protects both the buffers of the high level shifter and the latches of the low level shifter. The NFETs Mn7 and Mn8 are integrated in an isolated substrate biased to DECS.

whether the block has to be enabled or not, according to the selection signals. The realized logic function in (a) is:

$$\mathrm{SEL\_LV} = [(\mathrm{LX\_LV} \cdot \mathrm{LY\_LV}) + (\mathrm{SCTERA\_LV} \cdot \mathrm{ERASE\_LV})] \cdot \mathrm{LS\_LV} \qquad (5.1)$$

The block performs the logic AND of the predecoded logic signals LX_LV and LY_LV (sector enable signal). The block is also enabled when the sector erase mode is activated, i.e. SCTERA_LV high, as in such a case all the WLs need to be selected. The signal LS_LV is used to improve IP reusability and indicates the sector enable signal in future implementations. Subsequently, the dynamics of the output signals SEL_LV/SELN_LV is raised to the range [DECS VXS]. The signal VXSA is used to supply the high-low level shifter. A combinatory network generates the signals ANDCTRL_HV/ANDCTRLN_HV and CTRLU_HV/CTRLD_HV, outputs of the NAND block, from the high voltage signals SEL_HV/SELN_HV. The NAND block structure is duplicated for each combination of the LX_LV and LY_LV signals. In (b), the simpler architecture of the P block is illustrated. It includes a logic combinatory circuit and a H-L level shifter. In this case however, the logic function is expressed by:

$$\mathrm{SEL\_LV} = \overline{\overline{\mathrm{ERASE\_LV}} \oplus [(\mathrm{P\_LV} + \mathrm{SCTERA\_LV} \cdot \mathrm{ERASE\_LV}) \cdot \mathrm{LS\_LV}]} \qquad (5.2)$$

The architecture of the H-L level shifter is similar to the one in (a), while the transistor

size increases since the output load varies.

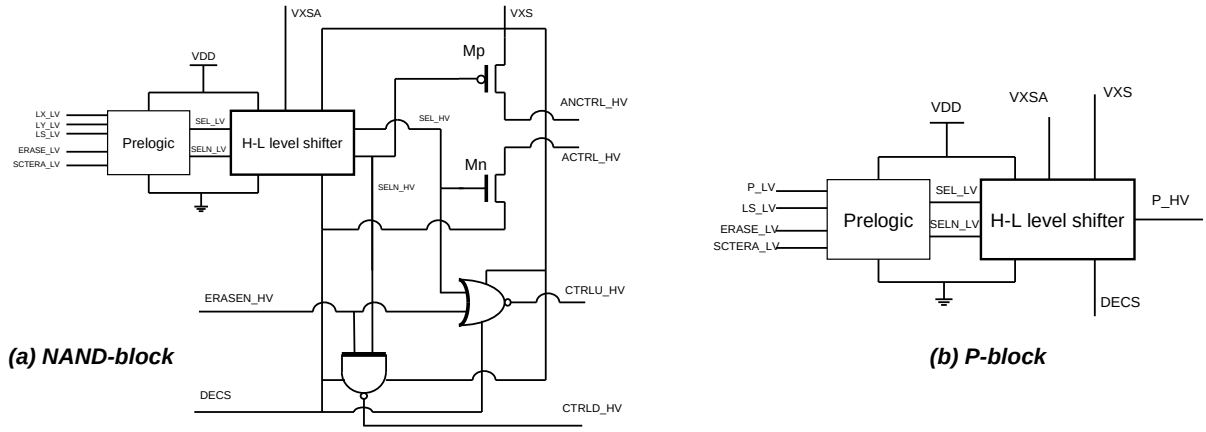


Figure 5-17 – Architecture of the NAND (a) and P (b) blocks in the WL selector to realize the signals in Table 5.3.

(iii) BL decoder

BL voltage	R	P	E
Selected	0.7 V	3/4.5 V	Z
Unselected	Z	Z	Z

Table 5.4 – Summary of the BL voltage DC specifications during read (R), program (P) and erase (E) operations. High impedance condition is indicated with Z.

The structure of column BL decoder is simpler due to the fact that the voltage to be applied on the BLs is limited to the range [0.7V 4.5V] as reported in Table 5.4. The BLs of the unselected devices are normally left floating (high impedance Z state). The decoding scheme is illustrated in Figure 5-18. Similarly to the WL decoder, a hierarchical structure composed of Main Bit Lines (MBL) and Local Bit Lines (LBL) can be defined, connecting global and local decoding levels. Since in read and program modes addressing is done by groups of 10 bits, the entire set of 640 LBLs can be divived in 10 groups, each IO including 64 LBLs. The 64 lines are then divided in 8 subgroups of 8 LBLs. The selection of a LBL in the group is performed using the local BL decoder, composed of 8 HV pass-gates controlled by signals YO_LV<0:7>. Only one of such pass-gates is turned in conductive state; all the other devices are off. The drains of the NFETs are directly connected to the LBL, while their source is connected to the global BL decoder via the MBLs. The global decoder completes the connection path from the selected LBL to the IO, selecting one of the 8 MBLs with signals YM_LV<0:7>. The architecture of the global BL decoder is identical to the local one, but load and degradation considerations need to be taken into account when sizing the pass-gates.

All the BLs can be addressed in group of 10 bits using the address bits ADD_LV<9:14>. The address signals are divided in two groups and predecoded to obtain signals YM_LV<0:7> and YO_LV<0:7>. Subsequently, the dynamic range of the decoding signals is increased from [0 VDD] to [0 VY], where VY is an external voltage varying from 3V to 7V depending on the operating mode. The signals YM_HV<0:7> and YO_HV<0:7> are used to bias the HV pass-gates of the global and local BL decoders, respectively. During read operation, the cell is decoded enabling the pass-gates corresponding to its address and a read voltage is applied on the IO. The current flowing into the device is measured on the IOs. Similarly, in program mode, the HV bias is forced on the IO. In erase mode, the BL decoder is disabled and all the LBLs have high impedance.

At layout level, the design of the column decoders is often problematic as the pass-gates, one for each column, must fit in the cell pitch. This cannot be achieved as the pitch of the cell is normally much smaller than the minimum size of an HV device. To relax this issue, the local decoder is divided in two parts that are laid out above and below the NVM memory bank, allowing the design of the transistor in the pitch of two or more cells.

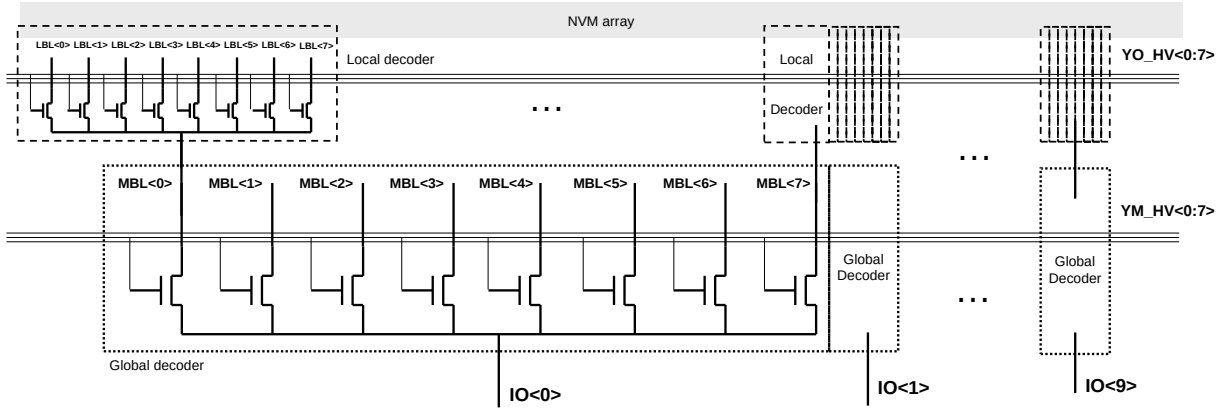


Figure 5-18 – Hierarchical architecture of the BL decoder connecting the 640 LBLs to the 10 IOs.

(iv) HV switches

The HV management system is constituted by five HV switches, realizing the HV path from the external pads to the WL and BL decoders. The set of switches is controlled by LV signals from the combinatory logic block and is designed to be reusable into memory arrays including multiple sectors. Thus, each switch is also controlled by the sector enable signal LS_LV. The set of switches includes:

- the VXS switch which connects the externally available VX voltage to the WL decoder, providing the HV signal VXS to the WL decoder. This signal is only generated in read/program modes and applied to the selected WLs. In erase mode, VXS is grounded and applied to the unselected WLs;

- the VXSA switch which connects the VX voltage to the WL decoder, providing its supply voltage VXSA; the HV bias VX is always transferred to VXSA when the sector is selected;

- the ERASEN_HV switch which increases the dynamic range of the logic signal ERASE_LV to [DECS VXS], as required in the NAND blocks of the WL decoder (see Figure 5-17);

- the SL/IPW switch, which is enabled in erase mode only, to provide a conductive path from the VBULK external signal to the source and isolated p-well of the matrix array;

- the DECS switch, transferring the negative voltage provided on the VNEG node to the signal DECS to be applied to the unselected WLs in read mode and to selected WLs in erase operation.

The design challenges are not high for the first four switches which require simple level shifter schemes similar to those in Figures 5-15 and 5-16. Conversely, the integration of the DECS switch requires more attention as, in such a case, the dynamic range of the control voltage needs to be shifted from [0 VDD] to [VNEG GND] avoiding reliability issues. Furthermore, the FETs constituting the low level shifter need to be integrated in individual isolated substrates.

The W/L ratio of the output drivers for all the switches should be high enough to guarantee current drive capabilities. In particular the SL switch should be able to deliver a current that is sufficient to program 10 cells in parallel. Additionally, minimal L dimensions are generally avoided to increase immunity to reliability. Indeed, the circuits constituting the HV management system should be robust to sustain long periods of HV biasing, corresponding to the total cumulative period of flash programming and erasing.

(v) Logic block

The logic combinatory block includes minimal functionality to enable address pre-decoding, operating mode switching and control over the LV signals enabling the different HV blocks. The operating mode is controlled using the signals MODE_LV<0:1> determining whether the array is in idle, read, program or erase operation. Sector erase functionality is enabled raising the SCTERA_LV signal. The set of address pre-decoders is enabled when the user signals the correctness of the address using the external signal DECODENABLE_LV. The discharge of the HV nodes after an erase is a very important issue due to the high parasitic capacitances associated with the source and drain nodes. A fast discharge is undesirable as it could result in destructive latch-up issues. Consequently, a functionality has been added to slowly ground the LBLs and the source of the array after an erase operation.

5.3.2 Characterization

(i) Testing methodology

A Verigy 93000 single density P1000 SOC ATE tester model has been used to characterize and test the chip functionality. The read process is the most critical, as in absence of sense-amplifiers, the IO voltage has to be forced and the read current measured. The erase mode consists of a high–level digital pattern that performs a sector erase operation. The program mode adopts a high level digital pattern to interface with a pattern generator. A low–level algorithmic memory pattern cycles through the addresses and array topology. A clock cycle with 1MHz base rate has been used during all the testing. In order to generate the high voltage biases beyond the range of the tester, simple amplification logic on the probe card has been adopted for all the appropriate signals. This circuitry has been designed to be quickly powered up and powered down and simply multiplies the incoming signal by a predefined factor.

(ii) DC and transient results

Static DC testing is performed to verify the absence of current leakage in all operating modes. Subsequently, the HV voltage management system and the basic functionality of the decoding circuits is tested. The device is placed in READ mode by controlling the MODE_LV<0:1> signals and a single address is characterized varying the HV biases VX and VY and measuring the cell current in the cell. A BL read voltage of 0.7V is used. In read mode, the cell access time is reduced by firstly applying the HV biases to supply the decoding circuitry and subsequently decoding the address of the cell (*hot* switch). Since the HV biases are relatively low in this operating regime, the signal can be maintained high for the entire read operation without causing degradation of the state of the cell. The cell currents for the different signals of the IO<0:9> bus measured as a function of the HV voltage bias VX applied on the decoded WLs, are illustrated in Figure 5-19(a). An average threshold voltage of approximately 2.5V is found with a large spreading in the stored charge due to process variability as previously shown in Chapter 2. A similar measurement has been performed in Figure 5-19(b), assessing the cell current as a function of the decoder supply voltage VY. For decreasing VY, the pass gates of the BL decoder close and the current decreases. A VY of 3V is sufficient to bias the BL decoder. Similar characterization has been performed for another address obtaining different level of currents due to variability (c-d). In (a) and (c), VY is 5V, while in (b) and (d) VX is 5V.

Transient characterization has been performed for a limited range of addresses to analyse the current variation with respect to program/erase time, and to verify the functionality of the HV switches. The cells are initially programmed using a progressive program algorithm (Figure 5-1 in Chapter 3) and measuring the IO current after each program pulse. The program pulse is applied entering in PROGRAM mode using the MODE_LV control signals. Subsequently, the address is decoded and the DECODENABLE_LV signal raised, while keeping the HV biases grounded. In this way, the path from the HV external biases to the WLs and BLs of the flash cells is created before biasing the HV pins to reduce

degradation and better control the applied pulses on the terminals of the cell (*cold* switch).

The exponential decrease of the current with time reported in is in good agreement with previously shown programming dynamics. Only 5 bits are programmed in parallel, using a 1010101010 mask on the IO bus. Erase dynamics is also shown in Figure 5-20(b).

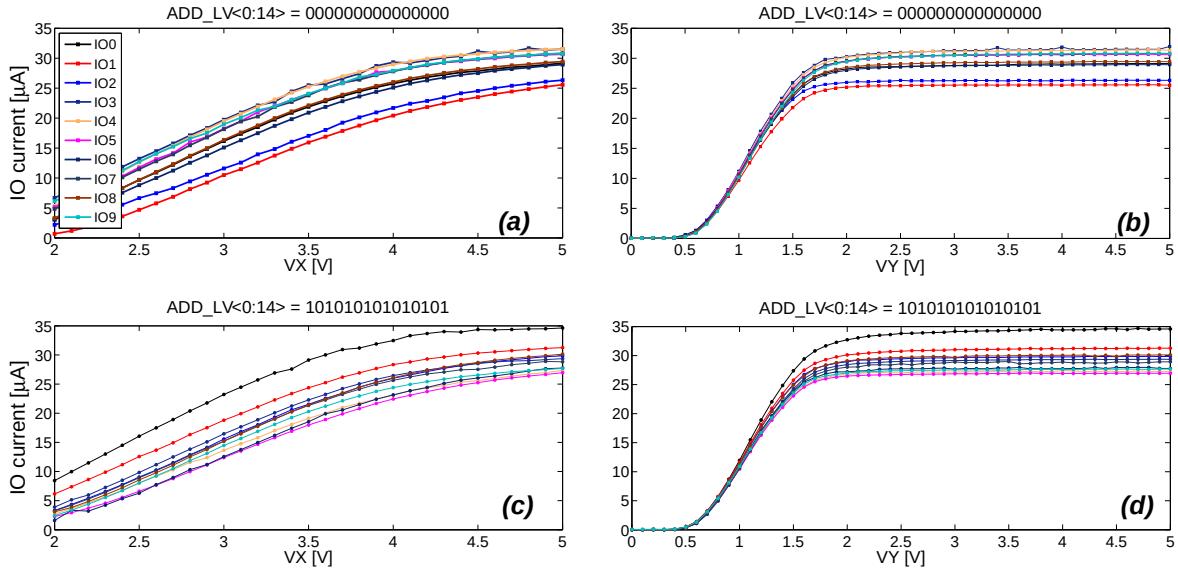


Figure 5-19 – (a) Measured IO current as a function of the HV supply voltage VX applied on the selected WLs and supply the WL decoder. (b) Measured current as a function of the VY bias supplying the pass-gates in the BL decoder.

Sector erase functionality has been verified for a limited set of 16 addresses. In Figure 5-21, the cell current is represented as a function of the address index and as a function of the IO index. The cell current in the virgin state is measured in the top-left corner of the array. Process variability effect are also visible in this case. In the following plot, all the cells of the sector have been programmed at $V_{IO} = 4.2V$ for $4\mu s$. The current in the programmed cells drastically decreases. Progressively applying $500\mu s$-long sector erase sequences, the cell current increases, reaching values comparable to the initial state. An erase voltage of (VNEG=-8V; VBUL=+8V) has been used. The cells have been read in DMA with VX=5V, VY=5V and VNEG=-2V.

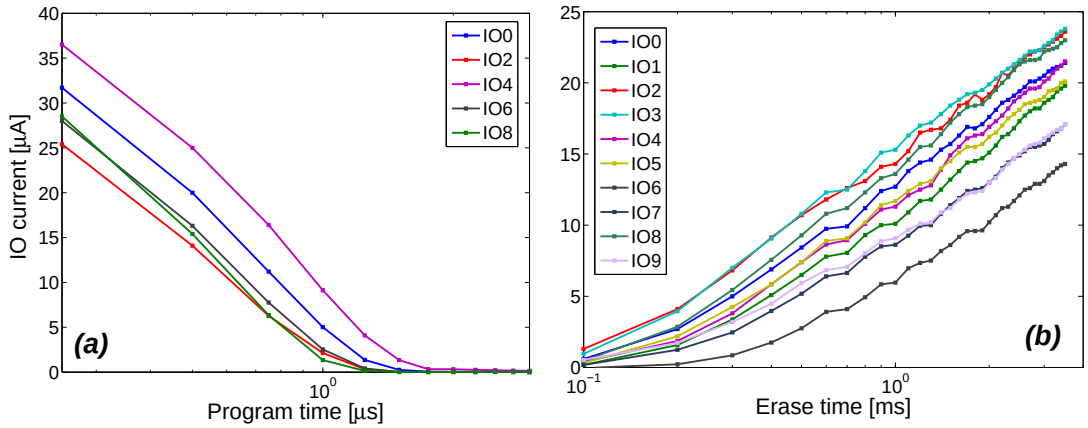


Figure 5-20 – Measured IO current as a function of the total programming (a) and erase (b) time. In (a) an IO mask 1010101010 has been applied, and consequently only the currents on the programmed IOs are shown. Programming is performed at VX=8.5V, VY=7V, V_{IO} = 3.8V. In (b), page erase is performed at VX=5V, VBULK=8V, VNEG = -8V, for a given address, and the current is measured. In both cases, read is performed at VX=5V, VY=5V, V_{IO}=0.7V.

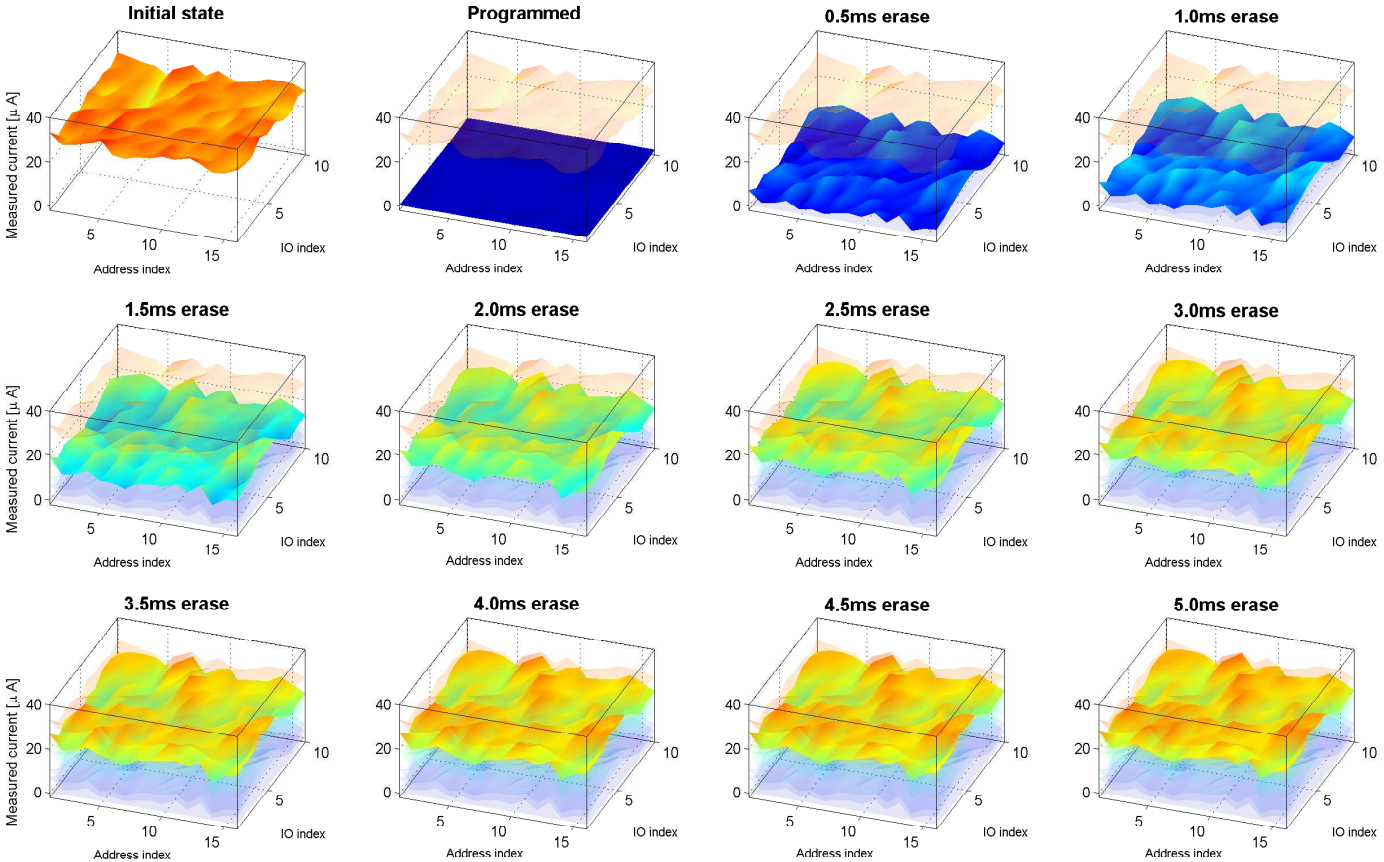


Figure 5-21 – Progressive sector erase operation: the IO currents are measured in read mode after each operation. In the plot at the top-left corner, the current of the virgin cells is plotted with respect to the address index and IO index. Subsequently, the entire sector is programmed and thus the currents of all the cells decrease to a value below the accuracy range of the instrument. In a second step, the currents progressively raise after the application of 500μs pulses.

(iii) Complete memory array testing

Having assessed the basic functionality of the memory in all the operating modes (read, programming, erase, sector erase), different program and erase patterns have been defined and applied to the entire range of addresses to verify the functionality of the decoding scheme and analyse statistical V_{th} distributions. A pattern is represented by a map defining which cells need to be programmed/erased. The tester scan all the WLs and BLs of the array programming/erasing the memory following this scheme. After the application of the pattern, all the cells are read and the correctness of the decoding scheme is verified.

A sector erase operation is initially applied on a fresh device to erase all the entire matrix. Subsequently, a diagonal program pattern is used to program the cells on the main diagonal of the memory sector (BL index equal to the WL index). The remaining cells are left in the erased state. Figure 5-22 shows in colour the IO currents of all the cells read after the diagonal program pattern application. Each point in the image corresponds to a specific cell, with colour amplitude coding. Higher currents are indicated in white and correspond to erased cells. Programmed devices show no current flow and are represented in black. The left Figure is divided in two parts representing the two memory sub-sectors. The main diagonal is shown starting from the middle of the plot, due to the logical ordering of the BLs. The right image shows a zoom of the first 64x64 cells, where the cells in the diagonal have been correctly programmed.

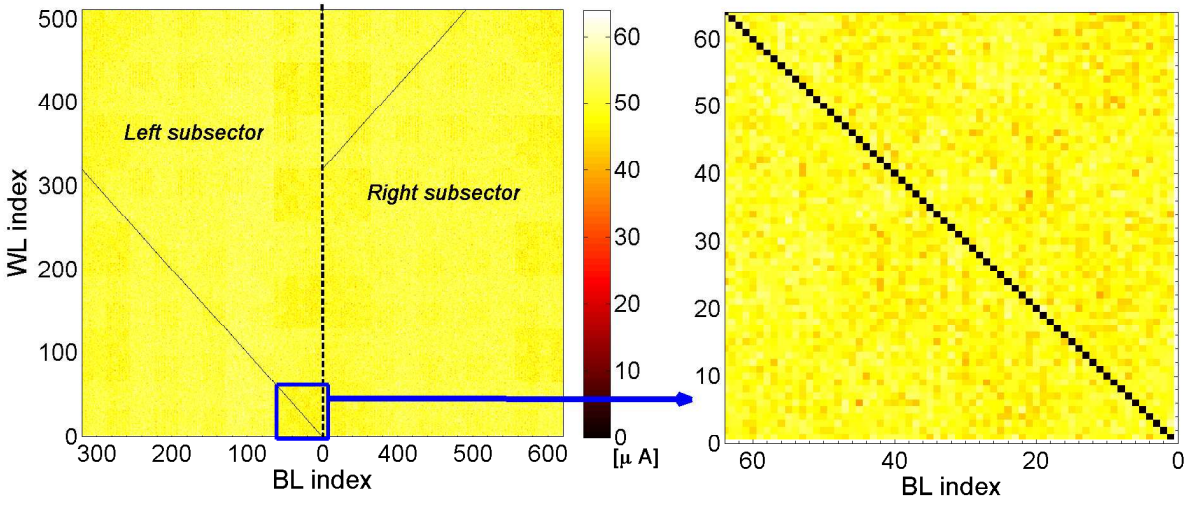


Figure 5-22 – Measured current versus BL and WL indexes for all the flash cells in the 40KB memory sector after the application of a diagonal pattern. Erased devices show high drain currents and are indicated in yellow. Good uniformity is obtained. On the right, a limited range of 64x64 devices is represented.

The good uniformity in the measured current of erased devices is also confirmed for other patterns. In Figure 5-23 the entire sector has been erased and programmed with a checker-board pattern. The read current is then measured for all the devices. In (a),

an address range of 64x64 cells is indicated while in (b) a random position is taken in the address range, and 128x128 cells are shown. In (c) the distribution of the measured currents is shown. The Gaussian distribution indicates higher mean with respect to Figure 5-19. This is due to the fact that the conductive cells have undergone two erase procedures and are consequently over-erased. During read operation, the unselected WLs have been biased at -2V to avoid leakage effects on the read currents.

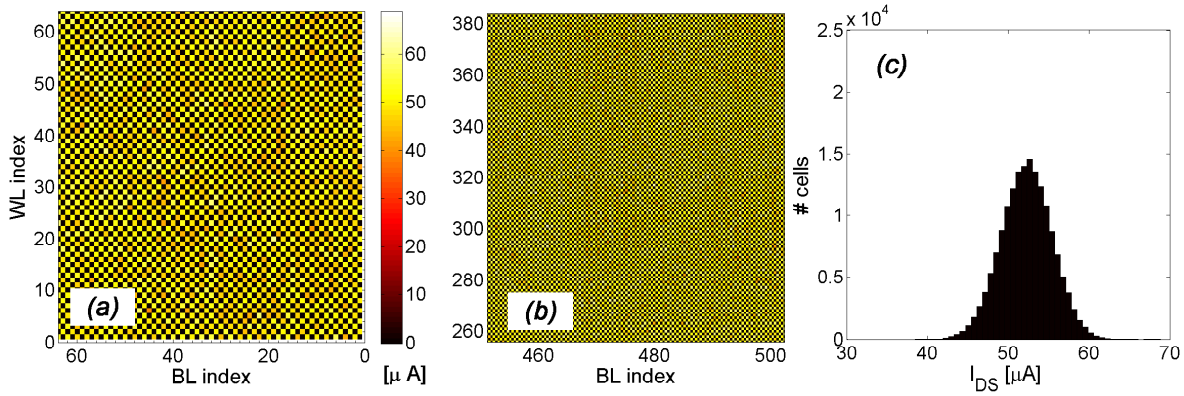


Figure 5-23 – Measured current versus BL and WL indexes for a 64x64 portion (a) and 128x128 portion (b) of the memory sector after the application of a checkerboard pattern. In (c), the distribution of the measured read current for the erased devices is presented.

Subsequently, after a new sector erase sequence, the sector is programmed with the IO stripe pattern; results are presented in Figure 5-24. Only IOs with odd index have been programmed. Since the BL range is divided in 10 blocks of 64 cells per IO, black stripes corresponding to the programmed devices are identified. The current of over-erased devices is further increased.

Figure 5-25 illustrates the two different current distributions that are found for the erased state. The distribution inside the blue box indicates the cells which have undergone a recent program sequence and only a single erase after it. The cells of this distribution are represented by the yellow/red cells in (a) and correspond to the programmed devices in Figure 5-23(a). The cells that belong to the distribution inside the red box are overerased and highlighted in (c) in yellow/red. They correspond to the bright spots in Figure 5-23(a) and have high read currents.

Word stripe and inverse word stripe patterns have also been tested by programming alternated WLs. Results in Figure 5-26 still indicate the presence of both the IO stripe pattern and checkerboard patterns. The double Gaussian distribution appears less pronounced as less cells are in an over-erased state. After this sequence, the inverse WL stripe pattern preceded by a sector-erase is applied. The WL stripe patterns consecutively applied cycle all the cells of the memory array and restore the Gaussian distribution of erased cells (Figure 5-27).

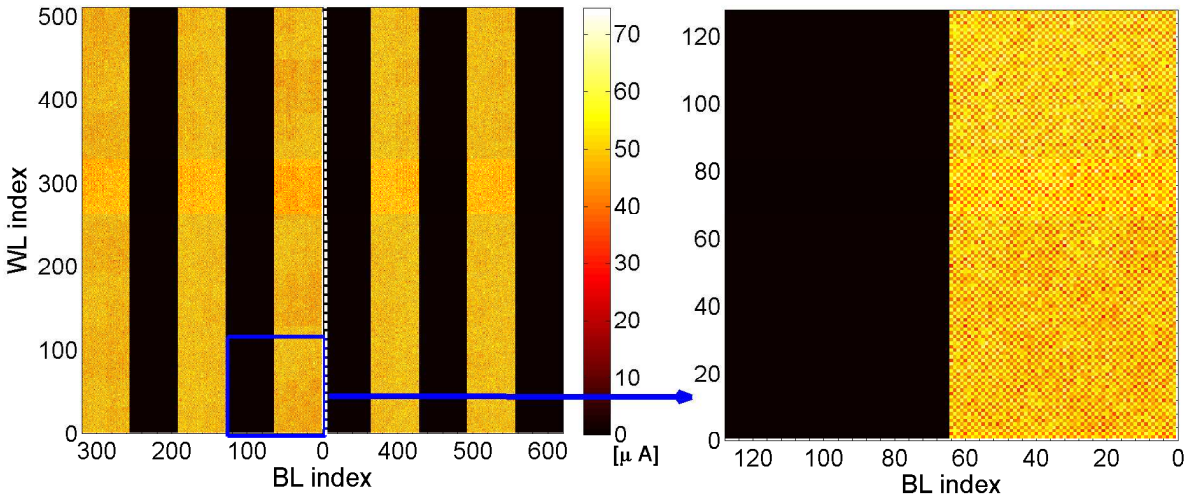


Figure 5-24 – Same setup as Figure 5-22 but after the application of the IO stripe pattern.

