Carrier injection and degradation mechanisms in advanced NOR Flash memories

Alban Zaka

UNIVERSITÉ DE GRENOBLE

THÈSE EN COTUTELLE INTERNATIONALE
POUR OBTENIR LE GRADE DE
DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE
ET DE L'UNIVERSITÀ DEGLI STUDI DI UDINE
Spécialité: Nano Electronique / Nano Technologies
Arrêté ministériel : 7 août 2006

PRÉSENTÉE PAR
ALBAN ZAKA

THÈSE DIRIGÉE PAR **GEORGES PANANAKAKIS** ET
CODIRIGÉE PAR **LUCA SELMI**
PRÉPARÉE AU SEIN DE **STMICROELECTRONICS**,
DE **L'INSTITUT DE MICROÉLECTRONIQUE, ELECTROMAGNETISME ET PHOTONIQUE - LABORATOIRE D'HYPERFRÉQUENCE ET DE CARACTÉRISATION**
DANS L'ÉCOLE DOCTORALE:**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE ET TRAITEMENT DU SIGNAL (EEATS)**
ET AU LABORATOIRE **DIPARTIMENTO DI INGEGNERIA ELETTRICA, GESTIONALE E MECCANICA**

Carrier injection and degradation mechanisms in advanced NOR Flash memories

THÈSE SOUTENUE PUBLIQUEMENT LE **23 JANVIER 2012**,
DEVANT LE JURY COMPOSÉ DE:

M. PHILIPPE DOLLFUS
DR, IEF, Paris - Président

M. TIBOR GRASSER
PR, TUWien, Vienne, Autriche - Rapporteur

M. CLAUDIO FIEGNA
PR, ARCES, Bologna, Italie - Rapporteur

M. GEORGES PANANAKAKIS
PR, IMEP-LAHC, Grenoble - Directeur de thèse

M. LUCA SELMI
PR, DIEGM, Udine, Italie - Directeur de thèse

M. RAPHAEL CLERC
MCF, IMEP-LAHC, Grenoble - Co-encadrant

M. PIERPAOLO PALESTRI
MCF, DIEGM, Udine, Italie - Co-encadrant

M. DENIS RIDEAU
ING, STMicroelectronics, Crolles - Co-encadrant

UNIVERSITÁ DEGLI STUDI DI UDINE

Dipartimento di Ingegneria Elettrica Gestionale e Meccanica
Dottorato in Ingegneria Industriale e dell'Informazione
– XXIV Ciclo –

Tesi di Dottorato

Carrier injection and degradation mechanisms in advanced NOR Flash memories

Alban ZAKA

23 gennaio 2012

Commissione Esaminatrice:
Dr. Ing. Philippe DOLLFUS, Presidente
Prof. Tibor GRASSER, Esaminatore
Prof. Claudio FIEGNA, Esaminatore
Prof. Georges PANANAKAKIS, Tutor (Fr)
Prof. Luca SELMI, Tutor (It)
Assoc. Prof. Raphael CLERC, Co-Tutor (Fr)
Assoc. Prof. Pierpaolo PALESTRI, Co-Tutor (It)
Dr. Ing. Denis RIDEAU, Co-Tutor

Contents

1 Introduction **1**
1.1 Flash memory context . . . 1
1.1.1 Market and evolution . . . 1
1.1.2 NAND and NOR memories . . . 2
1.1.3 Stand-alone vs. embedded memories . . . 5
1.1.4 The NOR scaling issues . . . 5
1.2 Scope of the thesis . . . 6
1.3 Organization of the thesis . . . 7
Bibliography . . . 8

2 Comparison between hot carrier injection models **11**
2.1 Modeling Framework . . . 12
2.1.1 The Boltzmann Transport Equation . . . 12
2.1.2 Band structure . . . 13
2.1.3 Scattering mechanisms . . . 15
2.2 Models description . . . 18
2.2.1 The Monte Carlo approach . . . 18
2.2.2 The Lucky Electron Model . . . 20
2.2.3 The Fiegna Model . . . 22
2.2.4 The Spherical Harmonics Expansion method . . . 23
2.3 Models benchmarking . . . 24
2.3.1 Homogeneous Case . . . 24
2.3.2 Non-homogeneous Case . . . 28
2.3.2.1 Distributions and non-local correction . . . 28
2.3.2.2 Gate current . . . 32
2.3.3 Summary . . . 37
2.4 Conclusions . . . 38
Bibliography . . . 39

3 Semi-analytic approach for hot carrier modeling **47**
3.1 Model presentation . . . 48
3.1.1 General overview . . . 48
3.1.2 Inputs . . . 49
3.1.3 Model description . . . 50
3.1.4 Post processing . . . 56
3.1.5 Summary . . . 61
3.2 Model with optical phonons . . . 62
3.2.1 Transport characteristics . . . 62
3.2.2 Distribution functions and generation rates . . . 64
3.2.3 Perpendicular fluxes and injection efficiencies . . . 69

3.2.4 Conclusions . . . 73
3.3 Analysis of the main features . . . 74
3.3.1 The impact of band structure . . . 74
3.3.1.1 Analytic dispersion relations . . . 74
3.3.1.2 Parabolic bands and local expression . . . 77
3.3.2 The role of the backscattering . . . 81
3.3.3 Conclusions . . . 83
3.4 Additional scattering mechanisms . . . 85
3.4.1 Electron Electron Scattering . . . 85
3.4.1.1 Scattering rates and implementation . . . 85
3.4.1.2 Results . . . 89
3.4.2 Impact Ionization . . . 93
3.4.3 Conclusions . . . 94
3.5 Conclusions . . . 95
Bibliography . . . 95

4 Comparison between measurements and modeling results 99
4.1 Measurements . . . 100
4.1.1 Motivation and methodology . . . 100
4.1.2 Cell description . . . 101
4.1.3 Measurement setup and extraction methodology . . . 103
4.1.4 Experimental errors . . . 109
4.2 Hot carrier injection regime . . . 111
4.2.1 TCAD and Monte Carlo simulations . . . 111
4.2.1.1 Structure calibration . . . 111
4.2.1.2 Calibration of the MC model . . . 113
4.2.1.3 Comparison with measurements for Flash cells . . . 114
4.2.2 Simulations using the 1D semi-analytic approach . . . 118
4.2.2.1 Description of the Charge Sheet Model . . . 118
4.2.2.2 Potential correction . . . 119
4.2.2.3 Comparison with I_g/I_d measurements . . . 121
4.3 Drain disturb regime . . . 123
4.3.1 The drain disturb phenomenon . . . 123
4.3.2 Simulation methodology . . . 125
4.3.3 Device optimization and comparison with measurements . . . 127
4.4 Conclusions . . . 131
Bibliography . . . 131

5 Modeling the cell degradation 137
5.1 Cell endurance . . . 138
5.1.1 Endurance characteristics . . . 138
5.1.2 Experimental analysis . . . 139
5.1.2.1 Characterization of the equivalent transistor . . . 139
5.1.2.2 Characterization of the Flash cell . . . 141

5.2 Interface state modeling 147
5.2.1 Historical background 147
5.2.2 Microscopic modeling framework 149
5.2.2.1 Rigorous approach 150
5.2.2.2 Possible approximation 151
5.2.2.3 Discussion 153
5.2.3 Application to 65nm technology 155
5.3 Conclusions 158
Bibliography 160

General conclusions **165**

List of publications **167**

A Probability Scheme **169**

B Perpendicular Flux Calculation **173**

Chapter 1

Introduction

This chapter introduces the background of the thesis. Section 1.1 gives a general overview of the Flash memory context starting from the market evolution to Flash memory architecture and associated issues. Section 1.2 defines the scope of this thesis which finally allows in section 1.3 to introduce the organization and the content of the following chapters.

1.1 Flash memory context

1.1.1 Market and evolution

The extraordinary development of the consumer electronics market has driven the continuous growth of the semiconductor industry in the last two decades. Most of the innovative products we have witnessed in the last years have been made possible by the ever-increasing storage capacity and flexibility provided by the memories, which is one of the most active segments with more than 20 % of the $ 300 billion worldwide semiconductor market [WSTS 2010]. Schematically, two families of memory products have been developed to meet the rising demand. On one hand, volatile memories (in which the information should be often refreshed and is definitely lost when unpowered) whose most representative product is Dynamic Random Acces Memory (DRAM) which occupies almost 60 % of the memory market share [iSupply 2010]. DRAM is present in numerous widely used consumer-end products such as PCs, smartphones and tablets. On the other hand, non-volatile memories (where there is no need to refresh the information which is kept even after the memory is unpowered) have seen its market share constantly increasing to reach a stable 35 % in the last years [Yinug 2007]. This evolution was propelled by two fundamental aspects Flash memories were able to confer to electronic devices: mobility and miniaturization. Beside these qualities, Flash memories have gained interest also because specific technology developments has been well integrated with many CMOS roadmap achievements.

The non-volatile memory sector has witnessed many evolutions in the past decades. Starting with the *Read Only Memory* (*ROM*), in which the content was defined at the fabrication steps and no subsequent modification was possible, the industry moved on to *Erasable Programmable ROM* (*EPROM*) where the memory state was changed electrically and by UV exposure. More flexibility was achieved by finally adopting *Electrically EPROM* (*EEPROM*) memories where all the operations were performed electrically. In this context, the Flash memory, which is an

EEPROM memory, was first introduced by Toshiba in the early '80s [Masuoka 1984]. In the early '90, Intel was the first semiconductor company to exclusively focus on Flash memories and leave the EPROM segment. By that time it controlled already 75 % of the Flash memory market, which roughly represented 1 % of the total semiconductor market. Several years later, AMD, Fujitsu, Atmel and SGS-Thomson entered the competition and fragmented the Flash industry until the late '90. Attracted by the high selling prices of Flash memory, other competitors initially present in the DRAM sector switched production to Flash memory and increased the production capacity. The most prominent example is Samsung which in 2005 already controlled one third of the Flash memory market. As a consequence, the avarage selling price of Flash memory decreased and has since lead to competitor consolidation or partnerships such as the Intel Micron Flash Technology (IMFT) joint venture, the cooperation between SanDisk and Toshiba, the joint-venture between AMD and Fujitsu giving birth to Spansion. The most recent exemple in this field was the birth of Numonyx as a joint-venture between STMicroelectronics and Intel which was then sold to Micron.

1.1.2 NAND and NOR memories

Among the Flash memories presently commercialized, the NAND and NOR memories take the overwhelming part of the market share. They have been designed to fulfill complementary needs which have been declined in different applications in the recent years. NOR-memories are used for *code storage* and *code and data storage* applications where a fast random access and a high code integrity is needed. Domains as different as mobile (cellular phones), networking (modem), automotive (cars), consumer products (PC, DVD, set top box, printer ...) have already integrated such memories. Instead, NAND memories are mainly used in *data storage* segments where low-cost and high-density are the major requirements. Some of the most popular products integrating NAND memory are the USB drives, the digital cameras, notebooks and smartphones in which large amounts of data should be written in a short time.

As a result, NAND and NOR memories come with different circuit architectures which provide the required qualities [Cappelletti 1999], [Brewer 2008]. Figure 1.1 shows that both memory types are organized in columns (*bitlines, BL*, connected to the drain terminals) and rows (*wordlines, WL*, connected to the gate terminals). In the NOR-case, the cells in the same bitline are connected in parallel as the source of each device is electrically accessible and grounded, while in the NAND-case all the cells in the bitline are connected in series. Hence, in the NOR architecture there are twice more metal lines (source and drain) with respect to the NAND architecture (drain only) which straightaway presents a higher density.

Broadly speaking, the structure of both cells is presently almost the same as it is derived from a standard nMOS transistor where the gate stack has been modified (Figure 1.2) to include two polysilicon areas (the *floating gate* and the *control gate*) which are separated from each other by an Oxide-Nitride-Oxide (ONO) layer. The

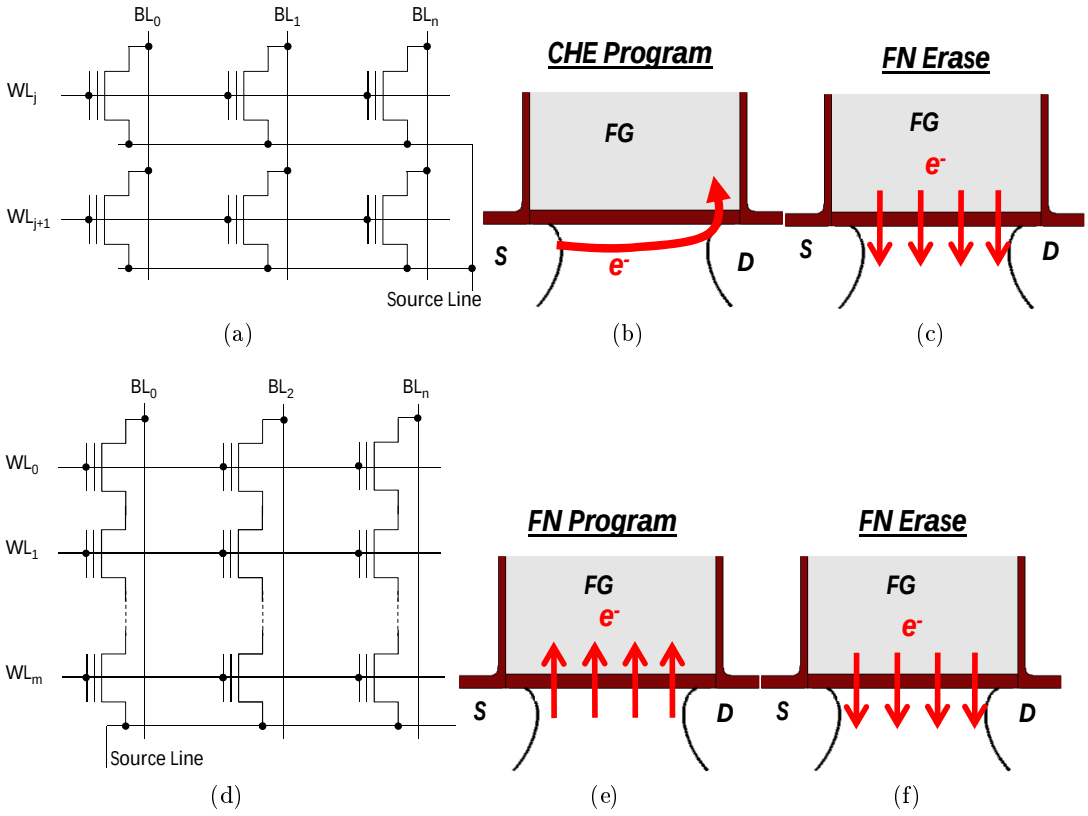


Figure 1.1: Schematic of the NOR *(a)* and NAND *(d)* array organization and sketch of the physical mechanisms involved during program and erase phases for each of NOR and NAND memories, respectively *(b, c)* and *(e, f)*.

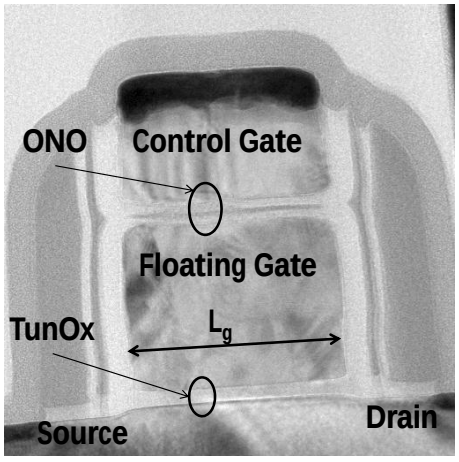


Figure 1.2: TEM image along the channel direction of a typical NOR Flash memory cell.

tunnel oxide further separates the channel from the floating gate which cannot be electrically addressed. This area is of critical importance for the device as it stores

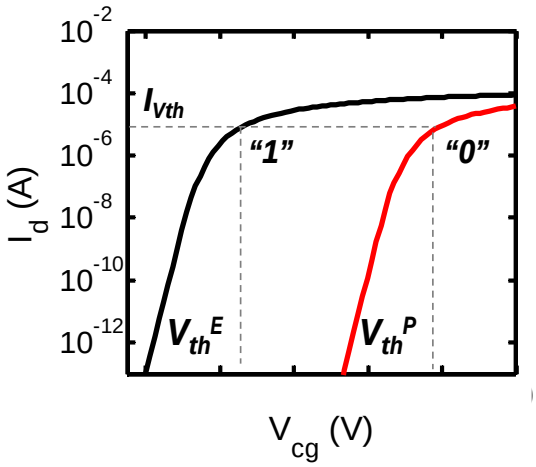


Figure 1.3: Typical current-voltage characteristics obtained after program and erase operations which respectively define the logic "0" and "1" states.

the electrons whose quantity determines the threshold voltage of the cell. Hence, the basic operating regimes of the cell include the *Program* and *Erase* phases, during which the electrons are put inside or pulled out of the floating gate, respectively, and the so-called *Read* phase during which the cell state is sensed. In classical 1-bit cells, two states corresponding to the "0" (after program) and "1" (after erase) logic states are distinguished based on the threshold voltage as shown in Figure 1.3. For both NAND and NOR cells, erase operation is performed on a whole sector (a large Fnumber of cells) using the Fowler-Nordheim (FN) tunneling mechanism (Figure 1.1*c, f*). While NAND cells are programmed using the same mechanism (Figure 1.1*e*), NOR cells use instead the Channel Hot Electron (CHE) injection phenomenon to put electrons into the floating gate (Figure 1.1*b*). In the latter case, the application of a high drain voltage during programming requires a rigorous control of the channel transport in the lateral direction in order to minimize short channel effects, current leakages and avoid the punch-through phenomenon and the drain turn-on effect. For this reason, the smallest dimension in NOR is usually the width direction and the NOR cells feature systematically longer gate lengths compared to NAND devices in which the relevant charge transport occurs in the vertical direction. For example, physical gate lengths of 100 and 25 nm are respectively predicted for NOR and NAND cells in 2012 [ITRS 2010]. The smaller size of the NAND cell further increases the storage capacity of NAND architecture and subsequently decreases the cost per bit. The NAND architecture features other advantages as well compared to the NOR case such as a lower programming time and lower active power consumption. Although CHE mechanism is intrinsically faster than FN, NOR cells are programmed once at a time while multiple NAND cells are often programmed simultaneously. Furthermore, the presence of a high channel current for CHE and its relatively small efficiency contribute to increase the consumption. However, the single-bit programming scheme adopted in the case of the NOR memories, guaran-

tees almost 100 % of the bits. This avoids the presence of error correction codes which are commonly integrated with NAND memories in order to verify the stored information and to correct it if needed. Furthermore the direct access to single bits provides a fast read time avoiding any complex circuitry in contrast to the NAND architecture.

1.1.3 Stand-alone vs. embedded memories

A closer look at the final products integrating Flash memories reveals that the above pros/cons and features of both memories have been used in two main directions. Indeed, we either find *stand-alone* Flash memory products where the Flash-related components occupy almost the total chip area, or *embedded* architectures where Flash memory is physically integrated into a host logic device (microcontrollers, application-specific integrated circuits, ...) on the same silicon substrate, in which the Flash memory is intended to add features to the system [Brewer 2008]. On the one hand, stand-alone memories require high density, low cost and high writing speed, which are achieved by adopting a simple wafer process and by minimizing the cell size. On the other hand, high performance and low cost are mainly required for embedded memories, with the low-cost criteria being not as critical as in the case of stand-alone memories. The direct interface with the host logic in the same chip and the customized memory array configuration allow the embedded Flash memory to achieve higher system speeds. The design flexibility allows the realization of a system on chip which can possibly lead to a lower system cost. Furthermore, the elimination of input/output buffers used in the case of stand-alone memories reduce the dissipated power and the number of pads and connections which globally help increase the system reliability. However, the accommodation of the Flash memory with the logic host increases the process complexity and possibly the test cost. Furthermore, as embedded memories are mostly application oriented, additional design time and cost is required with respect to standard stand-alone memories [Brewer 2008]. Finally, based on the previous considerations, we note that most of stand-alone memories are composed of NAND memories, while NOR memories are mainly used in embedded configurations.

1.1.4 The NOR scaling issues

The continuous shrinking of the device dimensions brings up significant technological problems when applied to planar bulk NOR memories. Indeed, in order to reduce the gate length and have a good electrostatic control of the channel, the tunnel oxide is thinned down, the effective substrate doping is increased and the source/drain junction depths are decreased [Lu 2009]. However, unwanted gate leakage occurring in the off-state, due to bulk oxide defects, imposes a minimal SiO_2 thickness of around 8-9 nm [Song 2003], [Park 2004], [Servalli 2005] in order to keep the stored electrons for a 10-year standard retention period. Hence, doping is the only way to control short channel effects and especially avoid punch-through at high drain

voltage. Indeed, whenever a cell is programmed, the other unselected cells sharing the same bitline (c.f. Figure 1.1) should have a low leakage current. Thus, minimal gate lengths of 120 nm with good punch-through characteristics down to 100 nm [Servalli 2005] have been obtained after adopting aggressive channel and drain doping profiles. Beside the electrostatic considerations, the optimization of such profiles is crucial to achieve good dynamic program performances in the CHE regime. A ΔV_{th} of 4-5 V should be reached in less then 1 μs when applying drain voltages of around 4 V. In this sense, adopting steep channel/drain junctions goes in the right direction. However, such junctions are more affected by the *drain disturb*, during which band-to-band tunneling [Ielmini 2006] on unselected cells yields an unwanted leakage current and a charge loss in the floating gate. Furthermore, all the injected carriers into the floating during CHE or drain disturb phases give rise to oxide degradation and thus must be minimized.

1.2 Scope of the thesis

The joint study of all the previously cited phenomena is a complex task. First, a good 2D/3D description of device electrostatics is required. Then, various models governing these phenomena should be available. In this context, Technology Computer Aided Design (TCAD) provides the necessary environment allowing the study of all these aspects and can give valuable insight for device optimization. The central and starting point of NOR memory optimization concerns CHE during program operation. However, several limits of traditional TCAD injection models (Fiegna Model *FM* [Fiegna 1991] and the Lucky Electron Model *LEM* [Hasnat 1997]) have been underlined [Fischetti 1995]. Their predictivity is thus questionable. In the meanwhile, the Monte Carlo (MC) method has been established and accepted as an accurate reference for hot carrier injection problems [Bude 2000], [Ghetti 2003], [Palestri 2006]. Indeed, it properly simulates carrier transport by accounting for the relevant scattering mechanisms in the full band structure of silicon. Unfortunately, such approach is still computationally expensive and is hardly applicable for every-day use in industry. The choice of the modeling approach thus depends on the trade-off between accuracy and computational burden. In this sense, the Spherical Harmonics Expansion (SHE) method has been recently re-considered and presented as a good candidate to solve this dilemma [Hong 2010], [Jin 2011]. However, an in-depth evaluation of such method and its applicability on state-of-the-art memories has not yet been performed. In particular, the electron-electron scattering effect needs to be evaluated in the context of constant reduction of operating voltages.

The assessment of all these models requires adopting a combined micro- and a macroscopic scale approach, respectively represented by the carrier distribution function and the terminal currents. Accurate experiments and a rigorous procedure for comparison with them is needed for such an evaluation. It should be first verified that the cell electrostatics is well reproduced before measuring the injection current. The measurement of the small gate currents is a delicate procedure and an estimation

of the associated errors should accompany the results. Therefore, the cell and its equivalent transistor are commonly employed. Finally, the same mixed approach (experimental and simulation, cell and test structure) should be as well adopted for the drain disturb and degradation phenomena in which small amounts of carriers have a considerable impact on the device performances.

The scope of this thesis is to investigate the hot carrier injection into the floating gate of Flash memory cells and some of the associated degradation mechanisms from a simulation and an experimental perspective. The simulation studies of the following pages rely on the guiding principle that a proper microscopic transport description is the first step in order to obtain sound macroscopic quantities. For this reason, much of the simulation work will be around the evaluation of the carrier energy distribution function as a function of the position, length and bias. Various device-level simulation approaches will be employed, benchmarked and developed in this thesis in the objective of calculating an *accurate* distribution function. At the same time, extensive characterization, including different kinds of measurements, will be employed to evaluate the hot carrier injection and the associated degradation as well as to assess the model predictions for the Flash cell or in simplified test structures.

1.3 Organization of the thesis

The results of this thesis have been condensed in four chapters whose content is briefly summerized in the following.

- Chapter 2 will present the hot electron injection models currently available in the TCAD environment: FM, LEM and SHE. Extensive comparisons with a reference Monte Carlo simulator will be given for the main figures of merit of the injection regime such as the distribution function and the gate current density along the channel as well as the injection efficiency. These comparisons will quantify the accuracy of each approach.

- Chapter 3 will introduce a new 1D semi-analytic model able to reproduce the electron distribution functions along the channel for various device lengths and bias conditions. The insight provided by the previous chapter will reveal particularly useful in developing a flexible and efficient approach which allows to study the impact of the most relevant hot carrier scattering mechanisms and band structure aspects. Hence, this chapter will discuss the role of each of the inelastic electron-phonon and electron-electron scattering, the impact ionization processes as well as the impact of the band structure through the extensive comparisons with rigorous Monte Carlo simulations.

- Chapter 4 will present an experimental study of CHE injection during the program operation and a subsequent comparison with the results obtained from the Monte Carlo, the Spherical Harmonics Expansion method and the

new semi-analytic model. Static and dynamic measurements have been extensively applied to asses the validity extent of these models in a broad range of gate lengths and bias configurations using a 65 nm technology. The last part of this chapter will be devoted to the introduction, the analysis and the optimization of the hot hole injection phenomena occurring during the drain disturb phase.

- Chapter 5 will finally present an analysis of the endurance characteristics degradation due to oxide defects and a modeling approach for the generation of the interface states as an important component of the oxide defects. First, the observed reduction of the programming window during cycling has been experimentally investigated with the purpose of separating the impact of the defects on each of the program, erase and read phases. Then, the precise knowledge of the distribution function has been coupled to a microscopic model for interface defect generation such as to provide a global framework able to reproduce the defects' density along the channel and their impact on the macroscopic transport parameters.

Bibliography

[Brewer 2008] J.E. Brewer and M. Gill. Nonvolatile memory technologies with emphasis on flash: a comprehensive guide to understanding and using nvm devices, volume 8. Wiley-IEEE Press, 2008. (Cited on pages 2 and 5.)

[Bude 2000] J.D. Bude, M.R. Pinto and R.K. Smith. *Monte Carlo simulation of the CHISEL flash memory cell.* IEEE Transactions on Electron Devices, vol. 47, no. 10, pages 1873–1881, 2000. (Cited on pages 6 and 31.)

[Cappelletti 1999] P. Cappelletti. Flash memories. Springer Netherlands, 1999. (Cited on pages 2, 73 and 140.)

[Fiegna 1991] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi and B. Ricco. *Simple and efficient modeling of EPROM writing.* IEEE Transactions on Electron Devices, vol. 38, no. 3, pages 603 –610, March 1991. (Cited on pages 6, 22, 24, 49 and 66.)

[Fischetti 1995] M.V. Fischetti, S.E. Laux and E. Crabbe. *Understanding hot electron transport in silicon devices: Is there a shortcut?* Journal of Applied Physics, vol. 78, no. 2, pages 1058 –1087, July 1995. (Cited on pages 6, 18, 19, 20, 21, 22, 28, 29, 93, 94, 113 and 114.)

[Ghetti 2003] A. Ghetti. *Hot-electron induced MOSFET gate current simulation by coupled silicon/oxide Monte Carlo device simulation.* Solid-State Electronics, vol. 47, no. 9, pages 1507–1514, 2003. (Cited on pages 6 and 113.)

[Hasnat 1997] K. Hasnat, C.F. Yeap, S. Jallepalli, S.A. Hareland, W.K. Shih, V.M. Agostinelli, A.F. Tasch and C.M. Maziar. *Thermionic emission model of electron gate current in submicron NMOSFETs.* IEEE Transactions on Electron Devices, vol. 44, no. 1, pages 129–138, 1997. (Cited on pages 6 and 23.)

[Hong 2010] S.M. Hong, G. Matz and C. Jungemann. *A deterministic Boltzmann equation solver based on a higher order spherical harmonics expansion with full-band effects.* IEEE Transactions on Electron Devices, vol. 57, no. 10, pages 2390–2397, 2010. (Cited on pages 6 and 24.)

[Ielmini 2006] D. Ielmini, A. Ghetti, A.S. Spinelli and A. Visconti. *A study of hot-hole injection during programming drain disturb in flash memories.* IEEE Transactions on Electron Devices, vol. 53, no. 4, pages 668–676, 2006. (Cited on pages 6, 124 and 125.)

[iSupply 2010] iSupply, 2010. (Cited on page 1.)

[ITRS 2010] International Technology Roadmap for Semiconductors ITRS, 2010. (Cited on page 4.)

[Jin 2011] S. Jin, S.M. Hong and C. Jungemann. *An Efficient Approach to Include Full-Band Effects in Deterministic Boltzmann Equation Solver Based on High-Order Spherical Harmonics Expansion.* IEEE Transactions on Electron Devices, no. 99, pages 1–8, 2011. (Cited on pages 6 and 24.)

[Lu 2009] C.Y. Lu, K.Y. Hsieh and R. Liu. *Future challenges of flash memory technologies.* Microelectronic Engineering, vol. 86, no. 3, pages 283–286, 2009. (Cited on page 5.)

[Masuoka 1984] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro and S. Tanaka. *A new flash E2PROM cell using triple polysilicon technology.* In International Electron Devices Meeting (IEDM) 1984, volume 30, pages 464 – 467, 1984. (Cited on page 2.)

[Palestri 2006] P. Palestri, N. Akil, W. Stefanutti, M. Slotboom and L. Selmi. *Effect of the gap size on the SSI efficiency of split-gate memory cells.* IEEE Transactions on Electron Devices, vol. 53, no. 3, pages 488–493, 2006. (Cited on pages 6, 19, 24, 30 and 62.)