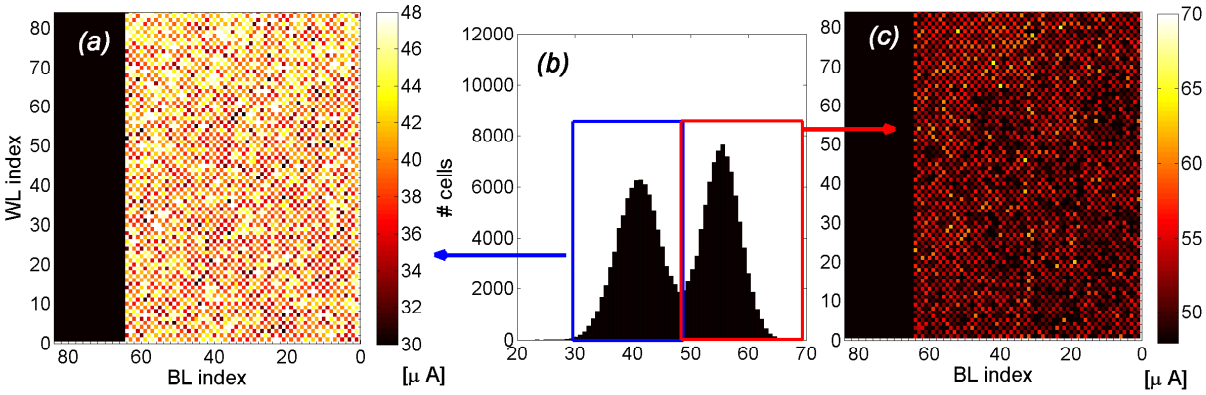


Figure 5-25 – Detail of the cell current for the first 85x85 devices after the application of the IO stripe pattern. In (b), the current distribution over the entire sector indicates the presence of two erased states. Some cells, in particular those which have not been programmed with the checker-board and diagonal patterns, have been erased several times in the previous testing sequences and present much higher currents (red cells in (c)). On the other hand, the devices corresponding to (a) and in blue have been erased only once after the last programming sequence. This effect is confirmed by the fact that both the distributions follow a complementary checker-board pattern.

5.4 Conclusion

In this chapter, the applications of the NVM-SPICE model to IC memory design have been described focusing on the evaluation of the program/erase performance metrics and on pulse analysis. In particular, several figure of merits have been evaluated as a function of the applied pulse configuration to show the model capabilities in assessing performance

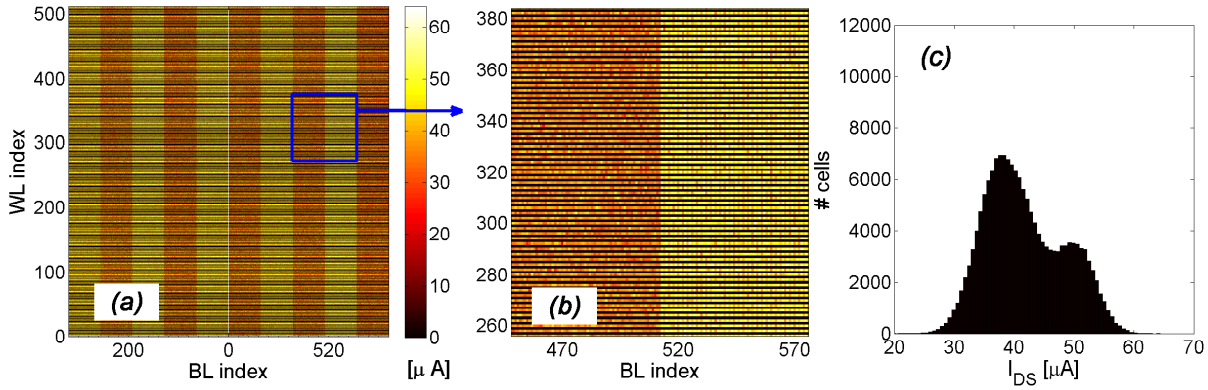


Figure 5-26 – Same setup as Figure 5-22, but after the application of the WL stripe pattern programming one entire WL over two. In (b) a limited set of cells in the matrix is shown. The IO stripe pattern is still recognizable in the over-erased devices and is highlighted in the double Gaussian distribution in (c).

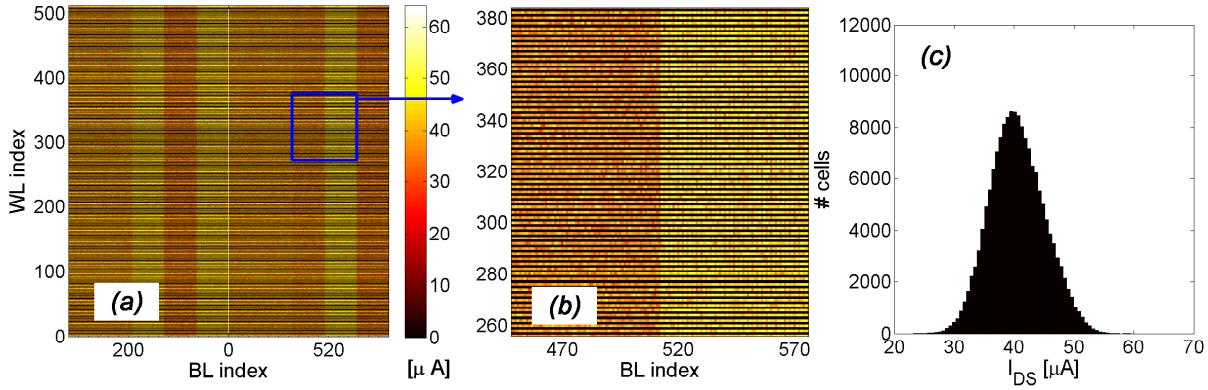


Figure 5-27 – Same setup as Figure 5-22 but programming with the inverse WL stripe pattern. The distribution in (c) is minimally skewed due to residual history effects.

tradeoffs for NVM design. Additionally, a methodology for pulse configuration extraction is established to improve cell performance depending on the applied biases. A 40KB memory sector has been designed, integrated and characterized. Compact parametric models for the BLs and WLs have been defined and can be used together with NVM-SPICE for decoder load and performance evaluation. HV circuit blocks including HV switches, level shifters and decoder have been described. The modes of operation are detailed, focusing on the read, program and erase sequences for address decoding and pulse application. Finally, characterization and testing results are reported to illustrate the functionality of the memory sector in the different operating regimes and for several test patterns.

Conclusion and outlook

Summary of the work

In this thesis, a complete and efficient design-oriented modelling methodology for flash memory devices has been presented. It includes an investigation of device electrostatics and charge trapping effects using numerical approaches, the formulation of a compact model for the flash device including key elements to simulate DC, transient and parasitic effects, and its application to IC memory design.

The flash memory device has been analysed by the means of TCAD modelling and the role of intra cell couplings investigated in detail. The dependence of the coupling coefficients on bias voltages and geometrical dimensions of the cell has been extracted with 2D and 3D simulations. Short-channel effects and the impact of velocity saturation on the characteristics of ultra-scaled devices have been highlighted. Furthermore, 3D AC analysis permitted the formulation and validation of an analytical model for the control gate-floating gate ONO capacitance, which is scalable along all the physical dimensions of the device. Using the charge balance paradigm, a complete surface potential-based compact model for the flash memory cell has been developed. The model takes into account the bias dependencies of the coupling coefficients to accurately compute the floating gate potential voltage in the cell and to reproduce the electrostatics of the device. Compact transient models validated on a general TCAD approach are used to reproduce the threshold voltage variation as a function of time during program and erase operation. The design and integration of flash test-structures in 65nm NOR technology and their complete characterization permitted the extraction of the model parameters in all operating conditions and demonstrated the excellent correlation between measurements and model simulation results.

The investigation of parasitic effects linked with device degradation and endurance required more attentive physical investigations. In particular, the effects of interface and oxide defects in CMOS oxides have been analysed with a multiphonon-assisted charge trapping approach, in AC, DC and transient regimes. The accuracy of the approach relies using a rigorous self-consistent calculation of the device electrostatics and permitted to extract trap concentration profiles for different stress conditions. A novel impedance model taking into account charge trapping effects in CMOS dielectrics permits to analyse the frequency response of the defects with respect to capacitance and conductance curves. The impact of stretch-out effects on the characteristics and oxide leakage by trap assisted tunneling are also investigated. Transient effects have been analysed criticizing the trap

equilibrium model adopted in AC. Charge pumping simulations are used to validate the approach in transient.

Parasitic effects have been subsequently added in the compact model of the flash device. The effects of device degradation have been decoupled separating the contribution on the electrostatics in DC read mode, and the reduction of program efficiency in transient regime. This permitted to develop an empirical approach for the simulation of device aging and degradation. Furthermore, complementary effects, such as BL disturb, device cross-talk and process variability, have been added to the compact model.

From a memory design perspective, the SPICE model has been integrated in a 65nm design kit for embedded low power CMOS. The model has been used to analyze cell performances in read and program/erase operations. In particular, parametric analysis has been used to evaluate the performances of the cell in program operation, taking into account several figures of merit. Parametric cells for entire BLs and WLs have been developed and can be used to evaluate the load of the decoder drivers. Finally, a 40KB memory testchip for Smart Card applications has been designed, integrated and functionally tested in all the regimes of operations.

Scientific contribution

The adopted numerical and compact modeling methodology addresses several points of significant scientific interest mainly focused in the first part of this work. In this work, the point of highest theoretical interest relies on the numerical Poisson-Schrödinger solver developed for the investigation of advanced CMOS technologies and fully detailed in Annex C. The model can be applied to evaluate the performances of a broad range of CMOS devices (from MOSFET structures to HKMG FD-SOI devices). The effects of heterostructure composition, interfacial layers, material band structures and strain on device electrostatics and tunneling currents can be analyzed. A key element of the solver consists of the multiphonon charge trapping model which has been applied to the analysis of the properties and the effects of the defects on MOSFET oxide stacks. Based on an accurate determination of the charge capture/emittion time constants in the oxide depth and as a function of the trap energy, the multiphonon-assisted charge trapping phenomena in DC, AC and transient regimes have been investigated. The novel impedance model rigorously takes into account the frequency, bias and temperature dependencies of the characteristics after a wide range of stress conditions. The interest in this case is also reinforced by the capability to investigate the regions of the dielectrics that could be characterized in the CGV analysis, charge pumping characterization and trap-assisted tunneling.

Concerning flash modeling, 3D TCAD analysis permitted the formulation of a novel and highly scalable model for the ONO capacitor, controlling the coupling of the cell and performances for all conditions of operation. The role of charge sharing, accurate overlap and fringing capacitance calculation and velocity saturation effects on the coupling coefficients permitted to investigate their bias and scaling dependencies. Finally, high accuracy has been achieved in DC and transient regimes over a large range of biases for all

standard operating regimes. A rigorous model based on NEGF has been used to analyse the barrier transparency and provides an excellent reference for other numerical or compact approaches.

The simulation of parasitic and long-term effects exploits empirical models built on considerations from numerical simulations. For device ageing, a novel methodology to separate the effects of degradation and distinguish the impact on the electrostatics and the program/erase efficiency variations, has been formulated and established. The approach reinforces the understanding of the charge trapping phenomena in flash devices. Correlation has been found between post-stress parameter measurements of the equivalent transistor device and of the floating gate cell. A novel empirical solution has been formulated, which is able to reproduce disturb effects on the cells in a wide range of stresses and bias conditions.

Industrial value and applications

The importance of this work in industry extends to various fields of technology development and IC design. The PS modelling tool has been adopted by STMicroelectronics to evaluate the performances of advanced CMOS technologies below the 32nm node. The tool permits to efficiently investigate new device structures, the impact of technology boosters on critical parameters (inversion capacitance, mobility, effective thickness, etc.) that determine strategical technology development. Specifically, within the STMicroelectronics-IBM joint project on eNVM 65nm technologies, process integration and TCAD teams adopted the model to understand physical mechanisms underlying degradation and process/structure optimization to reduce device ageing. Several features of the modeling tool could be also applied to the validation of numerical models included in standard commercial TCAD tools. The communication and collaboration between reliability, characterization, process integration and TCAD teams have been strongly improved, with a radically novel and physical perspective on defect understanding and characterization methodology.

TCAD modeling of flash devices and the analysis of coupling effects led to the development of simulation decks presently adopted by the modeling teams to support process integration. The isolation of scaling effects on cell couplings can be used to define a physical modeling and optimization methodology.

The major achievement in terms of industrial value consists of the Verilog-A implementation of the compact model NVM-SPICE for the fully comprehensive simulation of floating-gate devices. The model has been integrated in a 65nm design kit and delivered to IC designers in IBM and STMicroelectronics. The Verilog-A solution permits the code to be portable and compatible with different SPICE simulators. Applications of this model are quite relevant, as the designers not only can evaluate the cell characteristics in static DC regime (read conditions), but can also analyse the device performances in program/erase operation. This grants the possibility to resolve trade-offs in all the regimes of operations and apply programming/erasing algorithms to accurately control the final V_{th} distribution of the device. Two patent disclosures have been submitted using the compact model to find solutions for V_{th} distribution narrowing and program/erase performance optimization.

Statistical distributions and process corner simulations can be adopted for variability and simulation of the cell in worst–case conditions. Although the solutions to model disturb and ageing effects are presently empirical approaches, they can be adopted for the assessment of critical trade-offs in applications in harsh environments.

The design, integration and testing of the 40KB memory testchip for Smart Card applications allowed to demonstrate the direct application of the model in IC memory design. BL and WL models are valuable to avoid handling long SPICE simulations and permitting the sizing of decoder last stage drivers. Considerations on the power consumption of the cell in CHEI regime permit a more attentive optimization of the device in program. The testchip is presently used to evaluate the cell performances in the matrix environment and the degradation of the surrounding HV circuitry.

Outlook and perspectives

This work opens several perspectives and future outlooks over a wide range of modeling domains. Concerning numerical modeling, the developed PS solver could be extended to support 2D and 3D simulations of complete devices. The 1D slices indeed include ideal transport conditions across the dielectric, taking into account both TAT and direct tunneling components. A pseudo-2D Schrödinger approach could be adopted, simulating the device in transversal slices and computing the charge distribution. A finite difference solver can be used for the solution of the Poisson equation in 2D. Using full-band structure models, effort is currently undergoing to evaluate the carrier mobility with a rigorous physical Kubo-Greenwood approach. Surface scattering, optical/acoustical phonon scattering and Coulomb scattering contributions could be taken into account to analyze the mobility dependence with respect to the electrical field. Another feature that makes use of the charge trapping model consists of remote Coulomb effects to analyse mobility reduction due to defects at the interfaces. The multiphonon-assisted charge trapping model can be extended introducing metastate defect transitions or structural relaxation after capture events or during recovery. The creation of the defects could be modeled as well introducing defect creations models. From an application perspectives, investigations are undergoing related to complex dielectric stacks in sub 45nm technologies and considering new device architectures. The model is also being validated on single defect analysis experiments to assess probed regions with this technique and evaluate the trap capture/emission rates by reverse modeling.

Concerning the NVM-SPICE model, although Verilog-A implementation is portable on most of commercial SPICE simulators, a general C-based solution should be developed to further improve computational efficiency. This process and the model integration in the simulation engine is normally performed by EDA software vendors and requires the model code to be shared to the community. Furthermore, a request to the Compact Modeling Council for its adoption as industrial standard for the flash memory cell could be proposed. Non-local approaches for CHEI could be investigated using more computationally intensive MC simulations. While disturb, CHEI and endurance models could be improved in terms of

voltage and dimension scalability, it is expected that new features will be also introduced in future flash technologies. For instance, device and technology scaling requires the reduction of the thickness of the oxide tunneling layers that causes both retention and endurance issues. Therefore, models for these two aspects need to be revised in future technology nodes. The ONO dielectric stack could also be replaced by HighK materials to improve the coupling; gate disturb effects and TAT leakage through the ONO stack represent serious issues for retention. Therefore, with the continuous scalability of flash devices down to 32nm and 22nm technology nodes, new model solutions should be taken into account, accounting for quantum effects and device reliability.

The market of NVM and specifically flash memory appears increasing and novel consumer electronic devices emerge regularly. Following the market demands as well as Moore's law, industry develops flash memory in ultra deep submicron nodes in a short amount of time. Advanced modeling methodologies must be developed to respond to the consumer and manufacturers needs for performant flash memory devices in terms of integration density, technology and design efficiency, data safety, reliability, etc. They should be able to reproduce physical effects and their impact at design level in a compact and simulation-efficient manner. This thesis aims at contributing towards this ambitious and important goal.

Appendices

Appendix A

Surface-potential based models for MOSFETs

A.1 Surface potential analytical calculation

We recall from Chapter 2 the surface potential equation at position y along the channel:

$$\begin{aligned}(V_{FB}-V_{fb}-\psi_S(y))^2 &= \gamma^2\psi_T\left(e^{-\frac{\psi_S(y)}{\psi_T}}+\frac{\psi_S(y)}{\psi_T}-1+\right.\\ &\left.+e^{-\frac{2\psi_F}{\psi_T}}\left(e^{\frac{\psi_S(y)-V_C(y)}{\psi_T}}-\frac{\psi_S(y)}{\psi_T}-1\right)\right),\end{aligned} \tag{A.1}$$

The charge sheet model approximation introduced by Brews [17] and reprised by Tsividis [272] is applied, calculating the surface potential only at the source and at the drain locations. The drain current is thus given by:

$$I_{DS}=\mu\frac{W}{L}C_{OX}[F(L)-F(0)], \tag{A.2}$$

where μ indicates the carrier mobility in the channel and $F(y)$ is a function of the surface potential [272] expressed as:

$$\begin{aligned}F(y) &= (V_{FB}-V_{fb}-\psi_S(y))\psi_S(y)-\psi_T\psi_S(y)-\psi_T\gamma\sqrt{\psi_S(y)-\psi_T}\\ &\quad -\frac{1}{2}\psi_S(y)^2-\frac{2}{3}\gamma(\psi_S(y)-\psi_T)^{\frac{3}{2}},\end{aligned} \tag{A.3}$$

Closed form expressions of the surface potential have also been derived and are in use in most of compact surface potential-based models [28, 108, 273, 274]. In particular, Eq. 2.1 can be reduced to:

$$(V_{FB}-V_{fb}-u\psi_T)^2=\gamma^2\psi_T H(u), \tag{A.4}$$

where $u = \psi_S/\psi_T$, and taking $H(u)$ as:

$$H(u) = e^{-u} + u - 1, \tag{A.5}$$

While this approximation is well suitable in the majority of operating conditions, compact approaches require that continuity is preserved for all the bias conditions. Industrial surface potential compact models integrate a different expression for the function $H(u)$ valid also at flat-band conditions [109,275] ($\psi_S = 0$), where the traditional approaches face difficulties and the equation is ill-conditioned. In this view, the function takes the form of:

$$H(u) = e^{-u} + u - 1 + 2\Delta_n(\sinh(u) - u), \tag{A.6}$$

where $\Delta_n = e^{-(2\psi_F+V_C)/\psi_T}$. Indeed, this form does not affect the numerical calculation of ψ_S and permits the calculation also at flat band where the conditioning is valid, i.e. $(\partial\psi_S/\partial V_{FB})_{V_{FB}=V_{fb}} = 0$. Analytical solutions based on Taylor series approximation are commonly adopted to solve this equation in closed form [28, 109].

A.2 Intrinsic charges calculation

In semi-analytical approaches where $\psi_S(y)$ is solved at each position in the channel, the gate $q_g(y)$, bulk $q_b(y)$ and inversion $q_i(y)$ charge densities at position y in the channel are obtained using:

$$\begin{aligned}
q_g(y) &= sgn(\psi_S)\gamma C_{OX}\sqrt{\psi_T e^{-\frac{\psi_S}{\psi_T}} + \psi_S - \psi_T + e^{\frac{-2\psi_F}{\psi_T}}\left(\psi_T e^{\psi_S - \frac{V_C}{\psi_T}} - \psi_S - \psi_T e^{-\frac{V_C}{\psi_T}}\right)},\\
q_b(y) &= -sgn(\psi_S)\gamma C_{OX}\sqrt{\psi_S + \psi_T(e^{-\psi_S\psi_T} - 1)},\\
q_i(y) &= -q_g(y) - q_b(y),
\end{aligned} \tag{A.7}$$

where numerical integration over the channel length is performed to obtain the total terminal charge. Notice that the determination of the total gate charge Q_G is of critical importance as it is used in the charge balance equation for the implicit calculation of V_{FB}.

In a charge-sheet model, closed-form expressions for the terminal charges can be obtained, calculating the charges directly with [123]:

$$\begin{aligned}
Q_B &= -C_{OX}\frac{\gamma(A(V_{FB} - V_{fb}) - \gamma B - C)}{V_{FB} - V_{fb} - B - \gamma A},\\
Q_G &= C_{OX}\frac{(V_{FB} - V_{fb})(V_{FB} - V_{fb} - 2B - \gamma A) + D + \gamma C}{V_{FB} - V_{fb} - B - \gamma A},\\
Q_I &= (-Q_B - Q_G),
\end{aligned} \tag{A.8}$$

in which the coefficients:

$$\begin{aligned} A &= \frac{2}{3}\frac{\psi_S(0)+\sqrt{\psi_S(0)\psi_S(L)}+\psi_S(L)}{\sqrt{\psi_S(0)}+\sqrt{\psi_S(L)}}, \\ B &= \frac{\psi_S(0)+\psi_S(L)}{2}, \\ C &= \frac{2}{5}\frac{\psi_S(0)^2+\psi_S(0)^{1.5}\sqrt{\psi_S(L)}+\psi_S(0)\psi_S(L)+\sqrt{\psi_S(0)}\psi_S(L)^{1.5}+\psi_S(L)^2}{\sqrt{\psi_S(0)}+\sqrt{\psi_S(L)}}, \\ D &= \frac{1}{3}(\psi_S(0)^2+\psi_S(0)\psi_S(L)+\psi_S(L)^2), \end{aligned} \tag{A.9}$$

are expressed in closed form from the surface potentials at the source $\psi_S(0)$ and the drain $\psi_S(L)$.

In compact surface potential-based approaches simpler expressions are found. A common solution adopts the Symmetric Linearization Method (SLM) for the linearization of the bulk charge as a function of the voltage drop in the channel or as a function of ψ_S [276]. The bulk charge can thus be written as:

$$q_b(y) = -\gamma C_{OX}\left[\sqrt{\psi_0-\psi_T}+\alpha_b(\psi_S(y)-\psi_0)\right], \tag{A.10}$$

where ψ_0 is the reference surface potential at a specific point in the channel and α_b represents the linearization coefficient. While several choices could be made for ψ_0 and α_b, the scheme adopted in surface potential based compact models [109] imposes:

$$\begin{aligned} \psi_0 &\equiv \psi_m = \frac{\psi_S(0)+\psi_S(L)}{2}, \\ \alpha_b &\equiv \frac{\partial q_b}{\partial \psi_S}\Big|_{\psi_S=\psi_m}, \end{aligned} \tag{A.11}$$

A linearized expression can be found also for the inversion charge $q_i(y)$ assuming:

$$q_i(y) = C_{OX}(q_{im}+\alpha_m(\psi_S(y)-\psi_m)), \tag{A.12}$$

and having $\alpha_m = 1+\frac{\gamma}{2\sqrt{\psi_m-\psi_T}}$. The midpoint inversion charge normalized over C_{OX} is obtained with:

$$q_{im} = -\frac{\gamma\psi_T\Delta(\psi_m,V_{Cm})}{\sqrt{\psi_m-\psi_T+\psi_T\Delta(\psi_m,V_{Cm})}+\sqrt{\psi_m-\psi_T}}, \tag{A.13}$$

where V_{Cm} denotes the imref splitting at midpoint and:

$$\Delta(\psi_S,V_C) = \exp\left[\frac{\psi_S-2\psi_F-V_C}{\psi_T}\right]. \tag{A.14}$$

This latter quantity can be determined analytically solving Eq. A.1:

$$\Delta(\psi_m, V_{Cm}) = \frac{1}{2}[\Delta(\psi_S(0), V_C(0))] + \Delta(\psi_S(L), V_C(L))] - \frac{\psi}{4\gamma^2\psi_T}, \tag{A.15}$$

being $\psi = \psi_S(L) - \psi_S(0)$. Applying the Ward-Dutton partitioning scheme [277] and integrating along the length, the closed-form expressions of the terminal charges as a function of the inversion charge at the midpoint are obtained:

$$\begin{aligned}
Q_G &= C_{OX}\left(-q_{im} + \gamma\sqrt{\psi_m - \psi_T} + \frac{\psi^2}{12H}\right), \\
Q_B &= C_{OX}\left(-\gamma\sqrt{\psi_m - \psi_T} - \frac{(1-\alpha_m)\psi^2}{12H}\right), \\
Q_D &= \frac{1}{2}C_{OX}\left[q_{im} + \frac{\alpha_m\psi}{6}\left(1 - \frac{\psi}{2H} - \frac{\psi^2}{20H^2}\right)\right], \\
Q_S &= (-Q_G - Q_D - Q_B), \\
Q_I &= Q_S + Q_D,
\end{aligned} \tag{A.16}$$

considering $H = \psi_T - \frac{q_{im}}{\alpha_m}$. The expression of the drain-source current I_{DS} becomes:

$$I_{DS} = \frac{W}{L}\mu C_{OX}(q_{im} - \alpha_m\psi_T) \tag{A.17}$$

A.3 Mobility effects

The increase of electric field in advanced devices causes carrier scattering to become more and more probable at the Si/SiO_2 interface. Carrier mobility in the inversion layer is negatively affected by this phenomenon. Two effects are usually considered when modeling the mobility degradation, separating the dependence on normal fields (mobility reduction) and the dependence on lateral fields (velocity saturation) [122].

With the scaling down of MOSFET dimensions, quantum confinement strongly increases. Quantum-mechanical self-consistent approaches identified the increase of spacing between the sub-bands in band valleys as the main responsible of mobility reduction with the increase of the vertical field [278, 279]. However, such models require a large computational effort and therefore semi-empirical solutions should be investigated for adaptability in SPICE simulators.

Having calculated the charge density in the space-charge region q_{bm} and the inversion charge q_{im} in the middle of the channel, the vertical electrical field responsible of mobility reduction is determined with:

$$F_{eff} = \frac{q_{bm} + \eta q_{im}}{\epsilon_{Si}}, \tag{A.18}$$

where $\eta = 1/2$ for electrons [280]. Three mechanisms should be considered in analyzing the role of scattering at the Si/SiO_2 interface. Coulomb scattering mechanism is dominant in subthreshold regime and is due to Coulomb repulsion with charged sites at the oxide interface. The mobility reduction factor by Coulomb scattering μ_{cs} is thus proportional to the inversion charge and inversely proportional to the substrate doping since carrier screening in high doped surfaces prevents mobility degradation. Phonon scattering results from the effects of quantum vibrations of the crystal lattice and, experimentally, it can be seen that it implies a mobility factor $\mu_{ph} \propto F_{eff}^{-1/3}$. Finally, scattering due to surface roughness is a dominant phenomena in strong inversion conditions where the carrier distance from the interface determines the mobility variation. In such a case, for electrons one has $\mu_{sr} \propto F_{eff}^{-2}$. The three contributions are commonly taken into account using Matthiessen's rule [280], with the effective mobility μ calculated as:

$$\frac{1}{\mu} = \frac{1}{\mu_0} + \frac{1}{\mu_{cs}} + \frac{1}{\mu_{ph}} + \frac{1}{\mu_{sr}}, \tag{A.19}$$

More compact approaches exist [280], where semi-empirical parameters are introduced to model the dependencies on the vertical field and provide parameter extraction flexibility:

$$\mu = \frac{\mu_0}{1 + (M_{eu}F_{eff})^{T_{he}} + C_S q_{bm}^2/(q_{bm} + q_{im})^2 + G_R}, \tag{A.20}$$

being μ_0 the low-field mobility and the parameters M_{eu} and T_{he} empirical factors for the mobility degradation caused by surface roughness and phonon scattering of the effective vertical field. The factor C_S is introduced to model Coulomb scattering mechanisms. The factor μ_x describes non-universality effects and also empirically accounts for nonuniform doping. The term $G_R = \mu_0 \cdot (W/L) \cdot q_{im} \cdot R_S$ accounts for the series resistance R_S.

The carriers' drift velocity v in the channel saturates at high lateral fields reaching the value v_{sat}. To model this behavior, semi-empirical models are introduced, determining v with:

$$v = \frac{\mu \cdot \frac{\partial \psi_S}{\partial y}}{\sqrt{1 + \left(\frac{\mu}{v_{sat}} \cdot \frac{\psi_S}{\partial y}\right)^2}}, \tag{A.21}$$

where the lateral field is given by $F_{lat} = -\frac{\psi_S}{\partial y}$. The expression of the drain-source current in Eq. A.17 is thus modified with:

$$I_{DS} = \frac{W}{L}\frac{\mu}{G_{vsat}} C_{OX}(q_{im} - \alpha_m \psi_T). \tag{A.22}$$

The factor G_{vsat} is numerically expressed with:

$$G_{vsat} = \frac{G_{mob}}{L} \int_0^L \sqrt{1 + \left(\frac{\mu}{v_{sat}} \cdot \frac{\psi_S}{\partial y} \right)^2} \, dy \tag{A.23}$$

or analytically, assuming that the lateral electric field increases linearly along the channel and obtaining:

$$\begin{aligned} G_{vsat} &= \frac{G_{mob}}{2} \left[\sqrt{1 + \Gamma^2} + \frac{\ln(\Gamma + \sqrt{1 + \Gamma^2})}{\Gamma} \right], \\ \Gamma &= \frac{2\Theta_{sat}\psi}{\sqrt{G_{mob}}}, \end{aligned} \tag{A.24}$$

where $\Phi_{sat} = \mu_0/(v_{sat}L)$.

Appendix B

NVM–SPICE model extraction

The parameter extraction methodology follows a series of AC, DC and transient measurements where the device is fully characterized. In particular, the approach requires:

(*i*) calibration of the model of the equivalent transistor device in AC based on C_{GB} and C_{GC} capacitance measurements;

(*ii*) calibration of the model of the equivalent transistor device in DC conditions, based on $I_{DS}(V_{CS} = V_{FS}, V_{BS})$ curves in linear and saturation regimes;

(*iii*) calibration of the model of the equivalent transistor device in DC conditions, based on $I_{DS}(V_{DS}, V_{CS} = V_{FS})$ curves and on $I_{DS}(V_{CS} = V_{FS}, V_{DS})$ curves in saturation;

(*iv*) calibration of the model of extrinsic effects in the equivalent transistor device, including gate leakage $I_{GB}(V_{CS} = V_{FS})$, diode capacitance and leakage, gate-induced-drain-leakage (GIDL) on $I_{DS}(V_{DB})$ and weak-avalanche current on $I_B(V_{DS})$;

(*v*) calibration of the DC model for the flash device, enabling the ONO capacitance model and CBM, and extracting from $I_{DS}(V_{CS} = V_{FS})$ curves in read conditions, for erased and programmed states;

(*vi*) calibration of the transient erase model for the flash device, using transient $V_{th}(t, V_{CS} = V_{FS})$ dynamics starting from the programmed state;

(*vii*) calibration of the transient program model for the flash device, using transient $V_{th}(t, V_{CS} = V_{FS}, V_{DS})$ dynamics starting from the erased state;

(*viii*) calibration of the disturb model for the flash device, using transient $V_{th}(t, V_{CB}, V_{DB})$ dynamics;

(*ix*) calibration of the endurance model for the flash device under worst case conditions;

(*x*) definition of process corner and statistical variations from technology targets or statistical analysis.

Appendix B. NVM–SPICE model extraction

Appendix C

Advanced physics in ultrascaled devices with UTOXPP solver

C.1 Introduction

The downscaling of MOSFET dimensions has been a very successful process in improving the performances of CMOS devices as more and more aggressive ITRS requirements are targeted. Conventional scaling down of MOSFET dimensions faces physical and economical limits. The performances of novel solutions such as high-K dielectrics, mechanical strain and substrate orientation, or alternative device structures, such as fully depleted SOI devices, should be evaluated within the perspective of an industrial integration. The effects of these technological boosters on channel mobility and gate leakage have to be clearly quantified.

Technology Computer-Aided Design (TCAD) refers to the use of computer simulations to develop and optimize semiconductor devices. In state-of-the-art commercial TCAD device simulators [104], quantum effects are accounted for using the Density Gradient approximation, that well applies to traditional bulk devices, but risk of being inaccurate for advanced devices such as SOI structures. Moreover, emerging materials significantly challenge the conventional TCAD tools due to the lack of appropriate empirical model parameters calibration.

Among the new technological boosters of advanced devices, it is now well established that applied mechanical strain and channel orientation in substrate engineering can significantly improve carrier mobility. Moreover, alternative architectures, such as multi-gate MOSFETs and Fully-Depleted Silicon-On-Insulator (FD-SOI) devices are emerging. In addition, new materials and process steps need to be added to the manufacturing process flow at every new technological node, to be able to achieve the objectives of the ITRS roadmap.

A successful modeling approach should combine conventional TCAD simulation tools

with physically-based models. The investigation of physical phenomena and the calibration of empirical models implemented in commercial TCAD tools can be performed in this view. In response to the industrial need, new physically-based models have been developed in the academic world. However, their application to industrial environment is rarely effective due to their limited versatility, portability, support and user-friendly interfaces.

In this work, a new TCAD modeling tool is proposed for the investigation of advanced quantum effects, band structure models, quantum tunneling and multiphonon-assisted charge trapping (Section C.2). In Section C.3 case studies and model applications are provided. The model is presented also in [228, 281].

C.2 Model overview

State-of-the-art physically-based models featuring advanced tunneling models, band structure effects, and mobility calculation have been implemented into a 1D full band Poisson - Schrödinger solver named UTOXPP.

In Figure C-1, the features of the modeling tool are presented. The core of the model is represented by a 1D Poisson-Schrödinger solver. Full-band k.p and multi-band EMA models are used to calculate the material bandstructure of up to six different materials [282, 283]. This makes possible the investigation of heterostructure and strain effects present in modern devices and nanoscale technologies. The user has the possibility of calibrating the material parameters on more rigorous simulations, and a specific set of electrical and band structure parameters has been extracted from ab-initio studies for the common materials adopted in CMOS. In the full band models, the oxides have been treated as pseudo zinc-blende materials, adjusting the model parameters to match the band structure obtained from first principles [153, 284].

Semiclassical WKB and quantum tunneling models have been integrated for the calculation of the tunneling current through oxide stacks. An accurate determination of the barrier transmission in complex stacks requires the use of inelastic Non Equilibrium Green Function (NEGF) techniques [128], while a semi-analytical approach based on the WKB approximation can be adopted for simpler barriers [285]. A multiphonon-assisted non radiative trapping model has also been included and permits to accurately calculate the response of interface and oxide defects after electrical stress and extract the trap distribution in oxide layers [228].

The tool is used for DC, AC and transient analysis on CMOS devices. Additional optimization features include charge predication techniques to improve numerical convergence and reduce the simulation time on multicore CPU clusters. Multi-threading capabilities and job scheduling have also been included. For flash memory devices, UTOXPP includes a Newthon-Raphson approach to solve the charge balance equation on the floating gate node [9].

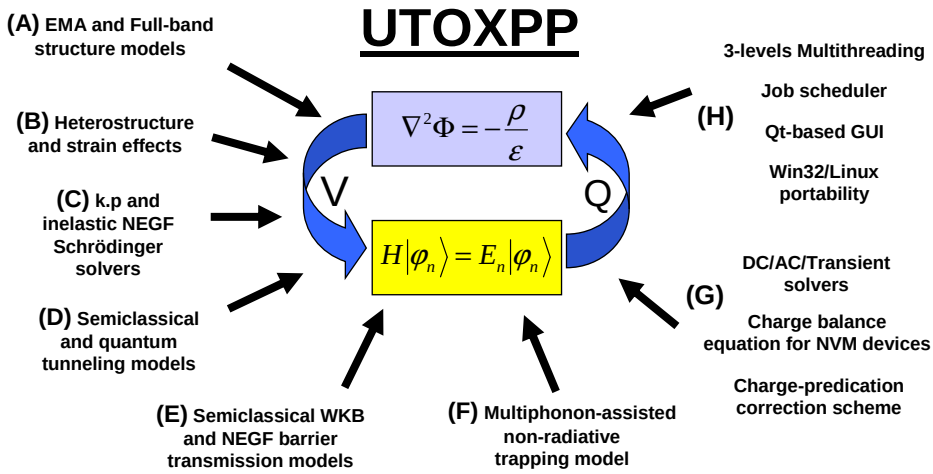


Figure C-1 – Building blocks of UTOXPP.

C.3 Case studies

C.3.1 Electrostatics in strained High-K Metal gate technology

The continuous technology improvements implying the reduction of device size to nanoscale dimensions requires a deep understanding and modeling of nano-scale processes. When complex hetero–structures including strained layers are introduced in the device, understanding nanoscale physical mechanisms and providing accurate predictions becomes critical for reducing technology development costs and improving product yield. All these properties are combined in modern devices, where the presence of atomic layers at the interfaces between different materials is becoming more and more important for controlling tunneling leakage currents and dissipated device power.

The models required for the investigations of such complex devices are based on first principles calculations and they are capable of reproducing band structure effects without fitting parameters. However, they are known to be limited to macro-cells with few (~ 100) atoms and thus they are not suitable for reproducing the electrostatics of the nanodevice. On the other hand, simpler and faster models based on effective mass approximations (EMA) can lack of accuracy or scalability, if a correct calibration methodology is not performed. The aggressive oxide scaling, the introduction of High-K oxide–stacks and mechanical strain have further increased the interest of models taking into account the role and the impact of material band structure and interfacial layers (IL) effects formed at the SiO_2/Si interface on electrical characteristics [223].

In this section, the impact of band structure and interfacial layers on electrical characteristics have been studied using self–consistent 1D Poisson–k.p–Schrödinger (PS) simulations. The importance of having accurately calibrated EMA models has been demonstrated,

taking into account quantum and band structure effects for the simulation of complex heterostructures. The test structures and the model validation methodology are described; in Section (i), band structure effects and the impact of strain in the structures are discussed, comparing model results with first principles simulations for extracting critical simulation parameters. In addition, the predictions of UTOXPP have been compared to 2D TCAD simulations. This study is reported in [228].

Both Metal/HK/P-Well and Metal/HK/N-Well devices have been integrated and characterized using a HP4284A LCR-meter for oxide capacitance measurements. The stacks are composed by TiN, HfO_2 and SiO_2 oxide over Si. Two different sets of oxide stacks have been studied varying the thickness of the SiO_2 layer (physical thicknesses t_{SiO_2} measured with ellipsometry ranging from 20Å to 35Å). The model has also been validated in the presence of a highly–strained layer (SL) using Si as a buffer in HighK/N-Well devices. Typical comparisons with excellent matching between measurements and simulations are shown in Figures C-2 and C-3 for different oxide stacks. In both the cases, a 5Å-thick IL has been modeled at the Si/SiO_2 interface, with a gradual linear variation of all the electrical and physical properties of the material. Figure C-2 also shows the comparison with 2D TCAD simulations, where quantum, band structure and IL effects are taken into account using a semi-empirical model (density gradient calibration).

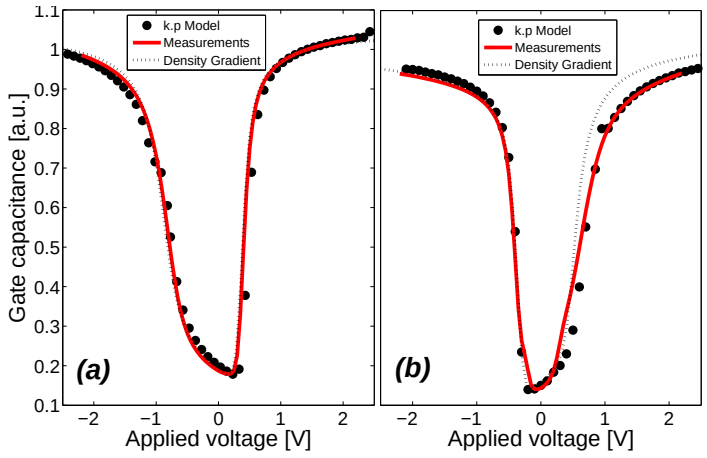


Figure C-2 – C-V curves on HighK/P-well (a) and HighK/N-well (b) structures with oxide stack having $t_{SiO_2} > 35$ Å(symbols: full band k.p model, dashed line: TCAD simulation using density gradient model, continuous line: measurements). The 5Å-thick interfacial layer has been modelled at the interface between SiO_2 and the substrate for both the devices.

(i) Band structure effects

Among the effects governing the electrostatics of the devices, the impact of band structure models of both the oxide and the semiconductor have been studied. The investigation of these effects has been carried out using the CV characteristic of the HighK/Nwell device shown in Figure C-3(b) as a reference.

Two advanced full band (FB) models (sp3d5s* tight binding (TB) model of [286] and 30-bands k.p model from [283]) are compared to simpler techniques based on multi band

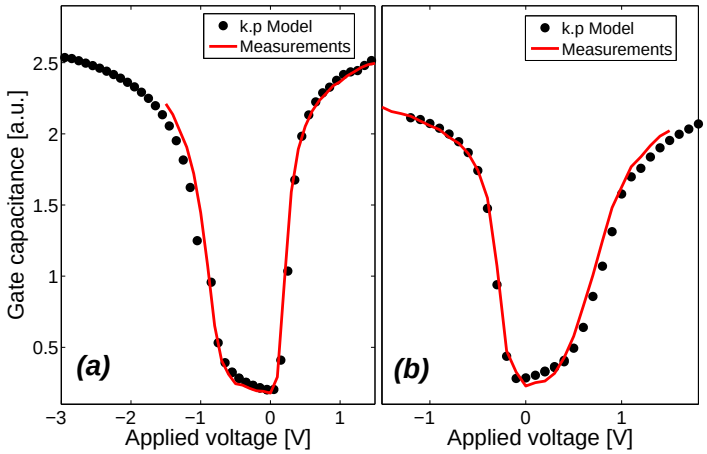


Figure C-3 – C-V curves on HighK/Pwell (a) and HighK/Nwel (b) structures with oxide stack having $t_{SiO_2} < 20$ Å(symbols: full band k.p model, line: measurements). The 5Å-thick interfacial layer has been modeled at the interface between SiO_2 and the substrate.

models (MBM) based on EMA. Figure C-4 shows the good matching achieved on the band diagram of SiO_2 between the ab-initio, FB and EMA simulations. In the FB models, the oxides have been treated as pseudo zinc-blende materials, adjusting the model parameters to match the band structure obtained from first principles [284], while for EMA an isotropic effective mass of $m_{ox} = 0.57$ for electrons has been used. The importance of having accurately calibrated EMA models relies in the reduced computational time of EMA models with respect to FB models.

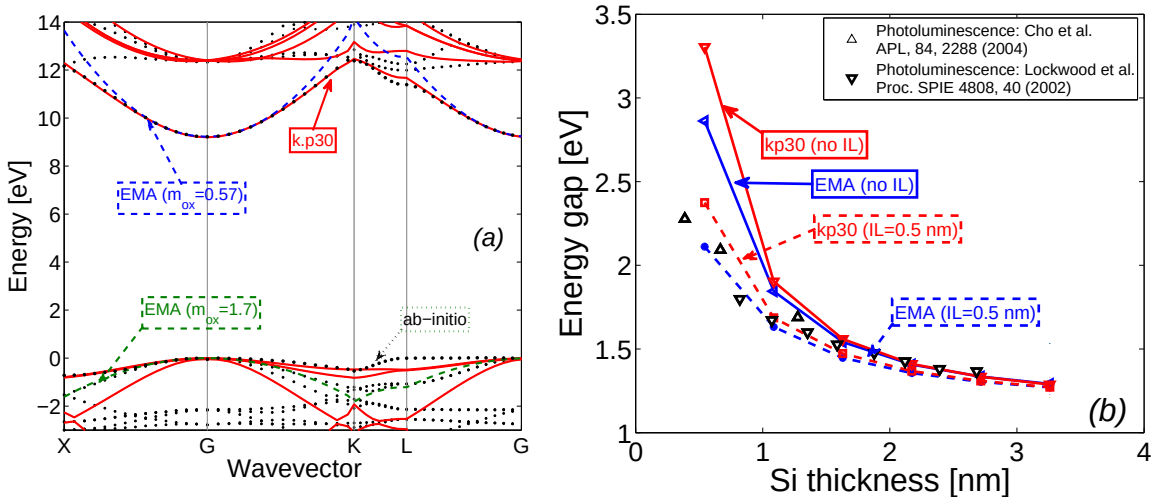


Figure C-4 – (a) Ab-initio band diagram of the SiO_2 material modelled with a β-cristobalite structure, FB k.p model results using a pseudo zinc-blende structure and EMA model comparison. (b) The relation between Si thickness and energy gap for k.p and calibrated EMA for a Si layer embedded in a SiO_2 buffer.

Typical outputs of UTOXPP are presented in the following. Figures C-5(a-c) show the calculated silicon band structure for h^+ using a 6-band k.p model for different lattice

orientations in a QW of 5nm surrounded by oxide layers. A 15-band k.p model has been adopted for the calculation of the band structure of e^- (Figure C-5(d)). These results can be used for the extraction of simpler band structure models suitable for compact approaches. EMA models have been calibrated in proximity of the confinement valleys Γ and Δ on the previously shown results. They are commonly adopted for electrical device simulations to reduce the computation effort (Figures C-5(e-f)). Additionally, strain effects on the bands can be taken into account and modeled in EMA and k.p using the Bir-Pikus approach [283]. The band structures of unstrained/strained germanium calculated with the 6-band k.p model are also shown in Figure C-6.

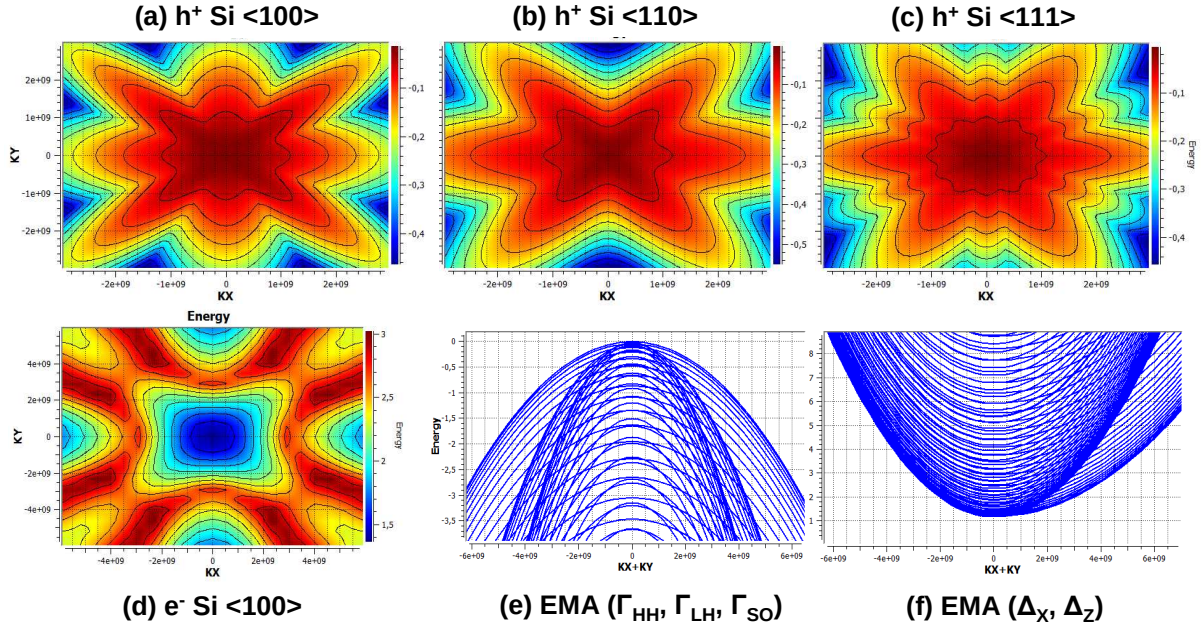


Figure C-5 – (a-c) Calculated silicon band structures for h^+ for different lattice orientations using the full band k.p model. In (d), silicon band structure calculated for e^- taking into account 30 bands with the full band k.p model. In (e-f), band structures of the Δ and Γ valleys in the effective mass approximation. All the structures have been calculated considering a 5nm SiO_2/Si/SiO_2 quantum well

The results on the device electrostatics of the two advanced models are compared to simpler 6-band k.p and 3-band EMA models (accounting for HH, SO, LH Γ-bands or Δ_X, Δ_Y and Δ_Z bands for electrons and holes respectively). Figure C-7(a) shows a comparison between C-V curves simulated in accumulation with a 15-band k.p and a 3-band EMA model for electrons, accounting for strain, heterostructure effects and mass variations [153]. A gradual 5Å-thick interfacial layer is modeled at the SiO_2/Si interface. The importance of modeling band deformation due to strain effects has been justified showing the large differences induced on the CV curve when the strain contribution is removed. Indeed, due to the presence of strain, heterostructure and interface properties, Δ_X and Δ_Y electrons exhibit a large band offset, with respect to the carriers in the Δ_Z valley which remain

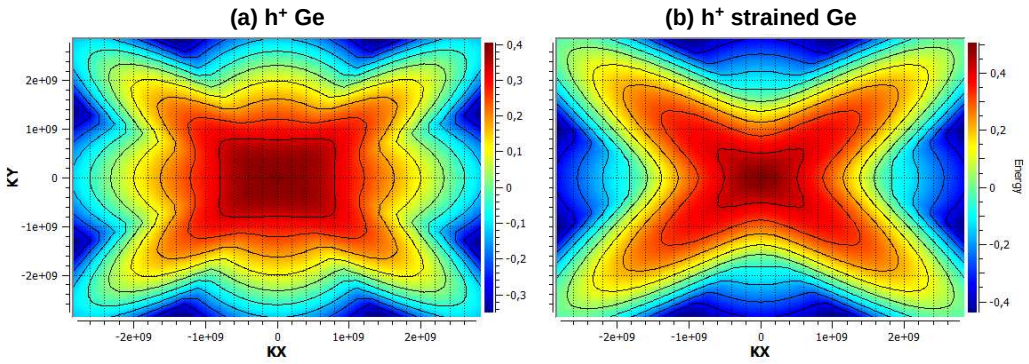


Figure C-6 – Germanium band structures calculated using the 6 bands k.p model for h^+ for unstrained (a) and biaxially-strained (b) material.

confined in the buffer. Figure C-7(b) shows the impact of FB models and strain on the accumulation charge and on the potential. Also in this case, a good alignment of the three models has been obtained.

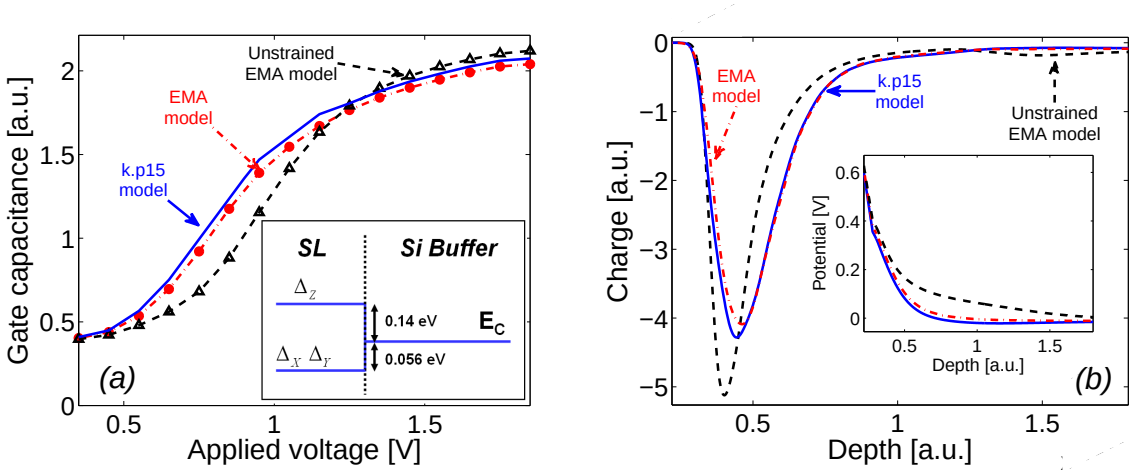


Figure C-7 – (a) C-V curves comparing a 3-band EMA model and a full band 15-k.p model for electrons in accumulation for the device shown in Figure C-3(b). In the inset, conduction and valence bands at $V = V_{FB}$. The curve is also compared to simulations on a device where the strain is not taken into account and important changes are observed. (b) Accumulation charge on the same device at $V = 1.85V$, comparing the 15-k.p model with the 3-band EMA model on both strained and unstrained devices.

In Figure C-8(a) the band structure obtained with full band k.p is compared to the one used for 3-band EMA. Also shown in Figure C-8(b) is the behavior of the wavefunction penetration into the oxide layer corresponding to the first level of energy of the carriers. The good alignment achieved for the three models, including TB results, confirms the

accuracy of the approach.

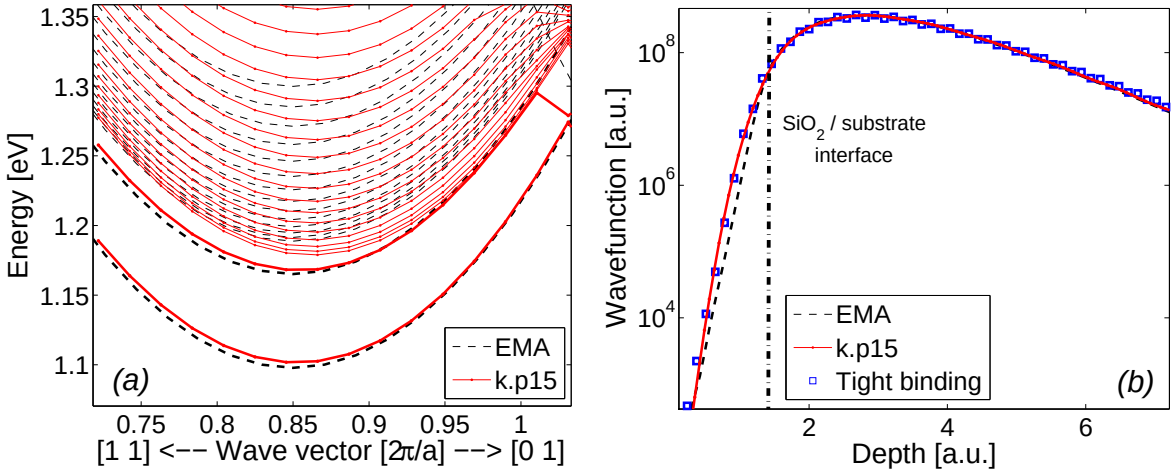


Figure C-8 – (a) Comparison between the band structures used in 15-k.p and 3-band EMA used for simulations in Figure C-3(b) with results from tight binding model in accumulation region. (b) The same wavefunction penetration has been obtained for all the models using the pseudo oxide band structure model of Figure C-4.

Similar studies have been performed in inversion, comparing a 30-band k.p (accounting for SO effect), a 6-band k.p and a 3-band EMA for holes considering all Γ-bands [287] and focusing on the influence on capacitances (Figure C-9), energy levels and waves (Figure C-10). In this case a full meshing on K in reciprocal space is performed for the accurate computation of the inversion charge. In Figure C-9, an oxide mass value of $0.45m_0$ has been used in the EMA model in order to amplify the small differences noticeable between the 30-k.p and EMA models after the ab-initio calibration. Smaller differences could be found with the fitted value of $1.7m_0$ obtained from ab-initio calculations. Also in this case the impact of strain on the valence bands has been analyzed and modeled in EMA using the analytical solution of the 6-k.p Bir-Pikus model at Γ [287]. Due to the reduced band splitting of the Γ valleys in inversion, the SL has a negligible impact on the inversion charge and has not been showed in the Figure.

(ii) Modeling of interfacial layers

The SiO_2/Si interface is modeled using a 5Å-thick interfacial layer region where the electro–chemical properties of the material are linearly varying. The variations of the band structure parameters used in the Schrödinger solver are also taken into account. Additionally 2D TCAD simulations have been performed using a commercial simulation package [288] to assess the predictions of the models implemented in TCAD tools (Figure C-11(a)). The density gradient model included in the simulation tool provides the control over two fitting parameters γ_e and γ_h, used for modulating the variation of electrons and holes Fermi levels, respectively. Figure C-11(b) and C-11(c) show how the variation of the IL thickness

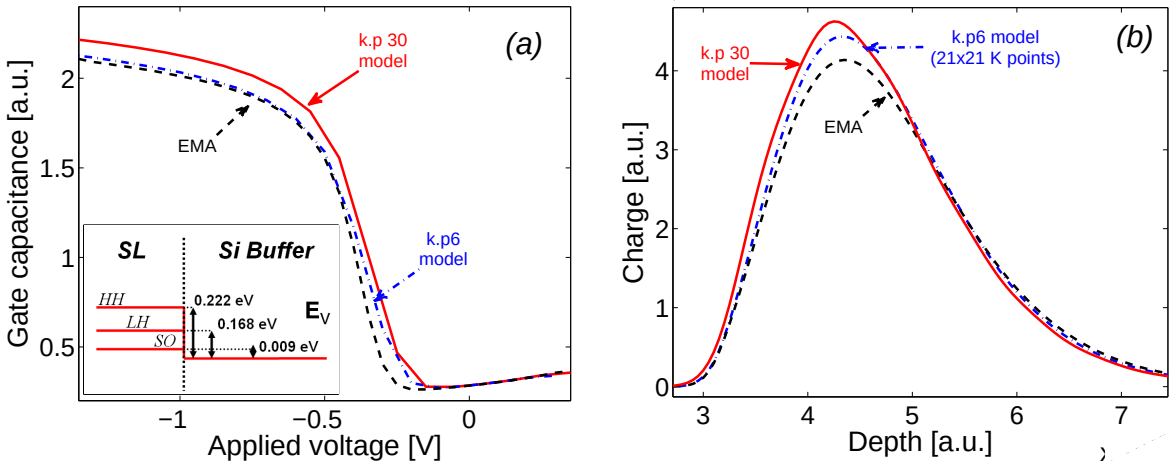


Figure C-9 – (a) C-V curves comparing a 3-band EMA model and a 30-k.p model (including a full meshing on K in reciprocal space) for holes in inversion on the device shown in Figure C-3(b). Additionally, the inset shows the impact of strain in the layer on the valence band. (b) Comparison of inversion charges calculated for two different bias voltages on the same device at $V_G = -1.15$V.

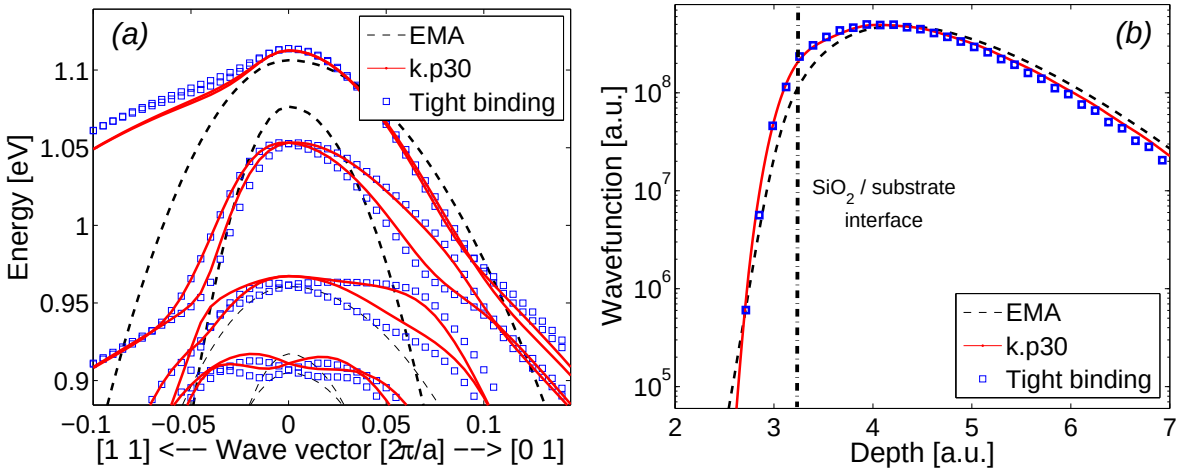


Figure C-10 – (a) Comparison between the band structures obtained with 30-k.p, 3-band EMA and sp3d5s* tight binding model in inversion region ($V_G = -1.15V$). (b) Wave function penetration in the oxide region obtained with the 3 models. Also in this case, to further illustrate the difference with EMA, $m_{ox} = 0.45m_0$ has been used leading to a smaller penetration in the oxide.

(ranging from 5Å to 10Å on Si [223]) is influencing both the capacitance and the inversion charge, affecting the threshold voltage of the device. It is worth mentioning that a simple variation of dielectric permittivity or oxide thickness is not sufficient to match the full CV curve for all the devices, since the model predictions will eventually fail in the depletion region.