[Park 2004] Chankwang Park, Sangpil Sim, Jungin Han, Chul Jeong, Younggoan Jang, Junghwan Park, Jaehoon Kim, Kyucharn Park and Kinam Kim. *A 70nm NOR flash technology with 0.049 mu;m2 cell size.* In VLSI Technology, 2004. Digest of Technical Papers. 2004 Symposium on, pages 238 – 239, june 2004. (Cited on page 5.)

[Servalli 2005] G. Servalli, D. Brazzelli, E. Camerlenghi, G. Capetti, S. Costantini, C. Cupeta, D. DeSimone, A. Ghetti, T. Ghilardi, P. Gulli*et al. A 65nm NOR flash technology with 0.042/spl mu/m/sup 2/cell size for high performance*

multilevel application. In Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International, pages 849–852. IEEE, 2005. (Cited on pages 5 and 6.)

[Song 2003] Y. Song, S. Lee, T. Kim, J. Han, H. Lee, S. Kim, J. Park, S. Park, J. Choi, J. Kim, D. Lee, M. Cho, K. Park and K. Kim. *Highly manufacturable 90 nm NOR flash technology with 0.081 mu;m2 cell size.* In VLSI Technology, 2003. Digest of Technical Papers. 2003 Symposium on, pages 91 – 92, june 2003. (Cited on page 5.)

[WSTS 2010] World Semiconductor Trade Statistics WSTS and Databeans, 2010. (Cited on page 1.)

[Yinug 2007] F. Yinug. *The Rise of the Flash Memory Market: Its Impact on Firm Behavior and Global Semiconducto Trade Patterns.* United States International Trade Commission, 2007. (Cited on page 1.)

Chapter 2

Comparison between hot carrier injection models

The modeling of hot carrier effects in semiconductor devices has been a constant concern since the development of the bipolar transistor in the second half of the previous century. Various modeling groups have tackled this issue under different assumptions resulting today in many available methodologies. Some of the latter have been progressively integrated into Technology Computer Aided Design (TCAD) tools which are commonly used to predict and optimize device performances in industry. The usefulness of each approach depends on the compromise resulting from the balance between the required degree of physical insight and computational burden.

Thus, within the context of the development of advanced embedded non-volatile memories, the hot electron injection models presently available in TCAD are investigated in this chapter. The scope of this analysis is two-fold. On one hand, the Lucky Electron Model, the Fiegna Model and the recently implemented Spherical Harmonics method are systematically compared to a well-established rigorous Monte Carlo simulations, taken as a reference throughout this work in order to determine the extent of validity of each model. On the other hand, this evaluation procedure sheds light on the physical reasons of failure or success for each method. Grasping the most important ingredients for an accurate modeling of hot carrier injection will turn out to be particularly useful in the next chapter.

This chapter is organized in three sections. The general framework of hot carrier models, including the band structure and the scattering mechanisms, is first presented in Section 2.1. The main assumptions and features of the models are then briefly introduced in Section2.2, before eventually comparing them under homogeneous and device conditions in Section 2.3.

2.1 Modeling Framework

This section aims to give the definitions of the main physical quantities involved in the modeling procedure, valid throughout this thesis. The framework of the hot carrier modeling is based on a semi-classical approach which main equation is first introduced in subsection 2.1.1. The fundamental ingredients of this equation include the band structure and the scattering mechanisms which are respectively treated in the following subsections 2.1.2 and 2.1.3. The notations and quantities defined in this section will be valid throughout the other chapters.

2.1.1 The Boltzmann Transport Equation

The carrier transport in silicon has been widely described under the semi-classical approximation governed by the Boltzmann Transport Equation (BTE):

$$\overbrace{\frac{\partial f\left(\vec{r},\vec{k},t\right)}{\partial t}}^{time\ term}+\overbrace{\vec{v}\cdot\vec{\nabla}_{\vec{r}}f\left(\vec{r},\vec{k},t\right)}^{diffusion\ term}-\overbrace{\frac{q\vec{E}}{\hbar}\cdot\vec{\nabla}_{\vec{k}}f\left(\vec{r},\vec{k},t\right)}^{drift\ term}= \\ -\underbrace{\int\left(1-f\left(\vec{r},\vec{k}',t\right)\right)S\left(\vec{r},\vec{k},\vec{k}'\right)f\left(\vec{r},\vec{k},t\right)d\vec{k}'}_{out-scattering\ term} \\ +\underbrace{\int\left(1-f\left(\vec{r},\vec{k},t\right)\right)S\left(\vec{r},\vec{k}',\vec{k}\right)f\left(\vec{r},\vec{k}',t\right)d\vec{k}'}_{in-scattering\ term} \quad (2.1)$$

The BTE is a conservation equation which can be derived after considering the incoming and outgoing carrier fluxes in the phase space [Lundstrom 2000]. This involves seven variables: x, y, z (assembled in $\vec{r}$: the real space position), k_x, k_y, k_z (assembled in $\vec{k}$: the momentum space position) and the time t. The unknown of the equation is the probability function $f(\vec{r},\vec{k},t)$, which represents the probability to find a carrier at a position $\vec{r}$ with a momentum $\vec{k}$ at a time t. The quantities $\vec{E}$, q and $\hbar$ are the electric field, the positive electron charge and the reduced Planck's constant, respectively, while $\vec{v}$ is the carrier group velocity defined as:

$$\vec{v}=\frac{1}{\hbar}\vec{\nabla}_{\vec{k}}\varepsilon(\vec{k}) \quad (2.2)$$

with $\varepsilon(\vec{k})$ being the dispersion relation (subsection 2.1.2). Finally, $S(\vec{r},\vec{k}',\vec{k})$ represents the transition rate for a carrier to instantaneously change its momentum from $\vec{k}'$ to $\vec{k}$ at the position $\vec{r}$.

The solution of Equation 2.1 provides the probability function f, which contains the relevant information of the carriers in the semiconductor. For instance the carriers concentration and the carriers mean velocity can be calculated by:

$$n(\vec{r}) = \frac{2}{(2\pi)^3} \int f(\vec{r},\vec{k}) d\vec{k} \tag{2.3}$$

$$\bar{v}(\vec{r}) = \frac{1}{n(\vec{r})} \frac{2}{(2\pi)^3} \int \frac{1}{\hbar} \vec{\nabla}_{\vec{k}} \varepsilon(\vec{k}) f(\vec{r},\vec{k}) d\vec{k} \tag{2.4}$$

However, solving the BTE is not an easy task because it is a non-linear integro-differential equation which includes a great number of variables. The non-linearities are introduced by the Pauli exclusion principle ($[1-f]$ term) and the calculation of some of the transition rate which require the preliminary knowledge of the probability function (e.g. carrier-carrier scattering). In most of the cases, a linear BTE neglecting both dependencies is used. This can further serve as a starting point for the solution of the non-linear equation [Jungemann 2003].

Section 2.2 briefly introduces some of the methods which are commonly used to solve the linear BTE. Although in some cases it is possible to explicitly consider the three momentum space components (Subsection 2.2.1), the probability function is generally plotted and analyzed as a function of the total energy $f(\varepsilon)$, after a summation over all the momentum directions. The latter approach is followed throughout this work accompanied with additional considerations on the isotropy of the probability function.

2.1.2 Band structure

The band structure is the relation between the momentum $\vec{k}$ values and the energy ε inside the elementary cell of the reciprocal space, called the First Brillouin Zone (FBZ). The dispersion relation $\varepsilon(\vec{k})$ inherently defines the electronic properties of the material. The calculation of the band structure requires solving the Hamiltonian of the system. In the electronic community, this is mainly performed using semi-empirical methods such as the Empirical Pseudopotential Method (EPM) [Chelikowsky 1976], the **k.p** method [Cardona 1966] and the tight binding method [Jancu 1998]. A comprehensive review of the methods and application to relevant semiconductor materials can be found in [Esseni 2011], [Rideau 2011]. Figure 2.1*a* reports the silicon full band structure for electrons along the main symmetry axes of the irreducible wedge in the FBZ.

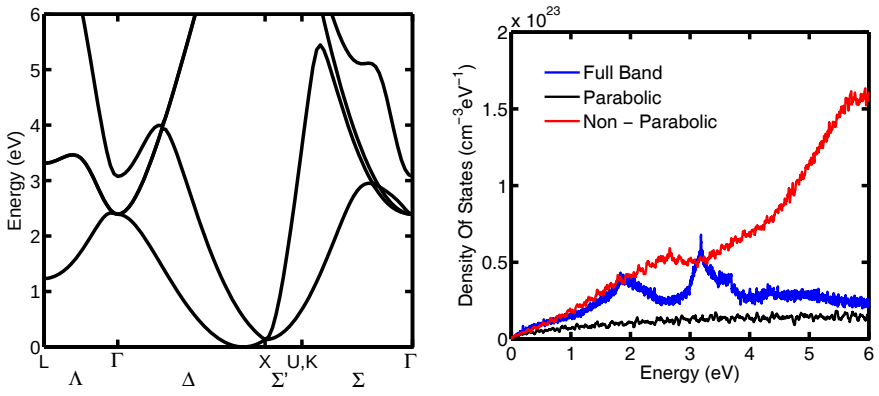
Knowing the dispersion relation, the density of states can be calculated by:

$$g(\varepsilon) = \frac{2}{(2\pi)^3} \int \delta\left(\varepsilon - \varepsilon(\vec{k})\right) d\vec{k} \tag{2.5}$$

In addition, the knowledge of the probability function and the density of state, allows to calculate the *distribution function* of the carriers as:

$$n(\vec{r},\varepsilon) = f(\vec{r},\varepsilon) \cdot g(\varepsilon) \tag{2.6}$$

The EPM calculation yields a numerical multi-branch conduction band (Figure 2.1a). However, near the equilibrium conditions, most of the carriers occupy the



(a) (b)

Figure 2.1: *a.* Silicon full band structure along the main high symmetry axes of the irreducible wedge, calculated with EPM method. *b.* Density of states obtained after the full band, the parabolic and the non-parabolic band structures, the latter including only the Δ valleys.

lowest available energies which show a local minimum situated at $k_{0x} = 0.85.2\pi/a_0$ in the Δ direction, with a_0 being the silicon lattice parameter equal to 5.43 Å. From the symmetries of the silicon crystal, six minima, called Δ *valleys*, are present in the FBZ, situated at $(\pm 0.85, 0, 0) \cdot 2\pi/a_0$, $(0, \pm 0.85, 0) \cdot 2\pi/a_0$, $(0, 0, \pm 0.85) \cdot 2\pi/a_0$. At these positions, the band structure can be locally approximated by ellipsoids which main axes are related to the transport masses m_x, m_y, m_z in the principal k_x, k_y, k_z directions:

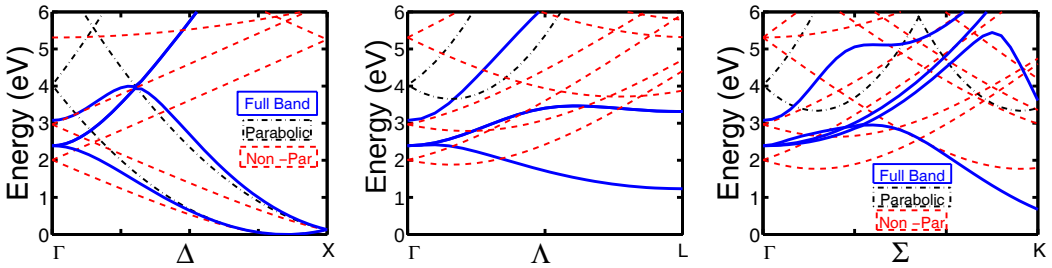
$$\varepsilon - \varepsilon_0 = \frac{\hbar^2}{2}\left[\frac{(k_x - k_{0x})^2}{m_x} + \frac{(k_y - k_{0y})^2}{m_y} + \frac{(k_z - k_{0z})^2}{m_z}\right] \tag{2.7}$$

where k_{0x}, k_{0y}, k_{0z} and ε_0 are the coordinates of a given minimum and its associated energy. As the ellipsoids show an axial symmetry around the main valley direction, only two masses, respectively the longitudinal m_l and the transverse m_t, are enough to characterize the ellipsoid. For instance, for a Δ valley situated at $(0.85, 0, 0)$, $m_x = m_l = 0.919 m_0$ and $m_y = m_z = m_t = 0.190 m_0$, with m_0 being the electron mass.

This approximation, also called *Parabolic Bands Approximation*, has been widely employed for modeling and simulation purposes. It indeed provides a simple analytical dispersion relation which is very useful for low energy electronic transport. However, at higher energies this approximation shows increased discrepancies with the full band structure calculated by EPM. The non parabolicity effects are usually modeled at first order by introducing a correction:

$$(\varepsilon - \varepsilon_0)\left[1 + \alpha(\varepsilon - \varepsilon_0)\right] = \frac{\hbar^2}{2}\left[\frac{(k_x - k_{0x})^2}{m_x} + \frac{(k_y - k_{0y})^2}{m_y} + \frac{(k_z - k_{0z})^2}{m_z}\right] \tag{2.8}$$

with α being the non parabolicity factor, usally equal to 0.5 eV^{-1} [Kane 1957] in silicon. Figure 2.2 compares the analytical expressions 2.7, 2.8 with the numerical full band structure along the three main high symmetry axes of the irreducible edge.



(a) (b) (c)

Figure 2.2: Full-Band, Parabolic and Non-Parabolic band structures represented along high symmetry paths of the Brillouin zone. The analytical expressions include all the Δ valley conduction bands up to 10 eV.

Both analytic expressions integrate the first 24 conduction bands coming from the Δ valleys of the FBZ and of the first adjacent cells (equivalent to band-folding). The obtained DOS is compared to the full band description on Figure 2.1*b*. Significant differences with the full band structure can be seen especially at high energies. All the presented bands will be used throughout this chapter and the following one.

2.1.3 Scattering mechanisms

The electrons in the conduction band continuously scatter with different entities in the device. In the semi-classical model, all scatterings are considered instantaneous and localized in real space. They modify the electrons momentum and energy according to the interaction type. While the effects of the most relevant interactions are given in this paragraph, the details of the calculation procedure for the scattering rates can be found in [Jacoboni 1983], [Lundstrom 2000], [Esseni 2011].

The *transition rate* $S(\vec{k}, \vec{k'})$ refers to the probability per unit time for a carrier to change its momentum from $\vec{k}$ to $\vec{k'}$. The computation of this quantity for different mechanisms is performed using the Fermi Golden Rule and is discussed in [Esseni 2011], for example. For practical purposes, it is important to define the probability for a carrier to scatter from $\vec{k}$ to any other momentum state. This quantity is called the *scattering rate* or equivalently the *relaxation rate* and is mathematically formulated as:

$$S(\varepsilon) = \frac{1}{\tau(\varepsilon)} = \frac{1}{(2\pi)^3}\int S(\vec{k}, \vec{k'})d\vec{k'} \tag{2.9}$$

Phonon Scattering

Phonons are particles representing the crystal lattice oscillations. They are one of the most important sources of scattering in electron devices. Due to the presence of two atoms of silicon per lattice unit cell, three *acoustic* and three *optical* phonon branches are present. Each of the branches is divided into one longitudinal and two degenerate transverse modes, whose dispersion relations can be found in [Kittel 1986]. Although the effects of the phonons are always referenced with respect to the bottom of the six Δ valleys (first conduction band), the transitions and selection rules are applied for all energies of all Δ valleys of the FBZ. In particular, two transition types are often distinguished:

- *intra-valley transition* : the electron remains in the same valley after the scattering. Due to selection rules, the transitions concerning the Δ valleys are assisted only by acoustic phonons [Esseni 2011]. In this case, the exchanged phonon momentum and energy is small [Lundstrom 2000], [Esseni 2011], hence at room temperature this transition is considered *elastic* (no energy relaxation).

- *inter-valley transition* : the scattering moves the electron either in the opposite Δ valley of the same axis or in one of the other Δ valleys ((g-type and f-type transitions, respectively, schematized in Figure 2.3). This transition can be assisted by either acoustic or optical phonons. The exchanged energies are in the order of the thermal energy and thus the transition is considered *inelastic*, through phonon emission or absorption.

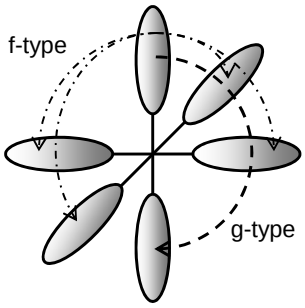


Figure 2.3: Schematic representation of the 6 Δ valleys in silicon and the associated f- and g-type intervalley transitions due to phonon scattering.

Under the isotropic approximation, a single effective coupling constant, also called *deformation potential*, is used to characterize the strength of the electron-phonon interaction. A deformation potential of $D_{ac} = 8.7\ eV$ is used for the intra-valley transitions. Table 2.1 summerizes the deformation potentials and the exchanged energies for the inter-valley transitions after [Jacoboni 1983].

Finally, the scattering rates of both intra- and inter-valley transitions can be written as:

Transition	Phonon	Energy [meV]	Def.Pot (D) [10^8 eV/cm]
inter-valley	TA-g	12	0.5
	LA-g	18.5	0.8
	LO-g	61.2	11
	TA-f	19	0.3
	LA-f	47.4	2
	LO-f	59	2

Table 2.1: The energy and the deformation potential of the phonons participating in the inter-valley transitions. The values are after [Jacoboni 1983]

$$S(\varepsilon)_{intra} = \frac{1}{\tau(\varepsilon)_{intra}} = \frac{4\pi^2 k_B T}{h\rho v_s^2} D_{ac}^2 \cdot g(\varepsilon) \tag{2.10}$$

$$S(\varepsilon)_{in.em} = \frac{1}{\tau(\varepsilon)_{in.em}} = \frac{h(N_{op}+1)}{2\rho\varepsilon_{in}} D_{in}^2 \cdot g(\varepsilon - \varepsilon_{in}) \tag{2.11}$$

$$S(\varepsilon)_{in.ab} = \frac{1}{\tau(\varepsilon)_{in.ab}} = \frac{hN_{op}}{2\rho\varepsilon_{in}} D_{in}^2 \cdot g(\varepsilon + \varepsilon_{in}) \tag{2.12}$$

ρ and v_s are the mass density and the sound velocity in silicon, while h, k_B, T bear their usual meaning. D_{ac} and D_{in} are the elastic acoustic and inter-valley deformation potentials, the latter being split into emission (em index) and absorption (ab index) processes with an energy exchange of ε_{in}. g is the electron density of states while N_{op} is the phonon number defined as $1/\left[\exp\left(\varepsilon/k_B T\right) - 1\right]$. These expressions show that the density of states plays a crucial role in determining the electron-phonon scattering rate, thus reinforcing the need to account for an accurate band structure. From Table 2.1, it can be seen that the g-type longitudinal optical phonon has the strongest deformation potential. Therefore, considering the quadratic dependence of the scattering rate on the latter, the *LO-g* phonon strongly determines the inter-valley transitions. Hence, in the following, the inelastic transitions will be associated to the optical phonons.

Impact Ionization

Impact ionization is another important scattering mechanism mainly affecting the high energy carriers. This process consists in a Coulomb interaction between an electron in the conduction band, called the *primary* electron, and another electron of the valence band, called the *secondary* electron, which results in the promotion of the latter in the conduction band and the generation of a secondary hole in the valence band. This endothermic process is triggered by the primary electron having an energy ε_{PRIM} higher than a given threshold. The energy threshold depends on the carrier momentum [Bude 1992],[Sano 1994], thus making the process anisotropic. However, it has been already shown that isotropic scattering rates, averaged over all

momenta having the same energy, well reproduce experimental data [Cartier 1993] with a threshold value close to the silicon band gap ($\varepsilon_{gSi} = 1.12eV$). Thus, it seems plausible that the anisotropy of the II process is hidden by electron-phonon scatterings which efficiently randomize the distribution function especially at high electron energies [Fischetti 1995].

Traditionally, the isotropic energy-dependent scattering rate have been provided either by the calculation of the scattering matrix elements [Kane 1967], [Bude 1992] or by empirical Keldysh-type expressions adjusted on experimental or simulation data [Thoma 1991], [Cartier 1993], [Kamakura 1994], [Jungemann 1996a].

Carrier-carrier scattering concerns the collision between two carriers and its main effect is the enhancement of the hot energy tail [Childs 1996], [Ghetti 1996], [Abramo 1996], [Fischer 1997], [Ghetti 2002], [Fixel 2008]. The energy-exchange between both carriers is a function of their initial momenta. Therefore, this process is inelastic and anisotropic. This mechanism becomes increasingly important with increased carrier concentration [Ferry 1999], thus particularly degrading the reliability of short-channel devices [La Rosa 2007].

Ionized impurity scattering is another mechanism affecting the carriers in the presence of external dopants in the silicon lattice. This interaction is generally considered elastic and anisotropic, mostly affecting the low-energy carriers [Lundstrom 2000].

Surface roughness scattering takes place at material interfaces, such as the Si/SiO_2 interface in a MOSFET device. It is considered an elastic and anisotropic process which impact is increasingly important towards strong inversion conditions.

2.2 Models description

The introduction of the most important quantities in the previous section opens the way to the presentation of the most widely used approaches to solve the BTE. The short summaries of the Monte Carlo approach in subsection 2.2.1 and of the TCAD-available models, namely the Lucky Electron Model, the Fiegna Model and the Spherical Harmonics Expansion method, respectively in subsections 2.2.2, 2.2.3 and 2.2.4, attempt to provide the most important features of these models. Extensive reading on the latter can be found in the proposed references.

2.2.1 The Monte Carlo approach

The Monte Carlo (MC) method is a stochastical approach to solve the BTE which involves the simulation and the monitoring of the trajectory of a large number of carriers. The trajectory of a carrier is composed of free-flight sequences, governed by the Newton's laws of motion and of scattering events described by quantum mechanical laws. The statistics gathered throughout the simulation allows to estimate the

probability function, the accuracy of which depends on the number of simulated carriers and on the simulation time. The carriers are simulated either simultaneously, giving rise to the so-called Ensemble MC [Fischetti 1988], or in sequence, leading to Single Particle MC [Bufler 2000]. Excellent descriptions of the methods and a wide range of applications can be found in [Lundstrom 2000], [Jungemann 2003], [Esseni 2011].

A first comprehensive review of the MC method for electron devices is given in [Jacoboni 1983]. The main inputs for the MC-based transport solvers are the band structure and the scattering rates in the semiconductor. The accurate study of the high-energy transport is made possible by the inclusion of the full-band structure which has a considerable impact on the DOS at high energy and consequently on the electron-phonon scattering rates [Tang 1983], [Fischetti 1988] (c.f. Section 2.1). The relevance of the latter rates together with the impact ionization rates is verified on various indicators of the carriers heating such as the velocity-field characteristics at low fields ($< 10^4 Vcm^{-1}$), the impact ionization coefficient vs. electric field and quantum yield vs. energy characteristics at high fields [Jungemann 2003]. Alongside these processes, carrier-carrier interactions and scatterings with ionized impurities need to be also included in the Full Band Monte Carlo (FBMC) simulations [Fischetti 1995].

The stochastic nature of the approach is conferred by the random choice of the free-flight duration, the scattering mechanism and the state after the scattering, all random choices being renewed at each time interval. The statistics at each time interval, which is a subdivision of the total simulation time, are collected either before the scattering [Jacoboni 1983] or at fixed time intervals [Fischetti 1988], depending on the approach. The MC is run independently for each bias condition; the starting point is often given by a hydrodynamic (HD) simulation which provides the initial guess for the potential and carrier density profile. The MC simulations can be further coupled with the Poisson equation to obtain Self Consistent (SC) simulations or on the contrary leading to Non Self Consistent (NSC) simulations which use the initial HD potential profile throughout the whole simulation time.

The ensemble MC used throughout this work [Palestri 2006] contains all the above-cited ingredients and will be considered as a reference in the following comparisons (c.f. subsection 2.3). In these simulations, the gate current is calculated as:

$$I_g(x) = \sum_i \frac{q w_i T\left(\varepsilon_\perp^i, x\right)}{\Delta t} \tag{2.13}$$

where i represents the particles hitting the interface at the x position during the time interval Δt with a statistical weight w_i. $T\left(\varepsilon_\perp^i, x\right)$ is the tunneling probability of a given particle hitting the interface with a given perpendicular energy $\varepsilon_\perp$. T is calculated using the transfer matrix approach [Ando 1987]. However, defining a perpendicular energy in a full-band structure is not an obvious task [Bufler 2005]. Throughout this thesis, the perpendicular energy in MC has been calculated by:

$$\varepsilon_{\perp} = \varepsilon_{tot} - \frac{\hbar^2 k_{\parallel}^2}{2m_{ins}} \tag{2.14}$$

This expression accounts for parallel momentum conservation at the interface [Fischetti 1995], where $m_{ins} = 0.5m_0$ is the electronic mass inside the oxide (in agreement with [Stadele 2003] for thick oxides), $k_{\parallel}$ is the parallel wave-vector referred to the Γ point and ε_{tot} is the particle's total energy with respect to the nearest conduction band minimum. Furthermore, barrier lowering due to image force has been also accounted for. It ought to be mentioned that Equation 2.14 and the expression proposed in [Bufler 2005] lead essentially to the same result [Jin 2009].

Finally, when calculating I_g, an adequate number of hot carriers should be available in order to obtain reliable results. However, the scattering processes tend to cool the carriers down, thus diminishing the hot carrier tail. As a consequence, statistical enhancement schemes are frequently employed to repopulate the high energy tail. In this work, the particles' weight in a given real space volume is periodically modified in order to obtain the same number of particles in all energy bins [Esseni 2011].

2.2.2 The Lucky Electron Model

The Lucky Electron Model (LEM), named after the pioneering work of Shockley [Shockley 1961], designates a family of *probabilistic* approaches aiming to model the substrate and the gate current. Shockley estimated the probability P for a carrier to travel a distance d without being subject to scattering, as:

$$P = \exp\left(-d/\lambda\right) = \exp\left(-\varepsilon/(qE\lambda)\right) \tag{2.15}$$

This equation can also be interpreted as the probability to gain an energy ε under the effect of a constant electric field E without scattering. Here, λ is the mean-free-path (MFP) between two consecutive interactions. As pointed out in [Shockley 1961], [Baraff 1964], this equation is valid under low-field conditions for electrons starting to accelerate from the bottom of the conduction band and consequently become *hot* electrons. The lucky electrons which gain an energy ε without scattering give rise to anisotropic distribution function at the considered energy as the carriers having scattered in the meanwhile are not included in the calculation [Jungemann 1996b]. In this 1D real space phenomenological approach, the band structure effects have been neglected.

During the '80, a LEM-based model for gate current (I_g) calculation in a 2D MOSFET structure has been proposed [Hu 1979], [Tam 1982], [Tam 1984], [Hu 1985]. It involves the probability for a carrier to gain enough kinetic energy while travelling from source to drain (P_1), the probability for the carrier to be redirected towards the Si/SiO_2 interface (P_2) and reach it without scattering (P_3) and the probability to overcome the oxide scattering-free (P_4). The succession of these four probabilistic events is depicted in Figure 2.4.

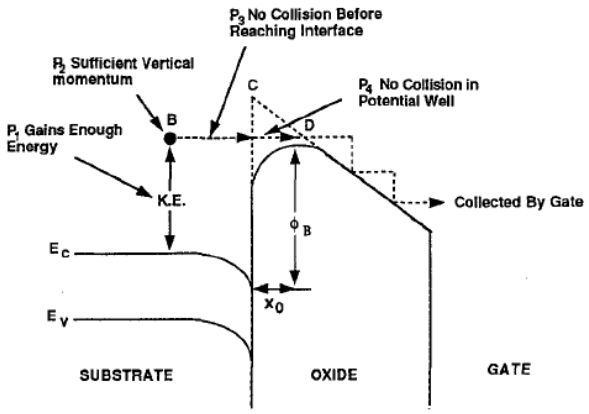


Figure 2.4: The sequence of four probabilistic events (P_1 to P_4) considered for electron injection into the gate [Tam 1984],[Hasnat 1996].

P_1, P_3 and P_4 are directly given from Equation 2.15, with λ representing the MFP of the inelastic collisions in the silicon for P_1 and P_3, and in the Oxide for P_4. P_2 bears instead the projection of the momentum in the normal-to-the-interface direction after a momentum-redirecting elastic collision:

$$P_2 = \frac{1}{2\lambda_r} \cdot \left(1 - \sqrt{\frac{\Phi_B}{\varepsilon}}\right) \tag{2.16}$$

with λ_r and Φ_B being the MFP of the elastic collision and the Si/SiO_2 barrier height as shown on Figure 2.4. The gate current is then given by:

$$I_g = \iint\limits_{L,W} \mathrm{d}x\,\mathrm{d}y \int\limits_{\Phi_B}^{\infty} J_n(x,y) P_1 P_2 P_3 P_4 \,\mathrm{d}\varepsilon \tag{2.17}$$

In this equation, $J_n(x,y)$ represents the channel current density at a given (x,y) position, while L and W are the device length and width, respectively. The lower integration bound, Φ_B, accounts for the classical image-force lowering effect and neglects the tunneling processes.

The merit of this expression is to allow for a rapid and efficient calculation of the gate current. However, its evaluation is based on constant MFPs which have been extracted under different 1D transport setups [Bartelink 1963], [Baraff 1964], [Crowell 1966], [Verwey 1975], [Ning 1977], [Cottrell 1979] using Equation 2.15. As the energy dependence of the mean free path is not accounted for [Goldsman 1990], it has been merely considered as a fitting parameter [Fischetti 1995]. An analogy between Equation 2.15 and the heated Maxwellian distribution was then proposed [Goldsman 1990], based on their similar exponential dependence. Furthermore, Equation 2.17 makes use of local values of the lateral electric field which implies that the carriers are in equilibrium with the field. In the present simulators [Synopsys 2010], the LEM-version by [Hasnat 1996] has been retained and

implemented.

However, there have been alternative LEM-based approaches. For instance, Meinerzhagen [Meinerzhagen 1988] applied the Shockley's expression by assuming that carriers move along the field lines which potential drop determines the total available kinetic energy. Troutman's approach [Troutman 1978] on the other hand, introduced an original way of treating the scattered carriers which can still contribute to the gate current contrarily to the retained LEM. However, all the considered approaches make use of the local field, potential or mean energy and do not explicitly consider the carrier history.