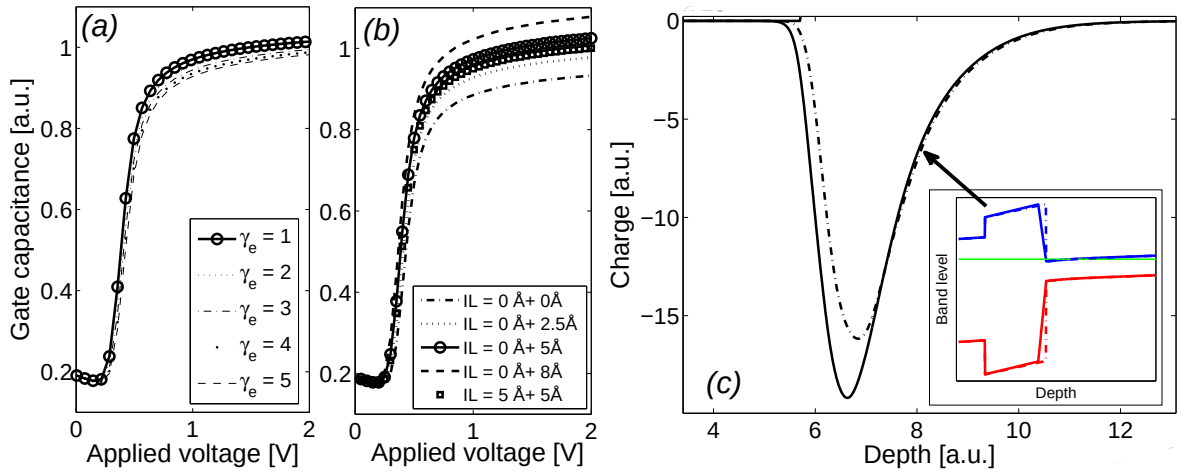


Figure C-11 – (a) TCAD simulation results on device shown in Figure C-2(a) using the density gradient model both in accumulation and inversion regions. (b) 3-bands EMA simulation results on device (a), showing the capacitance variation in inversion when the IL thickness is changed. (c) The physical reason of the capacitance increase with IL is due to the higher penetration of the wavefunctions in the IL and in the oxide, consequently increasing the total inversion charge. The same effect can be seen in accumulation. The inset shows how the band diagram is modified after the inclusion of a 5Å-thick IL.

C.3.2 The Non-Equilibrium Green Function Method applied to III-V compound structures

(i) Formalism

One of the most general and rigorous device modeling framework has been developed by Keldysh in 1965 and is provided by the so called non-equilibrium Green's function theory (NEGF). The introduction of quantum devices that require a fully quantum mechanical treatment, together with the recently available high computational capabilities, represented key factors in the recognition and diffusion of the method for quantitative and predictive analysis [128, 289, 290].

Accurate modeling of quantum nanodevices should include the capability to reproduce: interference effects, quantum mechanical tunneling, discrete energy levels due to quantum confinement and scattering mechanisms (mainly electron-phonon and electron-electron).

In this section, the general approach and the features included in the UTOXPP PS solver are detailed. Furthermore, the following sections present few test–cases conducted on structures based on III-V semiconductor compounds, providing insight on the highs and lows of the NEGF method.

A 1D system formed by grid/lattice points separated by a spacing Δx and coupled with their nearest neighbors only, is characterized by a tri-diagonal Hamiltonian:

$$(E - \Psi)H = 0 \Rightarrow$$

$$\begin{pmatrix} \bullet & \bullet & \bullet & & & & & & \\ & \bullet & \bullet & \bullet & & & & & \\ & & -t_{q-1,q-2} & E-\epsilon_{q-1} & -t_{q-1,q} & & & & \\ & & & -t_{q,q-1} & E-\epsilon_q & -t_{q,q+1} & & & \\ & & & & -t_{q+1,q} & E-\epsilon_{q+1} & -t_{q+1,q+2} & & \\ & & & & & \bullet & \bullet & \bullet & \\ & & & & & & \bullet & \bullet & \bullet \end{pmatrix} \begin{pmatrix} \bullet \\ \bullet \\ \Psi_{q-1} \\ \Psi_q \\ \Psi_{q+1} \\ \bullet \\ \bullet \end{pmatrix} = 0.$$

In this expression E is the carrier energy, Ψ_q is its wavefunction at the grid point q, while the diagonal and off-diagonal terms ϵ_q and $t_{q,q+1}$ indicate the potential and the interaction with the nearest neighbour points. In a uniform grid and for Hermitian Hamiltonians, one has:

$$\begin{aligned} t &= t_{q,q+1} = t_{q+1,q} = -\frac{\hbar^2}{2m_q \Delta x^2} \\ \epsilon_q &= V_q + \frac{\hbar^2}{m_q \Delta x^2} \end{aligned} \tag{C.1}$$

with V_q potential and m_q confinement effective mass at point q. The approach can be extended to any multi band EMA or k.p model. In case of full band approaches each element in the Hamiltonian is represented by a block matrix.

Any nanodevice can be divided in three regions: the active device region with a variable potential profile and the right/left semi-infinite leads/contacts having constant potential ϵ_R/ϵ_L applied, respectively. To solve the infinite-dimensional matrix representing the total Hamiltonian of the system, the influence of the right and left semi-infinite leads can be folded into the first and last points of the device region [290]. It can be demonstrated that after this modification the infinite system reduces to a finite n-dimensional system in the form:

$$A\Psi_D^{(L)} = i_L, \tag{C.2}$$

where the wave function $\Psi_D^{(L)}$ in the device due to waves incident from the left lead be obtained. In Eq. C.2, matrix A is:

$$A = EI - H - \Sigma_{lead}, \tag{C.3}$$

where Σ_{lead} is the self-energy matrix representing the influence of left and right leads on the device. Due to lead folding on the first and last point, only the first and last elements of the matrix are non zero:

$$\Sigma_{lead_{1,1}} = t_{d,l} e^{ik_l \Delta x} t_l^{-1} t_{l,d} = \Sigma_L(E), \tag{C.4}$$

$$\Sigma_{lead_{n,n}} = t_{d,r} e^{ik_r \Delta x} t_r^{-1} t_{r,d} = \Sigma_R(E), \tag{C.5}$$

where k_l/k_r and $t_{d,l}/t_{d,r}$ are the wavevectors and the interactions with the device points in the left/right contacts, respectively. The vector i_L in Eq. C.2 has only the first element non zero and equal to $-2it_{d,l}\sin(k_l\Delta x)$. Similar considerations could be applied for the waves incident from the right lead leading to:

$$A\Psi_D^{(R)} = i_R. \tag{C.6}$$

Since the Green's function G of the considered device can be written as:

$$AG = I, \tag{C.7}$$

one finds:

$$\begin{aligned} \Psi_D^{(L)} &= Gi_L, \\ \Psi_D^{(R)} &= Gi_R. \end{aligned} \tag{C.8}$$

The system of Eq. C.7 can be solved by inverting the matrix A or by using a recursive algorithm that only computes for the terms on the three diagonals of elements of G [290]. In particular, with the latter numerical technique suitable for inverting block-tridiagonal matrixes, we extended the EMA approach to full band structure models.

Additionally, in a ballistic system where scattering elements are not included, the determination of the wave function can be performed only knowing the first and last columns of G. The local density of states due to waves incident from left and right leads can be thus calculated for each grid and energy point having:

$$\begin{aligned} LDOS_L(q,E) &= \frac{G_{q,l}\Gamma_L(E)(G^\dagger)_{l,q}}{\pi}, \\ LDOS_R(q,E) &= \frac{G_{q,r}\Gamma_R(E)(G^\dagger)_{r,q}}{\pi}, \end{aligned} \tag{C.9}$$

with $G^\dagger$ the Hermitian conjugate of the Green's function and:

$$\begin{aligned} \Gamma_L(E) &= -2Im\left[\Sigma_L(E)\right], \\ \Gamma_R(E) &= -2Im\left[\Sigma_R(E)\right]. \end{aligned} \tag{C.10}$$

For inelastic systems, electron-phonon interaction should be considered to take into account scattering events. This leads to a matrix of in-scattering self-energies, whose element at point q and energy E can be expressed by:

$$\Sigma_{phq,q}^{in} = D_q\left[n_B(\hbar\omega_{ph})G_{q,q}^n(E-\hbar\omega_{ph}) + (n_B(\hbar\omega_{ph})+1)G_{q,q}^n(E+\hbar\omega_{ph})\right], \tag{C.11}$$

with n_B the Bose-Einstein occupation factor, $\hbar\omega_{ph}$ the phonon energy, D_q the deformation potential for electron-phonon scattering, and $G_{q,q}^n$ the electron correlation function at point

q. The latter quantity is correlated to the electron density in q with:

$$\begin{aligned} G^n(E) &= G(E)\Sigma^{in}(E)G^\dagger(E), \\ n_q(E) &= 2\frac{G^n_{q,q}(E)}{2\pi}, \end{aligned} \tag{C.12}$$

Similarly out-scattering self energies $\Sigma^{out}_{phq,q}$, where the hole correlation function G^p is considered, can be calculated. Introducing the scattering components in Eq. C.7, the Green's function G in the device is finally calculated solving:

$$[EI - H - \Sigma_{lead} - \Sigma_{Phonon}(E)]G = I, \tag{C.13}$$

It can be seen that due to the dependence of Eq. C.11 on the density, one needs to iteratively solve for the Green's function and self-energies. Subsequently, the carrier density calculated from Eq. C.12 is injected in Poisson's equation to determine the new potential profile. The process is iterated until convergence is achieved. The addition of phonon scattering in the Green's function equation presently remains a controversial subject due to the approximations introduced on determining ω_{ph} and D_q.

The current density flow from the left lead into the device is computed with:

$$J_L = \frac{q}{\hbar}\int \frac{dE}{2\pi} Tr[i(G_{1,1}(E) - G^\dagger_{1,1}(E))\Sigma^{in}_L(E) - G^n_{1,1}\Gamma_L(E)]. \tag{C.14}$$

In the phase coherent limit (absence of scattering) the conventional Landauer-Buttiker formula is valid and the transmission can be computed using:

$$T(E) = J(E)/(f_L(E) - f_R(E)), \tag{C.15}$$

where $f_L(E)$ and $f_R(E)$ are the Fermi-Diraq distributions in the left and right leads of the structure.

(ii) Quantum transmission of a potential barrier

In Chapter 2, we applied the NEGF method to the calculation of the barrier transmission of tunnel oxides and ONO stacks in flash structures. In this section, we highlight the flexibility of the tool to deal with III-V compounds and different quantum structures such as resonant tunneling diodes, quantum wells and potential barriers.

The trasmission of several potential barrier configurations and the dependencies on the barrier thickness and height has been computed using the NEGF solver. As a title of example, we consider a 50nm GaAs structure, with a 10nm $Al_{0.3}Ga_{0.7}As$ barrier in the middle. The multi-band structure of the three materials has been included in the material parameters definition. In particular, for electrons both the materials are characterized by the three valleys Γ, Δ and L which have been treated independently with an EMA for the analysis of the transmission and the LDOS contributions. Figure C-12(a) shows the

barrier profile for Γ for the simulated structure. The calculated conduction band offset is around 0.36eV. In Figure C-12, the calculated LDOS is also shown as a function of the position and energy for the three valleys. Resonance states, where the LDOS peaks, are strongly affected by the confinement masses and by the barrier height. Figure C-13(a) reports the transmission of the barrier where the three valleys act as independent channels in the conduction. In (b) the dependence on the barrier height for electrons in Γ is shown. Semi–classical calculations, such as the WKB approach, would lead to $T(E) = 1$ when the energy is higher than the potential step. In the calculated transmission, quantum resonances cause oscillations in the transmission for energies higher than the barrier height. It can be shown [241] that the exact analytical solution, valid for ideal square barriers when no potential is applied, excellently matches the numerical results.

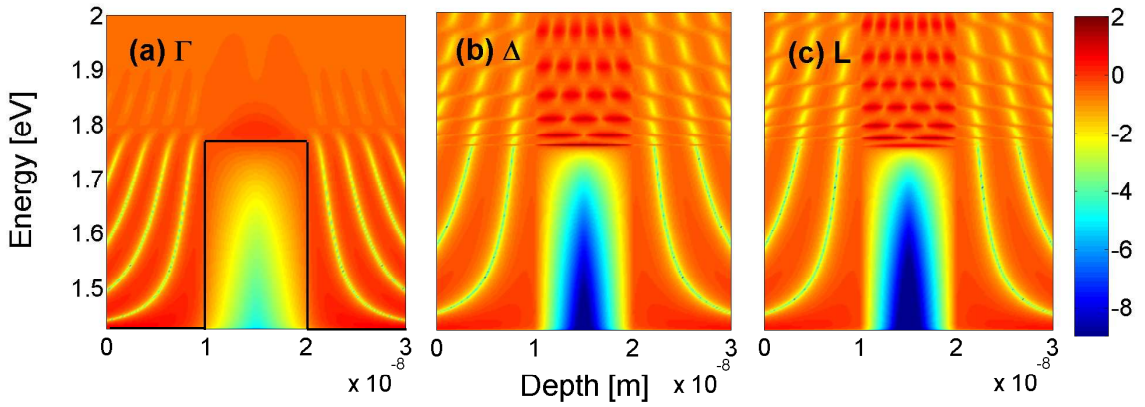


Figure C-12 – (a) In line, conduction barrier profile for electrons in Γ versus the position in the structure. Colour plot of the calculated local density of states in logarithmic scale for Γ (a), Δ (b) and L (c) electrons at room temperature in a structure having a 10nm-thick $Al_{0.3}Ga_{0.7}As$ barrier separating two GaAs regions. In (a) the barrier profile is also shown.

(iii) Quantum well confinement

Quantum confinement effects can be analysed as well in quantum well structures. In this view, $SiO_2/Si/SiO_2$ structures are simulated with the NEGF solver varying the width of the quantum well D from 1nm to 5nm. Figure C-14 shows the LDOS in the structure as a function of energy and depth in the device; quantum confinement of the electrons indicates the increase of the level separation when the width decreases. All the levels are characterized by the presence of a characteristic penetration of the wavefunction in the oxide which increases as energy raises.

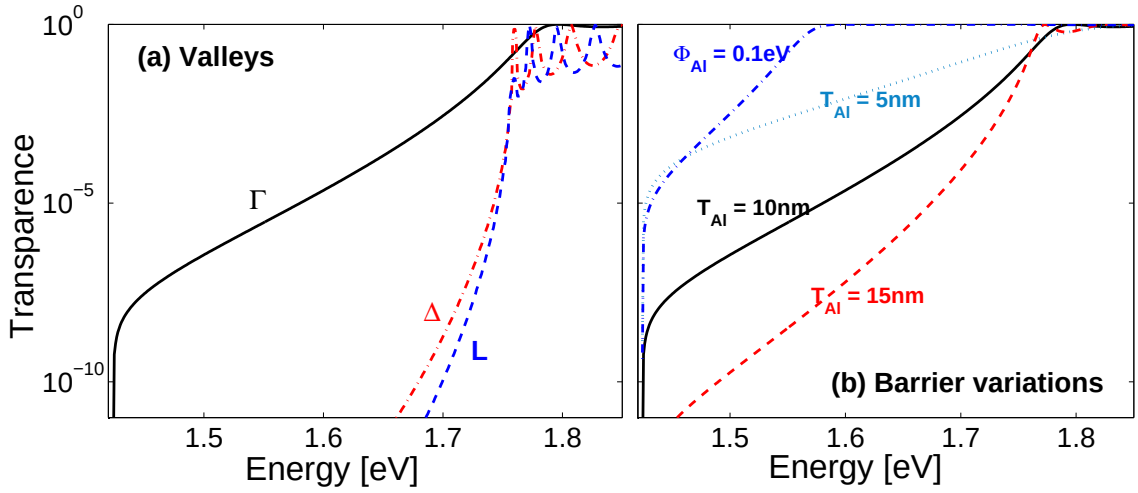


Figure C-13 – (a) Barrier transmission as a function of carrier energies for electrons in Γ, Δ and L electrons at room temperature. (b) Barrier height Ψ_{Al} and thickness T_{Al} dependence of the transmission for electrons in Γ.

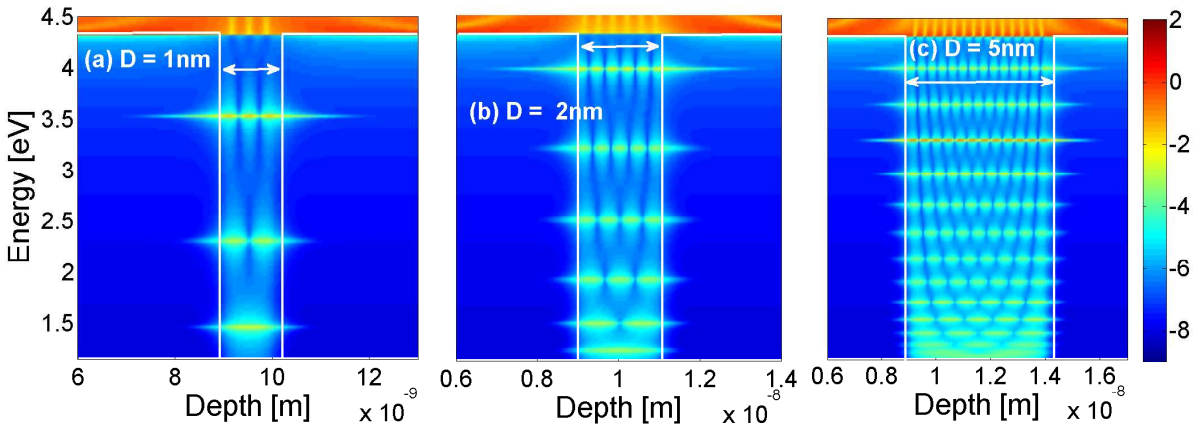


Figure C-14 – LDOS vs. depth and energy in logarithmic scale for three quantum well structures simulated with the NEGF solver for Δ_Z electrons and varying the width of the device.

(iv) Quantum resonances in Resonant tunnelling diodes

Resonant Tunnelling Diode (RTD) structures are investigated simulating a 30nm / 40nm / 30nm $Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}As$ structure for Γ electrons at different voltage bias conditions. The double barrier structure is the active intrinsic area of the device and is surrounded by two lightly n-doped GaAs regions forming the leads. The band offset is around 0.3eV for $Al_{0.3}Ga_{0.7}As/GaAs$ structures. Quantum confinement occurs in the GaAs layer. Tunnelling can occur only when the Fermi level of one of the leads is aligned with the quantized level in the GaAs layer. Strong quantum resonances are noticeable in

the barrier transmission at several voltage biases (Figure C-16(a)). This phenomenon is reflected on the tunneling current density in (b) where a peak is noticed at 0.2V when the Fermi level is aligned with the lowest quantum well resonance [291]. The current density is also dependent on the density of states of the right contact.

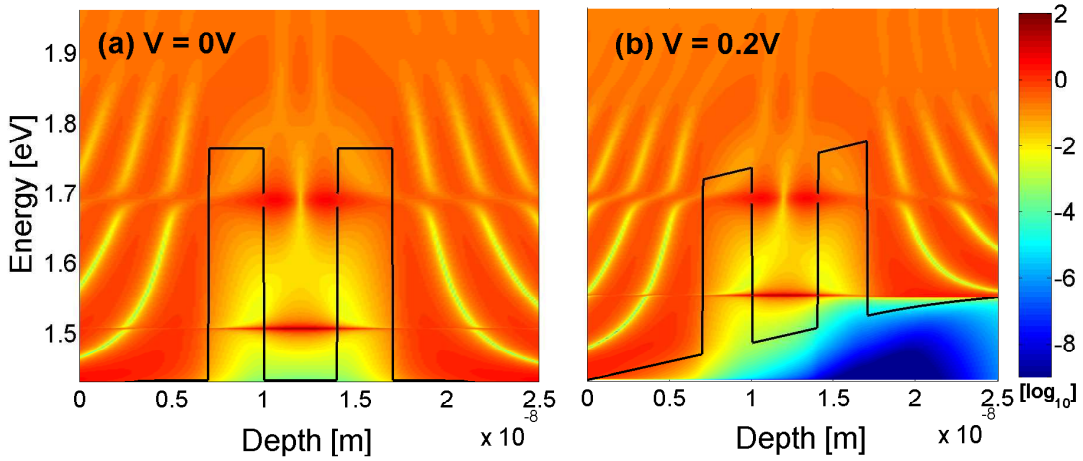


Figure C-15 – LDOS vs. depth and energy in a RTD structure simulated for two bias conditions with the ballistic NEGF model.

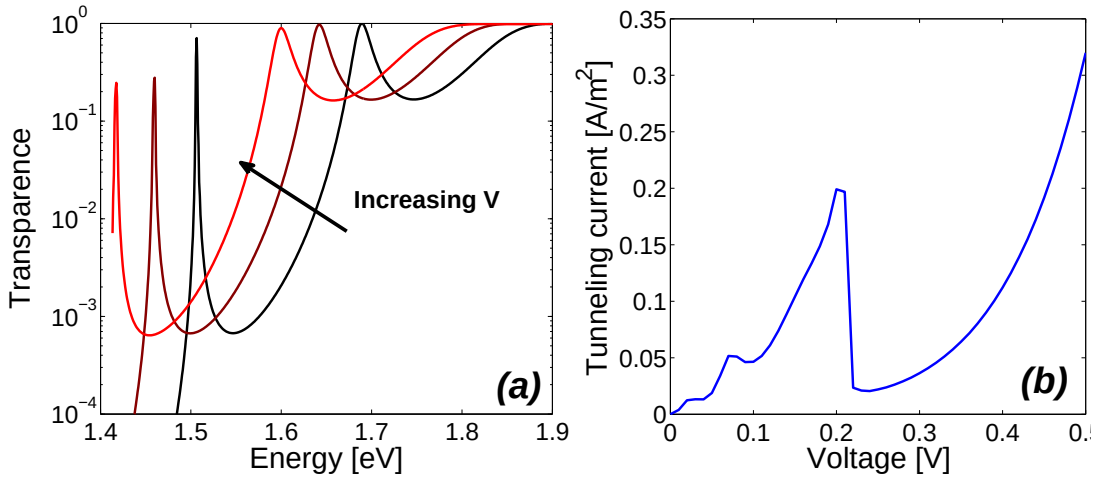


Figure C-16 – (a) Double-barrier transmission of electrons as a function of their energy for different bias voltage conditions. The two resonant states can be noticed. The tunneling current density in (b) peaks when the lowest quantum resonance is aligned with the Fermi level in the right contact. Ballistic elastic transport has been considered in this study. Similar considerations apply to inelastic trasport taking into account scattering effects.

C.4 Graphical User Interface

Given the complexity of the included physical models, where often the definition of a large number of parameters is required, a graphical user interface has been implemented to improve the user-friendliness of the tool. The interface, based on Qt C++ and OpenGL libraries [292], also makes possible the generation of high definition plots for reporting purposes. Some screenshots of the interface have been illustrated in the following.

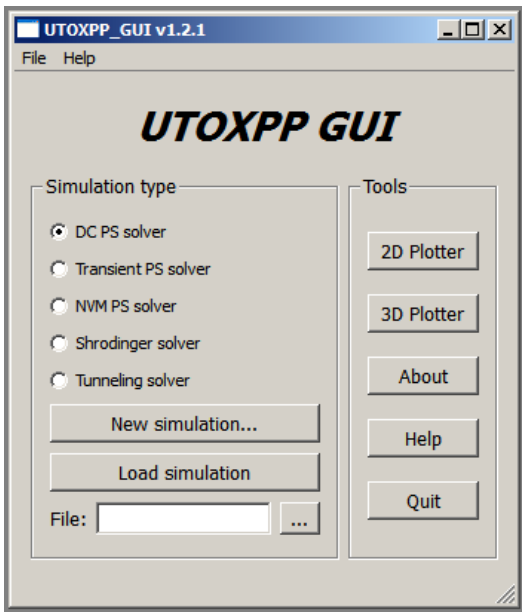


Figure C-17 – Screen-shot illustrating the main window of the UTOXPP graphical user interface which has been implemented to improve the user-friendliness of the tool. Portability is for both Win32 and X11 Linux/Unix systems.

C.5 Conclusion

A new TCAD modeling tool including advanced models for band structure and device simulation has been presented. It permits the investigation of the effects of technological boosters commonly used in advanced nanoscale CMOS technologies, the evaluation of quantum confinement and tunneling in complex heterostructures, the investigation of band structure deformation and device electrostatics in strained alloys and of impact of defects in oxide layers with a rigorous multiphonon approach. Including both a DC, AC and transient model, it permits to perform device simulations in most of the device operating conditions and can be used for validating or extracting simpler empirical models based on compact approaches. The graphical user interface strongly improves the user-friendliness and portability of the tool, while efficient techniques have been adopted for multithreading and parallel calculation.

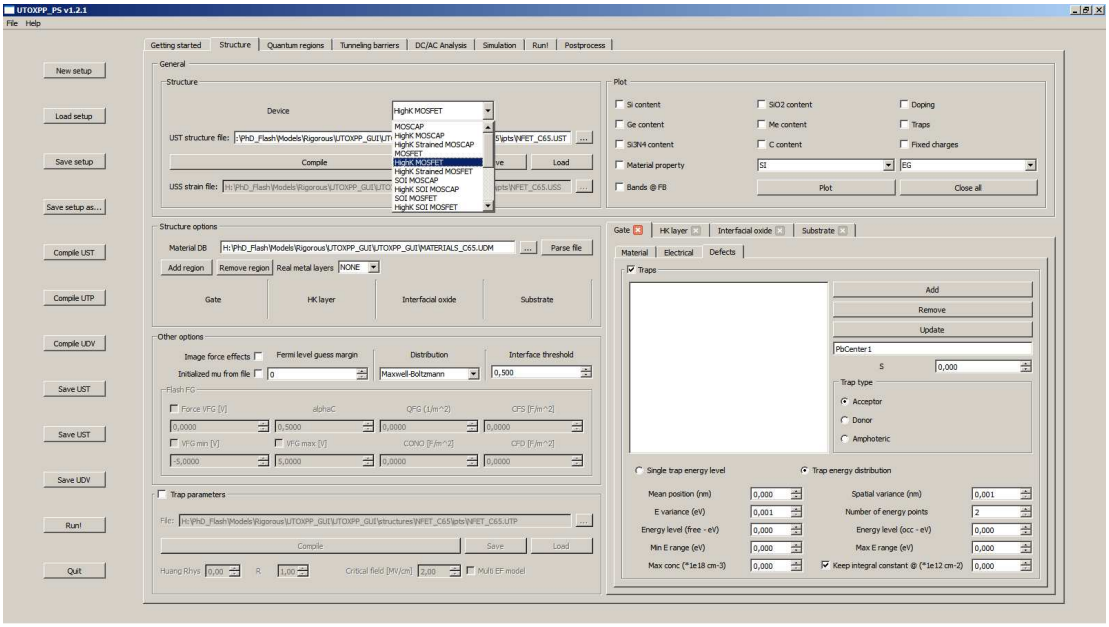


Figure C-18 – The user follows the modeling flow progressively defining the different parameters required. The first step requires selecting a structure configuration and defining the geometrical and electrical properties of the same. The predefined configurations include a wide range of devices, from MOS capacitors to HKMG FD-SOI or flash devices. A custom configuration can be defined by the user.

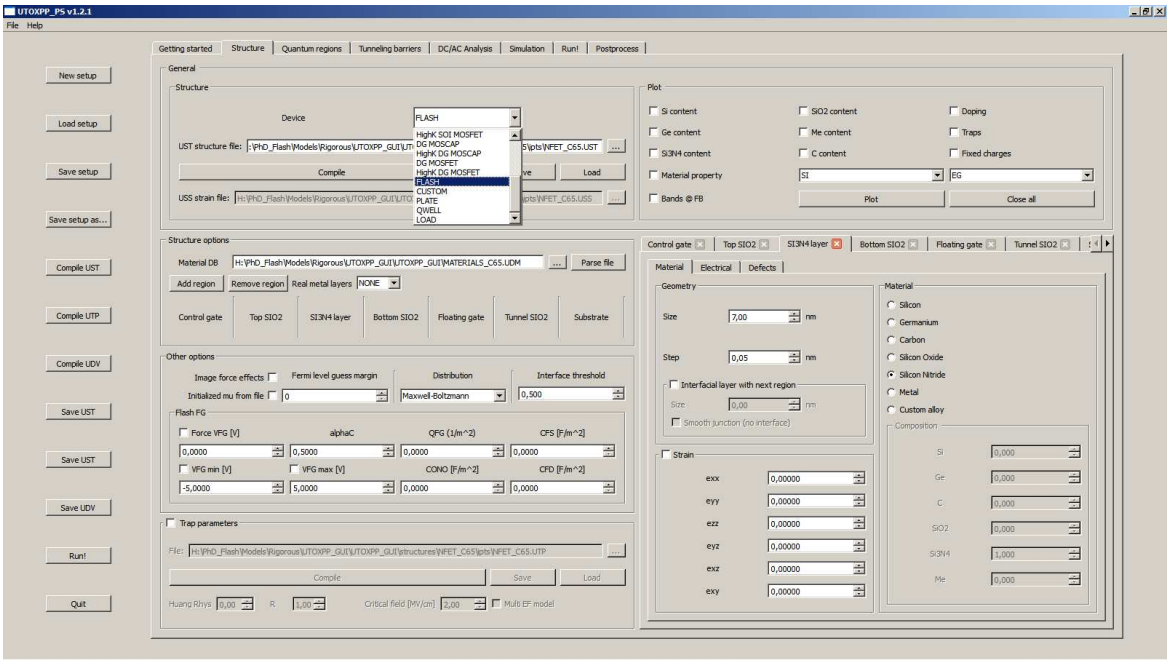


Figure C-19 – Physical dimensions, meshing, material properties, doping, fixed charges or strain can be defined for each layer.

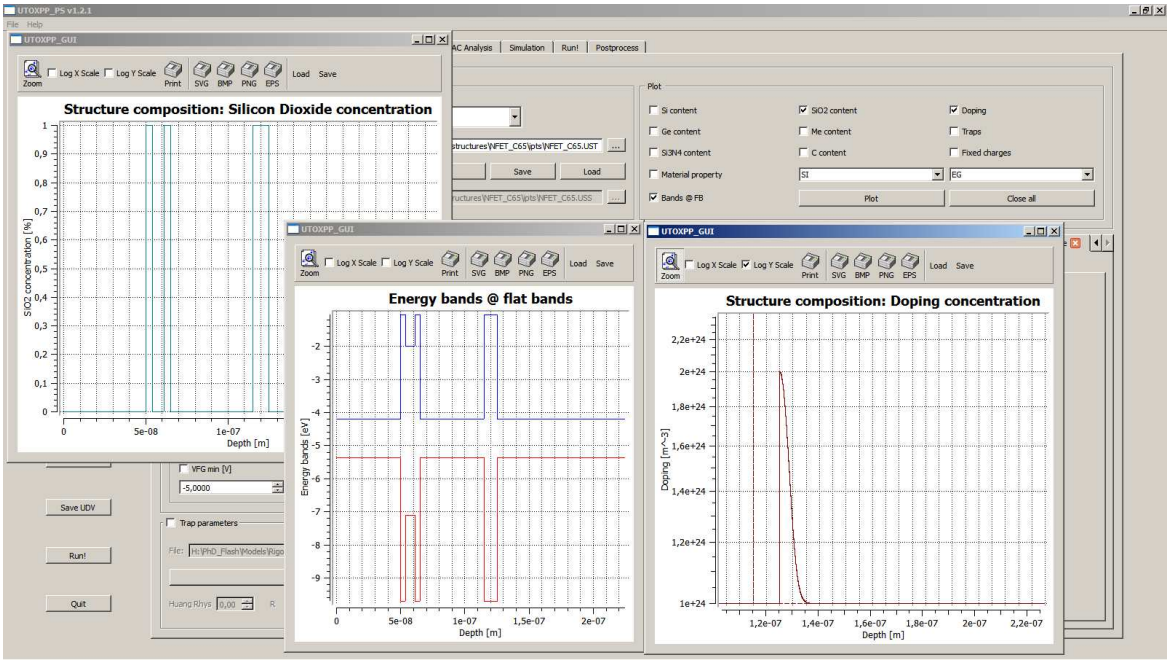


Figure C-20 – The user can review the structure properties and parameters by plotting the different quantities before launching a simulation. Material and strain properties in heterostructures can also be verified.

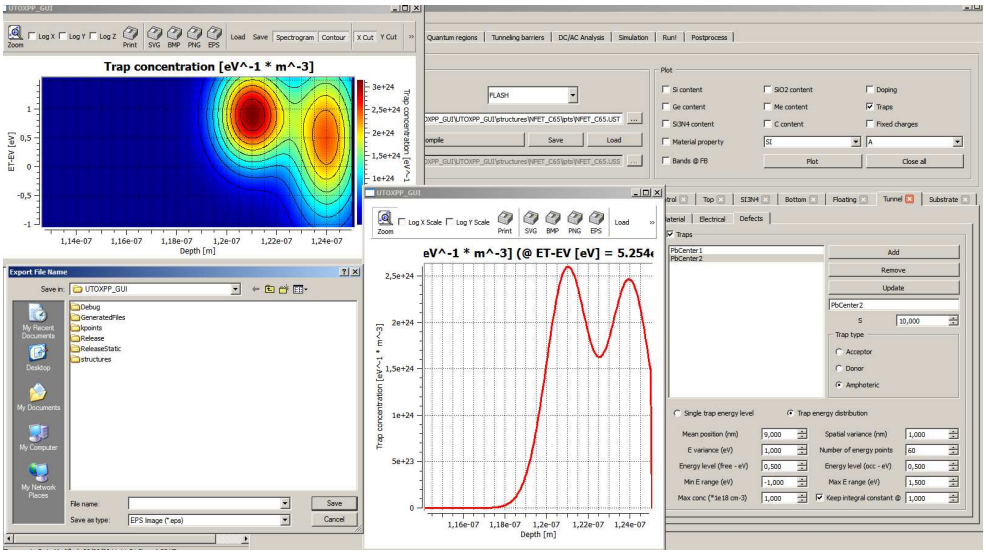


Figure C-21 – Spatial and energy distributions of defects can be defined and simulated with the MPA charge trapping model. In this example two Gaussian distributions have been represented. Surface colour plots as well as 2D cuts, help the user in the definition and calibration procedure.

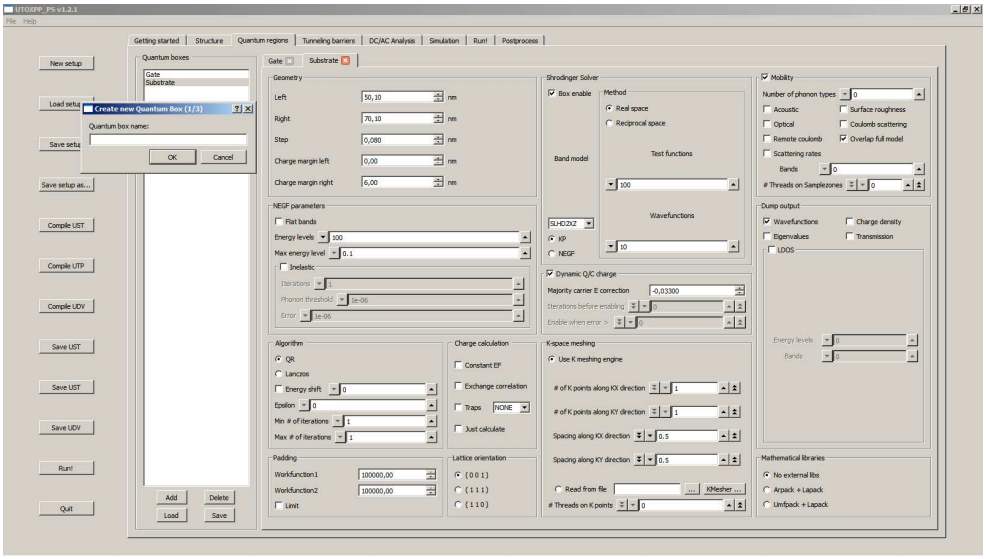


Figure C-22 – Having defined a device structure, the user describes the regions where the Schrödinger solver (k.p or NEGF) should compute the charge. Commonly regions where highly quantization occurs are declared, but the user might also be interested in calculating the quantum charge in the gate layer or across an entire energy barrier, for example when assessing the QBS tunneling current. Band structures models are also chosen, from EMA approaches to full-band k.p models.

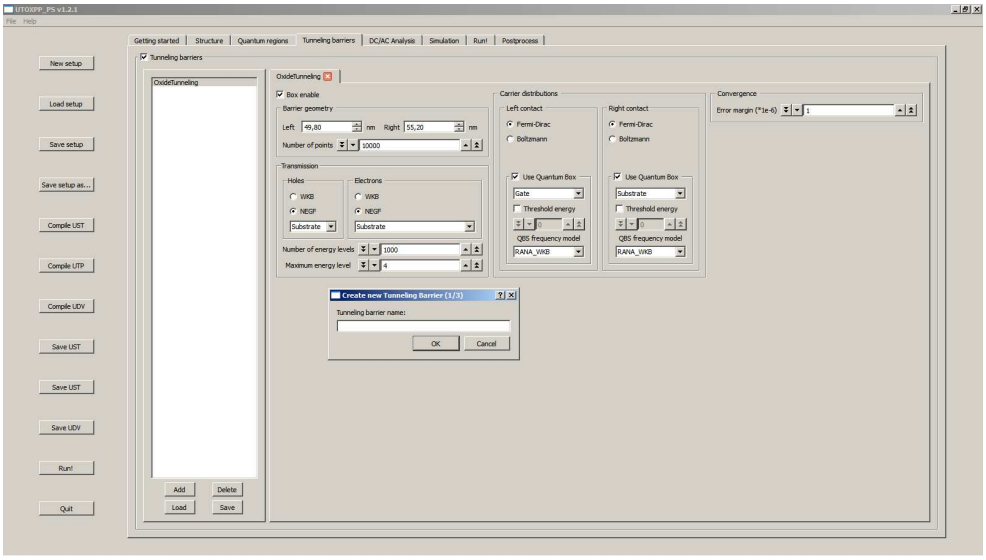


Figure C-23 – The tab *Tunneling barriers* is used to describe the potential barriers for quantum tunnelling current calculation. In this view, the user defines the geometry of the region and the meshing in space. A WKB or a NEGF model for the transmission of the barrier can be chosen. The models for the carrier distribution of the two reservoirs are chosen indicating if the QBS component should be considered in the calculation.

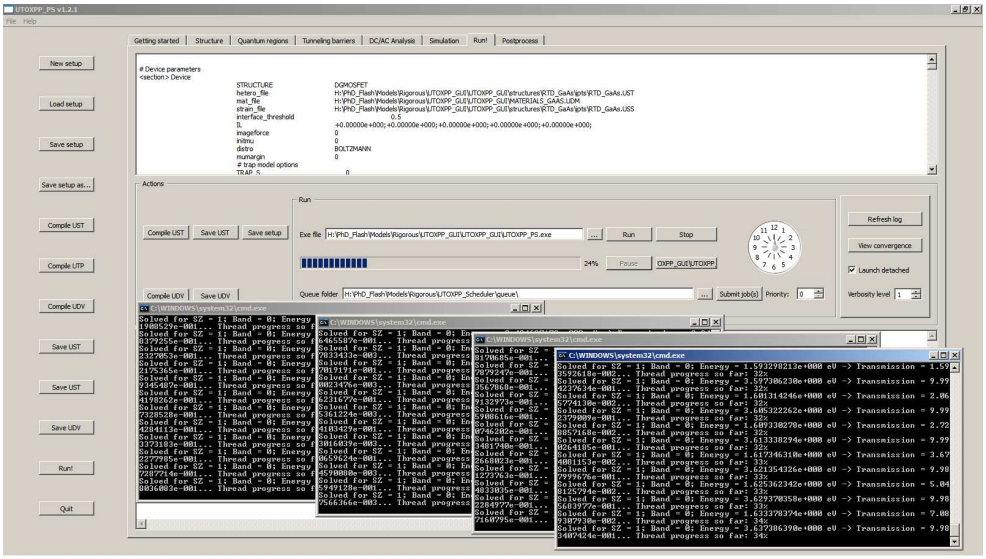


Figure C-24 – The tab *run* is used to review the parameters and launch the simulation. Multithreading calculation is supported on the bias points. In particular, in this example, the RTD structure is simulated for a voltage sweep from -1V to +1V with voltage step of 0.1V, and dividing the overall simulation in 4 subintervals which are handled by separate processes. Since a NEGF solution is needed for this structure, multithreading on the energy levels has been enabled at each voltage point. Two threads per window are running in parallel.

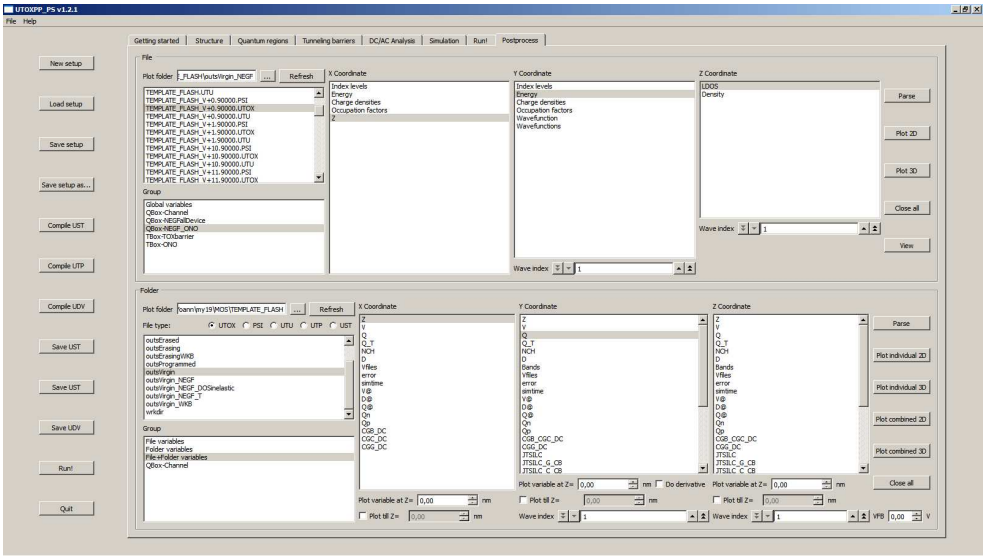


Figure C-25 – Data postprocessing is performed in tab *Postprocessing* where results for each simulated point can be plotted. Data that can be plotted include: charge and voltage distribution in the device, band diagram, electrical field, barrier transmission, carrier wavefunctions, quantization energy levels, etc.

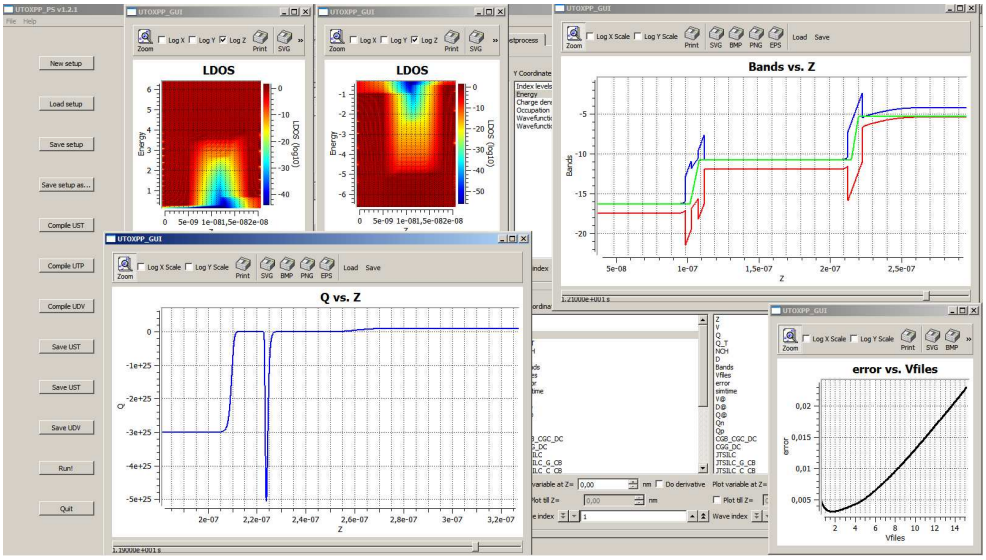


Figure C-26 – Output results could be postprocessed in a 2D form or in colour surface plots. In this graph, the LDOS calculated for e^- and h^+ on the tunneling oxide barrier are shown. Furthermore, the entire evolution of the band diagram and charge distribution could be shown for different bias conditions.

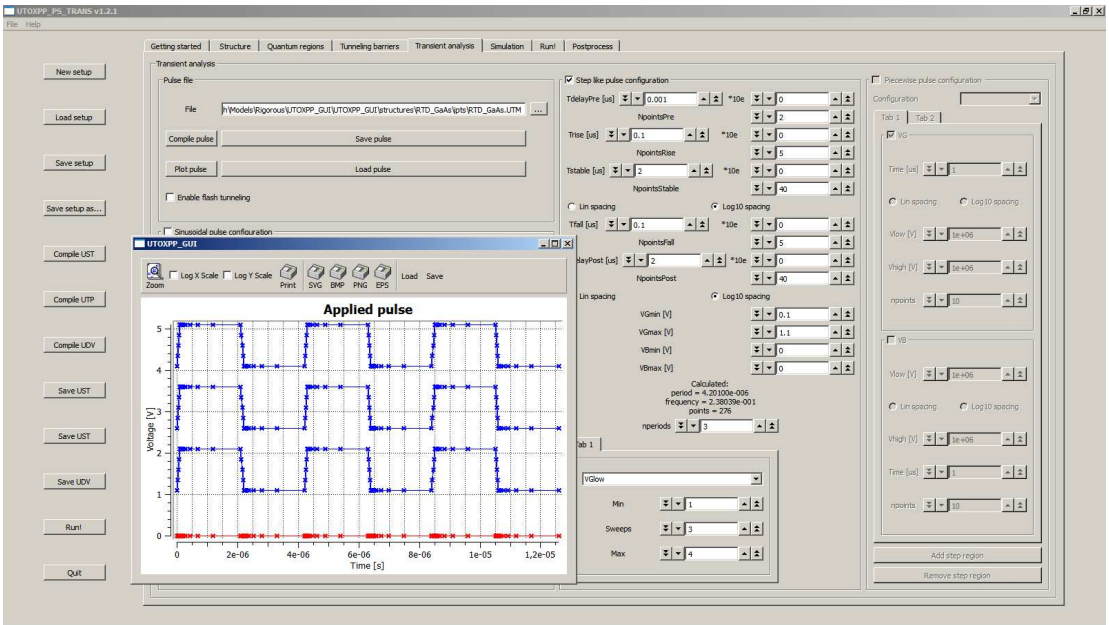


Figure C-27 – The GUI includes also a specific tool for defining voltage pulses and handling transient simulations. Both sinusoidal and square box pulses could be applied, controlling all the parameters, number of points and linear/logarithm repartitions on the voltage plateaux. Parameter sweeps could be also defined for charge pumping simulations.

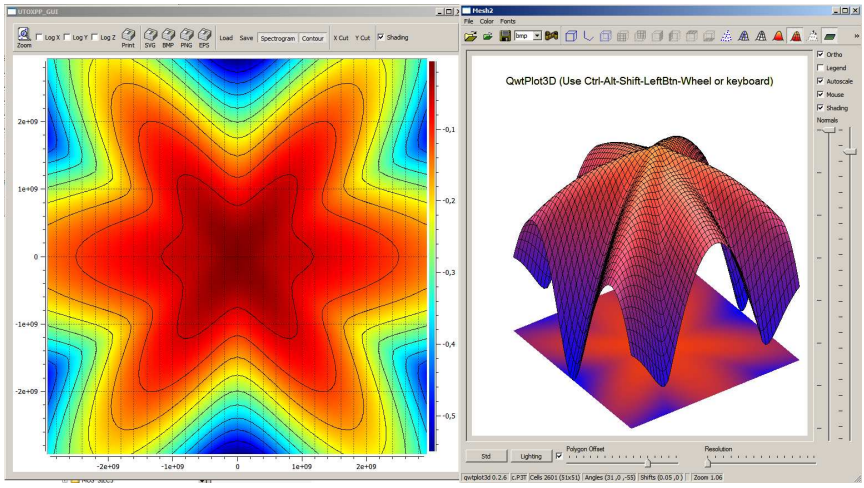


Figure C-28 – Open-GL libraries could be used to plot contour surface plots or visualize the band structure of the material in 3D.

Appendix C. Advanced physics in ultrascaled devices with UTOXPP solver

Appendix D

Charge trapping effects: a modeling study

D.1 Introduction

In this Annex, a modeling study has been performed to investigate the effects of model and trap distribution parameters on the electrical characteristics of a degraded device. The methodology based on the non-radiative multiphonon-assisted charge trapping model of Chapter 3 and from [75] is adopted to compare the parameter sensitivity of three widely diffused characterization methods: impedance analysis (Section D.2), TAT characterization (Section D.3) and MFCP extraction (Section D.4). This permits a simpler and effective extraction of trap concentrations in aged devices and a critical comparison of the methods.

D.2 AC analysis

The extension of the probed region determined from AC characteristics is firstly analysed to determine the impact of the trap model parameters, such as the Huang Rhys factor S and the critical field F_C in Eq. 3.7. In the following simulations, the $T_{OX} = 50$Å NMOS device of Figure 3-9(f) has been used as a reference. Figure D-1 indicates that the probed region in AC is strongly reduced towards the Si conduction and valence band edges for high values of the Huang-Rhys factor S of Eq. 3.8, while for values around 10 the region extension is maximum. The energy dependence is not changed for smaller values of S.

Similar considerations could be applied to the dependence on the critical field F_C. The increase of the critical field F_C causes a slight reduction of the maximum oxide depth at which defects can be scanned. This effect is highlighted in Figure D-2. The relevance of using a non-zero F_C for P_b centers can be questionable. Indeed, the exponential dependence of the rates on the field has been associated to deeper E' border defects [184]. However, due to self-consistence, the trapped charge and the wavefunction penetration in the oxide reduce the electric field at the interface and the prefactor in Eq. 3.7 tends to 1. Consequently, the influence of F_C on the CV probed region extension can be considered negligible for

interface defects in thick oxides.

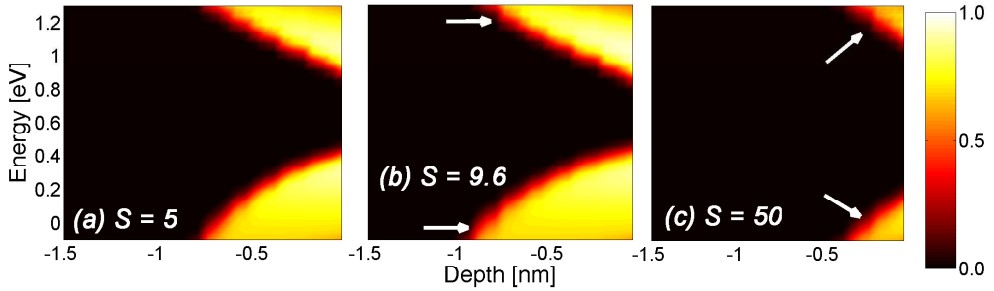


Figure D-1 – Probed region calculated using Eq. 3.33 at 1kHz and varying the value of the Huang-Rhys factor S from 5 to 50. While for S around the decade, only variations in depth are identifiable and defects placed in energy near the Si midgap can be scanned, higher values of this parameter drastically reduce the probed region and consequently the trap capacitive response. Arrows indicate the shrinking of the probed region for increasing values of S.

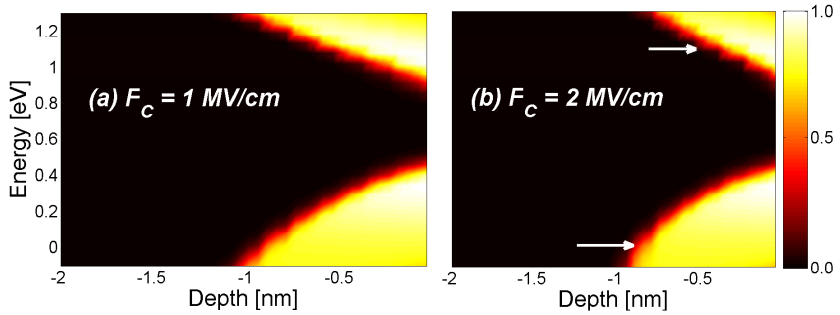


Figure D-2 – Probed region with AC technique as a function of oxide depth and trap energy, calculated using Eq. 3.33 at 1kHz and for different values of critical field F_C. Marginal variations in depth are visible. The Si/SiO_2 interface is placed at depth x=0nm, while the conduction and valence band edges of Si correspond to 0eV and 1.2eV, respectively. Arrows indicate the depth dependence of the probed region with F_C.

The region that can be probed using an AC pulse has also been calculated for three temperatures at ν = 1kHz, and evidenced how at high temperatures almost all midgap defects can be scanned (Figure D-3). Only the energy-dependence of the probed region distribution changes, while in oxide depth the increase is less pronounced.

Model parameters dependencies have also been investigated studying the effects of the critical field F_C and Huang-Rhys factor S in Eq. 3.7 on the channel and bulk trap capacitance components C_{GC}^T and C_{GB}^T. Marginal variations are noticed when F_C is reduced, while an increase of S reduces the frequency pole. This effect is associated to the reduction of the probed region towards the conduction/valence band edges previously shown in Figure D-1.

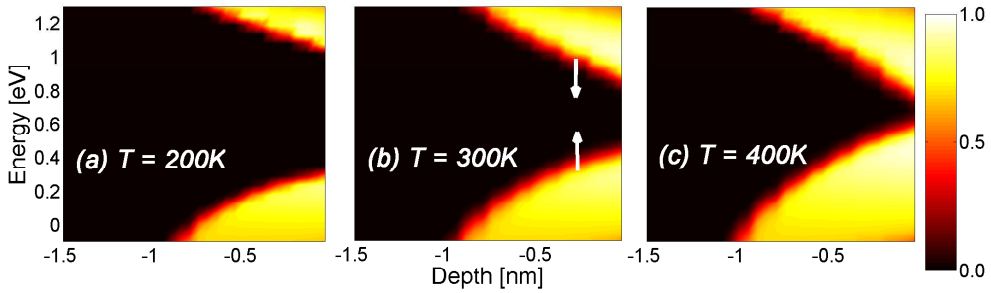


Figure D-3 – Temperature dependence of the probed region, simulated using a small-signal at 1kHz and plotted as a function of trap energy and trap position in the oxide. The Si/SiO_2 interface is placed at depth x=0nm, while the conduction and valence band edges of Si correspond to 0eV and 1.2eV, respectively. Arrows indicate the energy dependence of the probed region with temperature.

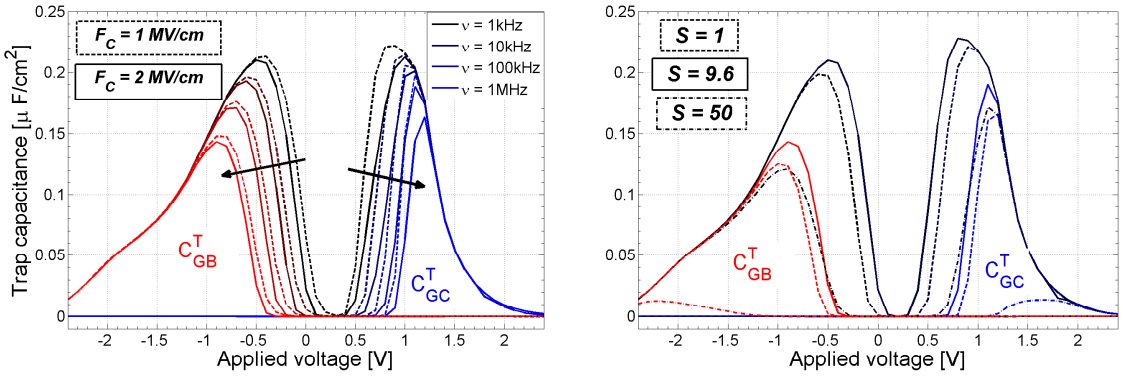


Figure D-4 – Simulated trap capacitances C^T_{GB} and C^T_{GC} on the device having $T_{OX} = 50$Åversus applied gate voltage and small-signal frequencies, calculated for different values of the critical field F_C (left) and Huang-Rhys factor S (right). In both the cases trap capacitive response increases from high (1MHz) to lower frequencies (1kHz). Arrows indicate the decrease of the response for increasing frequencies.

The dependence of the gate capacitances on trap distribution profiles and modeling parameters has been investigated adopting a bivariate Gaussian distribution to model the spatial and energetic trap distributions in the oxide layer:

$$N_T(x, E_T) = \frac{n_{T0}}{2\pi\sigma_x\sigma_E} \exp\left(-\frac{1}{2}\left[\frac{(x - x_T)^2}{\sigma_x^2} + \frac{(E_T - E_{T0})^2}{\sigma_E^2}\right]\right) \tag{D.1}$$

Each distribution parameter has been independently varied around the extracted reference values in Chapter 3. Considering the amphoteric nature of P_b center interface defects, for increasing values of N_T, the threshold voltage V_{th} and flat-band simulated V_{fb} voltage stretch-outs are subject to an increase. Simulations of the gate capacitance for three trap concentrations (Figure D-5) also show the increase of the parasitic frequency peaks on the

total gate capacitance C_{GG} for higher N_T.

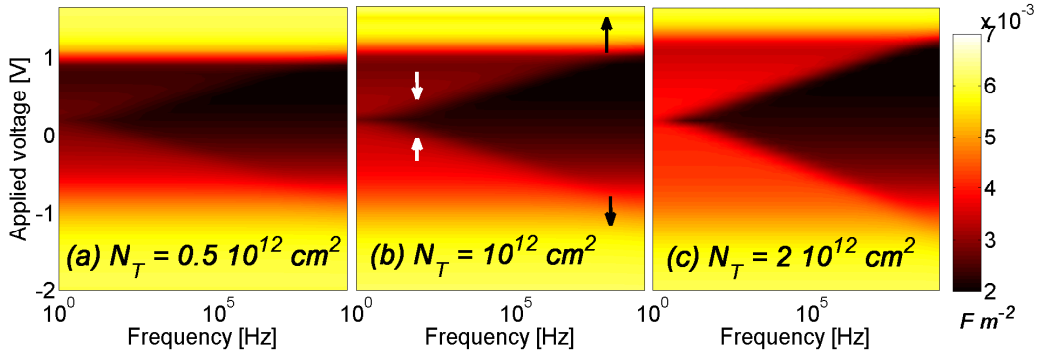


Figure D-5 – Simulated gate capacitance for the 50Å device plotted in color as a function of the applied gate voltage and the small signal frequency for different total trap concentrations. The increase of the defect concentration induces an increase of the stretch-out in both inversion and accumulation and an increase of the parasitic trap capacitances in depletion.

Defects located far away from the Si/SiO_2 interface having smaller C/E rates do not follow the small-signal component of the applied AC voltage and thus their response is visible only at low frequencies. In Figure D-6, the mean depth of the distribution x_T has been increased from 1Å up to 10Å from the Si/SiO_2 interface. Consequently, the trap parasitic component strongly decreases, i.e. lower frequencies are required to achieve the same frequency response of deeper poles. Since a total constant concentration of defects is maintained and considering the probed region minimally affected by the different defect charging in the dielectric, the trap concentration in the scanned region and their response decrease. Similar effects are also visible when the spatial variance σ_x of the distribution is increased from 1Å to 10Å. In this case, a larger spreading of the parasitic pole and of its frequency-dependence also occurs (Figure D-7). Indeed, when increasing the trap variance, the defects are more spread in the dielectric depth, resulting in the convolution in the trap response of a wide range of cut-off frequencies.

The mean position in energy E_{T0} affects both the position of the AC peak and its voltage dependence. Due to the energy dependence of the C/E rates [218], traps placed near the Si valence band influence the bulk capacitance in accumulation, while, for defects placed near the Si conduction band, a parasitic capacitance is added in weak inversion on the C_{GC} capacitance. Please notice that the pole frequency in this case is not affected. Effects similar to Figure D-7 are found when modifying the variance in energy, which spreads the frequency poles as previously shown for σ_X.

D.3 TAT analysis

The trap model parameters have a considerable impact on the TAT characteristics, due to their exponential effects on the capture/emission trapping rates. Given the variation of

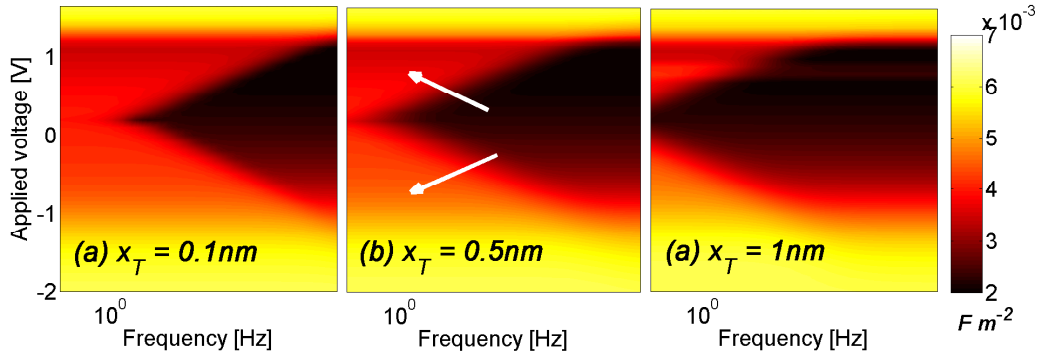


Figure D-6 – Simulated gate capacitance as a function of the applied gate voltage and the small signal frequency, varying the depth x_T of the trap concentration. Deeper defects respond to lower frequencies and consequently when the defect depth is increased the capacitance response is reduced.