2.2.3 The Fiegna Model

The Fiegna Model (FM)[Fiegna 1991] is an analytical approach to model the hot carrier injection into the gate. The model has been proposed following an analytical solution of the BTE obtained under homogeneous conditions [Cassi 1990]. Cassi and Ricco derived a closed-form expression for the probability function after neglecting the diffusion term in the stationary BTE expression and considering the emission of inelastic optical phonons as the only energy-loss mechanism.

Introducing a new non-parabolic dispersion relation,

$$\frac{\hbar^2 k^2}{2m} = a\varepsilon^b \tag{2.18}$$

where a and b are adjusted to best match the Kane's non-parabolic expression in different energy ranges, the obtained probability function finally obtained is:

$$f(\varepsilon) \propto \exp\left(-\kappa \frac{\varepsilon^3}{E^{1.5}}\right) \tag{2.19}$$

In this expression, the electric field (E) dependence (the 1.5 exponent) is adjusted after MC simulations in a homogeneous silicon slab, while κ accounts for the strength of the optical phonons and the band-structure effects. For practical purposes, κ is considered as a fitting parameter as already shown in [Fischetti 1995], [Zaka 2010] and used in [Zaka 2011].

Using the above probability function, Fiegna [Fiegna 1991] proposed to calculate the gate current as:

$$I_g = q \iint_{L,W} \mathrm{d}x\,\mathrm{d}y \int_{\Phi_B}^{\infty} f(\varepsilon) g(\varepsilon) v_\perp(\varepsilon)\,\mathrm{d}\varepsilon \tag{2.20}$$

where g and $v_\perp$ are the density of states and the normal-to-the-interface velocity, respectively. Both quantities are readily derived from the dispersion relation given in Equation 2.18. The image-force lowering effect has been considered in the Φ_B value. Similarly to the LEM approach (subsection 2.2.2), the gate current features a dependence on the local value of the electric field introduced by Equation 2.19.

The analytic solution of the BTE (Equation 2.1) has been the object of many works in the previous decades. These attempts often introduce approximations which restrict the field of applicability of the obtained probability function. In the same spirit of this approach, many authors have proposed similar [Goldsman 1988], [Hasnat 1997] and generalized [Sonoda 1996],[Grasser 2002] analytic closed-form expressions for the probability function. A short review of these methods can be found in [Grasser 2002]. The dependence on the effective field or the mean carrier energy in each method inevitably introduces other fitting parameters.

2.2.4 The Spherical Harmonics Expansion method

An alternative method for the solution of the BTE involves its projection in a spherical harmonics basis which reduces the dimensionality of the problem. The spherical and orthogonal harmonics form a complete set of normalized functions [Arfken 2005] thus enabling to expand the probability function for a constant wave-vector modulus:

$$f\left(\vec{r},\vec{k},t\right)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}f_l^m\left(\vec{r},k,t\right)Y_l^m(\theta,\phi) \tag{2.21}$$

The spherical harmonics $Y_l^m(\theta,\phi)$ of degree l and order m express the angular dependence of the distribution function in the momentum space via θ and ϕ. The objective of the approach is to find the coefficients $f_l^m\left(\vec{r},k,t\right)$. This is achieved by projecting the BTE in each of the basis functions resulting in a set of coupled differential equations [Ventura 1992], [Hennacy 1995]. The infinite set of equations is finally truncated at a given order.

One of the first attempts to use this approach dates back to the '60 [Baraff 1964], where the author used the Legendre polynomials, truncated at the 1^{st} order, as a special case of spherical harmonics involving a single angle dependence. The projection was generalized by Hennacy for both Legendre and spherical harmonics functions [Hennacy 1993], [Hennacy 1995] for an infinite number of functions and then applied to homogeneous silicon material (diffusion term neglected). The first application to a realistic 2D MOSFET device was performed using a 1^{st} order truncation to calculate the probability function along the channel [Ventura 1992], [Gnudi 1993]. An important requirement of this projection is the spherical symmetry of the dispersion relation. In fact, the above-cited references have used a many-band isotropic dispersion relation [Brunetti 1989], composed of parabolic and non-parabolic branches.

Vecchi [Vecchi 1998] proposed a Full-Band version of the approach by incorporating the DOS and the velocity calculated after the full-band description of silicon. This was indeed possible as the truncation at the 1^{st} order generates a second-order differential equation where both the DOS and the velocity explicitly appear. This is the version of the approach implemented in [Synopsys 2010]. The treatment of the collision operator in the right-hand side of the BTE is also facilitated by this choice as the major scattering mechanisms (phonon, impact ionization) are considered isotropic (no angle-dependence).

The main application of this approach was to have an estimation of the hot carrier population which can contribute to the gate current. This was done as a post-processing step after solving a classical macroscopic transport model (DD or HD). Assuming an isotropic distribution function, the gate current is thus given by [Jin 2009]:

$$I_g = -\frac{qg_v}{2} \iint\limits_{L,W} \left[\int\limits_0^{\infty} f(\varepsilon) g(\varepsilon) v(\varepsilon) \int\limits_0^1 T\left(\varepsilon - \frac{h^3 g(\varepsilon) v(\varepsilon) z}{8\pi m_{ins}}\right) \mathrm{d}z \, \mathrm{d}\varepsilon \right] \mathrm{d}x \, \mathrm{d}y \quad (2.22)$$

with g, v and T being the DOS, the group velocity and the tunneling probability while the g_v represents the valley degeneracy factor.

The self-consistent solution of the BTE via this approach has gained momentum in the very recent years due to the increase of the available computer memory. Nevertheless, several improvements are still necessary for this method to be comparable with the well-established MC approach. The necessity to account for direction-dependent effects has led to consider anisotropic multi-valley bands [Hong 2010], or directly the silicon full-band [Jin 2011]. Furthermore, it was shown that the 1^{st} order truncation is not accurate enough when the carriers transport becomes quasi ballistic. As the anisotropy increases, more SHE terms are needed [Jungemann 2006]. Therefore a generalization of the Vecchi's approach has been recently proposed by including the full-band structure and high-order terms [Jin 2011].

2.3 Models benchmarking

In this section, the ability of the previously presented models to reproduce hot carrier distributions is investigated. Subsection 2.3.1 presents such an evaluation in a uniform structure, corresponding to the bulk material case, while subsection 2.3.2 compares the models in realistic device conditions looking at different figures of merit. Finally, subsection 2.3.3 draws a summary of the models and highlights the main ingredients for a proper transport description.

2.3.1 Homogeneous Case

In order to ensure a fair comparison between the approaches, the probability function $f(\varepsilon, \vec{r})$ will be used for comparison instead of the distribution function $n(\varepsilon, \vec{r})$. This avoids additional discrepancies coming from the density of states. Furthermore, as the approaches have different degrees of approximations, the probabilities will be plotted normalized to their integral over energy. Figure 2.5 compares the probability functions obtained with the LEM [Hasnat 1996], FM [Fiegna 1991], Full-Band SHE [Synopsys 2010] and Full-Band MC [Palestri 2006], for different values of the electric field. SHE and MC simulations are performed for a carrier concentration of $10^{16} cm^{-3}$. The LEM and FM results have been obtained by setting the field value in Equations 2.15 and 2.19 to 10^4, 10^5 and $3 \cdot 10^5$ $V \cdot cm^{-1}$. MC simulations are

performed on a uniform $1-\mu m$ slab with periodic boundary conditions implying that the carrier distribution reaching the right electrode is re-injected at the left one. The result is a homogeneous carrier distribution throughout the slab. Instead, SHE simulations are performed on a uniform $10-\mu m$ slab where equilibrium distribution is imposed at the contacts (the carriers are thermalized). Although inhomogeneous distributions are obtained throughout the slab, for a sufficiently long distance the carriers will reach the equilibrium with the electric field. In this region, uniform transport conditions hold, thus leading to results comparable directly to those of the other methods.

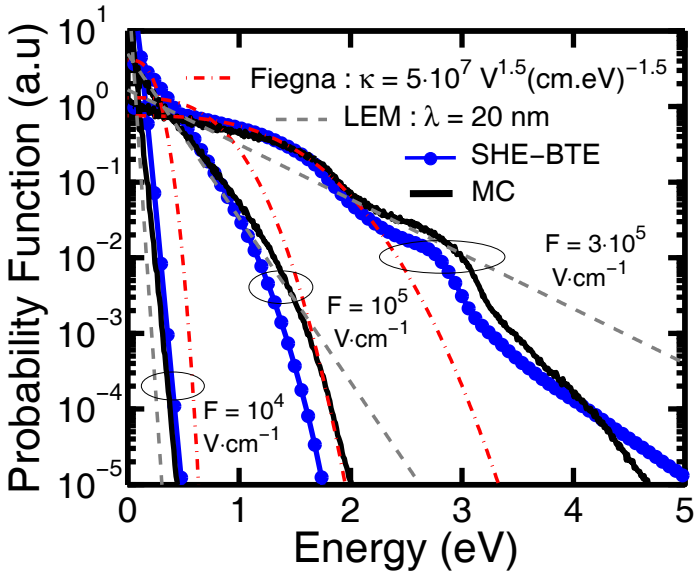


Figure 2.5: Probability functions simulated under different uniform electric fields with the full-band Monte Carlo method (MC), the 1^{st} order full-band Spherical Harmonics Expansion of the Boltzmann Transport Equation method (SHE-BTE), the Lucky Electron Model (LEM) and the Fiegna model. MC and SHE-BTE approaches include phonon scattering and impact ionization.

MC results of Figure 2.5 show that when the field increases from $10^4 V \cdot cm^{-1}$ to $3.10^5 V \cdot cm^{-1}$, the shape of the probability function becomes non-Maxwellian. The LEM (Equation 2.15) has been adjusted to best capture the MC distributions setting $\lambda = 20nm$. However, as the LEM features heated Maxwellian distributions due to a constant mean free path assumption [Hasnat 1996], it cannot reproduce the MC distributions at high fields relevant for carrier injection. Instead, the FM, which intrinsically shows a non-Maxwellian behaviour, much better agrees with the MC results after adjusting the model parameter κ. The best agreement has been obtained for $\kappa = 5 \cdot 10^7 V^{1.5} (cm \cdot eV)^{-1.5}$. It should be noticed that differently from the MC or the SHE method, the inelastic phonon scattering and the impact ionization processes are *virtually* included in a single fitting parameter for each of the models.

The best agreement with the MC has been reached by the SHE, confirming the results obtained by [Jin 2009]. The reasons for such an agreement rely on the incorporation of the full-band structure and on an accurate description of the scattering mechanisms. Among the inelastic ones, SHE includes a single optical phonon with an energy taken equal to 60 meV featuring the highest deformation potential D of Table 2.1. This ensures that most of the inelastic phonon scatterings are being accounted for. Furthermore, it includes an isotropic impact ionization process extracted after comparison with measurements [Jungemann 2003]. The scattering rates of these mechanisms are given in Figure 2.6. The optical phonons rates are calculated after equations 2.11 and 2.12, while the impact ionization rates for MC are taken from [Bude 1992]. Figure 2.6 shows that very similar rates are used in both simulators which justifies the results of Figure 2.5. In addition to these mechanisms, electron-electron scattering (EES) should be further considered as it plays an important role especially in device operation [Ghetti 1996]. This mechanism is presently included only in the MC approach. Figure 2.7 reports the effect of EES on the probability function for different electric fields and doping concentrations. An increase of the high-energy tail due to the energy transfer between electrons is observed at $10^5\ V \cdot cm^{-1}$. The tail increases with rising doping concentration, confirming the findings in [Childs 1996]. The EES effect is however not visible neither at low electric fields ($10^4\ V \cdot cm^{-1}$) where electrons are at equilibrium and no excess energy exchange occurs between them, nor for very high fields ($10^6\ V \cdot cm^{-1}$) where the important acceleration washes out the EES effect at high energies once the carriers reach the equilibrium with the electric field. This observations is valid for all the investigated free carrier concentrations.

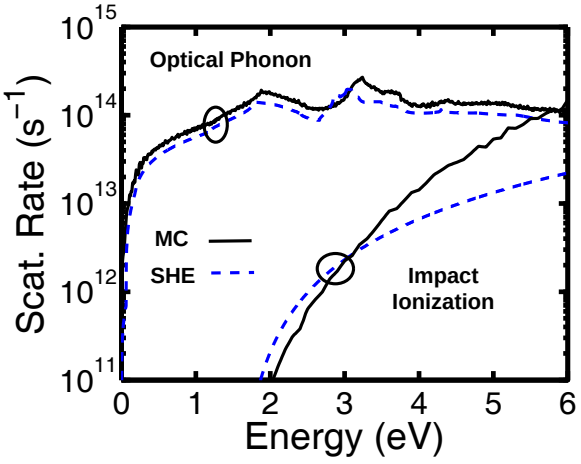


Figure 2.6: Optical phonon and impact ionization scattering rates used in the Full-Band Monte Carlo (lines, black) and Spherical Harmonics Expansion (dashed, blue).

In order to better visualize the effect of EES, instead of using periodic boundary conditions in the MC, thermalized carriers have been injected at the left side of the

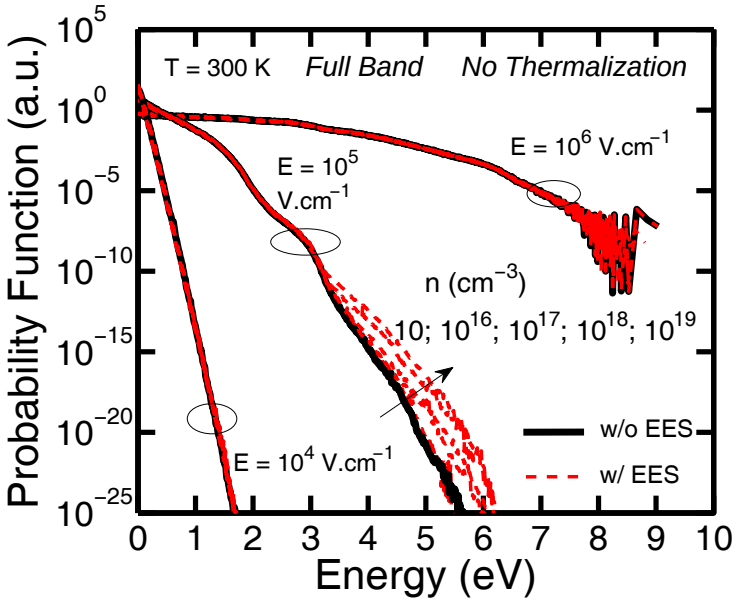


Figure 2.7: Occupation probability functions obtained with the full-band Monte Carlo under different uniform field conditions and doping concentrations. Solid lines feature phonon scattering and impact ionization processes while dashed curves additionally integrate electron-electron scatterings (EES) for different electron concentrations. Coulomb scatterings with ionized impurities ($N_D = n$) have been also included.

slab and then let be accelerated by the constant field. Such a setup allows to follow the establishment of equilibrium condition with the electric field. Such boundary conditions have been already used in literature [Childs 1996], [Abramo 1996]. Figure 2.8 reports the EES effect at two different positions of the slab.

At the beginning of the slab where the carriers are in out-of-equilibrium conditions, the effect of the EES is highly visible. But as the carriers move in the slab towards the other electrode under a constant field, they tend to reach the equilibrium condition: the result obtained near the end of the slab and the ones obtained under periodic boundary conditions are essentially the same. At this point the effect of EES has greatly diminished. Furthermore, after traveling this distance, the energy tail of the occupation probabilities becomes highly non-Maxwellian, as already shown on Figure 2.5. Further details on out-of-equilibrium distributions are given in the next subsection.

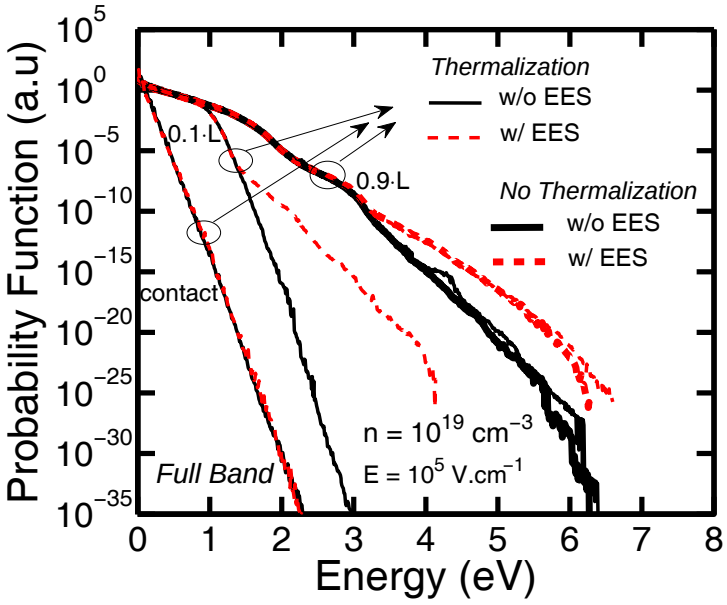


Figure 2.8: Probability functions obtained with the full-band Monte Carlo in a homogeneous slab of length $L = 1\mu m$ featuring a constant electric field of $10^5\ V \cdot cm^{-1}$. The distributions at the contact, at 0.1 $\mu m \cdot L$ and 0.9 $\mu m \cdot L$ have been reported when thermalizing contacts are used (label:*Thermalization*). The results of periodic boundary conditions under the same field/doping conditions are also reported (label:*No Thermalization*) for comparison.

2.3.2 Non-homogeneous Case

As pointed out by the literature [Fischetti 1995], [Hasnat 1996], [Zaka 2010]the LEM and FM are based on the assumption of local equilibrium of the carrier distribution with the electric field. Therefore, employing the LEM or FM in non-local conditions, e.g. for short device with rapidly varying field, is highly questionable. In this section, the ability of the FM and SHE to model hot carrier injection is thus evaluated in non-homogeneous conditions in terms of distributions (paragraph 2.3.2.1) and gate currents (paragraph 2.3.2.2).

2.3.2.1 Distributions and non-local correction

Two advanced eNVM technologies (Figure 2.9) have been used for this evaluation. The first (resp. second) technology features a 180 nm (resp. 140 nm) gate length (L_g), a 125 nm (resp. 100 nm) effective length (L_{eff}) and a 9.5 nm tunnel oxide thickness (T_{ox}). The considered devices also feature different doping profiles. The critical zone where injection occurs is located around the channel/LDD junction. Therefore our analysis will be focused on this area by choosing a position before the channel/Drain junction (Figure 2.9) near the Si/SiO_2 interface.

Figure 2.10 shows the probability functions obtained by Monte Carlo simulation

Devices	Dimensions (nm)		
	L_g	L_{eff}	T_{ox}
Flash (1)	180	125	9.5
Flash (2)	140	100	9.5

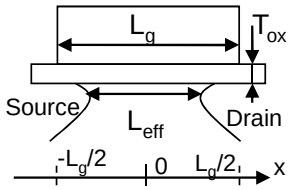


Figure 2.9: Main geometrical parameters of the simulated devices and their schematic drawing.

and the one predicted by the FM using the local field value of 500 kV/cm which is the electric field at the considered position in the simulated short channel device.

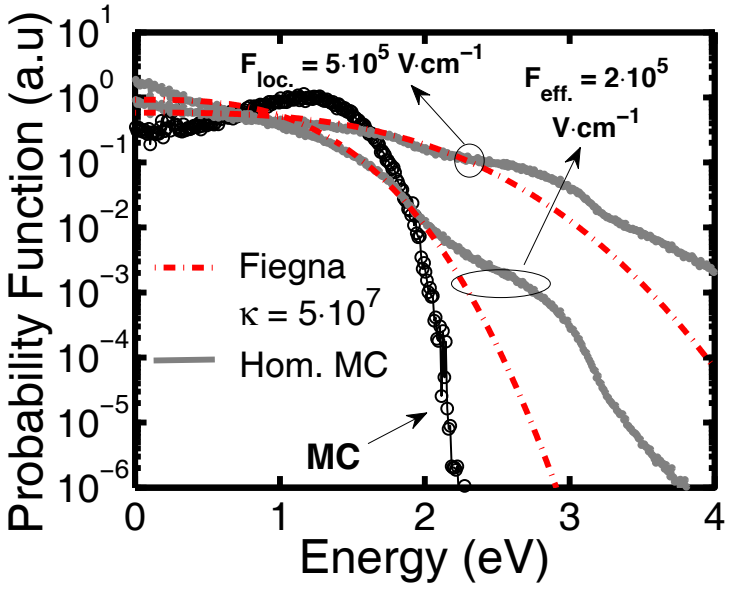


Figure 2.10: Normalized probability functions obtained with Monte Carlo (MC) simulation for the Flash(1) device at 20 nm from the drain. Homogeneous MC (Hom.MC) and FM distributions are also reported for the local and effective field values, respectively 500 kV/cm and 200 kV/cm.LEM has been omitted as the agreement with the homogeneous case is weak. The drain voltage V_d is 4.2 V and the floating gate voltage V_{fg} is 4.7 V. A similar plot has been obtained for Flash (2).

It can be seen in Figure 2.10 that there is no agreement between the FM and the MC (device condition) occupation probability function if the local field is used in the FM. This clearly indicates that the FM, which is a local model, fails to reproduce the non-locality of carrier transport consistently with the findings in [Fischetti 1995]. In order to keep using the same formalism, the local field is replaced by an *effective field* aiming to capture the non-local effects at 1^{st} order. This methodology involves results obtained under homogeneous conditions and is organized as follows:

1. The mean energy of the carriers (hereby referred to as inhomogeneous mean

energy) is calculated as a function of the position in the device, either from the MC distribution functions, the full HD transport model [Blotekjaer 1970] or a simplified HD model as the one proposed by Cook [Cook 1982].

2. From homogeneous MC simulations, the mean energy (hereby referred as homogeneous mean energy) is calculated as a function of the applied constant field either analytically or numerically, as shown in Figure 2.11.

3. For a given position in the device (therefore a given mean energy), the effective field is defined as the constant field simulated in step 2 which would lead to a homogeneous mean energy equal to the non-homogeneous one at that position.

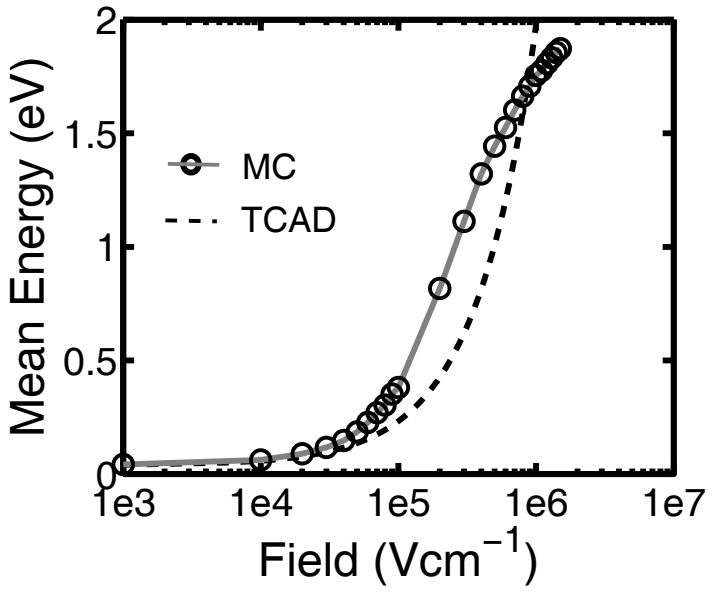
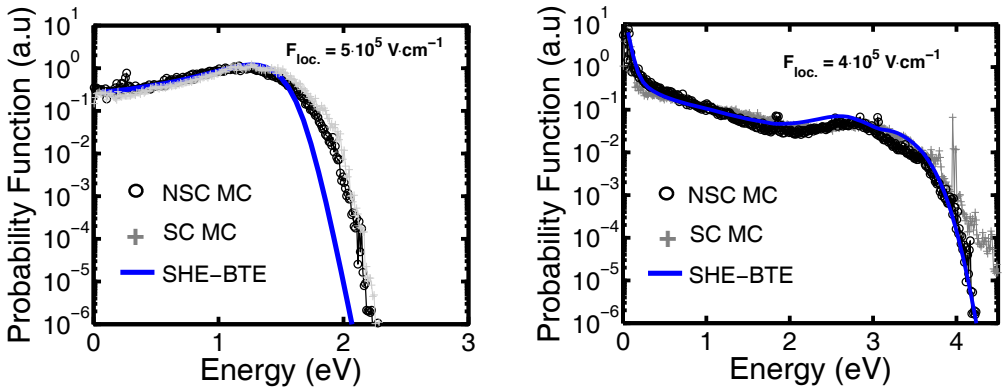


Figure 2.11: Carrier mean energy as a function of the applied (uniform) field obtained from the Full-Band MC simulator under uniform conditions [Palestri 2006] or from the analytical expression from the steady-state hydrodynamic model available in TCAD [Agostinelli 1994], [Synopsys 2010].

The validity of the effective field approach has been assessed by recalculating the probability function obtained with the FM using the procedure described above (leading to a $F_{eff} = 200$ kV/cm instead of 500 kV/cm in this case). The results displayed in Figure 2.10 show that even if the effective field procedure indeed reduces the difference between FM and device MC, the hot energy tail of the FM still overestimates the MC one. To investigate the origin of this discrepancy, the probability function obtained by MC simulation in the homogeneous case (hereby denoted as $Hom.MC$), using constant local field of 500 kV/cm and 200 kV/cm (effective field case) are also shown in Figure 2.10. It can be seen that even the tail of the homogeneous MC probability function in the effective field condition overestimates the

tail of the MC device (inhomogeneous) probability function. This difference illustrates the failure of the effective field correction procedure, and can be explained as follows. In homogeneous conditions, only scatterings limit the electron heating. On the contrary in the short device case, carrier heating is also limited by the finite voltage drop. This explains the sharp high energy tail in the inhomogeneous case. Thus, models based on local assumption of carrier transport cannot reproduce inhomogeneous probability functions, even when the effective field correction is applied.

Differently from the local models, the comparison between the MC and SHE probability functions, reported in Figure 2.12, shows that thanks to its intrinsically non-local nature, the SHE captures well the high energy tail of the distribution function. These conclusions have been confirmed by performing the comparison at two different positions along the channel, before and after the channel/drain junction. In addition, Figures 2.12a and 2.12b report both SC and NSC MC simulations. Although only NSC configuration should be used when comparing MC to the other models, it is interesting to notice that self consistency only slightly changes the probability functions for the considered cells and biases, in agreement with previous works [Jungemann 1996b], [Bude 2000], [Jungemann 2003].

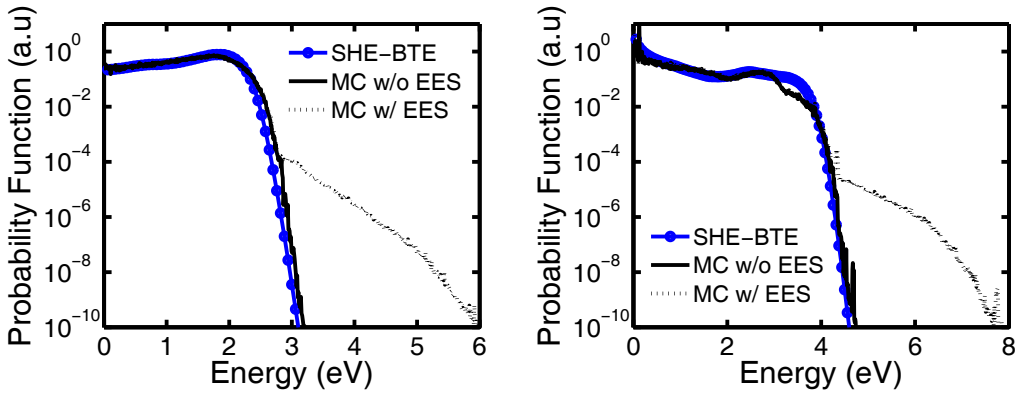


(a) (b)

Figure 2.12: Occupation probability functions obtained with full-band Monte Carlo (MC) and Spherical Harmonics Expansion (SHE BTE) for the Flash (1) device at 20 nm before the drain junction *(a)* and at 10 nm after the drain junction *(b)*. MC results are given for both Self Consistent (SC) and Non Self Consistent (NSC) cases. The curves have been normalized to give the same carrier concentration. The drain voltage V_d is 4.2 V and the floating gate voltage V_{fg} is 4.7 V.

However, as already pointed out in 2.3.1, the SHE approach does not include EES. The comparison with MC simulations in Figure 2.13 including such an interaction shows the enhanced high energy tail due to EES. As already shown in 2.3.1, the effect is particularly important in inhomogeneous out-of-equilibrium conditions, such as the ones inside the device under high drain voltage. The increased popu-

lation of the high energies will contribute to enhance the gate current in the bias configurations where either a repulsive vertical field is present ($V_{fg} < V_d$) or when the drain voltage is about the same as the Si/SiO_2 barrier height.



(a) (b)

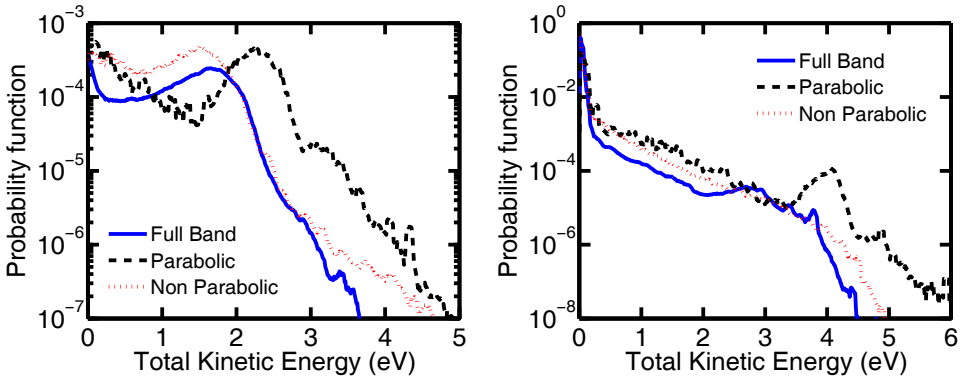
Figure 2.13: Occupation probability functions obtained with full-band Monte Carlo (MC) and Spherical Harmonics Expansion (SHE BTE) for the Flash (2) device at 15 nm before *(a)* and 10 nm after *(b)* the drain junction. MC results are given with and without the inclusion of Electron-Electron Scattering (EES). The curves have been normalized to give the same carrier concentration. The drain voltage V_d is 4.2 V and the floating gate voltage V_{fg} is 4.5 V.