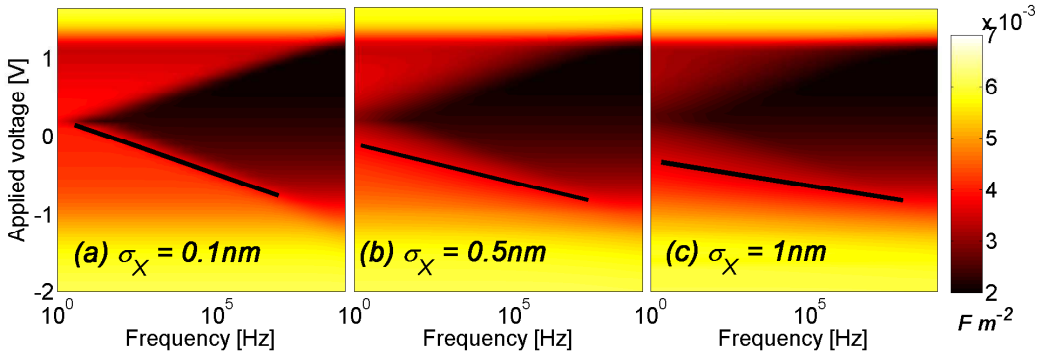


Figure D-7 – Same simulation setup as for Figure D-6, but increasing the variance of the spatial Gaussian distribution. In addition to the decrease of the trap capacitive response, the cut-off frequencies are more spread and the identification of a single pole is more difficult.

the probed region extension in the dielectric, the influence of the model parameters is also strongly dependent on the trap distribution profile.

Also in this case, the spatial/energetic profile is maintained constant and the impact of model parameters is analysed. Figure D-9 depicts the gate current components versus the applied gate voltage for different values of the critical field F_C (left) and the Huang-Rhys factor S (right). The critical field induces a variation of the voltage dependence of the TAT component: for high critical fields, and at high oxide fields, the capture/emission rates are exponentially reduced, lowering the total current contribution. For increasing values of S, the current exponentially increases due to the increase of the peak of the probed region with V_G.

The effects of the distribution profile variations are studied. The total trap concentration is maintained constant to achieve similar device electrostatics to all the variations. For the considered distribution profile of E' defects, the current is shown to increase when E_{T0}

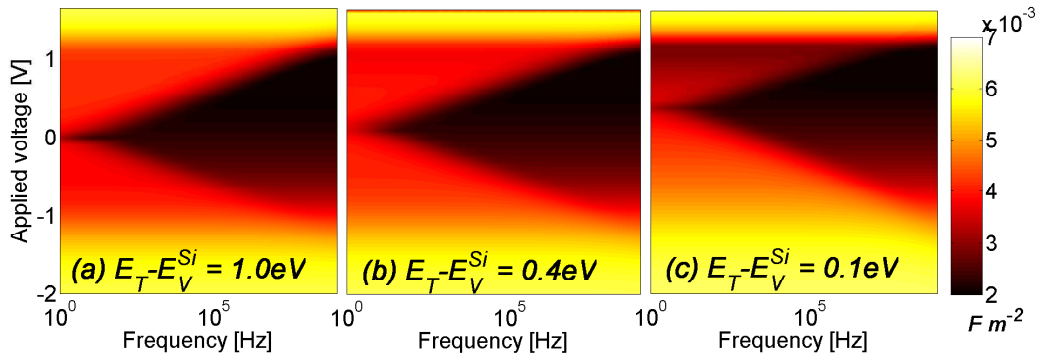


Figure D-8 – Same simulation setup as for Figure D-6, but varying the mean position E_{T0} of the distribution in energy. When defects are in proximity of the Si valence band the hole contribution in accumulation and depletion is dominant. Defects near the Si conduction band mostly contribute to the increase of the channel capacitance component in weak inversion.

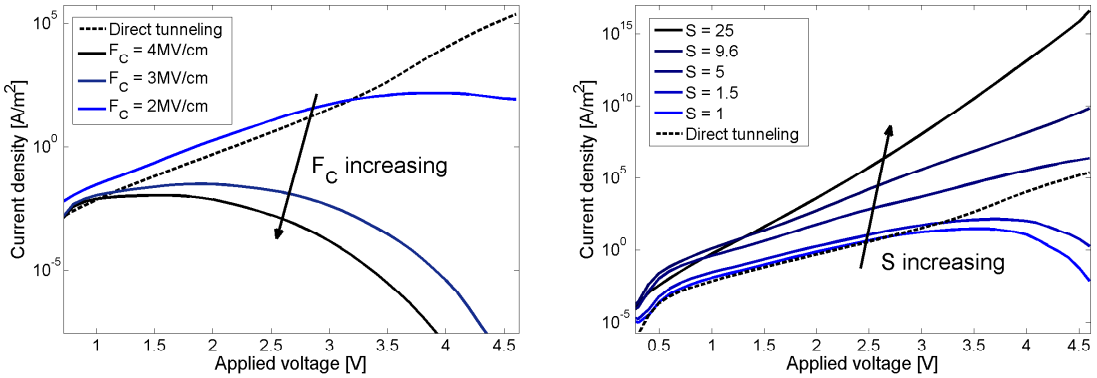


Figure D-9 – Simulated TAT current density vs. applied bias for different values of the critical field F_C (left) and Huang-Rhys factor (right).

is varied from 1.5eV to 0.9eV, as the distribution centers towards the probed region and a larger amount of defects participate to TAT (Figure D-10(left)). The defect depth increase indicated in the bottom plot also results in an increase of the TAT current, confirming that defects in the middle of the oxide have a major role in contributing to TAT. This last dependence is more pronounced for distributions having small spatial variances. The plots presented in Figures D-9 and D-10, where the influence of the parameters remains difficult to be interpreted, indicate how an accurate defect extraction should mostly rely on the determination of the probed region.

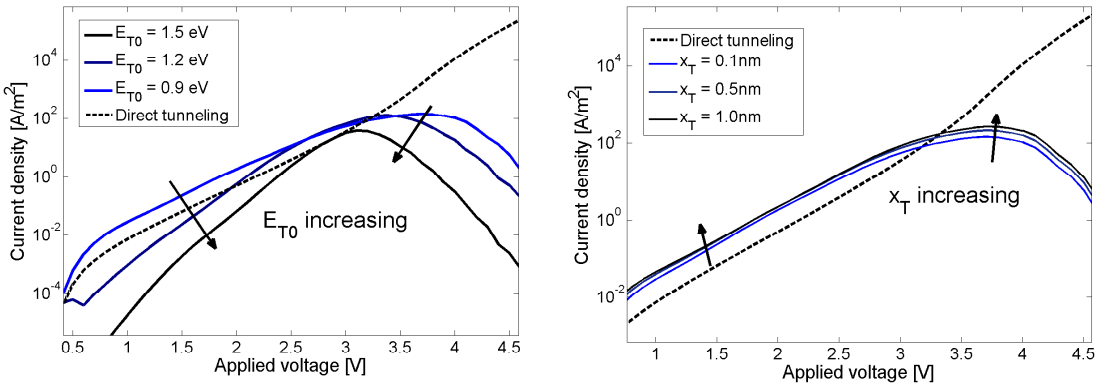


Figure D-10 – TAT current density vs. applied bias varying the mean energy and spatial position (left - E_{T0}; right - x_T of the defect distribution).

D.4 Charge pumping analysis

A development similar to the one presented in Sections D.2 and D.3 has also been performed considering multi-frequency charge pumping characterization. Also in this case, a device having 5nm of oxide thickness has been adopted as a reference for evaluating both the probed region and the charge pumping current in different bias configurations.

The CP probed region depends on the critical field F_C and on the Huang-Rhys factor S. An increase of F_C causes midgap C/E rates to decrease, with a corresponding reduction of the scanned area (Figure D-11). The influence of F_C on the probed region increases for deeper defects, due to the non-uniform field distribution in the oxide layer when traps are introduced. This is only taken into account when self-consistence in the device electrostatics is considered. The probed region variation that is represented in Figure D-12 for different values of S is associated to the increase of the C/E frequencies as well.

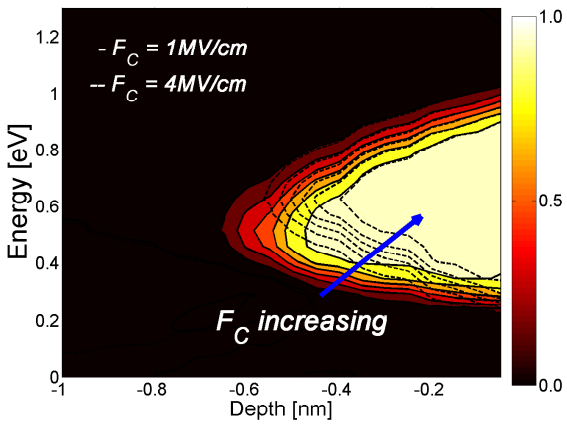


Figure D-11 – Dependence of the probed region in CP analysis with the critical field F_C plotted at $V_{low} = -1V$, ν=10kHz and $V_{sw} = 2.5V$. The Si/SiO_2 interface is placed at depth x=0nm, while the conduction and valence band edges of Si correspond to 0eV and 1.2eV, respectively.

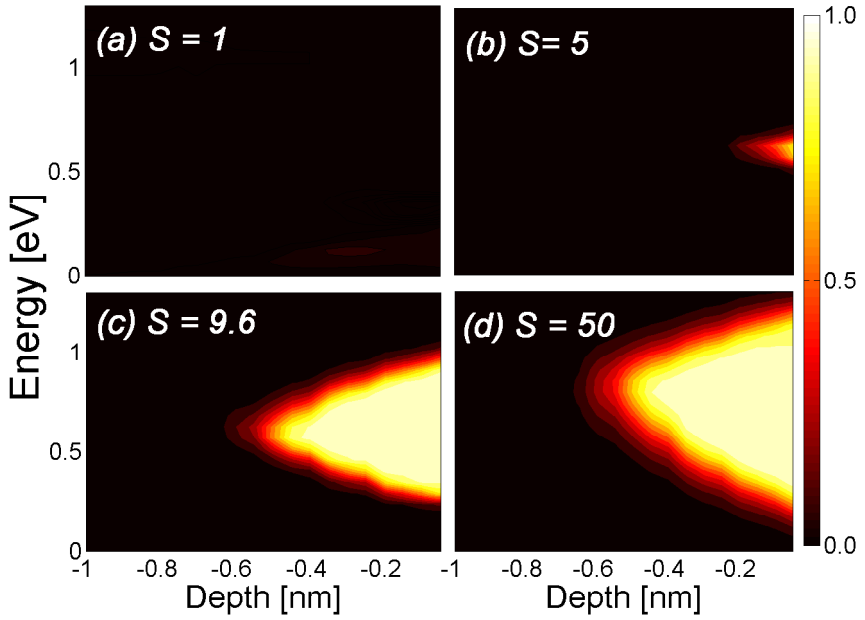


Figure D-12 – Dependence of the probed region in CP analysis with the Huang-Rhys factor S plotted at $V_{low} = -1V$, ν=10kHz and $V_{sw} = 2.5V$. The Si/SiO_2 interface is placed at depth x=0nm, while the conduction and valence band edges of Si correspond to 0eV and 1.2eV, respectively.

Figure D-13 shows the dependence of the charge pumping current with the trap spatial profile parameters: in the left plot, the increase of the spatial variance causes the smoothing of the trap pole frequency since deeper traps have lower C/E rates, while on the right the current variation with trap depth x_T is indicated. Consequently traps that are not included in the probed region do not contribute to the CP current, which strongly decreases. In Figure D-14 the position of the trap distribution profile in energy is varied from the edge of the Si conduction band to the edge of the Si valence band. The maximum response is achieved when the defect distribution is centered in the midgap. This is in accordance with the previously shown symmetry of the probed region in energy. Figure D-14 also shows the dependence on the total defect concentration. At higher concentrations the peak is more pronounced due to charging effects and self-consistency.

Large variations have been noticed on the charge pumping current simulated over a broad range of temperatures, and presented in Figure D-15. This dependence has been attributed to the strong variations of the scanned region extension versus temperature. Indeed for low and moderate voltages, the region is enlarged for increasing temperatures, while it tends to saturate at higher temperatures, due to the increased probability of C/E of carriers of the same type in proximity of E_C^{Si} and E_V^{Si}.

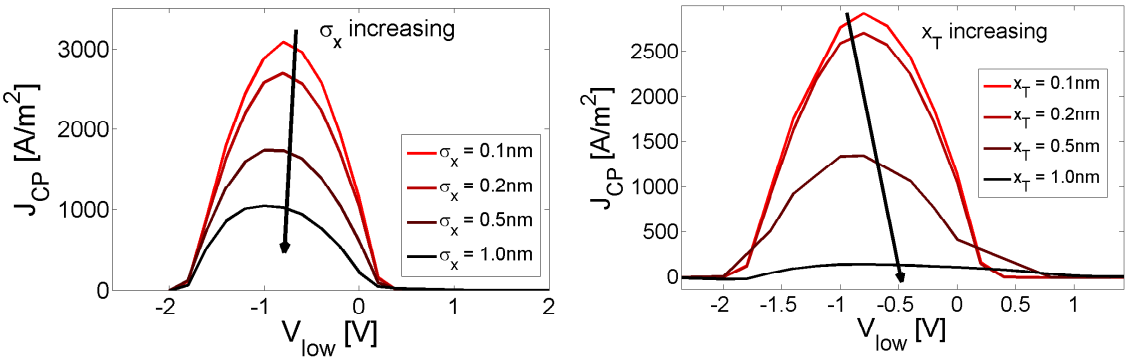


Figure D-13 – CP current vs. V_{low} varying the variance (left) and the mean position of the spatial distribution of defects (right). For these simulations $V_{sw} = 2.5V$ and ν =1kHz.

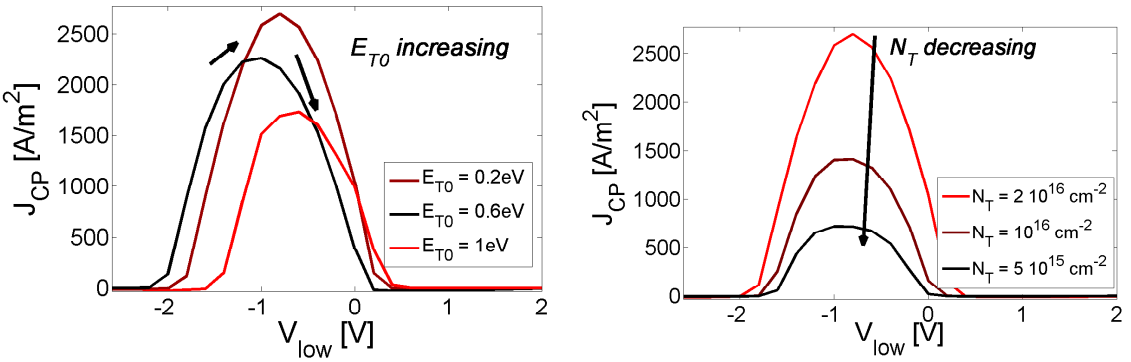


Figure D-14 – CP current vs. V_{low} varying the variance of the average energetic position of the profile (left) and the total defect concentration (right). For these simulations $V_{sw} = 2.5V$ and ν =1kHz.

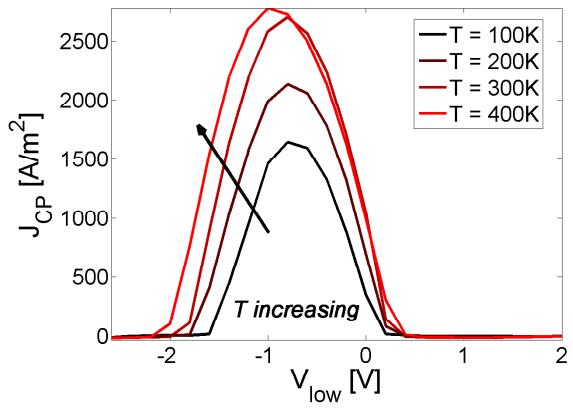


Figure D-15 – Charge pumping current vs. V_{low} as a function of temperature. Saturation occurs for some voltage biases when temperature increases, due to the extension of the probed region in the oxide layer which saturates at higher temperatures.

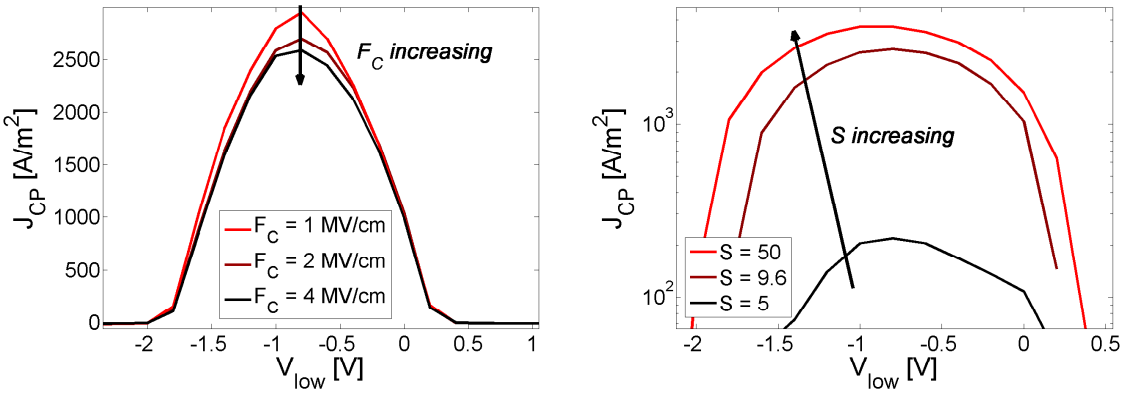


Figure D-16 – Dependence of the CP current vs. V_{low} voltage on the critical field F_C in (left, in linear scale) and Huang-Rhys factor S in (right, in logarithmic scale).

D.5 Conclusion

The analysis of the sensitivity to parameters represents an important aspect of a new model, in particular when their extraction is relatively complex, e.g. for the Huang Rhys factor S, or when they are derived from empirical phenomena, e.g. for the critical field F_C. For border defects, we have been able to conclude that the dependence on F_C is quite weak in thick oxides. Considering the empirical nature of this parameter, this result is encouraging, as the accuracy of the extracted defect profile would not significantly depend on the value of F_C. On the other hand, TAT current shows a large variation with both F_C and S and thus additional extraction of these two parameters need to be performed with alternative techniques. The relative importance of these parameters is expected to increase in future technologies exploiting novel oxide stacks.

Bibliography

[1] S. Lai, "Flash memories: Successes and challenges," *IBM Journal of Research and Development*, vol. 52, no. 4.5, pp. 529–535, 2008.

[2] S. Lai, "Non-volatile memory technologies: The quest for ever lower cost," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1–6, IEEE, 2008.

[3] G. Burr, M. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, L. Lastras, A. Padilla, *et al.*, "Phase change memory technology," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 28, p. 223, 2010.

[4] K. Natori and F. Masuoka, "Semiconductor memory device," 4 1980. US Patent 4,199,772.

[5] H. Iwahashi and F. Masuoka, "Semiconductor memory device." -, January 1987. US Patent 4,635,232.

[6] R. Barth, "Itrs commodity memory roadmap," tech. rep., ITRS, 2003.

[7] A. Cunha, M. Schneider, C. Galup-Montoro, C. Caetano, and M. Machado, "Unambiguous extraction of threshold voltage based on the transconductance-to-current ratio," *Proceedings of Nanotech*, vol. 3, pp. 139–42, 2005.

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

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

[10] C. Yih, Z. Ho, M. Liang, and S. Chung, "Characterization of hot-hole injection induced SILC and related disturbs in flash memories," *Electron Devices, IEEE Transactions on*, vol. 48, no. 2, pp. 300–306, 2001.

[11] D. Ielmini, A. Ghetti, A. Spinelli, and A. Visconti, "A study of hot-hole injection during programming drain disturb in flash memories," *Electron Devices, IEEE Transactions on*, vol. 53, no. 4, pp. 668–676, 2006.

[12] J. Watts, C. McAndrew, C. Enz, C. Galup-Montoro, G. Gildenblat, C. Hu, R. van Langevelde, M. Miura-Mattausch, R. Rios, and C. Sah, "Advanced Compact Models for MOSFETs," in *Technical Proceedings Workshop on Compact Modeling (WCM)*, pp. 3–12, 2006.

[13] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *Solid-State Circuits, IEEE Journal of*, vol. 22, no. 4, pp. 558–566, 1987.

[14] C. Hu, "BSIM model for circuit design using advanced technologies," in *VLSI Circuits, 2001. Digest of Technical Papers. 2001 Symposium on*, pp. 5–10, IEEE, 2001.

[15] W. Liu, *MOSFET models for SPICE simulation, including BSIM3v3 and BSIM4*. Wiley, 2001.

[16] H. Graaff and F. Klaassen, *Compact transistor modelling for circuit design*. Springer-Verlag, 1990.

[17] J. Brews, "A charge-sheet model of the MOSFET," *Solid-State Electron*, vol. 21, no. 2, pp. 345–355, 1978.

[18] H. Pao and C. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," *Solid-State Electronics*, vol. 9, no. 10, pp. 927–937, 1966.

[19] C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog integrated circuits and signal processing*, vol. 8, no. 1, pp. 83–114, 1995.

[20] M. Bucher, C. Lallement, C. Enz, F. Théodoloz, and F. Krummenacher, "The EPFL-EKV MOSFET Model Equations for Simulation," tech. rep., EPFL, 1998.

[21] J. Sallese, M. Bucher, F. Krummenacher, and P. Fazan, "Inversion charge linearization in MOSFET modeling and rigorous derivation of the EKV compact model," *Solid-State Electronics*, vol. 47, no. 4, pp. 677–683, 2003.

[22] C. Enz and E. Vittoz, *Charge-based MOS transistor modeling*. Wiley Online Library, 2006.

[23] M. Miura-Mattausch, N. Sadachika, D. Navarro, G. Suzuki, Y. Takeda, M. Miyake, T. Warabino, Y. Mizukane, R. Inagaki, T. Ezaki, *et al.*, "HiSIM2: Advanced MOSFET model valid for RF circuit simulation," *Electron Devices, IEEE Transactions on*, vol. 53, no. 9, pp. 1994–2007, 2006.

[24] R. Van Langevelde, A. Scholten, and D. Klaassen, "Recent enhancements of MOS Model 11," in *Workshop on Compact Modeling, NSTI-Nanotech 2004*, vol. 2, pp. 60–65, 2004.

[25] G. Gildenblat, C. McAndrew, H. Wang, W. Wu, D. Foty, L. Lemaitre, and P. Bendix, "Advanced compact models: gateway to modern CMOS design," in *Electronics, Circuits and Systems, 2004. ICECS 2004. Proceedings of the 2004 11th IEEE International Conference on*, pp. 638–641, 2004.

[26] G. Gildenblat, H. Wang, T. Chen, X. Gu, and X. Cai, "SP: An advanced surface-potential-based compact MOSFET model," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 9, pp. 1394–1406, 2004.

[27] G. Gildenblat, X. Li, H. Wang, W. Wu, R. van Langevelde, A. Scholten, G. Smit, and D. Klaassen, "Introduction to PSP MOSFET model," in *WCM*, pp. 19–24, 2005.

[28] G. Gildenblat, X. Li, H. Wang, W. Wu, A. Jha, R. van Langevelde, A. Scholten, G. Smit, and D. Klaassen, "[theory and modeling techniques used in the psp model,"

[29] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. Van Langevelde, G. Smit, A. Scholten, and D. Klaassen, "PSP: An Advanced Surface-Potential-Based MOSFET Model for Circuit Simulation," *Electron Devices, IEEE Transactions on*, vol. 53, no. 9, p. 1979, 2006.

[30] A. Bhattacharyya, "Modelling of write/erase and charge retention characteristics of floating gate eeprom devices," *Solid-state electronics*, vol. 27, no. 10, pp. 899–906, 1984.

[31] A. Kolodny, S. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of FG EEPROM cells," *Electron Devices, IEEE Transactions on*, vol. 33, no. 6, pp. 835–844, 1986.

[32] K. San, C. Kaya, D. Liu, T. Ma, and P. Shah, "A new technique for determining the capacitive coupling coefficients in flash EPROMs," *Electron Device Letters, IEEE*, vol. 13, no. 6, pp. 328–331, 1992.

[33] W. Choi and D. Kim, "A new technique for measuring coupling coefficients and 3-D capacitance characterization of floating-gate devices," *Electron Devices, IEEE Transactions on*, vol. 41, no. 12, pp. 2337–2342, 1994.

[34] 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," *Electron Device Letters, IEEE*, vol. 13, no. 11, pp. 566–568, 1992.

[35] R. Bez, E. Camerlenghi, D. Cantarelli, L. Ravazzi, G. Crisenza, S. Microelectron, and A. Brianza, "A novel method for the experimental determination of the coupling ratios in submicron eprom and flash eeprom cells," in *Electron Devices Meeting, 1990. Technical Digest., International*, pp. 99–102, 1990.

[36] B. Moison, C. Papadas, G. Ghibaudo, P. Mortini, and G. Pananakakis, "New method for the extraction of the coupling ratios in FLOTOX EEPROM cells," *Electron Devices, IEEE Transactions on*, vol. 40, no. 10, pp. 1870–1872, 1993.

[37] R. Duane, A. Concannon, P. O'Sullivan, and A. Mathewson, "Advanced numerical modelling of non-volatile memory cells," in *Solid-State Device Research Conference, 1998. Proceeding of the 28th European*, pp. 304–307, IEEE, 1998.

[38] L. Larcher, P. Pavan, L. Albani, and T. Ghilardi, "Bias and W/L dependence of capacitive coupling coefficients in FG memory cells," *Electron Devices, IEEE Transactions on*, vol. 48, no. 9, pp. 2081–2089, 2001.

[39] R. Fowler and L. Nordheim, "Electron emission in intense electric fields," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 119, no. 781, pp. 173–181, 1928.

[40] M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *Electron Devices, IEEE Transactions on*, vol. 15, no. 9, pp. 686–686, 1968.

[41] R. Forbes, "Refining the application of Fowler-Nordheim theory," *Ultramicroscopy*, vol. 79, no. 1-4, pp. 11–23, 1999.

[42] J. Sune, M. Lanzoni, R. Bez, P. Olivo, and B. Ricco, "Transient simulation of the erase cycle of floating gate EEPROMs," in *Electron Devices Meeting, 1991. IEDM'91. Technical Digest., International*, pp. 905–908, IEEE, 1991.

[43] M. Lanzoni, J. Sune, P. Olivo, and B. Ricco, "Advanced electrical-level modeling of EEPROM cells," *Electron Devices, IEEE Transactions on*, vol. 40, no. 5, pp. 951–957, 1993.

[44] S. Keeney, R. Bez, D. Cantarelli, F. Piccinini, A. Mathewson, L. Ravazzi, and C. Lombardi, "Complete transient simulation of flash EEPROM devices," *Electron Devices, IEEE Transactions on*, vol. 39, no. 12, pp. 2750–2757, 1992.

[45] K. Sonoda, M. Tanizawa, S. Shimizu, Y. Araki, S. Kawai, T. Ogura, S. Kobayashi, K. Ishikawa, T. Eimori, Y. Inoue, *et al.*, "Compact modeling of a flash memory cell including substrate-bias-dependent hot-electron gate current," *Electron Devices, IEEE Transactions on*, vol. 51, no. 10, pp. 1726–1733, 2004.

[46] L. Larcher, P. Pavan, S. Pietri, L. Albani, and A. Marmiroli, "A new compact DC model of FG memory cells without capacitive coupling coefficients," *Electron Devices, IEEE Transactions on*, vol. 49, no. 2, pp. 301–307, 2002.

[47] R. Bouchakour, N. Harabech, P. Canet, J. Mirabel, P. Boivin, and O. Pizzuto, "A new physical-based compact model of floating-gate eeprom cells," *Journal of Non-Crystalline Solids*, vol. 280, no. 1-3, pp. 122–126, 2001.

[48] R. Bouchakour, N. Harabech, P. Canet, P. Boivin, and J. Mirabel, "Modeling of a floating-gate eeprom cell using a charge sheetapproach including variable tunneling capacitance and polysilicon gatedepletion effect," in *Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on*, vol. 4, 2001.

[49] A. Maure, P. Canet, F. Lalande, B. Delsuc, and J. Devin, "Flash Memory Cell Compact Modeling Using PSP Model," in *Behavioral Modeling and Simulation Workshop, 2008. BMAS 2008. IEEE International*, pp. 45–49, IEEE, 2008.

[50] R. Tsu and L. Esaki, "Tunneling in a finite superlattice," *Applied Physics Letters*, vol. 22, p. 562, 1973.

[51] D. Cassi and B. Ricco, "An analytical model of the energy distribution of hot electrons," *Electron Devices, IEEE Transactions on*, vol. 37, no. 6, pp. 1514–1521, 1990.

[52] A. Abramo and C. Fiegna, "Electron energy distributions in silicon structures at low applied voltages and high electric fields," *Journal of applied physics*, vol. 80, no. 2, pp. 889–893, 1996.

[53] K. Sonoda, M. Yamaji, K. Taniguchi, C. Hamaguchi, and S. Dunham, "Moment expansion approach to calculate impact ionization rate in submicron silicon devices," *Journal of applied physics*, vol. 80, p. 5444, 1996.

[54] C. Hu, "Lucky-electron model of channel hot electron emission," in *Electron Devices Meeting, 1979 International*, vol. 25, pp. 22–25, IEEE, 1979.

[55] S. Tam, P. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFETs," *Electron Devices, IEEE Transactions on*, vol. 31, no. 9, pp. 1116–1125, 1984.

[56] A. Gehring, *Simulation of Tunneling in Semiconductor Devices*. PhD thesis, University of Wien, 2003.

[57] A. Ghetti, A. Hamad, P. Silverman, H. Vaidya, and N. Zhao, "Self-consistent simulation of quantization effects and tunneling current in ultra-thin gate oxide MOS devices," in *Simulation of Semiconductor Processes and Devices, 1999. SISPAD'99. 1999 International Conference on*, pp. 239–242, IEEE, 1999.

[58] S. Jin, A. Wettstein, W. Choi, F. Bufler, and E. Lyumkis, "Gate current calculations using spherical harmonic expansion of Boltzmann equation," in *Simulation of Semiconductor Processes and Devices, 2009. SISPAD'09. International Conference on*, pp. 1–4, IEEE, 2009.

[59] T. Grasser, H. Kosina, and S. Selberherr, "Influence of the distribution function shape and the band structure on impact ionization modeling," *Journal of Applied Physics*, vol. 90, p. 6165, 2001.

[60] T. Grasser, H. Kosina, M. Gritsch, and S. Selberherr, "Using six moments of Boltzmann transport equation for device simulation," *Journal of Applied Physics*, vol. 90, p. 2389, 2001.

[61] T. Grasser, H. Kosina, C. Heitzinger, and S. Selberherr, "Characterization of the hot electron distribution function using six moments," *Journal of Applied Physics*, vol. 91, p. 3869, 2002.

[62] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi, and B. Ricco, "Simple and efficient modeling of EPROM writing," *Electron Devices, IEEE Transactions on*, vol. 38, no. 3, pp. 603–610, 1991.

[63] L. Larcher and P. Pavan, "A New Analytical Model of Channel Hot Electron (CHE) and CHannel Initiated Secondary ELectron (CHISEL) Current Suitable for Compact Modeling," in *Modeling and Simulation of Microsystems*, pp. 738–741, 2002.

[64] S. Satoh, G. Hemink, K. Hatakeyama, and S. Aritome, "Stress-induced leakage current of tunnel oxide derived from flashmemory read-disturb characteristics," *Electron Devices, IEEE Transactions on*, vol. 45, no. 2, pp. 482–486, 1998.

[65] E. Rosenbaum and L. Register, "Mechanism of stress-induced leakage current in MOS capacitors," *Electron Devices, IEEE Transactions on*, vol. 44, no. 2, pp. 317–323, 1997.

[66] S. Takagi, N. Yasuda, and A. Toriumi, "A new IV model for stress-induced leakage current includinginelastic tunneling," *Electron Devices, IEEE Transactions on*, vol. 46, no. 2, pp. 348–354, 1999.

[67] S. Takagi, N. Yasuda, and A. Toriumi, "Experimental evidence of inelastic tunneling in stress-induced leakage current," *Electron Devices, IEEE Transactions on*, vol. 46, no. 2, pp. 335–341, 1999.

[68] D. Ielmini, A. Spinelli, A. Lacaita, A. Modelli, and D. e Informazione, "A statistical model for SILC in flash memories," *Electron Devices, IEEE Transactions on*, vol. 49, no. 11, pp. 1955–1961, 2002.

[69] D. Ielmini, A. Spinelli, M. Rigamonti, and A. Lacaita, "Modeling of SILC based on electron and hole tunneling. II. Steady-state," *Electron Devices, IEEE Transactions on*, vol. 47, no. 6, pp. 1266–1272, 2002.

[70] F. Jiménez-Molinos, A. Palma, F. Gamiz, J. Banqueri, and J. Lopez-Villanueva, "Physical model for trap-assisted inelastic tunneling in metal-oxide-semiconductor structures," *Journal of Applied Physics*, vol. 90, no. 7, pp. 3396–3404, 2001.

[71] F. Jiménez-Molinos, A. Palma, A. Gehring, F. Gámiz, H. Kosina, and S. Selberherr, "Static and transient simulation of inelastic trap-assisted tunneling," in *Proc. 14th Workshop on Modeling and Simulation of Electron Devices*, pp. 65–68, 2003.

[72] L. Terman, "An investigation of surface states at a silicon/silicon oxide interface employing metal-oxide-silicon diodes," *Solid-State Electronics*, vol. 5, no. 5, pp. 285–299, 1962.

[73] E. Nicollian and J. Brews, *MOS/metal oxide semiconductor/physics and technology*. 1982.

[74] P. Masson, J. Autran, M. Houssa, X. Garros, and C. Leroux, "Frequency characterization and modeling of interface traps in $HFSi_xO_y/HFO_2$ gate dielectric stack from a capacitance point-of-view," *Applied Physics Letters*, vol. 81, pp. 3392–3394, 2002.

[75] D. Garetto, Y. Mamy-Randriamihaja, A. Zaka, D. Rideau, A. Schmid, Jaouen, Herve, and Y. Leblebici, "Modeling of stressed mos oxides using a multiphonon-assisted quantum approach - part i: impedance analysis," *Accepted to IEEE Transactions of Electron Devices*, vol. 59-3, pp. –, 2011.

[76] B. Ricco, G. Gozzi, and M. Lanzoni, "Modeling and simulation of stress-induced leakage current in ultrathin SiO< sub> 2</sub> films," *Electron Devices, IEEE Transactions on*, vol. 45, no. 7, pp. 1554–1560, 2002.

[77] B. Ricco, G. Gozzi, and M. Lanzoni, "Modeling and simulation of stress-induced leakage current inultrathin SiO_2 films," *Electron Devices, IEEE Transactions on*, vol. 45, no. 7, pp. 1554–1560, 1998.

[78] J. De Blauwe, J. van Heudt, D. Wellekens, G. Groeseneken, and H. Maes, "SILC-related effects in flash E2PROM's-Part I: A quantitative model for steady-state SILC," *Electron Devices, IEEE Transactions on*, vol. 45, no. 8, pp. 1745–1750, 1998.

[79] J. de Blauwe, J. van Heudt, D. Wellekens, G. Groeseneken, H. Maes, and L. IMEC, "Silc-related effects in flash e 2 prom's-part ii: Prediction of steady-state silc-related disturb characteristics," *Electron Devices, IEEE Transactions on*, vol. 45, no. 8, pp. 1751–1760, 1998.

[80] L. Larcher, A. Paccagnella, and G. Ghidini, "Gate current in ultrathin MOS capacitors: a new model of tunnelcurrent," *Electron Devices, IEEE Transactions on*, vol. 48, no. 2, pp. 271–278, 2001.

[81] L. Larcher, A. Paccagnella, and G. Ghidini, "A model of the stress induced leakage current in gate oxides," *Electron Devices, IEEE Transactions on*, vol. 48, no. 2, pp. 285–288, 2001.

[82] L. Larcher, S. Bertulu, and P. Pavan, "SILC effects on E2PROM memory cell reliability," *Device and Materials Reliability, IEEE Transactions on*, vol. 2, no. 1, pp. 13–18, 2002.

[83] L. Larcher, "Statistical simulation of leakage currents in MOS and flash memory devices with a new multiphonon trap-assisted tunneling model," *Electron Devices, IEEE Transactions on*, vol. 50, no. 5, pp. 1246–1253, 2003.

[84] L. Larcher and P. Pavan, "Statistical simulations to inspect and predict data retention and program disturbs in Flash memories," in *Electron Devices Meeting, 2003. IEDM'03 Technical Digest. IEEE International*, pp. 7–3, IEEE, 2003.

[85] L. Vandelli, A. Padovani, L. Larcher, R. Southwick, W. Knowlton, and G. Bersuker, "Modeling temperature dependency (6–400K) of the leakage current through the SiO2/high-K stacks," in *Solid-State Device Research Conference (ESSDERC), 2010 Proceedings of the European*, pp. 388–391, IEEE, 2010.

[86] L. Vandelli, A. Padovani, L. Larcher, G. Bersuker, J. Yum, and P. Pavan, "A physics-based model of the dielectric breakdown in HfO_2 for statistical reliability prediction," in *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pp. GD–5, IEEE, 2011.

[87] L. Vandelli, A. Padovani, G. Bersuker, D. Gilmer, P. Pavan, and L. Larcher, "Modeling of the Forming Operation in HfO2-Based Resistive Switching Memories," in *Memory Workshop (IMW), 2011 3rd IEEE International*, pp. 1–4, IEEE, 2011.

[88] C. Chen and T. Ma, "Direct lateral profiling of hot-carrier-induced oxide charge and interface traps in thin gate mosfet's," *Electron Devices, IEEE Transactions on*, vol. 45, no. 2, pp. 512–520, 1998.

[89] A. Brand, K. Wu, S. Pan, and D. Chin, "Novel read disturb failure mechanism induced by flash cycling," in *Reliability Physics Symposium, 1993. 31st Annual Proceedings., International*, pp. 127–132, IEEE, 1993.

[90] G. Campardo, R. Micheloni, and D. Novosel, *VLSI design of nonvolatile memories*. Springer, 2005.

[91] K. Hass and D. Cox, "Level shifting interfaces for low voltage logic," in *9th NASA Symposium on VLSI Design*, pp. 3–1, 2000.

[92] D. Pan, H. Li, and B. Wilamowski, "A low voltage to high voltage level shifter circuit for mems application," in *University/Government/Industry Microelectronics Symposium, 2003. Proceedings of the 15th Biennial*, pp. 128–131, IEEE, 2003.

[93] T. Davies Jr, "Dual stage level shifter for low voltage operation," January 2005. US Patent 6,838,924.

[94] J. Lim, J. Ha, W. Jung, Y. Kim, and J. Wee, "A novel high-speed and low-voltage cmos level-up/down shifter design for multiple-power and multiple-clock domain chips," *IEICE Transactions on Electronics E - Series C*, vol. 90, no. 3, p. 644, 2007.

[95] M. Kumar, S. Arya, and S. Pandey, "Level shifter design for low power applications," *International journal of computer science & information Technology (IJCSIT)*, vol. 2 (5), pp. 124–132, 2010.

[96] C. Tran, H. Kawaguchi, and T. Sakurai, "Low-power high-speed level shifter design for block-level dynamic voltage scaling environment," in *Integrated Circuit Design and Technology, 2005. ICICDT 2005. 2005 International Conference on*, pp. 229–232, IEEE, 2005.

[97] F. Pan and T. Samaddar, *Charge pump circuit design.* McGraw-Hill Professional, 2006.

[98] J. Starzyk, Y. Jan, and F. Qiu, "A dc-dc charge pump design based on voltage doublers," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 3, pp. 350–359, 2001.

[99] W. Rhee, "Design of high-performance cmos charge pumps in phase-locked loops," in *Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on*, vol. 2, pp. 545–548, IEEE, 1999.

[100] M. Ker, S. Chen, and C. Tsai, "Design of charge pump circuit with consideration of gate-oxide reliability in low-voltage cmos processes," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 5, pp. 1100–1107, 2006.

[101] G. Palumbo, D. Pappalardo, and M. Gaibotti, "Charge-pump circuits: power-consumption optimization," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 49, no. 11, pp. 1535–1542, 2002.

[102] R. Pelliconi, D. Iezzi, A. Baroni, M. Pasotti, and P. Rolandi, "Power efficient charge pump in deep submicron standard cmos technology," *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 6, pp. 1068–1071, 2003.

[103] A. Richelli, L. Mensi, L. Colalongo, P. Rolandi, and Z. Kovacs-Vajna, "A 1.2-to-8v charge-pump with improved power efficiency for non-volatille memories," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 522–619, IEEE, 2007.

[104] Synopsys, "Synopsys-sentaurus, r z-2010.03," *SProcess / SDevice*, release Z-2010.03.

[105] D. Garetto, D. Rideau, A. Schmid, H. Jaouen, and Y. Leblebici, "Surface potential-based compact model for embedded flash devices oriented to ic memory design," in *Submitted to Ultimate Integration of Silicon, 2012. ULIS 2012. 12th IEEE International Conference on*, 2012.

[106] CMC, "Compact Model Council."

[107] F. Van de Wiele, "A long-channel MOSFET model," *Solid-State Electronics*, vol. 22, no. 12, pp. 991–997, 1979.

[108] R. van Langevelde and F. Klaassen, "An explicit surface-potential-based MOSFET model for circuit simulation," *Solid-State Electronics*, vol. 44, no. 3, pp. 409–418, 2000.

[109] H. Wang, T. Chen, and G. Gildenblat, "Quasi-static and nonquasi-static compact MOSFET models based on symmetric linearization of the bulk and inversion charges," *Electron Devices, IEEE Transactions on*, vol. 50, no. 11, pp. 2262–2272, 2003.

[110] D. Garetto, E. Dornel, D. Rideau, W. F. Clark, A. Schmid, S. Hniki, C. Tavernier, H. Jaouen, and Y. Leblebici, "Analytical and compact models of the ono capacitance in embedded non-volatile flash devices," in *Proceedings of ESSDERC 2009 - poster session*, 2009.

[111] D. Goren, M. Zelikson, T. Galambos, R. Gordin, B. Livshitz, A. Amir, A. Sherman, and I. Wagner, "An interconnect-aware methodology for analog and mixed signal design, based on high bandwidth (over 40 GHz) on-chip transmission line approach," in *DATEC, 2002. Proceedings*, pp. 804–811, 2002.

[112] D. Goren, M. Zelikson, R. Gordin, I. Wagner, A. Barger, A. Amir, B. Livshitz, A. Sherman, Y. Tretiakov, R. Groves, *et al.*, "On-chip interconnect-aware design and modeling methodology, based on high bandwidth transmission line devices," in *Proceedings of the 40th annual Design Automation Conference*, pp. 724–727, ACM, 2003.

[113] G. Pólya and G. Szegő, *Isoperimetric inequalities in mathematical physics*. No. 27, Princeton Univ Pr, 1951.

[114] S. Ramanujan, "Modular equations and approximations to π," *Quart. J. Math*, vol. 45, pp. 350–372, 1914.

[115] L. Fox and R. Sankar, "The regula-falsi method for free-boundary problems," *IMA Journal of Applied Mathematics*, vol. 12, no. 1, p. 49, 1973.

[116] D. Garetto, A. Zaka, V. Quenette, D. Rideau, E. Dornel, O. Saxod, W. F. Clark, M. Minondo, C. Tavernier, Q. Rafhay, R. Clerc, A. Schmid, Y. Leblebici, and H. Jaouen, "Embedded non-volatile memory study with surface potential based model," in *Technical Proceedings Workshop on Compact Modeling (WCM)*, 2009.

[117] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S. Lo, G. Sai-Halasz, R. Viswanathan, H. Wann, S. Wind, *et al.*, "CMOS scaling into the nanometer regime," *Proceedings of the IEEE*, vol. 85, no. 4, pp. 486–504, 1997.

[118] C. Wu, S. Yang, H. Chen, F. Tseng, and C. Shih, "An analytic and accurate model for the threshold voltage of short channel MOSFETs in VLSI," *Solid-state electronics*, vol. 27, no. 7, pp. 651–658, 1984.

[119] Quenette, V. and Rideau, D. and Clerc, R. and Tavernier, C. and Retailleau S. and Jaouen, H., "Dynamic charge sharing modeling for surface potential based models," in *Technical Proceedings Workshop on Compact Modeling (WCM)*, 2009.

[120] V. Quenette, *Approche physique pour la modelisation compacte des dispositifs MOSFETS avances*. PhD thesis, Universite Paris XI - Institut d'electronique fondamentale, 2010.

[121] P. Klein, K. Hoffmann, and B. Lemaitre, "Description of the bias dependent overlap capacitance at LDDMOSFETs for circuit applications," in *Electron Devices Meeting, 1993. Technical Digest., International*, pp. 493–496, 1993.

[122] C. Jacoboni, C. Canali, G. Ottaviani, and A. Alberigi Quaranta, "A review of some charge transport properties of silicon," *Solid-State Electronics*, vol. 20, no. 2, pp. 77–89, 1977.

[123] C. Galup-Montoro and M. Schneider, *MOSFET modeling for circuit analysis and design*. World Scientific Pub Co Inc, 2007.

[124] A. Roy, C. Enz, and J. Sallese, "Source–drain partitioning in MOSFET," *Electron Devices, IEEE Transactions on*, vol. 54, no. 6, pp. 1384–1393, 2007.

[125] E. Anemogiannis, E. Glytsis, and T. Gaylord, "Bound and quasibound state calculations for biased/unbiased semiconductor quantum heterostructures," *Quantum Electronics, IEEE Journal of*, vol. 29, no. 11, pp. 2731–2740, 1993.

[126] B. Majkusiak, "Gate tunnel current in an MOS transistors," *Electron Devices, IEEE Transactions on*, vol. 37, no. 4, pp. 1087–1092, 1990.

[127] D. Ko and J. Inkson, "Matrix method for tunneling in heterostructures: Resonant tunneling in multilayer systems," *Physical Review B*, vol. 38, no. 14, p. 9945, 1988.

[128] S. Datta, "Nanoscale device modeling: the Green's function method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253–278, 2000.

[129] R. Clerc and G. Ghibaudo, *Etude des effets quantiques dans les composants CMOS à oxyde de grille ultra mince, Modélisation et Caractérisation*. PhD thesis, INPG, 2001.

[130] X. Gu, T. Chen, G. Gildenblat, G. Workman, S. Veeraraghavan, S. Shapira, and K. Stiles, "A surface potential-based compact model of n-MOSFET gate-tunneling current," *Electron Devices, IEEE Transactions on*, vol. 51, no. 1, pp. 127–135, 2004.

[131] K. Hasnat, C. Yeap, S. Jallepalli, S. Hareland, W. Shih, V. Agostinelli, A. Tasch, and C. Maziar, "Thermionic emission model of electron gate current in submicron NMOSFETs," *Electron Devices, IEEE Transactions on*, vol. 44, no. 1, pp. 129–138, 1997.

[132] A. Zaka, D. Garetto, D. Rideau, P. Palestri, J. Manceau, E. Dornel, Q. Rafhay, R. Clerc, Y. Leblebici, C. Tavernier, *et al.*, "Characterization and modelling of gate current injection in embedded non-volatile flash memory," in *Microelectronic Test Structures (ICMTS), 2011 IEEE International Conference on*, pp. 130–135, IEEE, 2011.

[133] A. Zaka, P. Palestri, Q. Rafhay, R. Clerc, M. Iellina, C. Rideau, D. Tavernier, G. Pananakakis, H. Jaouen, and L. Selmi, "An efficient non local hot electron model accounting for electron-electron scattering," *Submitted to Electron Devices, IEEE Transactions on*, vol. -, pp. –, 2011.

[134] K. Hasnat, C. Yeap, S. Jallepalli, W. Shih, S. Hareland, V. Agostinelli Jr, A. Tasch Jr, and C. Maziar, "A pseudo-lucky electron model for simulation of electron gatecurrent in submicron NMOSFET's," *Electron Devices, IEEE Transactions on*, vol. 43, no. 8, pp. 1264–1273, 1996.

[135] A. Gnudi, D. Ventura, G. Baccarani, and F. Odeh, "Two-dimensional MOSFET simulation by means of a multidimensional spherical harmonics expansion of the Boltzmann transport equation," *Solid-state electronics*, vol. 36, no. 4, pp. 575–581, 1993.

[136] R. Powell and C. Berglund, "Photoinjection studies of charge distributions in oxides of MOS structures," *Journal of Applied Physics*, vol. 42, no. 11, pp. 4390–4397, 1971.

[137] C. Berglund and R. Powell, "Photoinjection into sio2: Electron scattering in the image force potential well," *Journal of Applied Physics*, vol. 42, no. 2, pp. 573–579, 1971.

[138] N. Arora, *MOSFET models for VLSI circuit simulation: theory and practice*. Springer-Verlag New York, Inc., 1993.

[139] R. Duane, Q. Rafhay, M. Beug, and M. van Duuren, "Intrinsic mismatch between floating-gate nonvolatile memory cell and equivalent transistor," *Electron Device Letters, IEEE*, vol. 28, no. 5, pp. 440–442, 2007.

[140] T. Lin, K. Soejima, J. Takahashi, C. Hung, K. Liou, and R. Wan, "Advanced program verify for page mode flash memory," May 5 1998. US Patent 5,748,535.

[141] P. Song, "Program verify and erase verify control circuit for EPROM/flash," Nov. 26 1996. US Patent 5,579,262.

[142] P. Cappelletti, B. Ricco, and D. Esseni, "Controlled hot-electron writing method for non-volatile memory cells," Jan. 9 2001. US Patent 6,172,908.

[143] D. Fleetwood, P. Winokur, R. Reber, T. Meisenheimer, J. Schwank, M. Shaneyfelt, and L. Riewe, "Effects of oxide traps, interface traps, and border traps on metal-oxide-semiconductor devices," *Journal of Applied Physics*, vol. 73, no. 10, pp. 5058–5074, 2009.

[144] T. Tsuchiya, "Trapped-electron and generated interface-trap effects in hot-electron-induced MOSFET degradation," *Electron Devices, IEEE Transactions on*, vol. 34, no. 11, pp. 2291–2296, 1987.

[145] D. Fleetwood, S. Pantelides, and R. Schrimpf, *Defects in Microelectronic Materials and Devices*. CRC, 2008.

[146] H. Wong, M. White, T. Krutsick, and R. Booth, "Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFET's," *Solid-state electronics*, vol. 30, no. 9, pp. 953–968, 1987.

[147] A. van der Wel, E. Klumperink, E. Hoekstra, and B. Nauta, "Relating random telegraph signal noise in metal-oxide-semiconductor transistors to interface trap energy distribution," *Applied Physics Letters*, vol. 87, no. 18, pp. 183507–183507, 2005.

[148] I. Chen, S. Holland, and C. Hu, "Electrical breakdown in thin gate and tunneling oxides," *Solid-State Circuits, IEEE Journal of*, vol. 20, no. 1, pp. 333–342, 1985.

[149] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to flash memory," *Proceedings of the IEEE*, vol. 91, no. 4, pp. 489–502, 2003.

[150] S. Lombardo, B. De Salvo, C. Gerardi, and T. Baron, "Silicon nanocrystal memories," *Microelectronic engineering*, vol. 72, no. 1-4, pp. 388–394, 2004.

[151] K. Brower and S. Myers, "Chemical kinetics of hydrogen and (111) $Si-SiO_2$ interface defects," *Applied Physics Letters*, vol. 57, no. 2, pp. 162–164, 1990.

[152] D. Griscom, "Hydrogen model for radiation-induced interface states in SiO 2-on-Si structures: a review of the evidence," *Journal of electronic materials*, vol. 21, no. 7, pp. 763–767, 1992.

[153] D. Garetto, Y. Randriamihaja, D. Rideau, E. Dornel, F. William, A. Schmid, V. Huard, H. Jaouen, and Y. Leblebici, "Small signal analysis of electrically-stressed oxides with poisson-schroedinger based multiphonon capture model," in *Computational Electronics (IWCE), 2010 14th International Workshop on*, pp. 1–4, IEEE, 2010.

[154] K. Brower, S. Myers, A. Edwards, N. Johnson, C. Van de Walle, and E. Poindexter, "Comment on" Electron paramagnetic resonance of molecular hydrogen in silicon"," *Physical review letters*, vol. 73, no. 10, pp. 1456–1456, 1994.

[155] T. Nishi, G. Medvedkin, Y. Katsumata, K. Sato, and H. Miyake, "Electron paramagnetic resonance and photoluminescence study of defects in CuGaSe 2 single crystals grown by the traveling heater method," *Jpn. J. Appl. Phys., Part 1*, vol. 40, no. 1, pp. 59–63, 2001.

[156] E. Poindexter, G. Gerardi, M. Rueckel, P. Caplan, N. Johnson, and D. Biegelsen, "Electronic traps and Pb centers at the Si/SiO2 interface: Band-gap energy distribution," *Journal of applied physics*, vol. 56, no. 10, pp. 2844–2849, 1984.

[157] W. Futako, N. Mizuochi, and S. Yamasaki, "In situ ESR Observation of Interface Dangling Bond Formation Processes During Ultrathin SiO_2 Growth On Si(111)," *Phys. Rev. Lett.*, vol. 92, p. 105505, Mar 2004.

[158] S. Karna, H. Kurtz, A. Pineda, W. Shedd, and R. Pugh, "Point defects in Si-SiO2 systems: current understanding," *Defects in SiO2 and related dielectrics: science and technology, Kluwer Academic Publishers*, pp. 599–615, 2000.

[159] K. Brower, "Kinetics of H_2 passivation of P_b centers at the (111) $Si-SiO_2$ interface," *Physical Review B*, vol. 38, no. 14, p. 9657, 1988.

[160] E. Cartier, J. Stathis, and D. Buchanan, "Passivation and depassivation of silicon dangling bonds at the Si/SiO_2 interface by atomic hydrogen," *Applied Physics Letters*, vol. 63, no. 11, pp. 1510–1512, 2009.

[161] E. Cartier and J. H. Stathis, "Hot-electron induced passivation of silicon dangling bonds at the $Si(111)/SiO_2$ interface," *Applied Physics Letters*, vol. 69, no. 1, pp. 103–105, 1996.

[162] J. Campbell, P. Lenahan, C. Cochrane, A. Krishnan, and S. Krishnan, "Atomic-scale defects involved in the negative-bias temperature instability," *Device and Materials Reliability, IEEE Transactions on*, vol. 7, no. 4, pp. 540 –557, 2007.

[163] A. Stesmans, M. Jivanescu, S. Godefroo, and M. Zacharias, "Paramagnetic point defects at sio[sub 2]/nanocrystalline si interfaces," *Applied Physics Letters*, vol. 93, no. 2, p. 023123, 2008.

[164] E. Poindexter, P. Caplan, B. Deal, and R. Razouk, "Interface states and electron spin resonance centers in thermally oxidized (111) and (100) silicon wafers," *Journal of Applied Physics*, vol. 52, no. 2, pp. 879–884, 1981.

[165] M. Cook and C. White, "Hyperfine interactions of the P_b center at the SiO_2/Si (111) interface," *Physical review letters*, vol. 59, no. 15, pp. 1741–1744, 1987.

[166] M. Cook and C. White, "Hyperfine interactions in cluster models of the Pb defect center," *Physical Review B*, vol. 38, no. 14, p. 9674, 1988.

[167] B. Tuttle, "Hydrogen and Pb defects at the (111) $Si - SiO_2$ interface: an ab-initio cluster study," *Physical Review B*, vol. 60, no. 4, p. 2631, 1999.

[168] R. Weeks, "The many varieties of E' centers: a review," *Journal of non-crystalline solids*, vol. 179, pp. 1–9, 1994.

[169] R. Weeks, "Paramagnetic resonance of lattice defects in irradiated quartz," *Journal of Applied Physics*, vol. 27, no. 11, pp. 1376–1381, 1956.

[170] D. Griscom, E. Friebele, K. Long, and J. Fleming, "Fundamental defect centers in glass: Electron spin resonance and optical absorption studies of irradiated phosphorus-doped silica glass and optical fibers," *Journal of Applied Physics*, vol. 54, no. 7, pp. 3743–3762, 1983.

[171] M. Jani, R. Bossoli, and L. Halliburton, "Further characterization of the e_1 center in crystalline SiO_2," *Physical Review B*, vol. 27, no. 4, p. 2285, 1983.

[172] F. Feigl, W. Fowler, and K. Yip, "Oxygen vacancy model for the E1'center in SiO2," *Solid State Communications*, vol. 14, no. 3, pp. 225–229, 1974.

[173] P. M. Lenahan and P. V. Dressendorfer, "Hole traps and trivalent silicon centers in metal/oxide/silicon devices," *Journal of Applied Physics*, vol. 55, no. 10, pp. 3495–3499, 1984.

[174] C. Shen, M. Li, C. Foo, T. Yang, D. Huang, A. Yap, G. Samudra, and Y. Yeo, "Characterization and physical origin of fast Vth transient in NBTI of pMOSFETs with SiON dielectric," in *Electron Devices Meeting, 2006. IEDM'06. International*, pp. 1–4, IEEE, 2006.

[175] A. Neugroschel, G. Bersuker, and R. Choi, "Applications of dciv method to nbti characterization," *Microelectronics Reliability*, vol. 47, no. 9-11, pp. 1366 – 1372, 2007. 18th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis.

[176] T. Aichinger, M. Nelhiebel, and T. Grasser, "A Combined Study of p-and n-Channel MOS Devices to Investigate the Energetic Distribution of Oxide Traps After NBTI," *Electron Devices, IEEE Transactions on*, vol. 56, no. 12, pp. 3018–3026, 2009.

[177] J. Zhang and W. Eccleston, "Positive bias temperature instability in MOSFETs," *Electron Devices, IEEE Transactions on*, vol. 45, no. 1, pp. 116–124, 1998.

[178] V. Huard, M. Denais, F. Perrier, N. Revil, C. Parthasarathy, A. Bravaix, and E. Vincent, "A thorough investigation of MOSFETs NBTI degradation," *Microelectronics Reliability*, vol. 45, no. 1, pp. 83–98, 2004.

[179] T. Grasser, W. Gos, and B. Kaczer, "Dispersive transport and negative bias temperature instability: Boundary conditions, initial conditions, and transport models," *Device and Materials Reliability, IEEE Transactions on*, vol. 8, no. 1, pp. 79–97, 2008.

[180] M. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A comprehensive model for PMOS NBTI degradation: Recent progress," *Microelectronics Reliability*, vol. 47, no. 6, pp. 853–862, 2007.

[181] T. Grasser and B. Kaczer, "Evidence that two tightly coupled mechanisms are responsible for negative bias temperature instability in oxynitride MOSFETs," *Electron Devices, IEEE Transactions on*, vol. 56, no. 5, pp. 1056–1062, 2009.

[182] T. Grasser, H. Reisinger, W. Goes, T. Aichinger, P. Hehenberger, P. Wagner, M. Nelhiebel, J. Franco, and B. Kaczer, "Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, pp. 1–4, IEEE, 2010.

[183] K. Jeppson and C. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices," *Journal of Applied Physics*, vol. 48, no. 5, pp. 2004–2014, 1977.

[184] V. Huard, "Two independent components modeling for negative bias temperature instability," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 33–42, IEEE, 2010.

[185] B. Ridley, "Multiphonon, non-radiative transition rate for electrons in semiconductors and insulators," *Journal of Physics C: Solid State Physics*, vol. 11, p. 2323, 1978.

[186] B. Ridley, "On the multiphonon capture rate in semiconductors," *Solid-State Electronics*, vol. 21, no. 11-12, pp. 1319–1323, 1978.

[187] D. Goguenheim and M. Lannoo, "Theoretical calculation of the electron-capture cross section due to a dangling bond at the Si (111)-SiO_2 interface," *Physical Review B*, vol. 44, no. 4, pp. 1724–1733, 1991.

[188] M. Fischetti, "Generation of positive charge in silicon dioxide during avalanche and tunnel electron injection," *Journal of Applied Physics*, vol. 57, no. 8, pp. 2860–2879, 1984.

[189] M. Satter *et al.*, "Modeling effects of interface traps on the gate CV characteristics of MOS devices on alternative high-mobility substrates," *Solid-State Electronics*, vol. 54, no. 6, pp. 621–627, 2010.

[190] G. Brammertz, H. Lin, K. Martens, D. Mercier, C. Merckling, J. Penaud, C. Adelmann, S. Sioncke, W. Wang, M. Caymax, *et al.*, "Capacitance–Voltage Characterization of GaAs–Oxide Interfaces," *Journal of the Electrochemical Society*, vol. 155, p. H945, 2008.