Finally, the impact of the band-structure is studied using the MC simulator. Figure 2.14 shows the probability functions obtained after parabolic, non-parabolic and full band description (c.f. subsection 2.1.2). Notice the enhanced hot carrier population predicted by parabolic bands for both channel positions. Instead, the non parabolic ones predict much closer results to the full band case. The results on the gate current are shown in the next paragraph.

2.3.2.2 Gate current

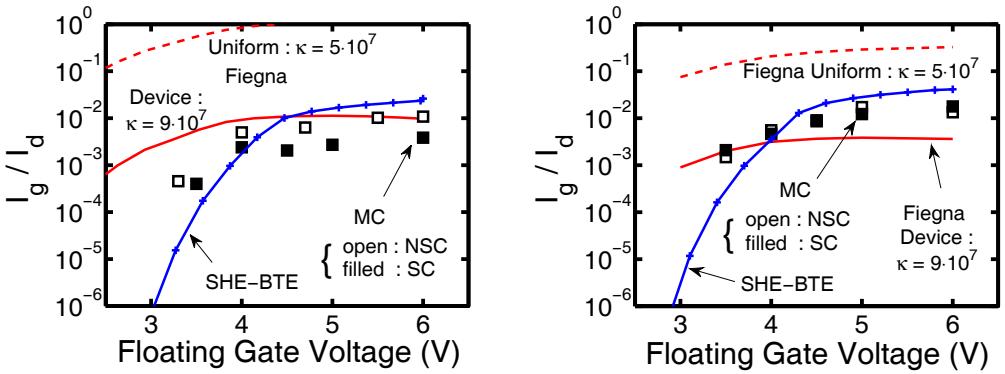
The accuracy of the FM and SHE models is now evaluated in terms of gate current. To this aim, the unit-less gate to drain current ratio, hereafter called *the injection efficiency*, is chosen as the indicator expressing the intrinsic properties of the cell. The obtained results with FM, SHE and MC are compared in Figure 2.15 as a function of the floating gate voltage.

SC and NSC MC simulations are reported. The results differ by less than a factor of ≈ 2 (consistently with the very similar distributions for NSC and SC simulations shown in Figure 2.12), thus showing little impact of self consistency on our cells for the studied injection conditions. For the considered devices, which have different doping profiles and junctions depths, self-consistency seems slightly more important for the longest cell. However only NSC should be considered for comparison with FM and SHE. The results show that the current ratio calculated



(a) (b)

Figure 2.14: Occupation probability function vs. kinetic energy obtained after self consistent Monte Carlo simulations using a full band, a parabolic and a non parabolic dispersion relation. The results are plotted for Positions 1 *(a)* and 2 *(b)* of Figure 2.9 for the Flash (2) device at $V_d = 4.2$ V and $V_{fg} = 5$ V.



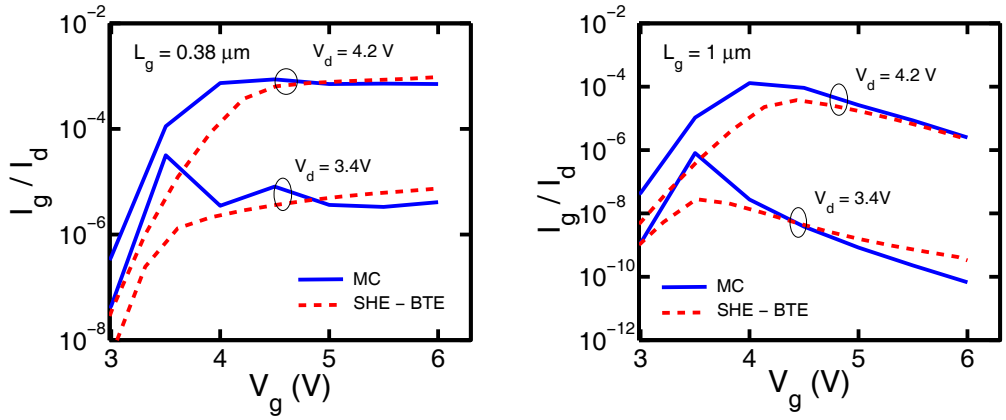
(a) (b)

Figure 2.15: Injection efficiency as a function of the floating gate voltage obtained by Monte Carlo (MC) simulations, the Fiegna Model and the Spherical Harmonics Expansion method, at $V_d = 4.2$ V for Flash (1) and Flash (2) device (resp. 2.15a and 2.15b). MC results are given for both Self Consistent (SC) and Non-Self Consistent (NSC) cases. The Fiegna's κ-value giving a better agreement in terms of injection effciency is different with respect to the one used in Figures 2.5, 2.12 ($\kappa = 9 \cdot 10^7 V^{1.5} (cm \cdot eV)^{-1.5}$ in device conditions, $\kappa = 5 \cdot 10^7 V^{1.5} (cm \cdot eV)^{-1.5}$ for homogeneous ones).

with the FM, calibrated to reproduce MC simulation in the homogeneous case (i.e. $\kappa = 5 \cdot 10^7 V^{1.5} (cm \cdot eV)^{-1.5}$, see Figure 2.5), does not reproduce the MC results. However, a better agreement can be obtained by refitting the model with $\kappa =$

$9 \cdot 10^7 m^{3/2} eV^{-3/2}$ (Figure 2.15a), illustrating the flexibility of the FM approach. Such flexibility is however affecting the results for a shorter device (Figure 2.15b), no longer perfectly matching MC results. Therefore a technology dependent calibration is needed for this model.

Regarding the SHE results, Figure 2.15 shows an overall good agreement with the MC simulations at the investigates biases. The MC and SHE simulations have been carried out including the image force effect and the parallel momentum conservation. Furthermore, in both approaches the same values for the energy barrier and the tunneling masses are used. A further comparison between both approaches is reported on Figure 2.16. Thus, despite the isotropic assumption made in the SHE method concerning the distribution function, contrarily to the MC approach, a good agreement has been obtained as a function of the gate length and drain bias without any adjustment procedure (Figure 2.16).

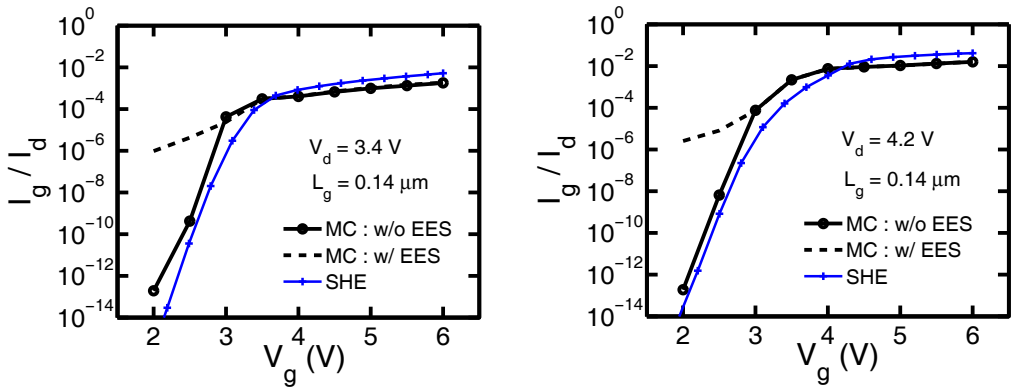


(a) (b)

Figure 2.16: Injection efficiency as a function of the floating gate voltage obtained by non-self consistent Monte Carlo (MC) simulations and the Spherical Harmonics Expansion (SHE-BTE) method, at $V_d = 4.2$ V and 3.4 V for devices featuring $0.38\mu m$ and $1\mu m$ gate length (resp. 2.16a and 2.16b).

However, at low gate voltages ($V_g < V_d$) the agreement with MC is no longer satisfactory (Figure 2.17). In this operating regime, the vertical field becomes repulsive for electrons and only those located in the high energy tail having a sufficient energy can overcome the barrier. As already shown on Figures 2.8 and 2.13, such a tail is highly enhanced by EES, which consequently increases the injection efficiency. A similar situation is found for drain voltages lower than the Si/SiO_2 barrier. This comparison shows that the SHE method lacks an important ingredient (i.e. EES) at low voltages.

In addition, Figure 2.18 reports the impact of barrier lowering and parallel momentum conservation effects on the injection efficiency. The barrier lowering (dashed curve) increases the injection efficiency particularly in the low voltage regime where



(a) (b)

Figure 2.17: Injection efficiency as a function of the floating gate voltage obtained by non-self consistent Monte Carlo (MC) simulations and the Spherical Harmonics Expansion (SHE) method for two drain voltages V_d=3.4 and 4.2 V, respectively reported in *a* and *b*. MC simulations with and without Electron-Electron Scattering (EES) have been shown, while SHE-BTE does not feature EES.

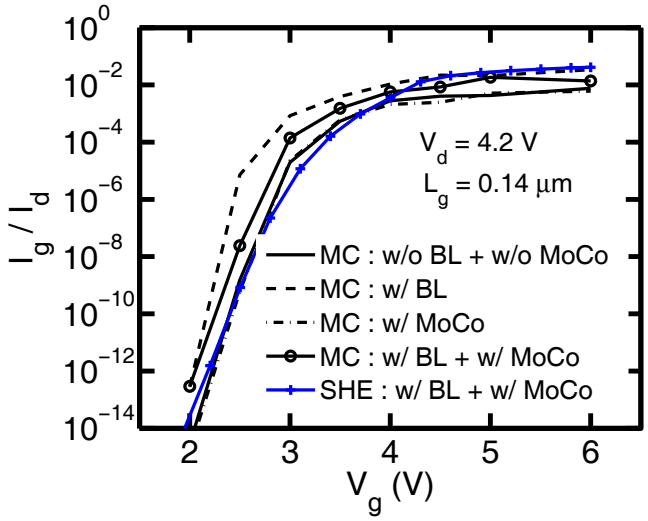


Figure 2.18: Injection efficiency as a function of the floating gate voltage obtained by non-self consistent Monte Carlo (MC) simulations and the Spherical Harmonics Expansion (SHE-BTE) method. MC simulations with and without Barrier Lowering (BL) and parallel Momentum Conservation (MoCo) are shown, while SHE-BTE simulation includes both of them.

any barrier modification leads to significant differences due to the exponential high-energy tail. When momentum conservation is not enforced (MC), the perpendicular energy is estimated by [Bufler 2005]:

$$\varepsilon_{\perp} = \varepsilon_{tot} \cdot \frac{k_{\perp}^2}{k_{tot}^2} \tag{2.23}$$

where $k_{\perp}$ and k_{tot} are respectively the perpendicular and the total wave-vector momentum of the particle at the interface, referred to the closest Δ valley minimum. Comparison with the case in which parallel momentum has been enforced (dot-dashed curve) shows that the injection efficiency is only slightly impacted. Both effects are summed for the curves featuring symbols.

In addition to the above-mentioned ingredients the impact of the band structure on the injection efficiency is also investigated. Figure 2.19 reports such ratio obtained in the case of full-band, parabolic and non-parabolic bands at $V_d = 4.2V$. A visible increase of the injection in the parabolic case is shown in agreement with the enhanced carrier population at high energies (c.f. Figure 2.14). Closer results are reported in the case of non-parabolic bands.

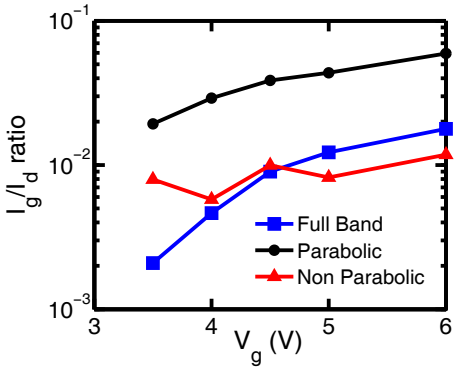
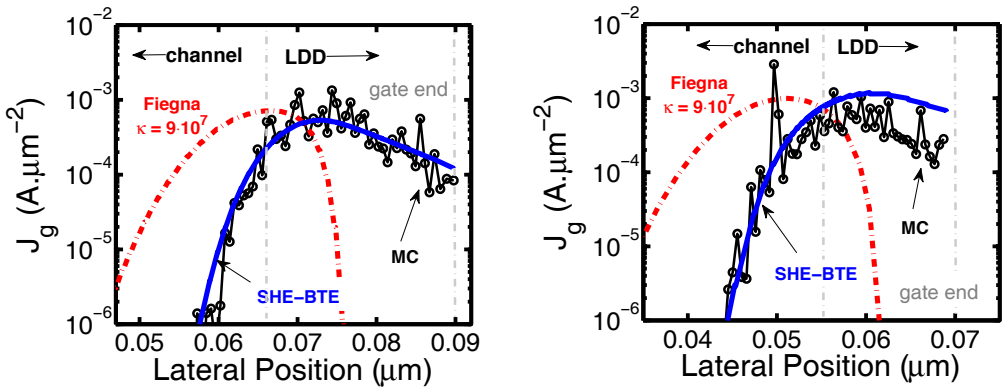


Figure 2.19: Injection efficiency vs. floating gate voltage obtained after self consistent Monte Carlo simulations using a full band, a parabolic and a non parabolic dispersion relation for Flash (2) device at V_d = 4.2 V.

Finally, the gate current density along the channel is also an important figure of merit as it is relevant for MOSFET and Flash memory degradation models [Doyle 1997], [La Rosa 2007]. Figure 2.20 therefore plots this quantity, obtained by MC, SHE and FM for both Flash devices (see Figure 2.9). In Figure 2.20a it can be seen that the non-local FM does not reproduce the sharp injection peak near the junction due to the rapidly varying field and mean energy decrease in the LDD. Therefore a higher (resp. lower) current density is predicted in the channel (resp. LDD), although the peak-value can still be calibrated. On the other hand, the SHE approach well captures the shape of the gate current density. The same trends have been obtained for the Flash (2) device, as shown on Figure 2.20b.



(a) (b)

Figure 2.20: Gate current density as a function of the position along the channel, obtained by Monte Carlo (MC) without EES, the Fiegna Model (with the new fitting obtained on Figure 2.15) and the Spherical Harmonics (SHE BTE) method for the Flash (2) and Flash (1) devices (resp. 2.20a and 2.20b) with $V_d = 4.2$ V and $V_{fg} = 5$ V.

2.3.3 Summary

Table 2.2 draws a condensed assessment of the general characteristics of the models described in section 2.2 in view of the comparisons performed in section 2.3.

Model	Band Structure	Scattering Mechanism	Transport Description	Real Dimension	CPU
Monte Carlo	Full Band	Ph./II/EES	Non-Local	2D	12h/bias
LEM	Parabolic	Ph./II: λ	Local	2D/3D	0
Fiegna	Non Parabolic	Ph.: κ	Local	2D/3D	0
SHE	Iso. Full Band	Ph./II	Non-Local	2D/3D	1h/I_gV_g

Table 2.2: Classification of the models under the main criteria discussed throughout this chapter. *Ph.*, *II*, *CCS* acronyms stand for Phonon, Impact Ionization and Carrier-Carrier Scattering processes.

The Monte Carlo approach is the most complete one as it accounts for non-local transport and proper treatment of the scattering mechanisms in a full band description. The computational time is its maim limitation, which also hampers its use in 3D structures. On the contrary, the computational time counts among the advantages of the Fiegna and Lucky Electron Model, which show instead weaknesses for a proper transport description due to their intrinsic local nature and the simplified treatment of the scattering. Finally, the SHE method contributes to bridge the gap between Monte Carlo and the local models (Table 2.2), as it combines most of the advantages of both approaches, in particular, the non-local description of the

transport including phonon scattering and impact ionization, within a reasonable simulation time. However, the lack of carrier-carrier scattering in the SHE method makes it hardly usable for low voltage operating regimes.

2.4 Conclusions

This chapter has introduced the hot electron injection models commonly employed to predict gate current in non-volatile memories. All the considered models fall within a semi-classical modeling framework whose main ingredients, i.e. the band structure and the scattering mechanisms, and the associated quantities have been introduced in the first section. The definitions therein, valid throughout this thesis, has allowed to present the four investigated models, namely: the Monte Carlo (MC) approach, the Lucky Electron Model (LEM), the Fiegna Model (FM) and the Spherical Harmonics Expansion (SHE) method. The condensed description of the second section has highlighted the important assumptions and the relevant relations employed in each of these models which altogether offer a large panel of choices for eventually solving the Boltzmann Transport Equation. Thus, from analytic deterministic (FM) and probabilistic (LEM) to numerical deterministic (SHE) and stochastic (MC) solutions, the choice will be made on the balance between the required degree of physical insight and computational burden.

The last section of this chapter compares the results obtained from the above models under homogeneous (constant field, infinite material) and non-homogeneous (varying field, device dimensions) conditions. Local models obtained under homogeneous conditions, such as FM and LEM, are commonly used for device simulation by including an effective field correction. The investigation of the correction procedure has shown that microscopic quantities, such as the distribution function and the gate current density along the channel, are still inaccurately reproduced with respect to the well-established MC reference. However, the flexibility of such models as well as their fast execution are clear assets which allow to easily adjust these models with a technology node in terms of gate current. In addition, the benchmarking procedure showed that the SHE and MC results were very close for all the fields, device lengths and bias configurations investigated in this work. The SHE method is thus a very interesting approach as it offers a good description of carrier transport within a limited simulation time. However, the SHE method does not include the Electron-Electron Scattering (EES) mechanism which hampers the use of the approach for low voltage operation conditions.

Finally, the comparison of these approaches has provided a valuable insight for hot carrier transport modeling. Indeed, the inclusion of the full band structure, of various energy-exchange mechanisms (such as inelastic phonons, impact ionization and EES) and of the carrier path in the channel, is mandatory. Such considerations are put in practice in the next chapter with the establishment of a semi-analytic model.

Bibliography

[Abramo 1996] 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, page 889, 1996. (Cited on pages 18, 27, 49, 66 and 87.)

[Agostinelli 1994] V.M. Agostinelli, T.J. Bordelon, Xiaolin W., K. Hasnat, C.-F. Yeap, D.B. Lemersal, A.F. Tasch and C.M. Maziar. *Two-dimensional energy-dependent models for the simulation of substrate current in submicron MOSFET's.* IEEE Transactions on Electron Devices, vol. 41, no. 10, pages 1784 –1795, oct 1994. (Cited on page 30.)

[Ando 1987] Y. Ando and T. Itoh. *Calculation of transmission tunneling current across arbitrary potential barriers.* Journal of Applied Physics, vol. 61, no. 4, pages 1497–1502, 1987. (Cited on page 19.)

[Arfken 2005] G.B. Arfken and H.J. Weber. *Mathematical Methods For Physicists 6th edn.* 2005. (Cited on page 23.)

[Baraff 1964] G.A. Baraff. *Maximum Anisotropy Approximation for Calculating Electron Distributions; Application to High Field Transport in Semiconductors.* Physical Review, vol. 133, no. 1A, pages A26–A33, Jan 1964. (Cited on pages 20, 21 and 23.)

[Bartelink 1963] D.J. Bartelink, J.L. Moll and N.I. Meyer. *Hot-Electron Emission From Shallow pn Junctions is Silicon.* Physical Review, vol. 130, no. 3, page 972, 1963. (Cited on page 21.)

[Blotekjaer 1970] K. Blotekjaer. *Transport equations for electrons in two-valley semiconductors.* IEEE Transactions on Electron Devices, vol. 17, no. 1, pages 38–47, 1970. (Cited on page 30.)

[Brunetti 1989] R. Brunetti, C. Jacoboni, F. Venturi, E. Sangiorgi and B. Ricco. *A many-band silicon model for hot-electron transport at high energies.* Solid-State Electronics, vol. 32, no. 12, pages 1663–1667, 1989. (Cited on page 23.)

[Bude 1992] J. Bude, K. Hess and G.J. Iafrate. *Impact ionization in semiconductors: Effects of high electric fields and high scattering rates.* Physical Review B, vol. 45, no. 19, page 10958, 1992. (Cited on pages 17, 18, 26, 59, 93 and 114.)

[Bude 2000] J.D. Bude, M.R. Pinto and R.K. Smith. *Monte Carlo simulation of the CHISEL flash memory cell.* IEEE Transactions on Electron Devices, vol. 47, no. 10, pages 1873–1881, 2000. (Cited on pages 6 and 31.)

[Bufler 2000] F.M. Bufler, A. Schenk and W. Fichtner. *Efficient Monte Carlo device modeling.* IEEE Transactions on Electron Devices, vol. 47, no. 10, pages 1891 –1897, oct 2000. (Cited on page 19.)

[Bufler 2005] F.M. Bufler and A. Schenk. *On the Tunneling Energy within the Full-Band Structure Approach.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2005, pages 155–158. IEEE, 2005. (Cited on pages 19, 20, 35 and 59.)

[Cardona 1966] M. Cardona and F.H. Pollak. *Energy-Band Structure of Germanium and Silicon: The k·p Method.* Physical Review, vol. 142, pages 530–543, Feb 1966. (Cited on page 13.)

[Cartier 1993] E. Cartier, M.V. Fischetti, E.A. Eklund and F.R. McFeely. *Impact ionization in silicon.* Applied Physics Letters, vol. 62, no. 25, pages 3339–3341, 1993. (Cited on pages 18, 93, 113 and 114.)

[Cassi 1990] D. Cassi and B. Ricco. *An analytical model of the energy distribution of hot electrons.* IEEE Transactions on Electron Devices, vol. 37, no. 6, pages 1514–1521, 1990. (Cited on page 22.)

[Chelikowsky 1976] J.R. Chelikowsky and M.L. Cohen. *Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors.* Physical Review B, vol. 14, pages 556–582, Jul 1976. (Cited on page 13.)

[Childs 1996] P.A. Childs and C.C.C. Leung. *A one-dimensional solution of the Boltzmann transport equation including electron–electron interactions.* Journal of Applied Physics, vol. 79, page 222, 1996. (Cited on pages 18, 26, 27 and 87.)

[Cook 1982] R.K. Cook and J. Frey. *Two-dimensional numerical simulation of energy transport effects in Si and GaAs MESFET's.* IEEE Transactions on Electron Devices, vol. 29, no. 6, pages 970–977, 1982. (Cited on page 30.)

[Cottrell 1979] P.E. Cottrell, R.R. Troutman and T.H. Ning. *Hot-electron emission in n-channel IGFETs.* IEEE Journal of Solid-State Circuits, vol. 14, no. 2, pages 442–455, 1979. (Cited on page 21.)

[Crowell 1966] C.R. Crowell and S.M. Sze. *Temperature dependence of avalanche multiplication in semiconductors.* Applied Physics Letters, vol. 9, no. 6, pages 242–244, 1966. (Cited on page 21.)

[Doyle 1997] B.S. Doyle, K.R. Mistry and J. Faricelli. *Examination of the time power law dependencies in hot carrier stressing of n-MOS transistors.* IEEE Electron Device Letters, vol. 18, no. 2, pages 51–53, 1997. (Cited on pages 36 and 148.)

[Esseni 2011] D. Esseni, P. Palestri and L. Selmi. Nanoscale MOS Transistors: Semi-Classical Transport and Applications. Cambridge Univ Pr, 2011. (Cited on pages 13, 15, 16, 19 and 20.)

[Ferry 1999] D.K. Ferry, S.M. Goodnick and K. Hess. *Energy exchange in single-particle electron-electron scattering.* Physica B: Condensed Matter, vol. 272, no. 1-4, pages 538–541, 1999. (Cited on pages 18, 85 and 87.)

[Fiegna 1991] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi and B. Ricco. *Simple and efficient modeling of EPROM writing.* IEEE Transactions on Electron Devices, vol. 38, no. 3, pages 603 –610, March 1991. (Cited on pages 6, 22, 24, 49 and 66.)

[Fischer 1997] B. Fischer, A. Ghetti, L. Selmi, R. Bet and E. Sangiorgi. *Bias and temperature dependence of homogeneous hot-electron injection from silicon into silicon dioxide at low voltages.* IEEE Transactions on Electron Devices, vol. 44, no. 2, pages 288–296, 1997. (Cited on page 18.)

[Fischetti 1988] M.V. Fischetti and S.E. Laux. *Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects.* Physical Review B, vol. 38, no. 14, pages 9721–9745, 1988. (Cited on pages 19 and 113.)

[Fischetti 1995] M.V. Fischetti, S.E. Laux and E. Crabbe. *Understanding hot electron transport in silicon devices: Is there a shortcut?* Journal of Applied Physics, vol. 78, no. 2, pages 1058 –1087, July 1995. (Cited on pages 6, 18, 19, 20, 21, 22, 28, 29, 93, 94, 113 and 114.)

[Fixel 2008] D.A. Fixel and W.N.G. Hitchon. *Kinetic investigation of electron–electron scattering in nanometer-scale metal-oxide-semiconductor field-effect transistors.* Semiconductor Science and Technology, vol. 23, page 035014, 2008. (Cited on pages 18 and 87.)

[Ghetti 1996] A. Ghetti, L. Selmi, R. Bez and E. Sangiorgi. *Monte Carlo simulation of low voltage hot carrier effects in non volatile memory cells.* In International Electron Devices Meeting (IEDM) 1996, pages 379–382. IEEE, 1996. (Cited on pages 18 and 26.)

[Ghetti 2002] A. Ghetti. *Explanation for the temperature dependence of the gate current in metal-oxide-semiconductor transistors.* Applied Physics Letters, vol. 80, page 1939, 2002. (Cited on pages 18, 87 and 90.)

[Gnudi 1993] 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, pages 575–581, 1993. (Cited on pages 23 and 49.)

[Goldsman 1988] N. Goldsman and J. Frey. *Electron energy distribution for calculation of gate leakage current in MOSFETs.* Solid-State Electronics, vol. 31, no. 6, pages 1089–1092, 1988. (Cited on pages 23 and 114.)

[Goldsman 1990] N. Goldsman, L. Henrickson and J. Frey. *Reconciliation of a hot-electron distribution function with the lucky electron-exponential model in silicon.* IEEE Electron Device Letters, vol. 11, no. 10, pages 472–474, 1990. (Cited on page 21.)

[Grasser 2002] 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, page 3869, 2002. (Cited on page 23.)

[Hasnat 1996] K. Hasnat and C. Yeap. *A pseudo-lucky electron model for simulation of electron gate current in submicron NMOSFET's.* IEEE Transactions on Electron Devices, vol. 43, no. 8, pages 1264–1273, 1996. (Cited on pages 21, 24, 25, 28, 48, 66, 84 and 114.)

[Hasnat 1997] K. Hasnat, C.F. Yeap, S. Jallepalli, S.A. Hareland, W.K. Shih, V.M. Agostinelli, A.F. Tasch and C.M. Maziar. *Thermionic emission model of electron gate current in submicron NMOSFETs.* IEEE Transactions on Electron Devices, vol. 44, no. 1, pages 129–138, 1997. (Cited on pages 6 and 23.)

[Hennacy 1993] K.A. Hennacy and N. Goldsman. *A generalized Legendre polynomial/sparse matrix approach for determining the distribution function in non-polar semiconductors.* Solid-State Electronics, vol. 36, no. 6, pages 869–877, 1993. (Cited on page 23.)

[Hennacy 1995] K.A. Hennacy, Y.J. Wu, N. Goldsman and I.D. Mayergoyz. *Deterministic MOSFET simulation using a generalized spherical harmonic expansion of the Boltzmann equation.* Solid-State Electronics, vol. 38, no. 8, pages 1485–1495, 1995. (Cited on page 23.)

[Hong 2010] S.M. Hong, G. Matz and C. Jungemann. *A deterministic Boltzmann equation solver based on a higher order spherical harmonics expansion with full-band effects.* IEEE Transactions on Electron Devices, vol. 57, no. 10, pages 2390–2397, 2010. (Cited on pages 6 and 24.)

[Hu 1979] C. Hu. *Lucky-electron model of channel hot electron emission.* In International Electron Devices Meeting (IEDM) 1979, volume 25, pages 22–25. IEEE, 1979. (Cited on page 20.)

[Hu 1985] C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan and K.W. Terrill. *Hot-Electron-Induced MOSFET Degradation – Model, Monitor, and Improvement.* IEEE Journal of Solid-State Circuits, vol. 20, no. 1, pages 295 – 305, February 1985. (Cited on pages 20, 147 and 148.)

[Jacoboni 1983] C. Jacoboni and L. Reggiani. *The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials.* Review of Modern Physics, vol. 55, no. 3, pages 645–705, Jul 1983. (Cited on pages 15, 16, 17, 19 and 113.)

[Jancu 1998] J.-M. Jancu, R. Scholz, F. Beltram and F. Bassani. *Empirical* spds* *tight-binding calculation for cubic semiconductors: General method and material parameters.* Physical Review B, vol. 57, pages 6493–6507, Mar 1998. (Cited on page 13.)

[Jin 2009] S. Jin, A. Wettstein, W. Choi, F.M. Bufler and E. Lyumkis. *Gate Current Calculations Using Spherical Harmonic Expansion of Boltzmann Equation.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2009, pages 1 –4, 2009. (Cited on pages 20, 24, 26, 60 and 155.)

[Jin 2011] S. Jin, S.M. Hong and C. Jungemann. *An Efficient Approach to Include Full-Band Effects in Deterministic Boltzmann Equation Solver Based on High-Order Spherical Harmonics Expansion.* IEEE Transactions on Electron Devices, no. 99, pages 1–8, 2011. (Cited on pages 6 and 24.)

[Jungemann 1996a] C. Jungemann, S. Keith, F.M. Bufler and B. Meinerzhagen. *Effects of band structure and phonon models on hot electron transport in silicon.* Electrical Engineering (Archiv fur Elektrotechnik), vol. 79, no. 2, pages 99–101, 1996. (Cited on page 18.)

[Jungemann 1996b] C. Jungemann, R. Thoma and W.L. Engl. *A soft threshold lucky electron model for efficient and accurate numerical device simulation.* Solid-State Electronics, vol. 39, no. 7, pages 1079–1086, 1996. (Cited on pages 20, 31, 79 and 111.)

[Jungemann 2003] C. Jungemann and B. Meinerzhagen. Hierarchical device simulation: The monte-carlo perspective. Springer Verlag, 2003. (Cited on pages 13, 19, 26, 31 and 93.)

[Jungemann 2006] C. Jungemann, A.T. Pham, B. Meinerzhagen, C. Ringhofer and M. Bollhofer. *Stable discretization of the Boltzmann equation based on spherical harmonics, box integration, and a maximum entropy dissipation principle.* Journal of Applied Physics, vol. 100, no. 2, pages 024502–024502, 2006. (Cited on page 24.)

[Kamakura 1994] Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo and M. Takenaka. *Impact ionization model for full band Monte Carlo simulation.* Journal of Applied Physics, vol. 75, no. 7, pages 3500–3506, 1994. (Cited on pages 18 and 93.)

[Kane 1957] E.O. Kane. *Band structure of indium antimonide.* Journal of Physics and Chemistry of Solids, vol. 1, no. 4, pages 249–261, 1957. (Cited on page 15.)

[Kane 1967] E.O. Kane. *Electron scattering by pair production in silicon.* Physical Review, vol. 159, no. 3, page 624, 1967. (Cited on page 18.)

[Kittel 1986] C. Kittel and P. McEuen. Introduction to solid state physics, volume 4. Wiley New York, 1986. (Cited on page 16.)

[La Rosa 2007] G. La Rosa and S.E. Rauch. *Channel hot carrier effects in n-MOSFET devices of advanced submicron CMOS technologies.* Microelectronics Reliability, vol. 47, no. 4-5, pages 552 – 558, 2007. 14th Workshop on Dielectrics in Microelectronics (WoDiM 2006). (Cited on pages 18, 36, 151, 152 and 153.)

[Lundstrom 2000] M. Lundstrom. Fundamentals of carrier transport. Cambridge Univ Pr, 2000. (Cited on pages 12, 15, 16, 18, 19 and 52.)

[Meinerzhagen 1988] B. Meinerzhagen. *Consistent gate and substrate current modeling based on energy transport and the lucky electron concept.* In International Electron Devices Meeting (IEDM) 1988, pages 504–507. IEEE, 1988. (Cited on page 22.)

[Ning 1977] T.H. Ning, C.M. Osburn and H.N. Yu. *Emission probability of hot electrons from silicon into silicon dioxide.* Journal of Applied Physics, vol. 48, no. 1, pages 286–293, 1977. (Cited on page 21.)

[Palestri 2006] P. Palestri, N. Akil, W. Stefanutti, M. Slotboom and L. Selmi. *Effect of the gap size on the SSI efficiency of split-gate memory cells.* IEEE Transactions on Electron Devices, vol. 53, no. 3, pages 488–493, 2006. (Cited on pages 6, 19, 24, 30 and 62.)

[Rideau 2011] D. Rideau. Full band models for strained silicon and germanium devices. 2011. (Cited on page 13.)

[Sano 1994] N. Sano and A. Yoshii. *Impact ionization rate near thresholds in Si.* Journal of Applied Physics, vol. 75, no. 10, pages 5102–5105, 1994. (Cited on pages 17, 93 and 114.)

[Shockley 1961] W. Shockley. *Problems related to pn junctions in silicon.* Solid-State Electronics, vol. 2, no. 1, pages 35–60, 1961. (Cited on pages 20 and 48.)

[Sonoda 1996] K. Sonoda, S.T. Dunham, M. Yamaji, K. Taniguchi and C. Hamaguchi. *Impact ionization model using average energy and average square energy of distribution function.* Jpn. J. Appi. Phys. Vol, vol. 35, pages 818–825, 1996. (Cited on page 23.)

[Stadele 2003] M. Stadele, A. Sacconi F. Di Carlo and P. Lugli. *Enhancement of the effective tunnel mass in ultrathin silicon dioxide layers.* Journal of Applied Physics, vol. 93, no. 5, pages 2681–2690, 2003. (Cited on page 20.)

[Synopsys 2010] Synopsys. *Synopsys Sentaurus, release D-2010.12, SDevice simulators*, 2010. (Cited on pages 21, 23, 24, 30, 125 and 156.)

[Tam 1982] S. Tam, P.K. Ko, C. Hu and R.S. Muller. *Correlation between substrate and gate currents in MOSFET's.* IEEE Transactions on Electron Devices, vol. 29, no. 11, pages 1740–1744, 1982. (Cited on page 20.)

[Tam 1984] S. Tam, P.K. Ko and C. Hu. *Lucky-electron model of channel hot-electron injection in MOSFET's.* IEEE Transactions on Electron Devices, vol. 31, no. 9, pages 1116–1125, 1984. (Cited on pages 20, 21, 48 and 78.)

[Tang 1983] J.Y. Tang and K. Hess. *Impact ionization of electrons in silicon (steady state).* Journal of Applied Physics, vol. 54, no. 9, pages 5139–5144, 1983. (Cited on page 19.)

[Thoma 1991] R. Thoma, H.J. Peifer, W.L. Engl, W. Quade, R. Brunetti and C. Jacoboni. *An improved impact-ionization model for high-energy electron transport in Si with Monte Carlo simulation.* Journal of Applied Physics, vol. 69, no. 4, pages 2300–2311, 1991. (Cited on page 18.)

[Troutman 1978] R.R. Troutman. *Silicon surface emission of hot electrons.* Solid-State Electronics, vol. 21, no. 1, pages 283–289, 1978. (Cited on page 22.)

[Vecchi 1998] M.C. Vecchi and M. Rudan. *Modeling electron and hole transport with full-band structure effects by means of the spherical-harmonics expansion of the BTE.* IEEE Transactions on Electron Devices, vol. 45, no. 1, pages 230–238, 1998. (Cited on page 23.)

[Ventura 1992] D. Ventura, A. Gnudi, G. Baccarani and F. Odeh. *Multidimensional spherical harmonics expansion of Boltzmann equation for transport in semiconductors.* Applied Mathematics Letters, vol. 5, no. 3, pages 85–90, 1992. (Cited on page 23.)

[Verwey 1975] J.F. Verwey, R.P. Kramer and B.J. De Maagt. *Mean free path of hot electrons at the surface of boron-doped silicon.* Journal of Applied Physics, vol. 46, no. 6, pages 2612–2619, 1975. (Cited on page 21.)

[Zaka 2010] A. Zaka, Q. Rafhay, M. Iellina, P. Palestri, R. Clerc, D. Rideau, D. Garetto, E. Dornel, J. Singer, G. Pananakakis, C. Tavernier and H. Jaouen. *On the accuracy of current TCAD hot carrier injection models in nanoscale devices.* Solid-State Electronics, vol. 54, no. 12, pages 1669 – 1674, 2010. (Cited on pages 22 and 28.)

[Zaka 2011] A. Zaka, J. Singer, E. Dornel, D. Garetto, D. Rideau, Q. Rafhay, R. Clerc, J.-P. Manceau, N. Degors, C. Boccaccio, C. Tavernier and H. Jaouen. *Characterization and 3D TCAD simulation of NOR-type flash non-volatile memories with emphasis on corner effects.* Solid-State Electronics, vol. 63, no. 1, pages 158 – 162, 2011. (Cited on pages 22 and 143.)

Chapter 3

Semi-analytic approach for hot carrier modeling

The study of the widely used injection models in the previous chapter was not only valuable to define their ability to predict the gate current and its main features. Indeed, it was shown that even the less physically-based models can be adjusted in particular regimes of a given technology and yet not have solid physical basis. More importantly, the confrontation of these models with more advanced and accurate methods, has pointed out the most important mechanisms involved in these phenomena. The previous chapter showed that the Monte Carlo and the Spherical Harmonics Expansion methods are good candidates for such modeling. Both are used either inside or in conjunction with the TCAD environment. However, the price to pay for their accuracy is a longer computational time or the necessity to have a full 2D structure description. While this may not be a disadvantage in the scope of device optimization, it can reveal unsuitable in terms of compact modeling where fast, robust and predictable result are expected. In this environment, the physical description has to be performed with only few working variables.

Therefore, taking advantage of the previous chapter's results, and keeping in mind these latter constrains, a simple yet efficient 1D approach for hot carrier modeling has been developed (Figure 3.1). The general model and its main calculation steps are presented in Section 3.1. The model featuring optical phonons as the only scattering mechanism is next applied in Section 3.2, where the main figures of merit of hot carrier transport are compared with Full Band Monte Carlo simulations, considered as a reference in this work. All comparisons are carried out on large gate length and bias intervals, in order to closely evaluate the extent of the model's efficiency.

This naturally leads to Section 3.3 which presents a critical inspection of each of the main ingredients in this approach. Their impact on the main figures of merit is pointed out and alternative versions of the model are proposed. Finally, Section 3.4 presents the simulation of the electron-electron and impact ionization processes within the new approach in conjunction with the optical phonons, thus demonstrating the integration capacity of the most relevant scattering mechanisms affecting hot carrier transport into a semi-analytic approach.

3.1 Model presentation

The calculation of the carriers' distribution function requires an accurate description of hot carrier transport in the channel. This section describes how this task is performed with a model featuring only few parameters. A general overview of the approach is first given in Subsection 3.1.1, before successively describing the inputs of the model in Subsection 3.1.2, the core calculation of the model in Subsection 3.1.3 and the post-treatments in Subsection 3.1.4.

3.1.1 General overview

The proposed approach is inspired by the previously described *Lucky Electron Model* (LEM) [Shockley 1961], [Tam 1984], [Hasnat 1996]. The general methodology followed in this modeling approach is shown in Figure 3.1.

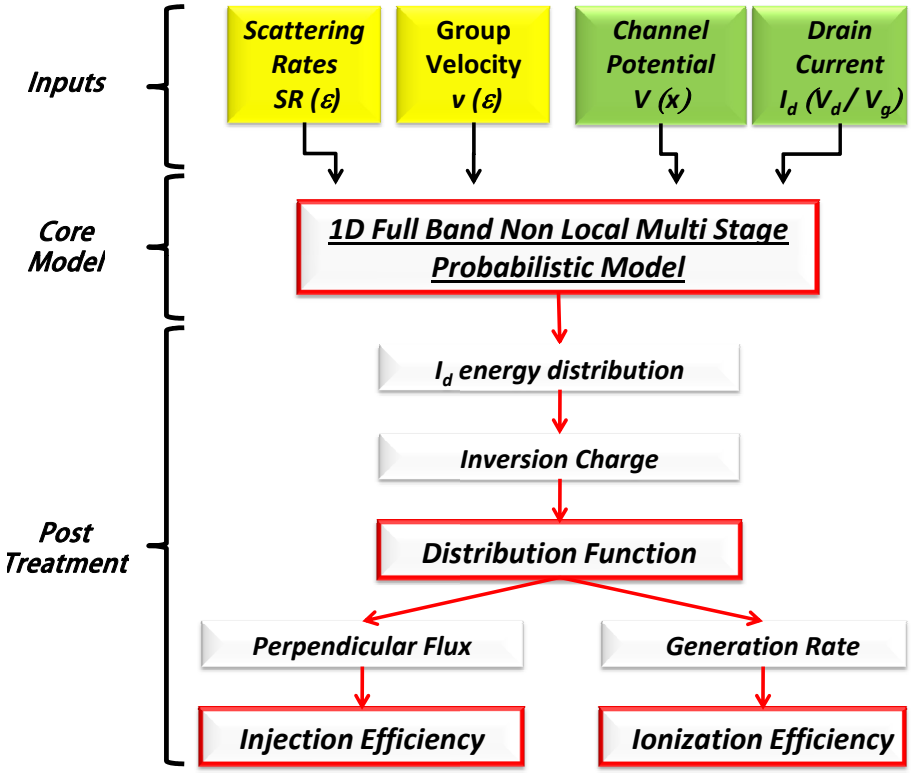


Figure 3.1: Overview of the 1D semi-analytic model including the major calculation steps. The model description is divided into the first two sections.

Three constitutive blocks, namely: the inputs, the calculation procedure of the core model and the post-treatment phase, are consecutively executed in order to evaluate the hot carrier effects. As an extension of the LEM, the new probabilistic approach includes a **non-local** and a **multi-step** treatment of the carrier transport while incorporating some **full band** electronic dispersion features of silicon .

3.1.2 Inputs

The approach requires four inputs, divided into two groups (Figure 3.1).

The first two inputs are the 1D potential along the channel $V(x)$, x being the direction parallel to the Si/SiO_2 interface, and the drain current I_d. Both quantities are bias dependent (V_d, V_g) and are first calculated by external simulators. Throughout this chapter, the potential along the channel is provided by 2D-TCAD simulations which can calculate the 2D potential $V(x, y)$, with y being the direction normal to the interface. In what follows, the channel potential used in the model is extracted at $y = 5$Å from the interface. For comparison purpose, the same 2D-TCAD potential profile is used for Monte Carlo simulations. Furthermore, in order to insure a fair comparison between the latter and the model (c.f. section 3.2), the drain current used in the model is provided by Monte Carlo simulations. For compact modeling integration, a Charge Sheet Model (CSM) has been used instead in order to provide these inputs more rapidly (c.f. Chapter 4 for the presentation of that model).

The model also requires quantities such as the scattering rates (SR) and the group velocity (v) expressed as a function of the carrier energy. These quantities are externally calculated only once after the silicon band structure approximation and are used as constant look-up tables during the simulations. In the full version of the model, three scattering mechanisms have been considered: inelastic optical phonons, impact ionization and electron-electron scattering. The interaction with optical phonons is however retained as the main electron thermalization mechanism playing a major role in shaping the electronic distribution in energy. Other approaches have adopted this simplification as well [Fiegna 1991, Gnudi 1993, Abramo 1996].

Figure 2.6 of the previous chapter reports the optical phonons scattering rate obtained in a full band description of silicon. The scattering rate, as well as the group velocity, are a direct consequence of the approximation used to model the silicon band structure. Figure 2.6 reports as well the energy-dependent impact ionization scattering rates. Finally, the electron-electron scattering rates show an increased complexity as they are shown to be also position-dependent. A detailed discussion is made in Section 3.4.

3.1.3 Model description

In this subsection, the partition of the drain current in various energy levels at each channel position is presented. The other quantities of interest, such as the carriers distribution function, will be extracted from the treatment of this sole quantity. Therefore, the description in this subsection will specifically concern the drain current distribution in energy, and terms such as *current flux*, *carrier flux* or simply *carrier*, will all refer to a portion of the drain current (unit: $A\mu m^{-1}eV^{-1}$). For this purpose, the direction along the channel and the total energy have been considered. A 2D system is thus obtained (Figure 3.2), where the channel position and the energy respectively constitute the abscissa and the ordinate.

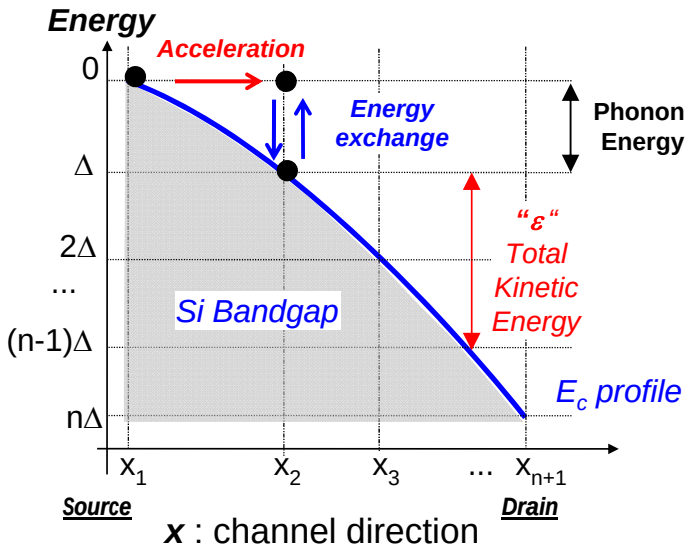


Figure 3.2: System definition including the major variables and processes.

The conduction band profile E_C separating the silicon band gap from the upper part of the diagram has also been schematically reported on the figure. The total electron kinetic energy (ε) is here defined as the remaining total energy after subtracting the potential energy E_C. Such a plot has been largely used when solving the BTE along the transport direction [Baranger 1984], [Sano 2004], [Jin 2008], [Lenzi 2008]. But, in the 1D-BTE formulation, the kinetic energy added to E_C in the plot as in the Figure 3.2 is only the longitudinal one. Instead, in the proposed approach, it is assumed that the accelerating force $\partial E_C/\partial x$ increases the total energy and not only the kinetic energy along x. In addition, the consideration of the total energy facilitates the treatment of the scattering mechanisms as the scattering rates of the latter are expressed as a function of the total energy (c.f. Chapter 2). Furthermore, in this context, the isotropic velocity is considered instead of the longitudinal one. With the carriers being conceptually free to move in any direction, the graphical representation of the carriers' longitudinal movement (Figure 3.2) can be seen as a longitudinal movement of two hemispheres defined by $k_x > 0$ (for the carriers travelling towards the Drain) and $k_x < 0$ (for the carriers travelling towards

the Source).

In this approach, the total kinetic energy increases as the carriers are accelerated under the lateral field. In addition, the carriers can exchange energy with the phonons either by emission (upward arrow) or absorption (downward arrow). The energy exchange is always performed by an equal quantity (Δ) representing the inelastic phonon energy (only the dominant optical phonon is considered here). Thus the phonon energy imposes a uniform energy discretization grid which in term implies a non-uniform channel discretization grid. The other energy-exchange mechanisms, impact ionization and electro-electron interaction, will also use this grid. For the sake of clarity, only interactions with optical phonons are considered in this first part, the other mechanisms being specifically treated in 3.4.

The intersections of both grids hence define the sites where the calculations are performed. In the spirit of the LEM, a ballistic probability is evaluated. The probability for a carrier *not to scatter* between x and $x + \Delta(x)$ is here given by the non-local expression:

$$P^{z}_{x \to x+\Delta x} = \exp\left(- \int_{x}^{x+\Delta x} \frac{SR^{z}[\varepsilon(x')]}{v[\varepsilon(x')]} \mathrm{d}x' \right) \tag{3.1}$$

where the z-index represents either the *Absorption*, *Emission* or *Total* interaction process with the optical phonons. The total scattering rate has been partitioned as:

$$SR^{Total} = SR^{Absorption} + SR^{Emission} \tag{3.2}$$

where the $SR^{Absorption}$ and $SR^{Emission}$ have been respectively calculated from Equations 2.12 and 2.11 of the previous chapter. In the first stage of the simulation, the elementary probabilities $P^{z}_{x \to x+\Delta x}$ are evaluated for each pair of adjacent horizontal sites. Their values depend on the site position, i.e. on the electron position and energy. The result of this calculation is shown on Figure 3.3 which illustrates the probability for a carrier to be ballistic, i.e. in this context, unaffected by the optical phonons, in a 0.14 μm gate-length device as it flows from a given position in the channel towards the drain. Two different positions (near the source and mid-channel) and two different kinetic energies have been chosen as starting points. In one case the carrier starts at the conduction band E_C (zero energy) and in the other at 1 eV above E_C. The plotted ballistic probability is simply obtained as the cumulative product of the elementary probabilities, obtained with Equation 3.1, up to the considered position.

It can be observed that the carriers starting with higher energy (dashed lines) tend to interact more with phonons, in agreement with the scattering rates of Figure 2.6. However, as the emission process is about one order of magnitude more efficient than the absorption process, the interactions mainly result in energy loss. Furthermore, as the field profile changes along the channel, it is clear that different probabilities are obtained as a function of the carrier position. Finally, the wiggles in the probability curves constitute a signature of the full band description of the

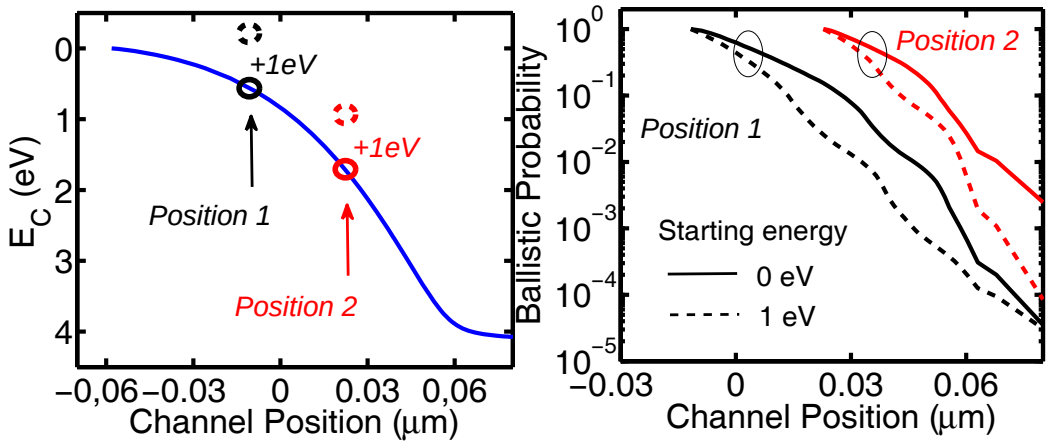


Figure 3.3: Non-local ballistic probabilities (right) for an electron starting at two different positions and energies (left).

scattering rates and the group velocity. Comparisons with other band structures are given in Section 3.3.

The ballistic probabilities of Figure 3.3 express the portion of the carriers that do not scatter. The remaining carriers are split between emission and absorption processes. Figure 3.4 shows the probabilities of each possible path for a carrier travelling from source to drain.

The flux is considered ballistic if neither absorption nor emission occurs (P^{BAL}); the probability ($1-P^{BAL}$) of scattering either by emission or absorption is expressed by considering a conditional probability scheme. A full derivation of the probabilistic events comprising impact ionization and the electron-electron scattering beside the phonon scattering process, is given in Annex A. The probability fluxes in Figure 3.4 result in:

$$P_m^{BAL} + P_m^{UP} + P_m^{DOWN} = 1 \tag{3.3}$$

with m representing either the source-drain (SD) or drain-source (DS) direction. The carriers flowing in the source to drain direction are originally provided by the source. In this approach, the carriers injected by the drain have been neglected.

After a scattering event, the carrier's energy and momentum has to be updated accordingly. The use of the total kinetic energy as the relevant energy in this system, simplifies this choice. Thus, the energy of the final state is already set by the exchange of a constant phonon energy Δ. Furthermore, as the phonon interaction is assumed isotropic [Lundstrom 2000], the final state momentum should be uniformly distributed in all directions. In the proposed 1D real-space approach, the carriers go either towards the drain or towards the source with an equal probability of 1/2, which respectively represent the $k_x > 0$ and $k_x < 0$ hemispheres. Considering that

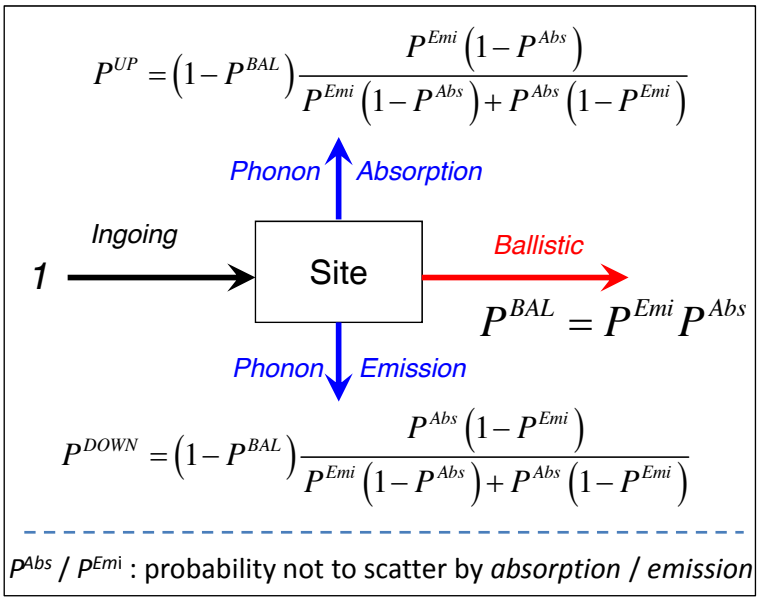


Figure 3.4: Graphical representation of the possible electron paths with their respective probabilities evaluated at each node of the system. P^{Emi}, P^{Abs} and P^{Total} are the probabilities for an electron not to emit, absorb or interact with optical phonons, respectively.

the final momentum state is proportional to the final density density of states, such splitting assumes that the same density of states is available for both hemispheres. Following this reasoning, an easy way to introduce backscattering has been devised which allows many of the carriers to be re injected in the system and to undergo more scattering events. The backscattered carriers are treated by using Equation 3.1 and enforcing the local flux conservation depicted in Figure 3.4, with horizontal fluxes being in the opposite direction. Such a description leads to consider eight fluxes for each site as depicted in Figure 3.5.

Fluxes a, e and c, f respectively represent the ballistic incoming and outgoing fluxes, while fluxes d, h and b, g respectively represent the out-scattering and in-scattering fluxes. All the fluxes are currents expressed in $A\mu m^{-1}eV^{-1}$. The first ones couple a given position with the adjacent ones, while the second ones couple a given energy level with the upper and lower level. The local relations in Figure 3.5 express the four outgoing fluxes as a function of the four incoming ones. Notice that flux c (f) is not supplied by the flux e (a). This translates the fact that no elastic scatterings have been integrated in the approach; the backscattered carriers travel in one of the adjacent energy levels after an inelastic collision.

All the above fluxes represent a portion of the drain current that should be conserved at the local level for a consistent modeling. Hence, the incoming fluxes equal the outgoing ones:

$$a + b + e + g = c + f + h + d \tag{3.4}$$

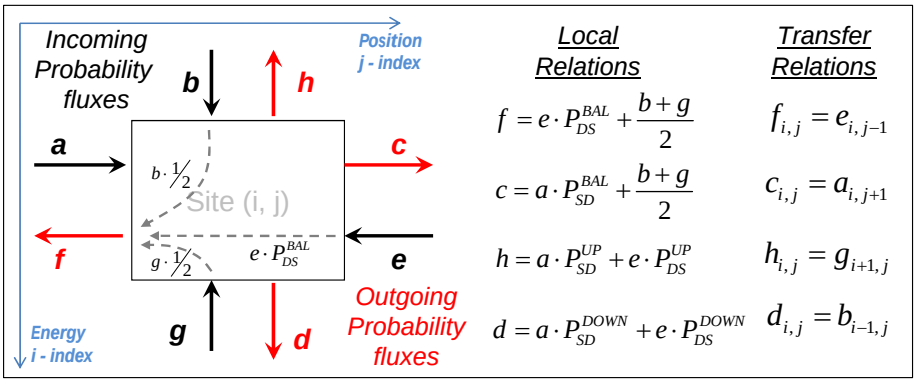


Figure 3.5: Schematic representation of the fluxes for each site and local and transfer relations among them. The contributions for the flux f have been drawn; similar picture is obtained for the other fluxes. The i, j indexes have been omitted for clarity in the local relations.

The transfer relations in Figure 3.5 finally insure the continuity in the 2D space till the boundaries are reached.

Boundary Conditions

Figure 3.6 summarizes the boundary conditions of the system.

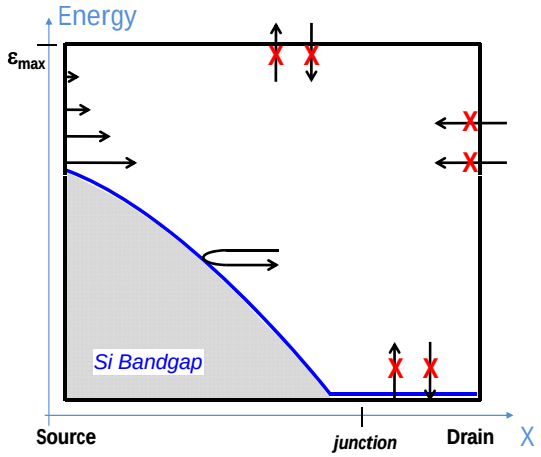


Figure 3.6: Schematic representation of the boundary conditions. Each arrow symbolizes a flux in the given direction; the cross on the arrows means that the considered flux is nil.