[191] G. Groeseneken, H. Maes, N. Beltran, and R. De Keersmaecker, "A reliable approach to charge-pumping measurements in mos transistors," *Electron Devices, IEEE Transactions on*, vol. 31, pp. 42 – 53, Jan. 1984.

[192] J. L. Autran, F. Seigneur, C. Plossu, and B. Balland, "Characterization of si/sio_2 interface states: Comparison between different charge pumping and capacitance techniques," *Journal of Applied Physics*, vol. 74, no. 6, pp. 3932–3935, 1993.

[193] N. Saks, "Measurement of single interface trap capture cross sections with charge pumping," *Applied physics letters*, vol. 70, p. 3380, 1997.

[194] D. Bauza, O. Ghobar, N. Guénifi, and S. Bayon, "Advanced Analysis of Silicon Insulator Interface Traps in MOSFET's with SiO_2 and HfO_2 as Gate Dielectrics," *ECS Transactions*, vol. 19, no. 2, pp. 19–54, 2009.

[195] W. Wu, B. Tsui, M. Chen, Y. Hou, Y. Jin, H. Tao, S. Chen, and M. Liang, "Transient Charging and Discharging Behaviors of Border Traps in the Dual-Layer HfO2/SiO2 High-κ Gate Stack Observed by Using Low-Frequency Charge Pumping Method," 2007.

[196] J. Brugler and P. Jespers, "Charge pumping in MOS devices," *Electron Devices, IEEE Transactions on*, vol. 16, no. 3, pp. 297–302, 1969.

[197] G. Groeseneken and H. Maes, "Basics and applications of charge pumping in submicron mosfet's," in *Microelectronics, 1997. Proceedings., 1997 21st International Conference on*, vol. 2, pp. 581–589, IEEE, 1998.

[198] M. Masuduzzaman, A. Islam, and M. Alam, "Exploring the capability of multifrequency charge pumping in resolving location and energy levels of traps within dielectric," *Electron Devices, IEEE Transactions on*, vol. 55, no. 12, pp. 3421 –3431, 2008.

[199] M. Masuduzzaman, A. Islam, and M. Alam, "Physics and mechanisms of dielectric trap profiling by Multi-frequency Charge Pumping (MFCP) method," in *Reliability Physics Symposium, 2009 IEEE International*, pp. 13 –20, 2009.

[200] J. Hao and J. Gao, "Characterization of oxide thin films using optical techniques," *Applied surface science*, vol. 253, no. 1, pp. 372–375, 2006.

[201] T. Grasser, H. Reisinger, P. Wagner, F. Schanovsky, W. Goes, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 16–25, IEEE, 2010.

[202] B. Kaczer, T. Grasser, J. Martin-Martinez, E. Simoen, M. Aoulaiche, P. Roussel, and G. Groeseneken, "NBTI from the perspective of defect states with widely distributed time scales," in *Reliability Physics Symposium, 2009 IEEE International*, pp. 55–60, IEEE, 2009.

[203] H. Reisinger, T. Grasser, and C. Schlunder, "A study of NBTI by the statistical analysis of the properties of individual defects in pMOSFETS," in *Integrated Reliability Workshop Final Report, 2009. IRW'09. IEEE International*, pp. 30–35, IEEE, 2010.

[204] M. Kirton and M. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/f) noise," *Advances in Physics*, vol. 38, pp. 367–468, 1989.

[205] M. Haider, J. Pitters, G. DiLabio, L. Livadaru, J. Mutus, and R. Wolkow, "Controlled coupling and occupation of silicon atomic quantum dots at room temperature," *Physical Review Letters*, vol. 102, no. 4, p. 46805, 2009.

[206] B. Grandidier and C. Delerue, "Deep insight into a quantum system: a single silicon dangling bond," tech. rep., 2010.

[207] R. Castagne and A. Vapaille, "Description of the SiO_2—Si interface properties by means of very low frequency MOS capacitance measurements," *Surface Science*, vol. 28, no. 1, pp. 157–193, 1971.

[208] K. Zaininger and G. Warfield, "Limitations of the mos capacitance method for the determination of semiconductor surface properties," *Electron Devices, IEEE Transactions on*, vol. 12, no. 4, pp. 179–193, 1965.

[209] F. Heiman and G. Warfield, "The effects of oxide traps on the MOS capacitance," *Electron Devices, IEEE Transactions on*, vol. 12, no. 4, pp. 167–178, 1985.

[210] L. Freeman and W. Dahlke, "Theory of tunneling into interface states," *Solid-State Electronics*, vol. 13, no. 11, pp. 1483–1503, 1970.

[211] W. Shockley and W. Read Jr, "Statistics of the recombinations of holes and electrons," *Physical Review*, vol. 87, no. 5, p. 835, 1952.

[212] H. Lakhdari, D. Vuillaume, and J. Bourgoin, "Spatial and energetic distribution of $Si-SiO_2$ near-interface states," *Physical Review B*, vol. 38, no. 18, pp. 13124–13132, 1988.

[213] D. Bauza and G. Ghibaudo, "New model for the characterization of bulk traps by current deep level transient spectroscopy in metal-oxide-semiconductor transistors," *Journal of applied physics*, vol. 70, no. 6, pp. 3333–3337, 1991.

[214] A. Palma, A. Godoy, J. Jimenez-Tejada, J. Carceller, and J. Lopez-Villanueva, "Quantum two-dimensional calculation of time constants of random telegraph signals in metal-oxide–semiconductor structures," *Physical Review B*, vol. 56, no. 15, pp. 9565–9574, 1997.

[215] D. Bauza and Y. Maneglia, "In-depth exploration of $Si-SiO_2$ interface traps in MOS transistors using the charge pumping technique," *Electron Devices, IEEE Transactions on*, vol. 44, pp. 2262 –2266, Dec. 1997.

[216] P. Masson, J.-L. Autran, and J. Brini, "On the tunneling component of charge pumping current in ultrathin gate oxide mosfets," *Electron Device Letters, IEEE*, vol. 20, pp. 92 –94, Feb. 1999.

[217] X. Garros, M. Casse, G. Reimbold, F. Martin, L. Brunet, F. Andrieu, and F. Boulanger, "Reliability concerns in High-K/Metal gate technologies," in *IC Design and Technology (ICICDT), 2010 IEEE International Conference on*, pp. 90–93, IEEE, 2010.

[218] D. Garetto, Y. Mamy-Randriamihaja, A. Zaka, D. Rideau, A. Schmid, Jaouen, Herve, and Y. Leblebici, "Analysis of defect cross sections using non-radiative mpa quantum model," *Accepted to Solid State Electronics*, vol. -, pp. –, 2011.

[219] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps ," 2010.

[220] B. Ridley, "The photoionisation cross section of deep-level impurities in semiconductors," *Journal of Physics C: Solid State Physics*, vol. 13, p. 2015, 1980.

[221] D. Goguenheim and M. Lannoo, "Theoretical and experimental aspects of the thermal dependence of electron capture coefficients," *Journal of Applied Physics*, vol. 68, no. 3, pp. 1059–1069, 1990.

[222] F. Jiménez-Molinos, F. Gámiz, A. Palma, P. Cartujo, and J. López-Villanueva, "Direct and trap-assisted elastic tunneling through ultrathin gate oxides," *Journal of Applied Physics*, vol. 91, no. 8, pp. 5116–5124, 2009.

[223] S. Markov, *Gate Leakage Variability in Nano-CMOS Transistors*. PhD thesis, University of Glasgow, 2009.

[224] P. Carrier, L. Lewis, and M. Dharma-Wardana, "Optical properties of structurally relaxed Si/SiO2 superlattices: The role of bonding at interfaces," *Physical Review B*, vol. 65, no. 16, p. 165339, 2002.

[225] F. Grunthaner, P. Grunthaner, R. Vasquez, B. Lewis, J. Maserjian, and A. Madhukar, "High-Resolution X-Ray Photoelectron Spectroscopy as a Probe of Local Atomic Structure: Application to Amorphous SiO_2 and the Si-SiO_2 Interface," *Physical Review Letters*, vol. 43, no. 22, pp. 1683–1686, 1979.

[226] A. Demkov and O. Sankey, "Growth Study and Theoretical Investigation of the Ultrathin Oxide $SiO_2 - Si$ Heterojunction," *Physical review letters*, vol. 83, no. 10, pp. 2038–2041, 1999.

[227] J. Ryan, P. Lenahan, T. Grasser, and H. Enichlmair, "Recovery-free electron spin resonance observations of nbti degradation," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 43–49, IEEE, 2010.

[228] D. Garetto, D. Rideau, E. Dornel, W. F. Clark, C. Tavernier, Y. Leblebici, A. Schmid, and H. Jaouen, "Modeling study of capacitance and gate current in strained highk-metal gate technology," in *13th International Nanotech Conference 2010*, 2010.

[229] W. Goes, M. Karner, V. Sverdlov, and T. Grasser, "Charging and discharging of oxide defects in reliability issues," *Device and Materials Reliability, IEEE Transactions on*, vol. 8, no. 3, pp. 491–500, 2008.

[230] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, "Understanding negative bias temperature instability in the context of hole trapping (Invited Paper)," *Microelectronic Engineering*, vol. 86, no. 7-9, pp. 1876–1882, 2009.

[231] T. Grasser and B. Kaczer, "Critical Modeling Issues in Negative Bias Temperature Instability," in *Proceedings of European Symposium on Reliability of Electron Devices*, European Symposium on Reliability of Electron Devices, 2009.

[232] T. Nguyen, G. Mahieu, M. Berthe, B. Grandidier, C. Delerue, D. Stiévenard, and P. Ebert, "Coulomb Energy Determination of a Single Si Dangling Bond," *Physical Review Letters*, vol. 105, no. 22, p. 226404, 2010.

[233] M. White and C. Chao, "Statistics of deep-level amphoteric traps in insulators and at interfaces," *Journal of applied physics*, vol. 57, no. 6, pp. 2318–2321, 1985.

[234] J. Zheng, H. Tan, and S. Ng, "Theory of non-radiative capture of carriers by multiphonon processes for deep centres in semiconductors," *Journal of Physics: Condensed Matter*, vol. 6, p. 1695, 1994.

[235] G. Watson, *A treatise on the theory of Bessel functions*. Cambridge Univ Pr, 1995.

[236] B. Ridley and M. Amato, "A model for the interpretation of measurements of photoionisation and capture cross sections associated with deep-level impurities," *Journal of Physics C: Solid State Physics*, vol. 14, p. 1255, 1981.

[237] M. Berthe, R. Stiufiuc, B. Grandidier, D. Deresmes, C. Delerue, and D. Stievenard, "Probing the carrier capture rate of a single quantum level," *Science*, vol. 319, no. 5862, p. 436, 2008.

[238] M. Berthe, A. Urbieta, L. Perdigão, B. Grandidier, D. Deresmes, C. Delerue, D. Stiévenard, R. Rurali, N. Lorente, L. Magaud, *et al.*, "Electron transport via local polarons at interface atoms," *Physical review letters*, vol. 97, no. 20, p. 206801, 2006.

[239] D. Garetto, Y. Mamy-Randriamihaja, D. Rideau, A. Schmid, Jaouen, Herve, and Y. Leblebici, "Comparing defect characterization techniques with non-radiative multiphonon charge trapping model," *Submitted to Journal of Computational Electronics*, vol. -, pp. –, 2011.

[240] D. Garetto, Y. Mamy-Randriamihaja, A. Zaka, D. Rideau, A. Schmid, H. Jaouen, and Y. Leblebici, "Ac analysis of defect cross sections using non-radiative mpa quantum model," in *Ultimate Integration of Silicon, 2011. ULIS 2011. 11th IEEE International Conference on*, 2011.

[241] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, "Theory of direct tunneling current in metal–oxide–semiconductor structures," *Journal of applied physics*, vol. 91, p. 1400, 2002.

[242] N. Klein and P. Solomon, "Current runaway in insulators affected by impact ionization and recombination," *Journal of Applied Physics*, vol. 47, no. 10, pp. 4364–4372, 1976.

[243] D. Goguenheim, D. Vuillaume, G. Vincent, and N. Johnson, "Accurate measurements of capture cross sections of semiconductor insulator interface states by a trap-filling experiment," *Journal of Applied Physics*, vol. 68, no. 3, pp. 1104–1113, 1990.

[244] T. Kang, M. Chen, C. Liu, Y. Chang, and S. Fan, "Numerical confirmation of inelastic trap-assisted tunneling (ITAT) as SILC mechanism," *Electron Devices, IEEE Transactions on*, vol. 48, no. 10, pp. 2317–2322, 2002.

[245] A. Stoneham, "Non-radiative transitions in semiconductors," *Reports on Progress in Physics*, vol. 44, p. 1251, 1981.

[246] Y. Mamy Randriamihaja, D. Garetto, V. Huard, D. Rideau, D. Roy, and M. Rafik, "New insights into gate-dielectric breakdown by electrical characterization of interfacial and oxide defects with reverse modeling methodology," in *Submitted to Reliability Physics Symposium Proceedings, 2012, IEEE International*, 2012.

[247] L. Skuja, T. Suzuki, and K. Tanimura, "Site-selective laser-spectroscopy studies of the intrinsic 1.9-ev luminescence center in glassy sio2," *Physical Review B*, vol. 52, no. 21, p. 15208, 1995.

[248] J. Stathis, "Percolation models for gate oxide breakdown," *Journal of applied physics*, vol. 86, p. 5757, 1999.

[249] S. Tous, E. Wu, and J. Sune, "A compact analytic model for the breakdown distribution of gate stack dielectrics," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pp. 792–798, IEEE, 2010.

[250] D. Bauza, "A General and Reliable Model for Charge Pumping Part I: Model and Basic Charge-Pumping Mechanisms," *Electron Devices, IEEE Transactions on*, vol. 56, no. 1, pp. 70 –77, 2009.

[251] D. Bauza, "A General and Reliable Model for Charge Pumping Part II: Application to the Study of Traps in SiO2 or in High-K Gate Stacks," *Electron Devices, IEEE Transactions on*, vol. 56, no. 1, pp. 78 –84, 2009.

[252] D. Bauza, "Rigorous analysis of two-level charge pumping: Application to the extraction of interface trap concentration versus energy profiles in metal–oxide–semiconductor transistors," *Journal of Applied Physics*, vol. 94, no. 5, pp. 3239–3248, 2003.

[253] P. Riess, R. Kies, G. Ghibaudo, G. Pananakakis, and J. Brini, "Reversibility of charge trapping and SILC creation in thin oxides after stress/anneal cycling," *Microelectronics Reliability*, vol. 38, no. 6-8, pp. 1057–1061, 1998.

[254] D. Garetto, Y. Mamy-Randriamihaja, A. Zaka, D. Rideau, A. Schmid, Jaouen, Herve, and Y. Leblebici, "Modeling of stressed mos oxides using a multiphonon-assisted quantum approach - part ii: transient effects," *Accepted to IEEE Transactions of Electron Devices*, vol. 59-3, pp. –, 2011.

[255] D. Garetto, A. Zaka, J.-P. Manceau, D. Rideau, E. Dornel, W. F. Clark, A. Schmid, H. Jaouen, and Y. Leblebici, "Characterization and physical modeling of endurance in embedded non-volatile memory technology," in *Proceedings of International Memory Workshop 2011*, 2011.

[256] D. Garetto, Y. Rideau, Denis Mamy-Randriamihaja, C. Tavernier, and H. Jaouen, "New insights into interfacial and oxide defects by electrical characterization with reverse modeling methodology," in *International Workshop on Simulation and Modeling of Memory devices*, 2011.

[257] M. Ferrier, R. Clerc, G. Ghibaudo, F. Boeuf, and T. Skotnicki, "Analytical model for quantization on strained and unstrained bulk nMOSFET and its impact on quasi-ballistic current," *Solid-state electronics*, vol. 50, no. 1, pp. 69–77, 2006.

[258] A. Allison, "The numerical solution of coupled differential equations arising from the schrödinger equation," *Journal of Computational Physics*, vol. 6, no. 3, pp. 378–391, 1970.

[259] J. Canosa and R. De Oliveira, "A new method for the solution of the schrödinger equation," *Journal of Computational Physics*, vol. 5, no. 2, pp. 188–207, 1970.

[260] A. Ghatak, R. Gallawa, and I. Goyal, "Accurate solutions to Schrodinger's equation using modified Airy functions," *Quantum Electronics, IEEE Journal of*, vol. 28, no. 2, pp. 400–403, 1992.

[261] A. Polyanin and A. Manzhirov, *Handbook of mathematics for engineers and scientists*. Chapman & Hall/CRC, 2007.

[262] Y. Park and D. Schroder, "Degradation of thin tunnel gate oxide under constant fowler-nordheim current stress for a flash eeprom," *Electron Devices, IEEE Transactions on*, vol. 45, no. 6, pp. 1361–1368, 1998.

[263] A. Zaka, J. Singer, E. Dornel, D. Garetto, D. Rideau, Q. Rafhay, R. Clerc, J. Manceau, N. Degors, C. Boccaccio, *et al.*, "Characterization and 3d tcad simulation of nor-type flash non-volatile memories with emphasis on corner effects," *Solid-State Electronics*, vol. 63-1, pp. 158–162, 2011.

[264] C. Guérin, *Etude de la dégradation par porteurs chauds des technologies CMOS avancées en fonctionnement statique et dynamique*. PhD thesis, Université de Provence d'Aix-Marseille, 2008.

[265] S. Rauch III, G. La Rosa, and F. Guarin, "Role of ee scattering in the enhancement of channel hot carrier degradation of deep sub-micron nmosfets at high vgs conditions," in *Reliability Physics Symposium, 2001. Proceedings. 39th Annual. 2001 IEEE International*, pp. 399–405, IEEE, 2001.

[266] C. Guérin, V. Huard, and A. Bravaix, "The energy-driven hot-carrier degradation modes of nmosfets," *Device and Materials Reliability, IEEE Transactions on*, vol. 7, no. 2, pp. 225–235, 2007.

[267] E. Kane, "Zener tunneling in semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 12, no. 2, pp. 181–188, 1960.

[268] R. K. Eguchi, "Non-volatile memory having a dynamically adjustable soft program verify," January 2010. US patent 7,649,782.

[269] S. Rajput and S. Jamuar, "Low voltage analog circuit design techniques," *Circuits and Systems Magazine, IEEE*, vol. 2, no. 1, pp. 24–42, 2002.

[270] K. Usami and M. Igarashi, "Low-power design methodology and applications utilizing dual supply voltages," in *Proceedings of the 2000 Conlerence on Asia and South Pacific Design Automation*, Published by the IEEE Computer Society, 2000.

[271] N. Otsuka and M. Horowitz, "Circuit techniques for 1.5-v power supply flash memory," *Solid-State Circuits, IEEE Journal of*, vol. 32, no. 8, pp. 1217–1230, 1997.

[272] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. McGraw-Hill, Inc., 1987.

[273] J. He, M. Chan, X. Zhang, and Y. Wang, "A physics-based analytic solution to the MOSFET surface potential from accumulation to strong-inversion region," *Electron Devices, IEEE Transactions on*, vol. 53, no. 9, pp. 2008–2016, 2006.

[274] T. Chen and G. Gildenblat, "An extended analytical approximation for the MOSFET surface potential," *Solid-state electronics*, vol. 49, no. 2, pp. 267–270, 2005.

[275] W. Wu, T. Chen, G. Gildenblat, and C. McAndrew, "Physics-based mathematical conditioning of the MOSFET surface potential equation," *Electron Devices, IEEE Transactions on*, vol. 51, no. 7, pp. 1196–1199, 2004.

[276] T. Chen and G. Gildenblat, "Symmetric bulk charge linearisation in charge-sheet MOSFET model," *Electronics Letters*, vol. 37, no. 12, pp. 791–793, 2001.

[277] D. Ward and R. Dutton, "A charge-oriented model for MOS transistor capacitances," *Solid-State Circuits, IEEE Journal of*, vol. 13, no. 5, pp. 703–708, 1978.

[278] P. Lin, J. Guo, and C. Wu, "A quasi-two-dimensional analytical model for the turn-oncharacteristics of polysilicon thin-film transistors," *Electron Devices, IEEE Transactions on*, vol. 37, no. 3 Part 1, pp. 666–674, 1990.

[279] M. Kondo and H. Tanimoto, "An accurate Coulomb mobility model for MOS inversion layer and its application to NO-oxynitride devices," *Electron Devices, IEEE Transactions on*, vol. 48, no. 2, pp. 265–270, 2001.

[280] R. Van Langevelde, A. Scholten, and D. Klaassen, "Physical background of MOS model 11," *Measurements*, vol. 6, p. 9, 2003.

[281] D. Garetto, D. Rideau, C. Tavernier, Y. Leblebiciand, A. Schmid, and H. Jaouen, "Advanced physics for simulation of ultrascaled devices with utoxpp solver," in *14th International Nanotech Conference 2011*, 2011.

[282] D. Rideau, M. Feraille, L. Ciampolini, M. Minondo, C. Tavernier, H. Jaouen, and A. Ghetti, "Strained Si, Ge alloys modeled with a first-principles-optimized full-zone k.p method," *Physical Review B*, vol. 74, no. 19, p. 195208, 2006.

[283] D. Rideau, M. Feraille, M. Michaillat, Y. Niquet, C. Tavernier, and H. Jaouen, "On the validity of the effective mass approximation and the Luttinger kp model in fully depleted SOI MOSFETs," *Solid State Electronics*, vol. 53, no. 4, pp. 452–461, 2009.

[284] *abinit v4.6.5 - http://www.abinit.org.*

[285] J. Dunham, "The Wentzel-Brillouin-Kramers method of solving the wave equation," *Physical Review*, vol. 41, no. 6, p. 713, 1932.

[286] Y. Niquet, D. Rideau, C. Tavernier, H. Jaouen, and X. Blase, "Onsite matrix elements of the tight-binding Hamiltonian of a strained crystal: Application to silicon, germanium, and their alloys," *Physical Review B*, vol. 79, no. 24, p. 245201, 2009.

[287] C. Chao and S. Chuang, "Spin-orbit-coupling effects on the valence-band structure of strained semiconductor quantum wells," *Physical Review B*, vol. 46, no. 7, pp. 4110–4122, 1992.

[288] Synopsys, *Synopsys Sentaurus, release C-2009.06, Sprocess and Sdevice simulators*, 2009.

[289] P. Vogl and T. Kubis, "The non-equilibrium GreenŠs function method: an introduction," *Journal of computational electronics*, vol. -, pp. 1–6, 2010.

[290] M. Anantram, M. Lundstrom, and D. Nikonov, "Modeling of nanoscale devices," *Proceedings of the IEEE*, vol. 96, no. 9, pp. 1511–1550, 2008.

[291] G. Klimeck, R. Lake, R. Bowen, W. Frensley, and T. Moise, "Quantum device simulation with a generalized tunneling formula," *Applied physics letters*, vol. 67, no. 17, pp. 2539–2541, 1995.

[292] *QT v4.6 - http://qt.nokia.com/.*

Curriculum Vitae

Garetto Davide

Address: 11, Rue Paul Janet - F-38100 – Grenoble (France)

Telephone: +33 (0)4 38 92 37 24 Mobile: +39 345 4331658

E-mail: david.garetto@fr.ibm.com

Nationality: Italian

Language(s): Italian (mother tongue), English (proficient user), French (advanced user)

Date of birth: 12.09.1984

Gender: Male

Work experience

Dates: October 2008 – December 2011 (expected)

Occupation or position held: **R&D engineer and PhD candidate in microelectronics**

Reference contact: Dr. Michael L. Kerbaugh, mkerbaug@us.ibm.com

Employer: IBM France – Systems and Technology Group
STMicroelectronics – 850, Rue Jean Monnet– 38926 Crolles (France)

Main activities and responsibilities:

In the framework of a joint research & development partnership between IBM STG and STMicroelectronics on derivative advanced CMOS technologies, I'm currently in charge of :

- the establishment of a methodology from numerical modeling to IC design, via the development, validation and delivery of compact device models in 65nm embedded flash memory technology;
- the development and benchmarking of novel physical TCAD modeling methods for ultra scaled devices from bulk CMOS to FinFET technologies;
- the complete design and prototyping of a 40KB memory sector test chip in embedded NVM technology;
- improve communication between the two companies and between various teams.

We developed a surface potential-based model for the flash memory cell with the purpose of providing a comprehensive physical understanding of the device operation suitable for performance optimization in memory circuit design. A robust algorithm is integrated to solve the charge balance on the isolated floating gate node taking into account parasitic couplings in the cells. The model takes into account all scaling effects in standard compact approaches and includes transient currents, parasitic disturbs and aging effects that determine the performances of the NVM cell. The physical understanding of quantum transport and non-radiative charge trapping phenomena has been achieved with rigorous advanced quantum modeling methods. We implemented the model in Verilog-A language for compatibility with common circuit simulators and validated on 3D TCAD simulations and measurement results on designed test structures.
In terms of the personal development, this position has allowed me to develop extended knowledge in a wide number of fields, ranging from first principles quantum modeling to characterization and product optimizations for IC design.

Dates: February 2008 – August 2008

Occupation or position held: **Research assistant in Phase Change Memories group**

Reference contact: Dr. Geoffrey W. Burr, burr@almaden.ibm.com

Employer: IBM Almaden Research Center – 650, Harry Road – 95120 San Jose – CA (USA)

Main activities and responsibilities:

During my internship in the Phase Change group at the IBM Almaden Research Center, I was in charge of:

- co-developing a customized finite-difference multiphysics modeling tool to investigate the writing of intermediate states for multi-level cell (MLC) PCM 3D structures;
- integrating "soft" crystallization and full nucleation numerical 3D models to explain the creation of partially crystallized polycrystalline grains justifying gradual resistivity variations of the phase change material;
- implementing a model for the access active device controlling the cell and integrated it into the self consistent modeling tool;
- participating in the improvement of algorithm convergence in presence of abrupt and wide resistivity variations.

Dates: May 2006 – July 2006

Occupation or position held	**Internship as C++ graphic programmer**
Employer	SEAC02 S.r.l. , via Avogadro 4, I-10100 Torino
Main activities and responsibilities	Development of a 3D editor in GLSL, OpenGL 2.0 and C++

Education and training

2008 - 2011	PhD candidate in electrical engineering, Microelectronic Systems Laboratory (LSM) - Swiss Federal Institute of Technology (EPF Lausanne)
2007 - 2008	Master's Degree in Micro and Nanotechnologies for integrated systems, Politecnico di Torino – ENSERG / INP Grenoble – EPF Lausanne, 110/110 cum laude
2003 – 2006	Bachelor Degree in Computer Engineering, Politecnico di Torino - 110/110 cum laude

Competences

Computer skills and competences

- FEM and FDTD Simulation: Ansys Multiphysics, Femlab Comsol, Synopsys Sentaurus suite
- Advanced skills in the use of Ansi C, C++, Java J2SE, HTML, QT GUI design, OpenGL, GLSL, Delphi, VerilogA, Matlab, Java Eclipse SDK, Microsoft Visual Studio IDE
- Other: Cadence Design Framework, Cadence Encounter, Calibre verification tools, Mentor Graphics ELDO, HSPICE, Agilent ICCAP, ModelSim, VHDL-AMS, SQL, 3D Studio Max
- Device characterization using Agilent UF300 fully automated testers

Achievements and awards

List of publications:

D. Garetto, et al. “Embedded non–volatile memory study with surface potential based model”, Workshop on Compact Modeling (WCM), 2009

D. Garetto, et al. “Analytical and compact models of the ONO capacitance in embedded non-volatile flash devices”, ESSDERC 2009 – fringe poster session, 2009

D. Garetto, et al. “Modeling study of capacitance and gate current in strained High-K Metal gate technology: impact of Si/SiO2/HK interfacial layer and band structure model”, 13th International Nanotech Conference and Expo 2010

D. Garetto, et al., AC analysis of defect cross sections using non-radiative MPA quantum model, Ultimate Integration of Silicon, 2011, 11th IEEE International Conference

D. Garetto, et al. “Small signal analysis of electrically-stressed oxides with PS-based multiphonon capture model”, 13th International Workshop on Computational Electronics 2010

D. Garetto, et al, “Analysis of defect cross sections using non-radiative MPA quantum model”, accepted for publication in Solid State Electronics, 2011

D. Garetto, et al., “Characterization and physical modeling of endurance in embedded non-volatile memory technology”, International Workshop on Non-Volatile Memories, 2011

D. Garetto, et al., “Advanced physics for simulation of ultrascaled devices with UTOXPP Solver”, 14th International Nanotech Conference 2011

D. Garetto, et al., “Modeling of stressed MOS oxide with multiphonon-assisted quantum approach - part I: AC impedance model”, submitted to IEEE Transaction of Electron Devices 2011

D. Garetto, et al., “Modeling of stressed MOS oxide with multiphonon-assisted quantum approach - part II: transient effects”, submitted to IEEE Transaction of Electron Devices 2011

D. Garetto, et al., “Comparing defect characterization techniques with non-radiative multiphonon charge trapping model", submitted to Journal of Computation Electronics, 2011

D. Garetto et al., “New insights into interfacial and oxide defects by electrical characterization with reverse modeling methodology”, International Workshop on Simulation and Modeling of Memory devices, 2011

G.W. Burr, M. J. Breiwisch, M. Franceschini, *D. Garetto,* K. Gopalakrishnan, B. Jackson, B. Kurdi, C. Lam, A. Lastras, A. Padilla, B. Rajendran, and S. Raoux, "Phase change memory technology", in J. Vac. Sci. Techn., 2010

Rideau, D.; Quenette, V.; Garetto, D.; Dornel, E.; Weybright, M.; Manceau, J.; Saxod, O.; Tavernier, C. & Jaouen, H., Characterization & modeling of gate-induced-drain-leakage with complete overlap and fringing model Microelectronic Test Structures (ICMTS), 2010 IEEE International Conference on, 2010, 210-213

A. Zaka, D. Garetto, D. Rideau, P. Palestri, J. Manceau, E. Dornel, Q. Rafhay, R. Clerc, Y. Leblebici, C. Tavernier, Characterization and modelling of gate current injection in embedded non-volatile flash memory, Microelectronic Test Structures (ICMTS), 2011 IEEE International Conference on, 2011

Zaka, A.; Singer, J.; Dornel, E.; Garetto, D.; Rideau, D.; Rafhay, Q.; Clerc, R.; Manceau, J.; Degors, N.; Boccaccio, C. & others, Characterization and 3D TCAD simulation of NOR-type flash non-volatile memories with emphasis on corner effects, Solid-State Electronics, Elsevier, 2011, 63-1, 158-162

Mamy Randriamihaja, Y.; Garetto, D.; Huard, V.; Rideau, D.; Roy, D. & Rafik, M., New insights into gate-dielectric breakdown by electrical characterization of interfacial and oxide defects with reverse modeling methodology, Submitted to Reliability Physics Symposium, IEEE International, 2012

Awards: *IBM PhD fellowship* 2010 ; *IBM PhD fellowship* 2011
Lead inventor in 3 patent disclosure submissions