Several boundary conditions have been implemented:

- Dirichlet conditions are imposed at the source side ($j = 1$) with an injecting (entering) drain current following a Maxwellian distribution in energy. The

drain current is normalized by the device width and expressed per unit of energy ($A.\mu m^{-1}.eV^{-1}$).

$$c_{i,1} = \frac{I_d \cdot \exp\left(-\frac{\varepsilon_i - \Delta}{k_B T}\right)}{W \cdot \int_0^{+\infty} \exp\left(-\frac{\varepsilon - \Delta}{k_B T}\right) d\varepsilon} \tag{3.5}$$

- The maximum energy in the system ε_{max} ($+\infty \longrightarrow \varepsilon_{max}$) is chosen in order to allow a sufficient description of the hot carrier tail and a reasonable computation time. When only optical phonons are considered in the system, $\varepsilon_{max} = V_d + 1$ eV has been considered as a good compromise, whereas when electron-electron scatterings are included in the system, ε_{max} values up to $V_d + 4$ eV should in many cases be considered which inevitably increase the simulation time. Furthermore, ε_{max} should be at least equal to the Si/SiO_2 barrier (3.1 eV). Whatever the value, no electrons can overcome this upper limit. Hence:

$$b_{i_{\varepsilon max},j} = h_{i_{\varepsilon max},j} = 0 \tag{3.6}$$

- In the adopted semi-classical perspective, no electrons can enter the silicon band gap. In the channel, this is expressed by:

$$a_{(i,j)_{BARRIER}} = f_{(i,j)_{BARRIER}} = d_{(i,j)_{BARRIER}} = g_{(i,j)_{BARRIER}} = 0 \tag{3.7}$$

where the BARRIER index designates the sites having zero kinetic energy located at the potential profile. Furthermore, it has been considered that the carriers which 'hit' the barrier, bounce back and are re injected into the system, thus giving rise to a modified expression for flux c :

$$c_{(i,j)_{BARRIER}} = b_{(i,j)_{BARRIER}} + e_{(i,j)_{BARRIER}} - h_{(i,j)_{BARRIER}} \tag{3.8}$$

Inside the LDD, the slowly-varying potential is considered flat and the boundary condition becomes:

$$d_{(i,j)_{BARRIER}} = g_{(i,j)_{BARRIER}} = 0 \tag{3.9}$$

Numerical Solution

The size of the obtained system is a function of the potential profile and the ε_{max} value. The V_d value impacts the potential fall across the channel and thus determines the available energy levels in the system given a constant energy discretization grid. For typical V_d values, when only optical phonons are considered,

the system reaches 3000-4000 nodes, each having eight unknown fluxes to be calculated. In this approach, the computational time does not depend on the device length as the discretization along x is always imposed by Δ (Figure 3.2), differently from the Monte Carlo or the Spherical Harmonics Expansion methods. Using the local and transfer relations on Figure 3.5 the number of unknowns per node can be brought down to two. Indeed, each flux can be expressed as a function of c, e fluxes, which in turn are functions of the adjacent sites' c, e fluxes. A linear system is thus obtained with a largely sparse matrix(Figure 3.7).

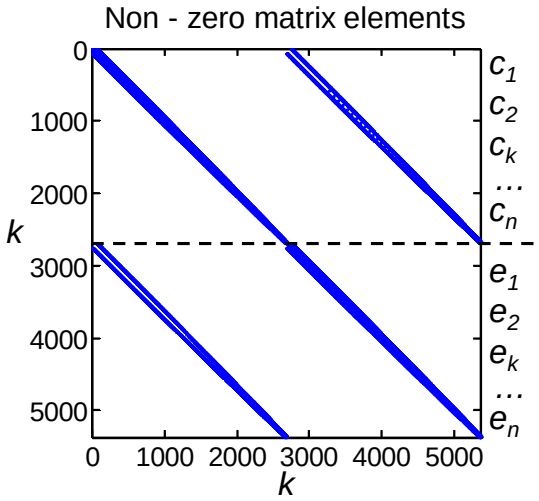


Figure 3.7: Non-zero elements of the linear system matrix. The unknowns c_i and e_i are arranged consecutively.

The computational burden is composed of the calculation of the elementary ballistic probabilities (Equation 3.1) and of the linear system solution. The approach is implemented in Matlab and a total simulation time in the range of 3-6 seconds using an Intel Core-2 3GHz processor is necessary for each bias condition.

3.1.4 Post processing

The model in the previous subsection allowed us to calculate the distribution of the drain current among the available energy levels in the system. In this subsection the transformations necessary to obtain the useful transport quantities shown in Figure 3.1, are presented.

First of all, the model gives access to the drain current ($A.\mu m^{-1}$) at each channel position. This is calculated as:

$$I_d(j) = \Delta\varepsilon \sum_{i=1}^{i_{max}} I_d(i,j) = \Delta\varepsilon \sum_{i=1}^{i_{max}} (c_{i,j} - e_{i,j}) \tag{3.10}$$

$\Delta\varepsilon$ is the constant energy grid spacing ($\Delta\varepsilon$ = 60 meV) and i_{max} corresponds to the index of the maximum energy in the system. The drain current should be

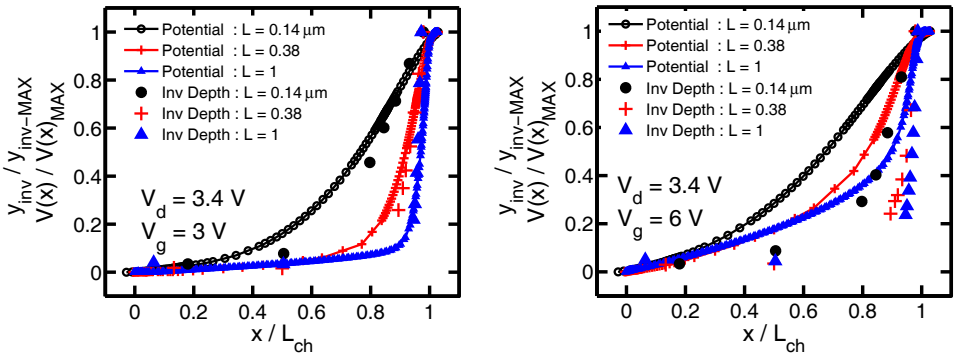
constant along the channel (independent of j); this will be verified in the next section. Next, the inversion charge (cm^{-2}) density is calculated along the channel:

$$N_{inv}(j) = \Delta\varepsilon \sum_{i=1}^{i_{max}} N_{inv}(i,j) = \Delta\varepsilon \sum_{i=1}^{i_{max}} \frac{c_{i,j} + e_{i,j}}{q \cdot v_{i,j}} \tag{3.11}$$

The sum of the fluxes is retained to reflect the total quantity of the carriers. Their sum is then weighted by the energy-dependent group velocity $v_{i,j}$ to obtain the number of carriers per unit of area.

In order to calculate the absolute value of hot carrier effects, the carrier density at given vertical positions should also be investigated. Hence, the inversion charge density of Equation 3.11 needs to be transformed in a volume density. Such a step requires the knowledge of the inversion depth y_{inv} along the channel. This certainly constitutes a challenging task for every compact-like approach. In fact, in such approaches, a charge sheet is assumed to be located at the interface [Brews 1978].

In order to illustrate this point we extracted the inversion depth along the channel from Monte Carlo simulations. Figure 3.8 reports such quantity as a function of the channel position for three channel lengths and two bias conditions.



(a) (b)

Figure 3.8: Normalized inversion depth and electrostatic potential along the relative channel position for different gate lengths and bias conditions. A three-fold decrease of the electrons density with respect to the value at the interface has been chosen as criterion to define the inversion depth from Monte Carlo simulations. All quantities are normalized to their respective maximum values. L_{ch} is the distance between the LDD extensions.

The inversion depth has been obtained as the distance from the Si/SiO_2 interface where the electron density is three times lower with respect to the maximum concentration (at the interface in this semi-classical approach). Using this definition, the inversion depth is about 1 nm at the source side for all devices and rises towards the drain as the device switches from strong inversion to depletion. The inversion depth has been extracted up to the beginning of the LDD region as the presence of the cold

carriers in the drain would inevitably interfere with the extraction. Figure 3.8 also reports the potential along the channel for each of the conditions. It can be seen, that the relative variation of the inversion depth is similar to the relative variation of the potential along the channel. This simple observation allowed us to integrate into the model the inversion depth variation from source to drain. However, the inversion depth value at the drain end, which is the most interesting part of the device in terms of hot carrier effects, depends on the gate length and on the bias condition. Figure 3.9 reports y_{inv} as a function of the channel length.

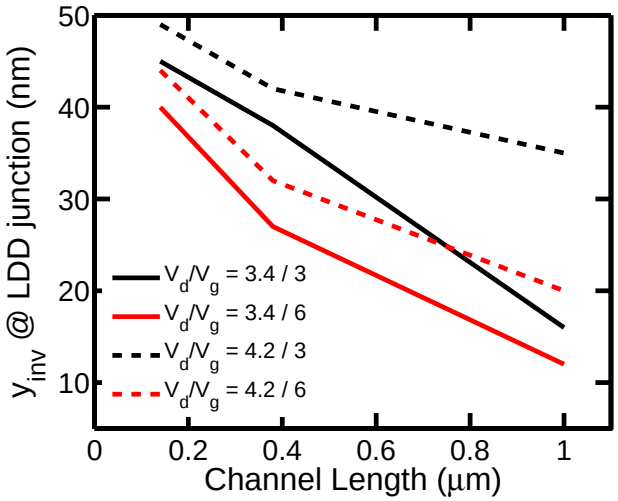


Figure 3.9: Inversion depth value extracted at the channel/LDD junction as a function of the channel length from Monte Carlo simulations. A three-fold decrease of the electron density with respect to the interface has been chosen as criterion to define the inversion depth.

The increase of the inversion depth with the scaling of the gate length is a consequence of the short the channel effects. As expected from 2D field distribution considerations, the inversion depth at the drain side increases with conditions enhancing the pinch-off region (increasing V_d and decreasing V_g).

Although a linear interpolation function could be used to express all the gate lengths, a constant value of 20 nm for the inversion depth at the drain y_{inv_D} has been instead used in this study. This choice has been made to prevent any cumbersome adjustments and to keep the approach as simple as possible. Therefore, the inversion depth (nm) has been varied from from $y_{inv_S} = 1\ nm$ (source) to $y_{inv_D} = 20\ nm$ (drain) following the channel potential variation using the following expression:

$$y_{inv}(j) = y_{inv_S} + (y_{inv_D} - y_{inv_S}) \frac{V(j)}{V_D} \tag{3.12}$$

with $V(j)$ and V_D being the potential variation along the channel and the maximum potential reached at the drain, respectively. The potential at the source $V(0)$

is set to zero by definition. Using Equation 3.11, the carrier density at a given channel position (cm^{-3}) is finally calculated as:

$$n(j) = \Delta\varepsilon \sum_{i=1}^{i_{max}} n(i,j) = \Delta\varepsilon \frac{\sum_{i=1}^{i_{max}} N_{inv}(i,j)}{y_{inv}(j)} = \Delta\varepsilon \frac{\sum_{i=1}^{i_{max}} \frac{c_{i,j}+e_{i,j}}{q \cdot v_{i,j}}}{y_{inv}(j)} \tag{3.13}$$

where the inversion charge density is assumed constant between $y = 0$ and y_{inv}. $n(i,j)$ elements with the same j-index, i.e. same position, make up the distribution function at the considered position. The latter is expressed in $cm^{-3}eV^{-1}$ and it has homogeneous dimensions as the product of the probability function times the density of states ($f \cdot g$).

The hot carrier effect evaluated in the following are the electron-hole generation by impact ionization and electron injection into the gate. Two different criteria are chosen to compute them in each case. For the sake of clarity and for an easier reading of the formulae, the line and column indexes of the described system, i and j, are respectively replaced with the continuous variables ε and x in the following. They bear of course the same exact meaning as described in the previous subsection.

First of all, the impact ionization generation rate G_{II} ($cm^{-3}s^{-1}$) along the channel is calculated at a given vertical position ($y = 5\ A$) as:

$$G_{II}(x) = \int_0^{\varepsilon_{max}} n(\varepsilon, x) \cdot S_{II}(\varepsilon) d\varepsilon \tag{3.14}$$

where $S_{II}(\varepsilon)$ is the impact ionization scattering rate as a function of the carrier energy [Bude 1992]. The use of $G_{II}(x)$ allows for a detailed investigation of the role of the distribution function shape along the channel. The macroscopic bulk current I_b ($A\mu m^{-1}$) can then be calculated as:

$$I_b = q \int_0^{L_g} \int_0^{\varepsilon_{max}} N_{inv}(\varepsilon, x) S_{II}(\varepsilon) d\varepsilon dx \tag{3.15}$$

The I_b-related quantities are readily obtained from the calculation of the distribution function. The carrier distribution as a function of the total carrier energy (ε), however cannot be used in a straightforward manner to calculate the gate current I_g because it would give too much current [Bufler 2005]. In the semi-classical approach, the gate current depends on the component of the carriers energy representative of the momentum component normal to the Si/SiO_2 interface. This energy is hereafter called the *normal energy* $\varepsilon_\perp$. A variable change must hence be performed on the distribution function. Furthermore, in order to obtain the gate current or current density, it seems natural to transform the concentration in a current flux. To calculate the current flux normal to the interface as a function of the normal energy $J_\perp(\varepsilon_\perp)$, we assumed:

- an isotropic distribution function

- an isotropic parabolic band structure

These assumptions are commonly made in other approaches as well [Jin 2009]. To project the distribution in the perpendicular direction, the normal flux corresponding to a Dirac delta carrier distribution in total energy ($h(\varepsilon_0)\delta(\varepsilon-\varepsilon_0)$, with h in units of cm^{-3}) is calculated at first. Under the above assumptions, a constant normal flux up to the considered energy is found:

$$h(\varepsilon_0)\delta(\varepsilon-\varepsilon_0) \Rightarrow J_\perp(\varepsilon_\perp) = q\frac{h(\varepsilon_0)}{2\sqrt{2m\varepsilon_0}}\Theta(\varepsilon_0-\varepsilon_\perp) \tag{3.16}$$

with Θ being the Heaviside function (c.f. Annex B for a full derivation of this expression). The distribution function in total energy is then discretized with a Dirac comb (spacing $\Delta\varepsilon_0$, index p) and eventually written as:

$$n(\varepsilon) = \sum_{p=0}^{p_{max}} n(p\Delta\varepsilon_0)\delta(\varepsilon-p\Delta\varepsilon_0) \tag{3.17}$$

A constant normal flux is calculated for each Dirac delta contribution as schematically shown on Figure 3.10.

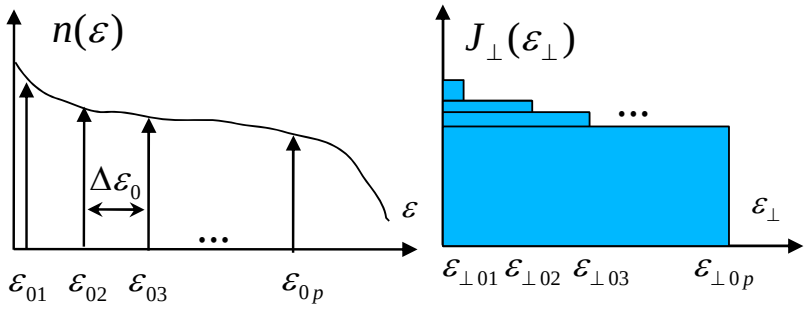


Figure 3.10: Drawing representing the calculation of the perpendicular flux as a function of the normal energy ($(J_\perp(\varepsilon_\perp))$) from the distribution function in total energy ($n(\varepsilon)$). The latter is sampled with a Dirac comb to which corresponds a constant function up to the considered energy.

The Dirac comb spacing $\Delta\varepsilon_0$ can be different from the energy grid spacing $\Delta\varepsilon$ used for distribution function calculation. Smaller values of $\Delta\varepsilon_0$ insure more accurate projections at the cost of a higher computation time. Throughout this study, $\Delta\varepsilon_0 = 20\ meV$ has been retained as a good compromise between the factors above. The normal flux for a given normal energy $\varepsilon_{\perp 0i}$ is then calculated as a sum over all the contributions coming from energies equal or greater than the considered energy:

$$J_\perp(\varepsilon_{\perp 0i}) = q\sum_{p=i}^{p_{max}} \frac{\Delta\varepsilon_0 \cdot n(p\Delta\varepsilon_0)}{2\sqrt{2mp\Delta\varepsilon_0}}\Theta(p\Delta\varepsilon_0-\varepsilon_{\perp 0i}) \tag{3.18}$$

m is the isotropic electron conduction mass in silicon equal to $0.26m_0$ and calculated as:

$$m = \frac{3}{\frac{1}{m_l} + \frac{2}{m_t}} \tag{3.19}$$

m_l and m_t is the longitudinal ($0.89m_0$) and transverse ($0.19m_0$) electron conduction mass in silicon, respectively. The use of a parabolic band approximation gives a simple but still an efficient way to separate the normal from the parallel carrier energy. As a matter of fact, such operation is impossible in the full-band description of silicon.

This leads us finally to the gate current calculation, performed by the classical expression:

$$I_g = \int_0^{L_g} J_g(x)\, dx = \int_0^{L_g} \int_0^{\varepsilon_{\perp max}} J_\perp\left(x, \varepsilon_\perp\right) \cdot T\left(\varepsilon_\perp\right)\, d\varepsilon_\perp \tag{3.20}$$

$J_g(x)$ is the gate current density along the channel and $T(\varepsilon_\perp)$ is the tunneling probability as a function of the normal energy. The latter is calculated in the WKB approximation in order to insure a fast calculation at each channel position.

3.1.5 Summary

In this section, a new probabilistic approach for hot carrier modeling has been introduced. Its main features and assumptions have been described and critically discussed. Many important transport ingredients such as the non-locality, the main scattering events and the full-band description have been included in the approach. In the purpose of proposing a simple yet efficient compact-oriented approach, only few transport variables are used by the model, which are the potential profile along the channel and the constant drain current flowing in the structure. The core of the model concerns the calculation of the distribution of the drain current as a function of the energy along the channel. The current fluxes thus obtained are subsequently used in a post processing procedure to extract all the desired hot carrier characteristics. The model is validated against Monte Carlo simulations in the next sections.

3.2 Model with optical phonons

The calculation procedure described in the previous section is hereafter examined step-by-step in order to compare the proposed model with Full Band Monte Carlo (FBMC) simulations [Palestri 2006]. Throughout this section, the model includes only scattering by optical phonons. As a first step, the current conservation along the channel will be verified in subsection 3.2.1 alongside the inversion charge and the electron density. Then, subsection 3.2.2 will present the distribution functions and the resulting impact ionization generation rates and efficiencies. Finally, subsection 3.2.3 will discuss the obtained gate current in terms of injection efficiency.

3.2.1 Transport characteristics

Before presenting the hot carrier effects, the validity of the proposed model in terms of carrier transport is first discussed. Among the basic transport characteristics, the drain current, the inversion charge and the total carrier density are considered and presented following the procedure of subsection 3.1.4.

Figure 3.11 reports the drain current ($A\mu m^{-1}$) along the channel, calculated by Equation 3.10, with respect to the current set point (dashed line), for different gate lengths and bias conditions.

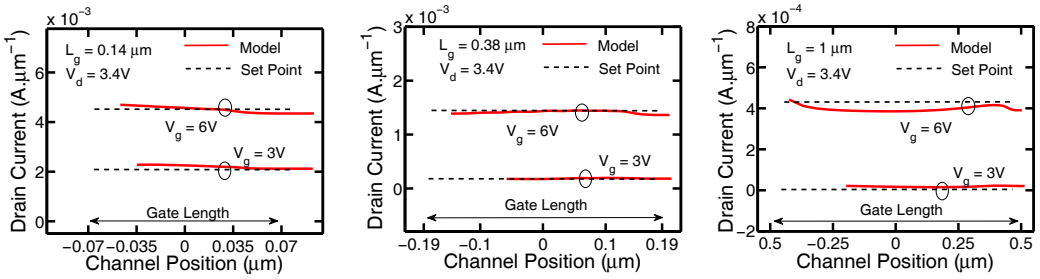
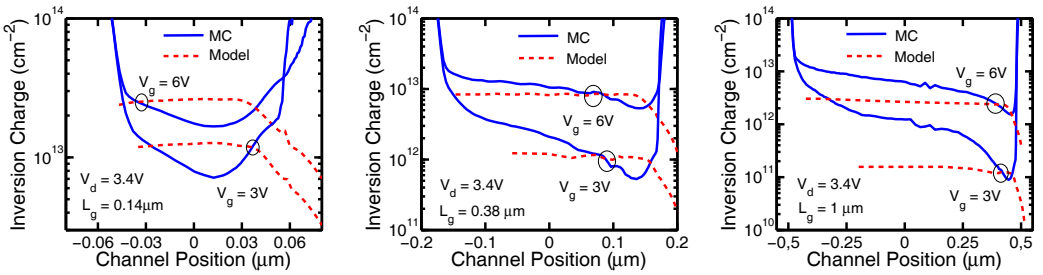


Figure 3.11: Simulated drain currents along the channel (red, plain) and their initial set point (black, dotted) for two gate voltages (3 and 6 V) and three gate length devices: 0.14 μm (a), 0.38 μm(b) and 1 μm (c).

A constant drain current within 5-10 % of the set point is obtained. These small differences are introduced by two aspects. On one hand, the current conservation condition (Equation 3.4) is valid for dense enough grids. On the other hand, as already mentioned, the treatment of the cold carriers at the potential barrier is somewhat simplistic. Hence, it is not surprising to find that the largest discrepancies are observed for the longest device (Figure 3.11 (c)) where the potential is slowly varying in the first half of the channel. This gives rise to a loose grid and a large cold carrier population.

Next, the inversion charge density (cm^{-2}), calculated after Equation 3.11, is compared with FBMC simulations for the same conditions as above. In the FBMC

simulator, the same optical phonon scattering rate as in the model (Figure 2.6 in Chapter 2) is included. However, FBMC includes a 3D momentum description and a 2D potential profile, in contrast with the 1D description of the semi-analytic model. In addition, the FBMC simulations include acoustic phonon scatterings, treated as elastic interactions, having an impact only on the momentum direction, thus contributing to obtain isotropic distributions.



(a) (b) (c)

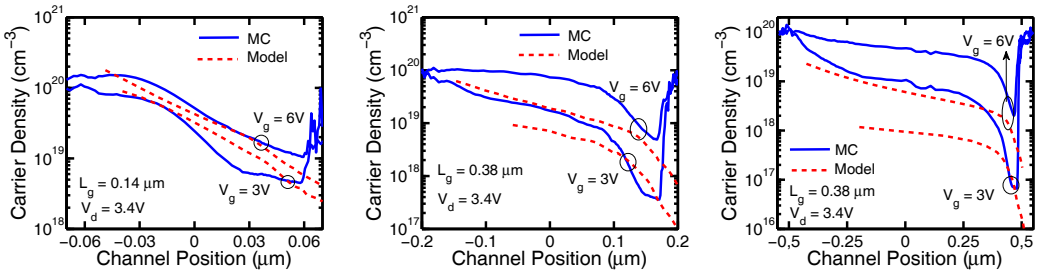
Figure 3.12: Comparison of the inversion charge simulated with the Model (red) and the Monte Carlo (blue) along the channel for two gate voltages (3 and 6 V) and three devices featuring L=0.14 μm (a), L=0.38 μm(b) and L=1 μm (c).

The profile of inversion charge density along the channel is well known and described elsewhere [Sze 2007]. The model's predictions are in a good qualitative agreement with FBMC concerning the shape inside the channel. They also stand within a factor of two for the shortest lengths, while the discrepancy is somehow more important for the longest device especially in the first half of the channel. Such discrepancy is not really surprising considering that this quantity should be obtained by taking into account the carriers' energy through the whole inversion depth. In this model, only a 1D cut close to the interface is considered. Furthermore, the calculation of this quantity in the model is performed based on the drain current injected at the source side. However, especially for the longest device, many more electrons are available especially in the source side which do not contribute to the drain current, and thus are not accounted for in the model.

Lastly, another discrepancy is observed at the drain, where a different qualitative trend with respect to FBMC is found. This is introduced by the lack of cold carriers modeling coming from the drain which increase the interface charge as simulated by the FBMC. The carriers injected from the source are instead continuously accelerated which results in a monotonic decrease of the model-simulated interface charge.

The last important macroscopic transport quantity is the total carrier concentration calculated after Equation 3.13. Figure 3.13 reports the comparison with FBMC simulations for the same conditions as above.

As the total carrier concentration is directly obtained from the interface charge density, the same comments as above apply here as well. In addition, the carrier



(a) (b) (c)

Figure 3.13: Comparison of the carrier density simulated with the Model (red) and the Monte Carlo (blue) along the channel for two gate voltages (3 and 6 V) and three gate length devices: 0.14 μm (a), 0.38 μm(b) and 1 μm (c).

concentration makes also use of the inversion depth along the channel calculated with Equation 3.12. Although, the carrier concentration is affected by the inversion depth value, it however stays within a factor of two with respect to FBMC simulations close to the channel/drain junction.

Overall, the comparison of macroscopic channel quantities has allowed us to verify the model's consistency in terms of transport. Considering the assumptions and pragmatic choices made for carrier transport modeling and reminding that the macroscopic quantities include both cold and hot carriers, the achieved results can be considered acceptable. The hot carrier population is the object of the following subsections which seek to demonstrate the validity of our approach by employing various figures of merit.

3.2.2 Distribution functions and generation rates

The carrier energy distribution, here-after called distribution function, is the key microscopic ingredient for hot carriers study. For this reason, particular attention has been paid in comparing this quantity as a function of the channel position for different device lengths and bias conditions. The distribution function in a given position is made up from the collection of $n(i, j)$ (Equation 3.13) having the same j-index, i.e. same column, same position. Figure 3.14 reports the distribution functions at four channel positions in a 0.14 μm channel device for a $V_d/V_g = 3.4/4.5V$ bias condition. Both simulations are performed after 2D-TCAD simulations: the FBMC uses the 2D potential in a Frozen Field configuration, while the proposed model uses a 1D potential profile obtained from a horizontal cut at 5 Å from the Si/SiO_2 interface. The FBMC curves are obtained after averaging the distributions calculated in a box featuring 1 nm height starting from the interface and several nm in the length direction. The same averaging length has been assumed in the case of the 1D model, although small differences may occur due to meshing differences.

Before commenting the model's results, it is important to make some general observations concerning the distribution functions. The shape of the distributions

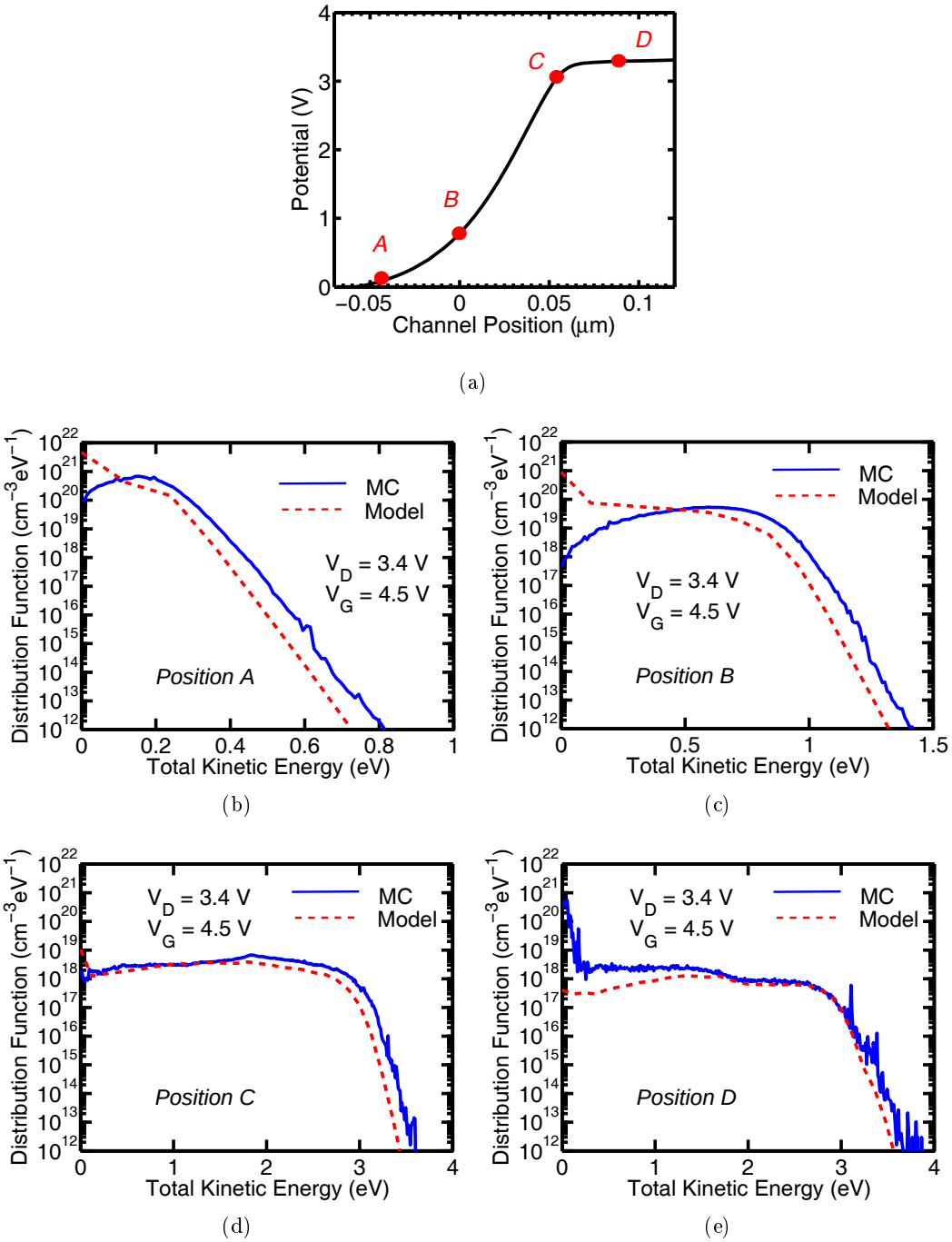


Figure 3.14: Comparison of the distribution functions obtained after Monte Carlo (MC) and the Model at the channel positions A, B, C, D as shown in (a), respectively corresponding to (b), (c), (d), (e).

closely depend on the "carriers' history", which is a direct function of the travelled distance and the potential drop the carrier is subject to till the considered position.

It is noticeable that all the curves exhibit a sort of 'plateau' which finishes around an energy relatively close to the potential drop at that position. At higher energy the concentration drops with a Maxwellian exponential tail which reflects the small amount of lucky carriers which have absorbed optical phonons [Abramo 1996]. This means that the potential drop determines the maximal available energy for most of the carriers. Hence, a suitable description of the plateau, especially at high energies, and of the Maxwellian tail is of capital importance for a consistent hot carrier effects modeling.

Figure 3.14 shows that the model has an excellent ability to reproduce the distributions of the reference Monte Carlo simulations as a function of the channel position over more than seven orders of magnitude. The carriers, especially the hottest ones, are correctly distributed throughout the channel and in the LDD region. However, the complexity of hot-carrier modeling requires a proper description of the distribution for a large range of biases and lengths. As a matter of fact, such questions have always been the Achilles' heel of the local models [Fiegna 1991], [Hasnat 1996]. In order to show the relevance and the accuracy of our approach, Figure 3.15 reports another comparison with FBMC for the same device over a large V_d bias range. Comparisons are performed near the channel/LDD metallurgical junction, where the maximum of the gate current injection occurs.

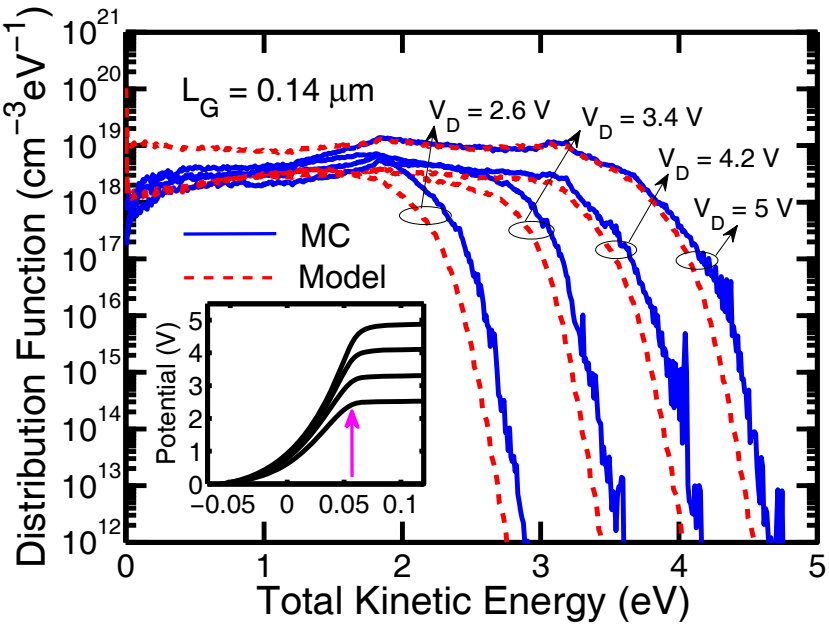


Figure 3.15: Comparison of electronic energy distributions simulated with the Model (red) and the Monte Carlo (blue) for different drain voltages ($V_D = 2.6, 3.4, 4.2, 5V$) at fixed gate voltage ($V_G = 4.5V$) and channel length ($L_G = 0.14\mu m$). The distributions are extracted at the channel/LDD junction as shown by the inset.

The comparison with Monte Carlo has been also extended to other gate lengths (Figure 3.16). A good agreement is still observed with the same model setup without

any ad-hoc adjustment.

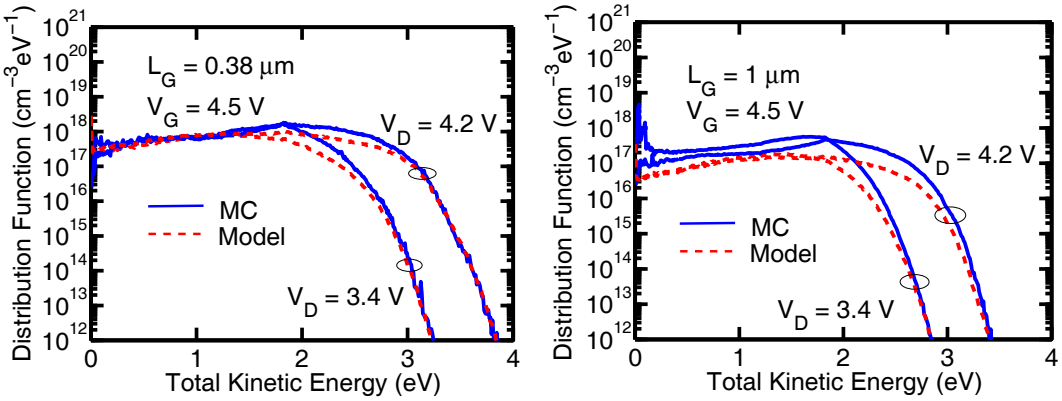
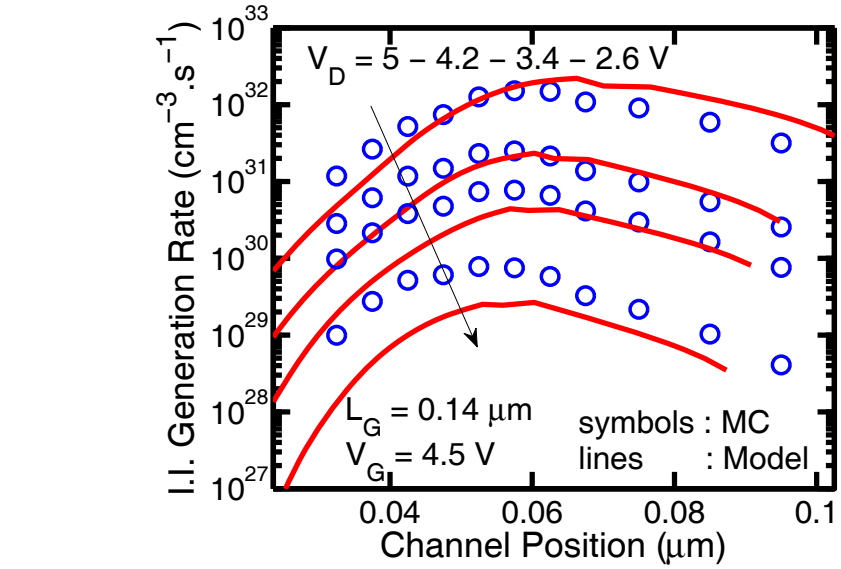


Figure 3.16: Comparison of electronic energy distributions simulated with the Model (red) and the Monte Carlo (blue) at different drain voltages ($V_D = 3.4, 4.2V$) and constant gate voltage ($V_G = 4..5V$) for devices featuring $0.38\mu m$ (a) and $1\mu m$ gate lengths. The distributions are extracted at the channel/LDD junction.

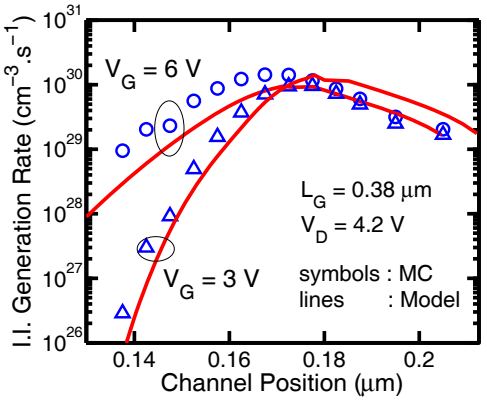
The accuracy of the simulated distribution function is evaluated next by considering electron-hole pair generation by impact ionization. First, the microscopic generation rate along the channel is calculated at 5 Å from the interface using Equation 3.14 for both the model and the FBMC. Figure 3.17 reports the calculated rates as a function of the channel position for different gate lengths and bias conditions. Drain and gate biases have been respectively varied for the smallest and the other investigated devices in order to show and emphasize the largest variations.

The increase of the drain bias, which brings hotter electrons in the device (c.f. Figure 3.15), leads to higher generation rates (Figure 3.17). The general bell-shape curve is conserved with the peak position being situated around the channel/LDD metallurgical junction. Its precise position and shape depends however on the V_d/V_g condition: when the width of the depletion at the junction is increased (when V_d increases or V_g decreases) the peak is shifted towards the LDD with the profile becoming sharper which is enhanced in the case of long devices. A very good qualitative and quantitative agreement has been achieved in all these cases with the differences limited within a decade.

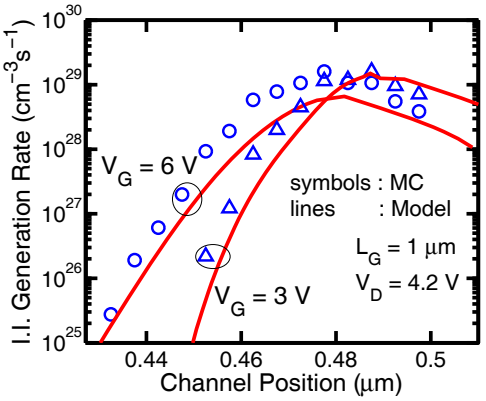
In the above comparisons, the whole distribution, including cold and hot carriers, has been used to calculate the generation rates. However, subsection 3.2.1 showed that a combination of the reduced model dimensionality and the lack of modeling accuracy for the coldest carriers had an impact on the absolute value of the carriers density. Therefore, the quantitative value of the generation rates is also affected. In order to investigate only the hot carriers and their transport in the channel, the distribution functions have been limited to 1eV and the remaining high energy part of the curve has been normalized by its carrier density, i.e. the integral under the



(a)



(b)



(c)

Figure 3.17: Comparisons between impact ionization generation rates along the channel at a depth of $5A$ from the interface calculated from the Model (lines) and the Monte Carlo (symbols) for different gate lengths: $0.14\mu m$ (a), $0.38\mu m$ (b) and $1\mu m$(c). Variations of the drain voltage (a) and the gate voltage (b, c) are performed while keeping the other terminal at a constant value.

curve. The considered distribution function is thus calculated as:

$$n_{norm/1eV}(\varepsilon, x) = \frac{n(\varepsilon, x)}{\int_1^{\varepsilon_{max}} n(\varepsilon, x) d\varepsilon} \tag{3.21}$$

This is as if we were assuming a constant electron density along the channel distributed in energy from 1eV up to the maximum energy following the simulated

distribution functions. The new generation rate becomes:

$$G_{II-norm/1eV}(x) = \int_{1}^{\varepsilon_{max}} n_{norm/1eV}(\varepsilon, x) \cdot S_{II}(\varepsilon) d\varepsilon \tag{3.22}$$

The new rates are reported on Figure 3.18. A very good quantitative agreement is obtained especially for the longest devices. For the shortest device, the model shows a slightly higher V_d dependence with respect to the FBMC. These results suggest that the model would perform even better if there were no errors on cold carriers in the distribution.

The integral of the position dependent generation rates determines the substrate current as calculated with Equation 3.15. In order to investigate the intrinsic efficiency of the impact ionization process in a given condition, the substrate current is normalized to the corresponding drain current. The unit-less I_b/I_d ratios, are reported in Figure 3.19 as a function of the gate voltage for different device lengths and bias conditions.

For a constant drain voltage, the normalized bulk current monotonously decreases with increasing gate voltage. The pinch-off region is indeed reduced for higher gate voltages which cools down the distribution function. While this is especially true for the longest cells, the shortest one features a rapidly varying potential throughout the entire channel, which in turn reduces the effect of the gate voltage change. In all cases, a very good quantitative agreement is obtained for all the investigated conditions.

Overall, these results demonstrate that the distribution function shape calculated by the non-local model is quite accurate. This conclusion has been consolidated by showing results covering an extended range of gate lengths and biases illustrating both microscopic and macroscopic indicators. All these results reveal that the main elements of the hot carrier transport have been correctly described.

3.2.3 Perpendicular fluxes and injection efficiencies

In this subsection the distribution functions are used to calculate the gate current. The first step is to calculate the particle flux impinging on the Si/SiO_2 interface, here-after denoted the perpendicular flux, as a function of the normal-to-the-interface energy component. Closely following the considerations of subsection 3.1.4 and Annex B, Figure 3.20 reports the perpendicular flux ($A\mu m^{-2}eV^{-1}$) as a function of the normal energy for two gate lengths and two channel positions. No projection in the perpendicular direction is required for the FBMC distributions as the normal fluxes are directly taken from the full band, 3D k-space simulation considering the number of particles impinging the interface as well as their normal energy.

This comparison naturally bears all the previously discussed uncertainties and errors. Notice for instance that the model-predicted low energy part of the mid-channel flux is not accurate in the case of the long device. This is largely due to the incorrect total number of particles predicted in this case. However, close to the

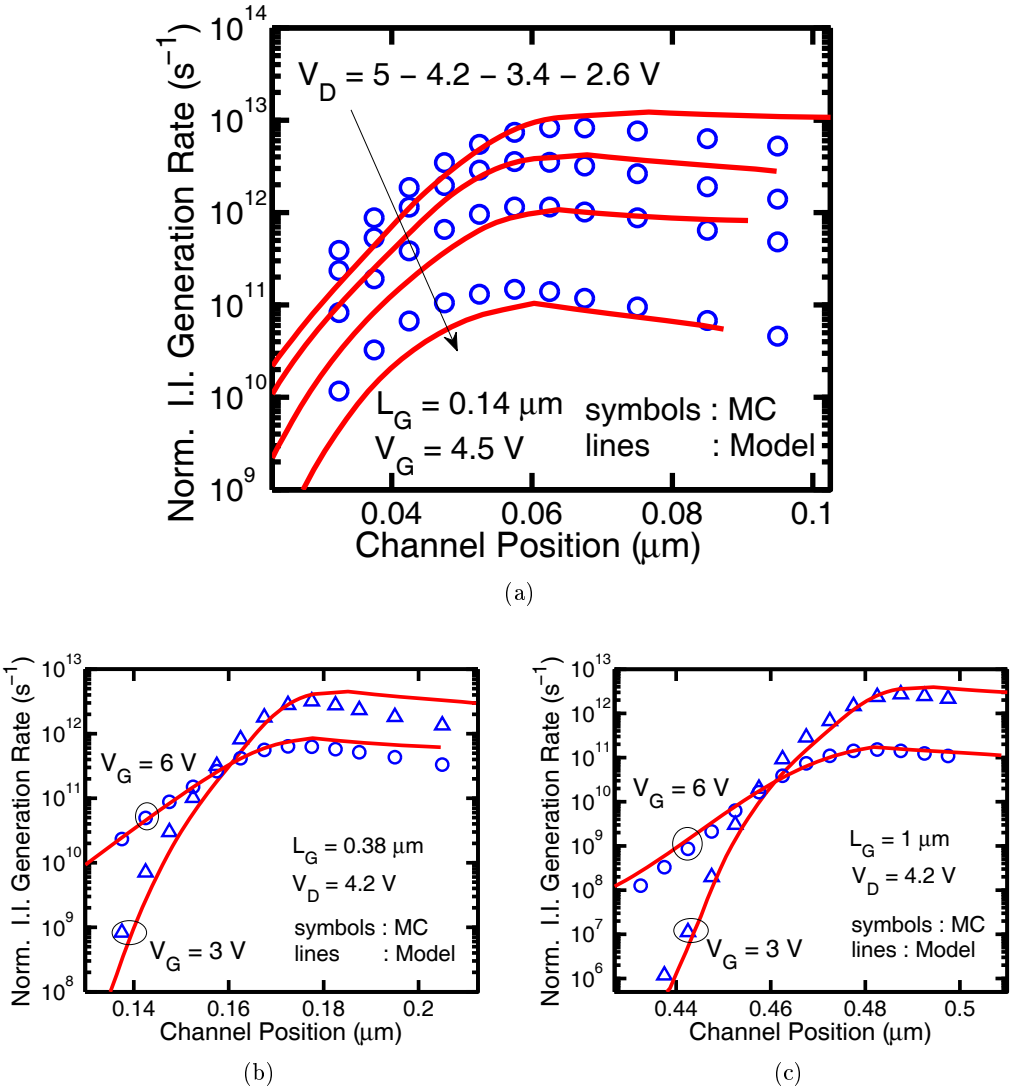
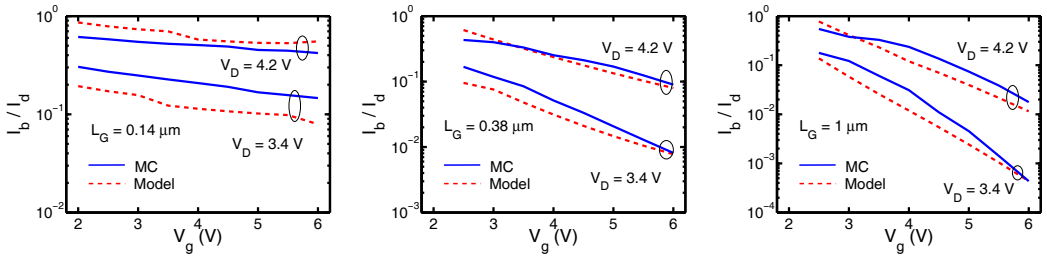


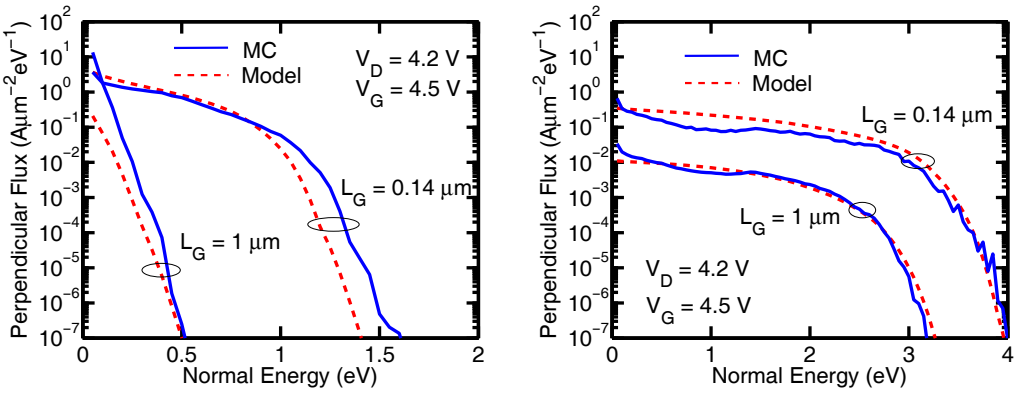
Figure 3.18: Comparisons between normalized impact ionization generation rates along the channel at a depth of $5A$ from the interface calculated from the Model (lines) and the Monte Carlo (symbols) for different gate lengths: $0.14\mu m$ (a), $0.38\mu m$ (b) and $1\mu m$(c). The normalizations are performed using Equation 3.22 and the same variations as in Figure 3.17 are presented.

drain where the injection occurs, a shape and amplitude in very good agreement with the reference FBMC is predicted by the model. This is due to the fact that on one hand, the particles number in this region is acceptably well captured by the model and on the other hand, their distribution in total energy was shown to be quite in-line with that predicted by FBMC. The smooth shape of the normal flux, introduced by the continuous summation over the energies, does not reflect all the features of the simulated FBMC flux. However, the discrepancy stays within an



(a) (b) (c)

Figure 3.19: Comparison between the normalized bulk current (I_b/I_d) vs. gate voltage obtained after Monte Carlo (MC) and the Model for two drain voltages (3.4 and 4.2V) and for three gate lengths: 0.14μm - a, 0.38μm - b, 1μm - c.



(a) (b)

Figure 3.20: Comparison between the perpendicular fluxes as a function of the normal energy obtained with Full Band Monte Carlo (MC) and the Model for two gate lengths (0.14 and 1μm) at a constant bias condition ($V_D/V_G = 4.2/4.5$). The fluxes are shown for the mid-channel (a) and channel/drain junction (b) positions.

acceptable interval.

The agreement between the model and FBMC confirms that the distribution functions are indeed highly isotropic due to randomizing scattering events taking place throughout the electron's path. As a matter of fact, the majority of the carriers scatter at least once in their trajectory for the gate lengths considered in this work. This validates the approach (subsection 3.1.3) considering isotropic distribution functions while locally considering ballistic probabilities.

The gate current (I_g) is finally calculated using Equation 3.20 which includes an integration of the gate current density (J_g) over the gate length. Figure 3.21 reports the gate current density along the channel and the injection efficiency as a function of the gate voltage for all the investigated devices. Similarly to the ionization efficiency, the injection efficiency is the unit-less ratio given by I_g/I_d.

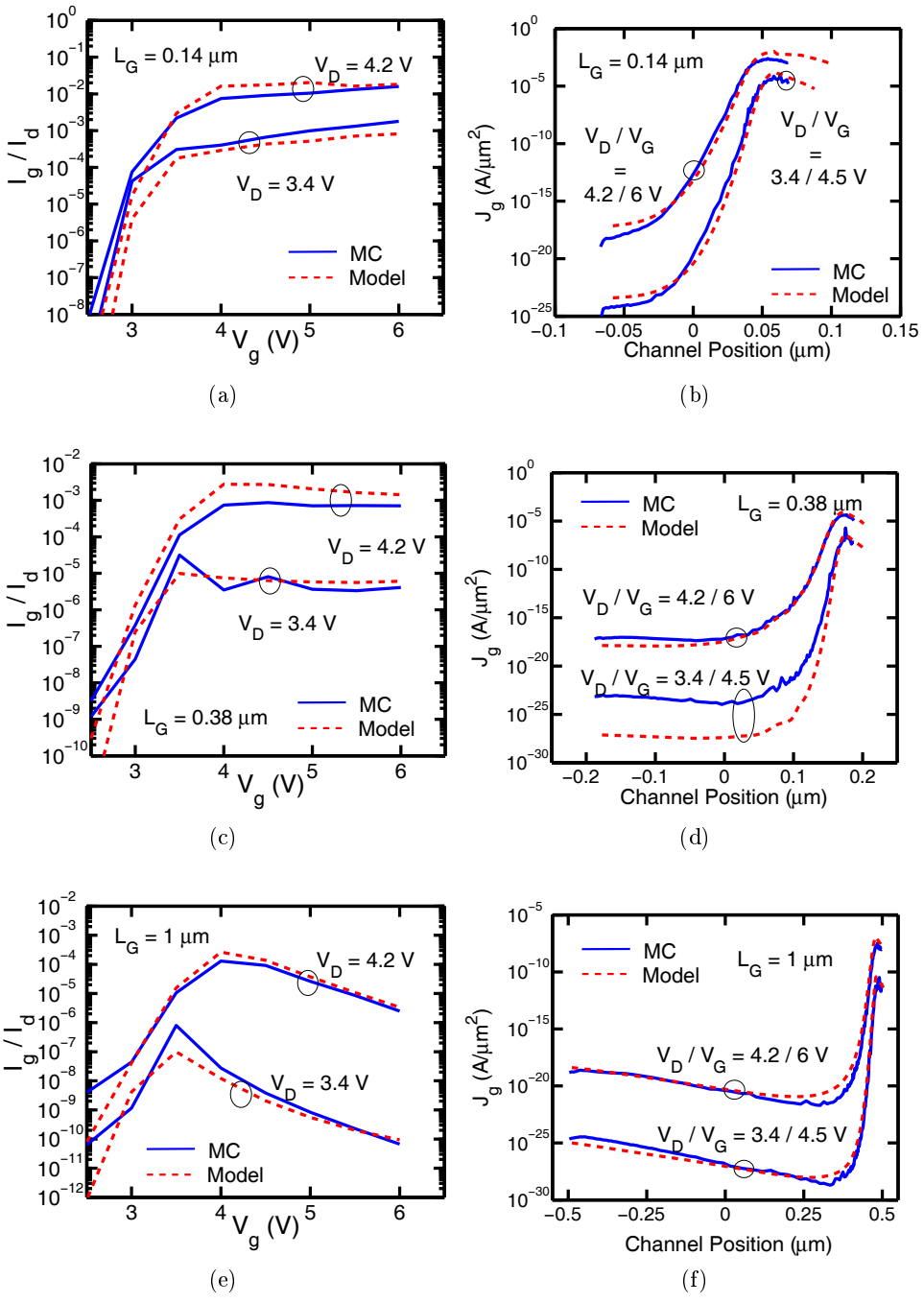


Figure 3.21: Comparison of the injection efficiencies (I_G/I_D) vs. gate voltage (a, c, e) and gate current densities along the channel (b, d, f) obtained with the Monte Carlo (MC) and the Model for different gate lengths and bias conditions.

The injection efficiency curves are composed of two parts which clearly reflect the distribution function's shape. For $V_g < V_d$ a strong exponential trend is ob-

served. On one hand, Figures 3.15, 3.16 show that the end of the 'plateau' at the metallurgical junction is situated around V_d. On the other hand, the electrons are subject to a repulsive electric field in the oxide which increases the potential barrier from ε_B to $\varepsilon_B + V_{dg}$. Therefore, under such bias conditions, the injected carriers come from the exponential tail. Hence, when V_g increases in this regime, the number of available carriers for injection increases exponentially.

When $V_g > V_d$, the carriers come from the 'plateau', while the tail of the distribution is essentially negligible. The injection efficiency expectingly increases with V_d. However different trends are observed for increasing V_g as a function of the the gate length. As a matter of fact, the classical *bell shape* efficiency, which was observed in older technologies [Eitan 1981], is observed for the longest device. Instead, a constantly increasing efficiency with increasing V_g is predicted for the smallest gate lengths. Although the lateral electric field at the drain side is reduced due to the V_g increase, its effect on the injection is over compensated by an increase of the lateral field in the middle of the channel [Cappelletti 1999]. The antagonistic effects seem to cancel out for the intermediate gate length where a rather stable injection efficiency is obtained.

In the proposed comparison, both the 'plateau' and the 'exponential tail' of the distribution functions are tested. A good match between the model and the FBMC is obtained for all devices and investigated biases. This is also confirmed by the current density along the channel, which closely follows the FBMC-predicted injection, by especially well capturing the injection peak position in addition to its amplitude. In terms of injection efficiency, the model stands within a factor of 3 from the FBMC. Despite the limitations and the uncertainties already discussed, our 1D Full-Band Non-Local Multi-Stage Probabilistic approach captures the main features of the hot carrier injection with good accuracy.

3.2.4 Conclusions

In this section, the developed 1D approach has been extensively compared with full band Monte Carlo simulations on a full range of gate lengths and bias conditions. The full chain from drain to gate current with its numerous intermediate steps has been closely investigated. Although the macroscopic transport quantities are non-negligibly affected by the simplified treatment of the cold carriers, the model proves to be very accurate near the drain in terms of distribution functions and perpendicular fluxes. The latter are used to calculate both the ionization and injection efficiency. A remarkable matching with Monte Carlo results at a current scale has been achieved, thus demonstrating the effectiveness of this approach to model hot carrier effects.

3.3 Analysis of the main features

The previous sections were dedicated to the description of our approach to hot carrier modeling and its evaluation. The objective of this section is to give a critical insight on the main features of the model which allowed us to achieve the discussed results. Hence, the relevance of the band structure description is first discussed in subsection 3.3.1. The same subsection contains as well the discussion on the non-locality, due to a close relation between these two aspects of the model. Subsection 3.3.2 lastly covers the role of the backscattered carriers in this multi-step approach.

3.3.1 The impact of band structure

The choice of the silicon band structure for transport simulation has a considerable impact on the model. Indeed, once the dispersion relation is chosen, the group velocity and the scattering rates are immediately calculated. The latter quantities are then used as inputs (Figure 3.1). The previous section described the model where a full band structure and optical phonon scattering were introduced. In this subsection, we examine the impact of a different description of the band structure taking into consideration the main figures of merit previously introduced. The first part of the subsection deals with analytical band structures, while in the second part of the section traditional local approximation is analyzed as a limit case when parabolic bands are used.

3.3.1.1 Analytic dispersion relations

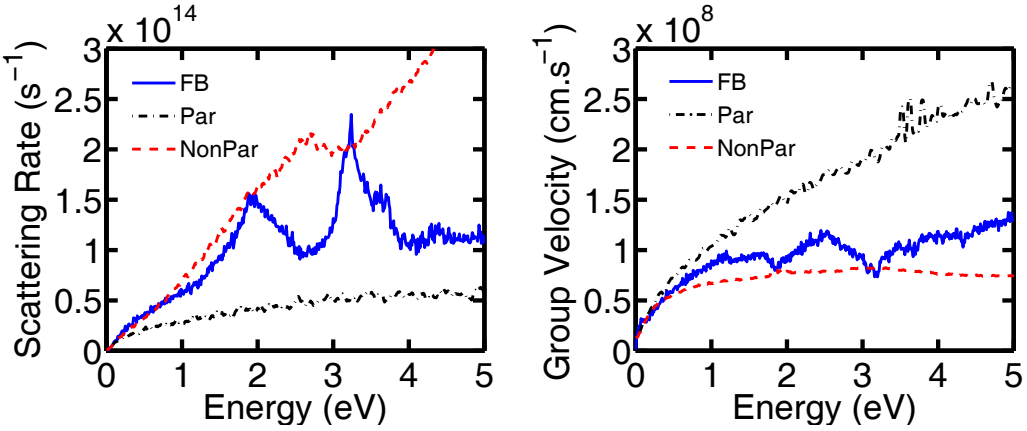
Among the most widely used band structures for silicon, the parabolic and non-parabolic ones certainly occupy a special place (c.f. Chapter 1). For convenience, below are reported their analytic expressions with quantities having their usual meaning.

$$\varepsilon = \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z}\right) \tag{3.23}$$

$$\varepsilon(1+\alpha\varepsilon) = \frac{\hbar^2}{2}\left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z}\right) \tag{3.24}$$

Figure 2.2 of Chapter 2 compares these expressions with the numerical full band structure along the three main high symmetry axes of the irreducible edge of the First Brillouin Zone (FBZ).

The group velocity and the density of states are then calculated, with the latter being used to compute the scattering rates with the optical phonons. Finally, an integration over the equi-energy surfaces is performed in order to build the tables of the group velocity and scattering rates as function of the kinetic energy. These quantities are reported on Figure 3.22.



(a) (b)

Figure 3.22: The scattering rates with optical phonons (a) and the electron group velocity (b) obtained from a full-band (FB), parabolic (Par) and non-parabolic (NonPar) conduction band description.

Lower scattering rates and higher group velocities are obtained with the parabolic expression, while closer values with respect to the full band description are obtained for the non-parabolic approximation. The energy dependent ratio of both quantities is closely related to the probability for a carrier to be ballistic (Equation 3.1). Figure 3.28 reports the latter probability for all the considered band structures in a specific case where the electron starts to accelerate near the mid-channel position in a 0.14 μm gate length device.

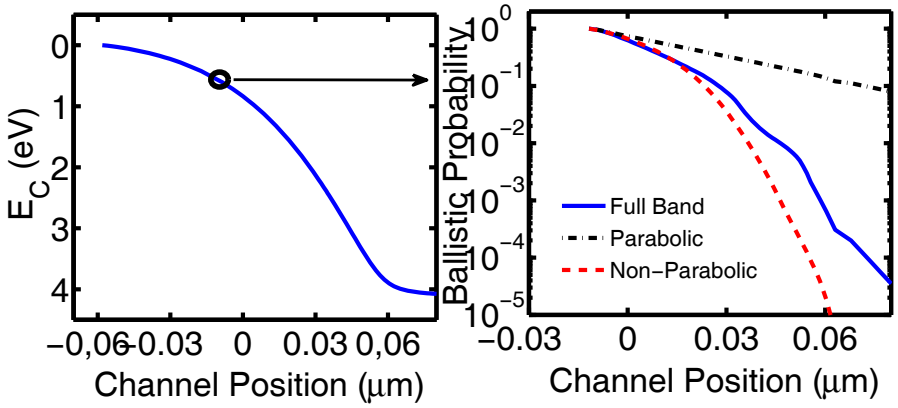


Figure 3.23: Non-local ballistic probabilities (right) obtained with the full-band (blue), parabolic (black) and non-parabolic (red) band structure for an electron starting at the middle of the channel (left).

The non-parabolic bands, compared to the parabolic ones, exhibit a better agree-

ment with the full bands. As a matter of fact, contrary to the other cases, parabolic bands show a straight exponential behaviour, which practically means that the carrier has the same probability to scatter regardless of its position in the system. This is a direct consequence of the square root dependence of both quantities on the carrier energy. The particular case of parabolic bands is treated in more details in the following part of this subsection.

The probability to be ballistic is then used to calculate the fluxes in the system. The described method and considerations made in the previous section are used to calculate the carriers densities along the channel (Figure 3.24) and their energetic distribution (Figure 3.25). Both quantities are compared with full band Monte Carlo simulations.

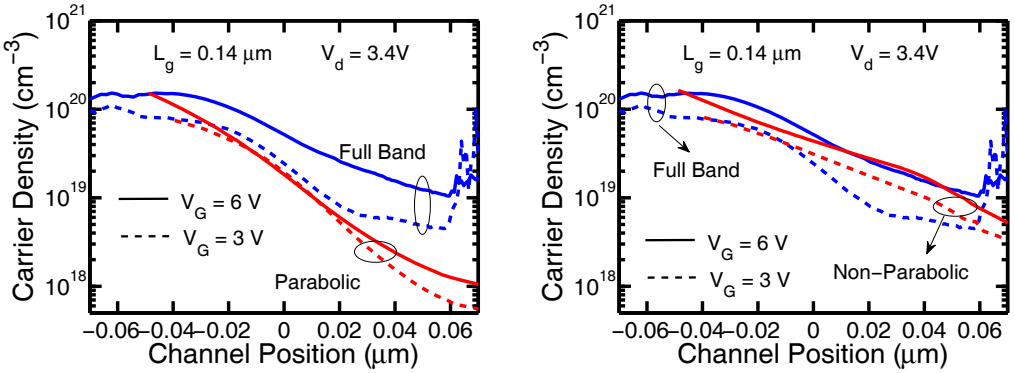
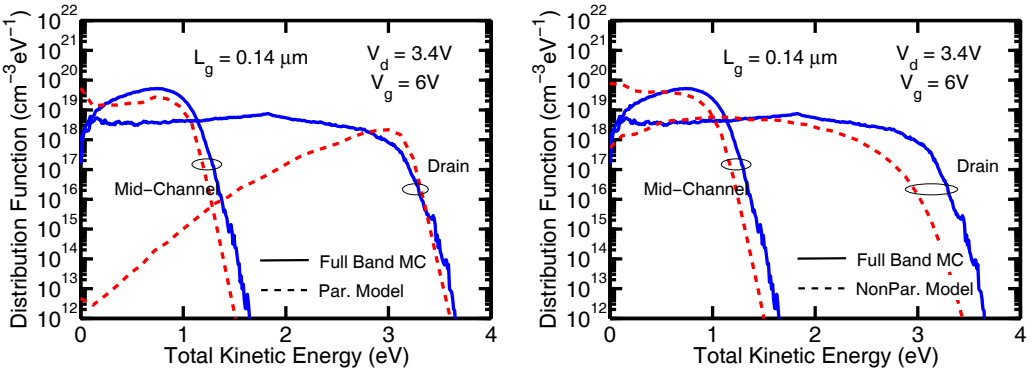


Figure 3.24: Carrier densities along the channel obtained with full-band Monte Carlo and the Model incorporating the parabolic (a) or the non-parabolic (b) band structure. Two gate voltages are shown ($V_G = 3/6V$) at a constant drain voltage $V_D = 3.4$.

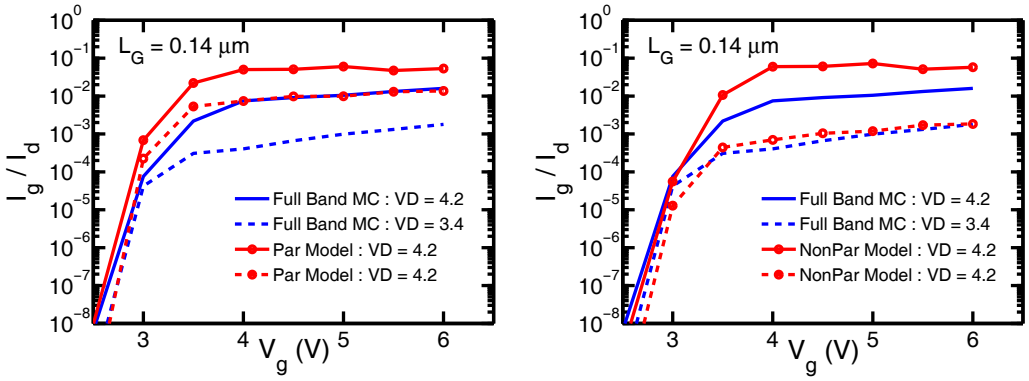
On one hand, the parabolic bands show lower carrier density along the channel, and in particular close to the channel/drain junction. On the other hand, the predicted distribution at the junction position shows an increased concentration of the carriers at high energies. Both results are in agreement with the high group velocity which reduces the carrier concentration (Equation 3.13) and the low scattering rate which increases the carriers' ballistic population (Figure 3.23). In the same perspective, the non-parabolic bands are quite close to the full band solution. This situation is also observed in the injection efficiencies reported on Figures 3.26, where a relatively good matching is observed for the non-parabolic bands, while rather different quantitative results are obtained for their parabolic counterpart.

In conclusion, the classical non-parabolic band structure, carefully extended for higher energies, turns out to be a good approximation for hot carrier injection modeling in the investigated bias conditions. The simpler parabolic case however, is not sufficient to capture the main elements of this phenomenon, due to very different scattering rates and group velocity with respect to a realistic description.



(a) (b)

Figure 3.25: Electronic distribution functions in total kinetic energy obtained with full-band Monte Carlo and the Model incorporating the parabolic (a) or the non-parabolic (b) band structure. Mid-channel and drain distributions are given for the 0.14 μm (a) gate length device.



(a) (b)

Figure 3.26: Comparison of the injection efficiencies as a function of the gate voltage obtained with full band Monte Carlo (MC) and the non-local Model incorporating parabolic (a) and non-parabolic (b) conduction band structures. Two drain voltages, 3.4 V (*dashed*) and 4.2 V (*plain*)) are shown for the 0.14 μm (a) gate length device.

3.3.1.2 Parabolic bands and local expression

One of the key features of this approach is the introduction of non-local expressions when calculating the scattering probabilities along the channel. The previous part of the subsection showed that the employed band structure has a significant impact on the distribution functions and the injection efficiencies. These quantities are affected by the choice of the bands which completely determine the ballistic (or scattering) probabilities a carrier is subject to. The scattering probability is calculated using the non-local expression (Equation 3.1) which assigns different el-

ementary ballistic probabilities to carriers at different energies. This results in a non-Maxwellian variation of the ballistic probability along the channel (Figures 3.3 and 3.23) for the full and non-parabolic band structures. However, a Maxwellian ballistic probability was obtained in the case of parabolic bands (Figure 3.23). Such a behaviour can also be predicted by the local probability expression used in the classical LEM approach [Tam 1984]:

$$P_{0\to x} = \exp -x/\lambda_{op} \tag{3.25}$$

where λ_{op} represents the mean distance travelled by a carrier between two successive interactions with optical phonons and hence called the *mean free path*. From Equation 3.1, λ_{op} is mathematically defined as:

$$\frac{1}{\lambda_{op}} = \frac{1}{\varepsilon}\int_0^{\varepsilon} \frac{SR_{op}(\varepsilon')}{v(\varepsilon')}\mathrm{d}\varepsilon' \tag{3.26}$$

A closer examination of the mean free path as a function of the carrier energy, obtained from Equation3.26, is reported in Figure 3.27 for all the investigated band structures.

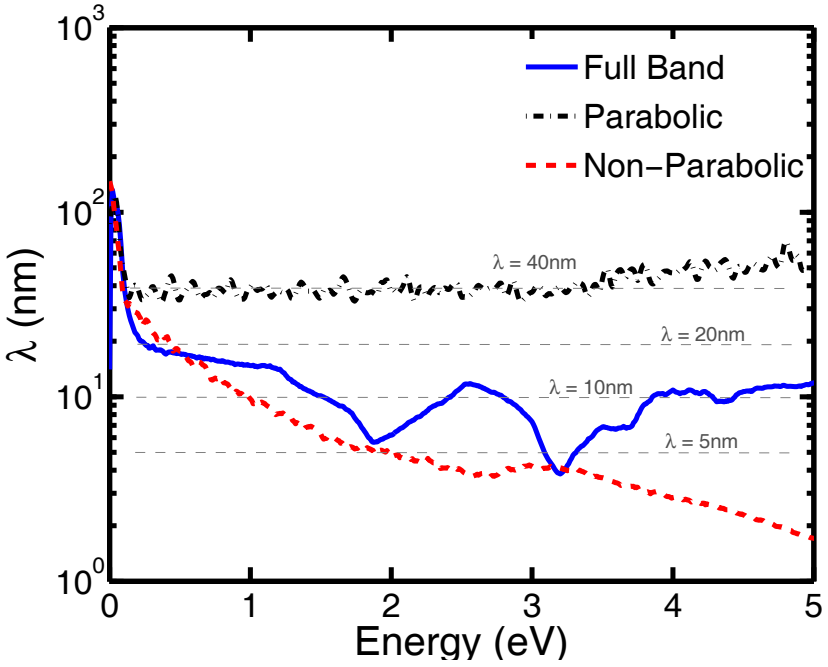


Figure 3.27: Optical phonon mean free path as a function of the carrier energy obtained after full, parabolic and non-parabolic band structure description.

First of all, the full band description shows that the mean free path varies over more than one order of magnitude in the considered range of energies and that this trend is qualitatively the same which is observed when adopting the non-parabolic description. On the contrary, a rather flat mean free path is obtained for parabolic

bands. These differences justify the non-Maxwellian and Maxwellian behaviour of the ballistic probability shown in Figure 3.23. Therefore, the use of a non-local relation when employing a parabolic band structure does not insure a realistic behaviour as the scattering rates and group velocity of the latter intrinsically limit the usefulness of introducing non-locality in the model.

Furthermore, Figure 3.27 shows that the parabolic approximation predicts a constant mean free path of around 40 nm, while for the relevant energies the other bands predict a mean free path in the 5-20 nm interval. To study the impact of such discrepancy we adopt a non-local calculation (Equation 3.1) of the ballistic probabilities according to three values of the mean free path. In order to insure a consistent calculation of the carriers concentration, distribution functions and gate current, a constant group velocity has been retained equal to 10^8 $cm.s^{-1}$, representative of the relevant energy interval (Figure 3.22). This in turn implies a constant scattering rate varying according to the mean free path value. Note that a different choice has been made in [Jungemann 1996] for the purpose of a conceptually similar comparison. There, the authors have fixed the scattering rate towards integrating the *Lucky Electron Model* approach in the Boltzmann Equation.

Figure 3.28 shows the ballistic probabilities along the channel obtained for the different mean free path values.

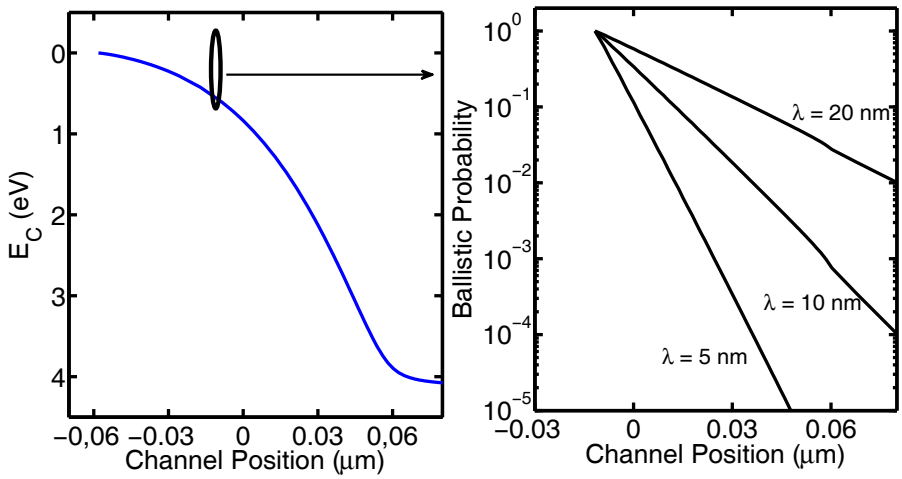


Figure 3.28: Non-local ballistic probabilities (right) obtained with different constant mean free paths for an electron starting at the middle of the channel (left).

The expected Maxwellian behaviour with different slopes are obtained. The increase of the mean free path implies less scatterings in the channel and therefore an increased ballistic probability is obtained. A direct consequence of the latter probabilities is shown on Figure 3.29. The distribution functions obtained with

constant mean free paths can be quite different from the Monte Carlo reference. When too many interactions are present in the system, because of a small mean free path, fewer electrons are found at high energy. The vice-versa holds also true. In this case, a constant mean free path of 10 nm seems to be a good compromise as it well approximates the energy-dependent mean free path case. It has to be emphasized that the best fitting constant λ_{op} is a strong function of device length, bias and sometimes position along the channel so that adopting a constant λ_{op} would have very limited predictive ability.

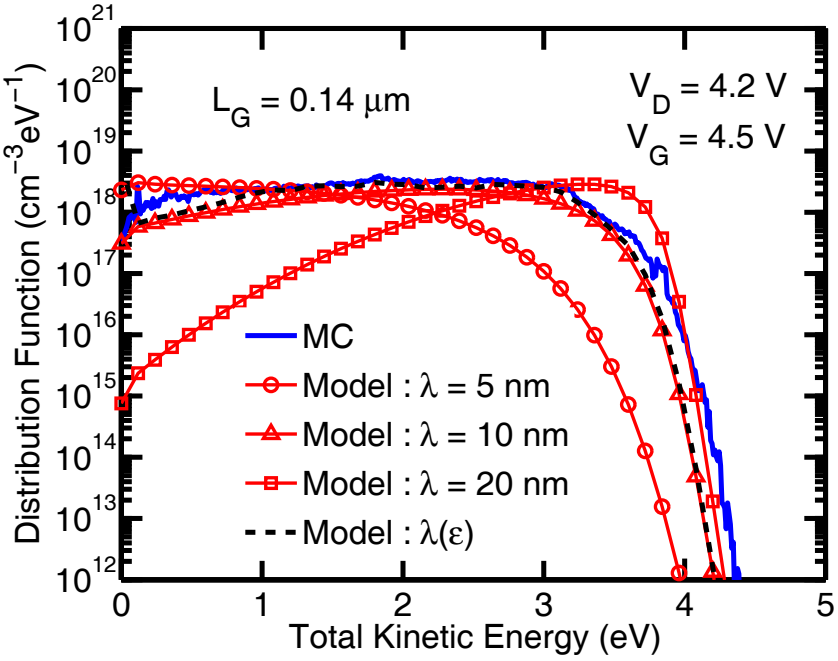
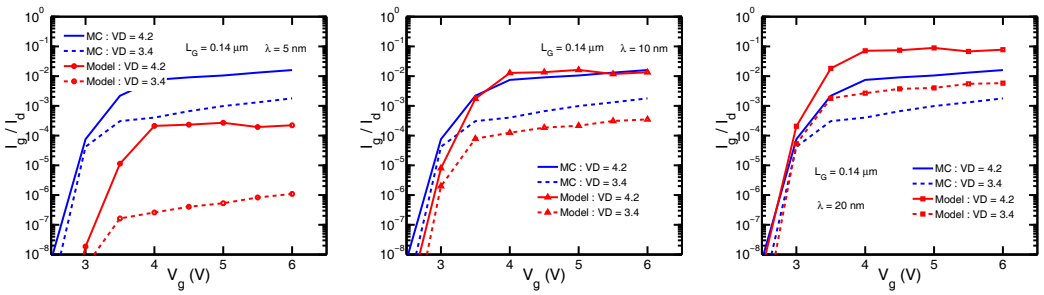


Figure 3.29: Distribution functions obtained with full band Monte Carlo (MC) and the Model featuring different mean free paths. The full band version of the Model including an energy-dependent mean free path is also given for comparison.

The injection efficiency calculated using the same λ_{op} values are reported in Figure 3.30. Although the shape of the efficiency curves is approximately the same in all models, the amplitude can be quite different. As expected from the distribution functions, the smaller the mean free path, the lower the gate current is. Moreover, the V_d dependence of I_g/I_d decreases with increasing mean free path.

Overall, in this special case the proposed semi-analytic model can be reasonably employed by using a constant mean free path of around 10 nm and an electron group velocity of 10^8 $cm.s^{-1}$. Such an approach is equivalent to a parabolic band structure approximation with adjusted electron effective masses.



(a) (b) (c)

Figure 3.30: Comparison of the injection efficiency vs. gate voltage for two drain voltages obtained with full band Monte Carlo (MC) and the Model featuring mean free paths of 5, 10 and 20 nm respectively shown on (a), (b), (c).

3.3.2 The role of the backscattering

The model described in the previous section included all the possible paths the carriers can follow in two dimensions (c.f. Figure 3.5). In particular the backscattered carriers moving in the drain to source direction were introduced (fluxes e, f). In the purpose of explicitly determining their impact in the model and in the perspective of considering simpler and faster versions of the model, the main figures of merit of the system without the e, f fluxes are considered. The system and the flux relations become:

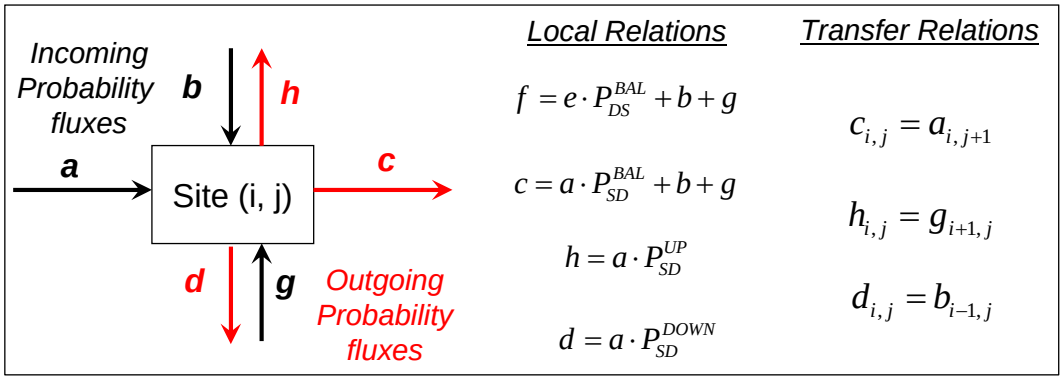


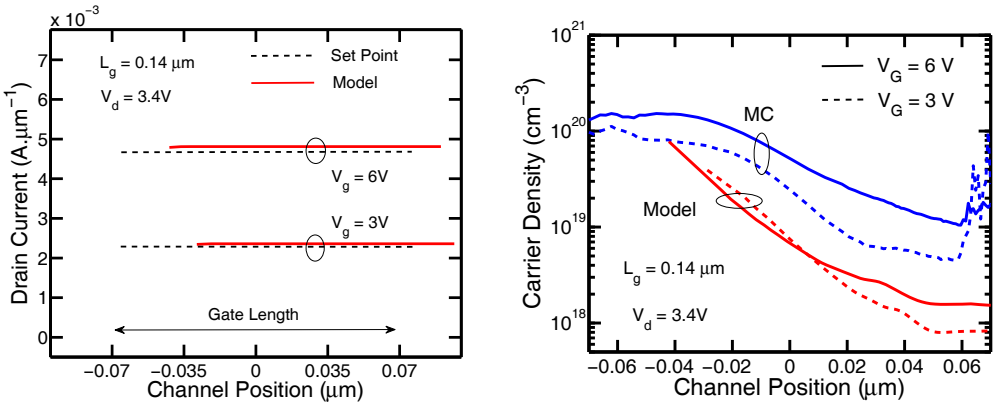
Figure 3.31: The simplified system with only source to drain and energy-exchange fluxes and their relations.

The same notations as in Figure 3.5 are used. The local relations are modified accordingly in order to insure the local flux conservation at each site:

$$a + b + g = c + d + h \tag{3.27}$$

The current along the channel, now composed only by the summation of flux c over the energy, is presented in Figure 3.32 *a*. The set point current is rigorously

reproduced. This result is slightly better than the one obtained in the full-flux case, as the boundary conditions at the barrier are easier to treat, i.e: all the carriers go from source to drain.



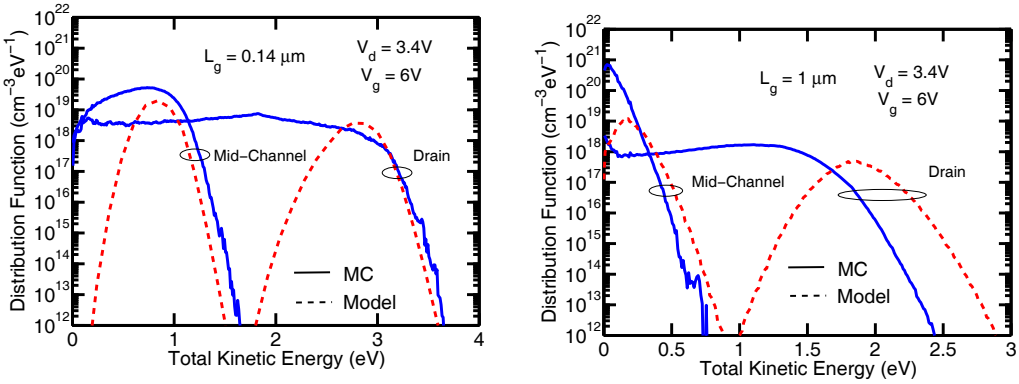
(a) (b)

Figure 3.32: Drain currents (a) and carrier densities (b) along the channel at two gate voltages ($V_G = 3/6V$) at a constant drain voltage ($V_D = 3,4V$).

Figure 3.32 *b* shows the carrier density along the channel calculated using the same method as the one described in the previous section. The predicted carrier density is well below the MC results, especially in the zone of interest near the channel/LDD metallurgical junction. The same result is obtained for all the gate lengths and bias conditions,

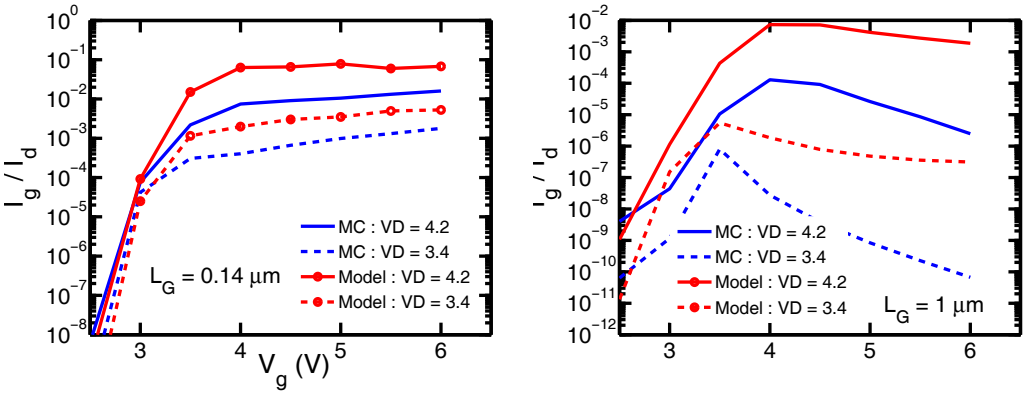
The current conservation in conjunction with a smaller carrier concentration along the channel would naturally lead to think to a distribution function mainly concentrated at high energy for which the group velocity is higher. This is confirmed by Figure 3.33 which shows the distributions obtained for the shortest and longest device at mid-channel and junction positions. With no backscattering included in the system, the low energy part of the distributions is almost completely depleted of carriers. Moreover, the high energy part is particularly peaked, with high values reached at high kinetic energy; the differences with the reference MC results are increased for the longest cell. Hence, the role of backscattering is crucial in redistributing the electrons toward low energies. Additional low energy paths for electrons are thus necessary towards a realistic description of transport. This is especially true for relatively long devices in which most of the electrons are in close to equilibrium conditions in a large fraction of the channel.

Finally, as expected from the distribution functions, although the qualitative trends are correctly captured the injection efficiencies in Figure 3.34 are higher than the FBMC results. A length-dependent correction factor should be introduced in case one wants to use such an approach which would hamper the predictive ability of such calculation.



(a) (b)

Figure 3.33: Electronic distribution functions in total kinetic energy obtained with full-band Monte Carlo and the non-local Model without the backscattering events, for 0.14 μm (a) and 1 μm (b) gate length devices. Mid-channel and drain distributions are shown for a particular bias condition.



(a) (b)

Figure 3.34: Comparison of the injection efficiency as a function of the gate voltage obtained with Monte Carlo (MC) and the non-local Model without backscattering events. Two drain voltages, 3.4 V (*dashed*) and 4.2 V (*plain*), and two gate-lengths, 0.14 μm (a) and 1 μm (b), are presented.

3.3.3 Conclusions

This section investigated the importance of the main features included in the new probabilistic model, namely: the *full-band* description, the *non-local* probability calculation and the *multi-step* processes in the presence of backscattered carriers. The first two features have been shown to be closely related to each-other via the scattering rates and the group velocity. The full or the non-parabolic band description should be used in conjunction with a non-local energy-dependent probability

calculation, whereas the parabolic approximation was shown to be equivalent to the constant mean free path case. In addition, it was shown that the absence of backscattered carriers in the system leads to hotter distribution functions. Thus accounting for complex and multiple injection paths is indeed at least as important as accounting for a realistic band structure.

These results show the intrinsic limits of the original LEM approach proposed by different authors such as, for instance [Hasnat 1996]. The obtained currents and distributions should always be empirically corrected as a function of the gate length or bias condition. As a consequence, this investigation justifies the inclusion of the above-mentioned features as mandatory ingredients towards a reliable hot carrier effects modeling.

3.4 Additional scattering mechanisms

The previous sections showed that optical phonons scatterings as well as a realistic band structure and carrier movement description, constitute essential components for hot carrier modeling. In addition to the phonons, the carriers are known to scatter via other mechanisms which also contribute to shape the distribution function. In this section, the inclusion of the Electron-Electron Scattering (subsection 3.4.1) and Impact Ionization (subsection3.4.2) as additional scattering mechanisms in the proposed modeling approach is presented. These mechanisms were turned off in the MC simulations shown previously.

3.4.1 Electron Electron Scattering

The implementation of the new scattering mechanism is first presented in paragraph 3.4.1.1 where special care is taken to explain the assumptions made towards a simplification of the simulation procedure. Paragraph 3.4.1.2 then discusses the results in terms of the main figures of merit and the range of validity of the proposed implementation.

3.4.1.1 Scattering rates and implementation

The inclusion of the electron-electron interaction is a complicated task in any transport model. The fundamental difficulty resides in the fact that the Boltzmann Transport Equation becomes non-linear. In order to calculate the collision rate of two specific electrons, their distribution function should be previously known, but at the same time the outcome of the collision affects the distribution function itself. Furthermore, the calculation of the scattering rates, is not a trivial task as it involves integrations over the state of the two involved particles before and after the interaction.

In the proposed semi-analytic model, an analytic formulation of the electron electron scattering rates has been used after Ferry [Ferry 1999]. The authors in the study have separated the total interaction into *energy loss* and *energy gain* processes, which can be readily implemented. For convenience, below is reported the derived energy gain or loss scattering rates corresponding to equations 8/9 of [Ferry 1999]:

$$\Gamma_{ee}^{a/e}(k) = \frac{nmq^4}{4\pi\varepsilon_{Si}^2\hbar^3 k}\left(\frac{m}{2\pi k_B T_e}\right)^{1/2}\int_{w_1}^{w_2} d\omega \int_{-\zeta}^{+\zeta}\exp\left[-\frac{\hbar^2}{8mk_BT_e}\left(q \pm \frac{2m\omega}{\hbar q}\right)^2\right]\frac{dq}{\left(q^2+q_D^2\right)^2} \quad (3.28)$$

where the + and − signs in the exponential correspond to the absorption (energy gain) and emission (energy loss) processes, respectively. In this expression, a given

primary electron with momentum k exchanges an energy w and changes its initial momentum by ζ, isotropically taken in the interval:

$$-\zeta = \sqrt{k^2 + \frac{2m\omega}{\hbar}} - k < \zeta < \sqrt{k^2 + \frac{2m\omega}{\hbar}} + k = +\zeta \tag{3.29}$$

The upper integration limit of the energy w_2 is ∞ in the case of absorption and $\varepsilon(k)/\hbar$ in the case of emission, while w_1 is zero for both cases. The carrier population is defined by its total density n and its mean temperature T_e. The other quantities in the expression have their usual meaning. Finally, the Debye length is defined as:

$$q_D = \sqrt{\frac{n\zeta^2}{\varepsilon_{Si} k_B T_e}} \tag{3.30}$$

Equation 3.28 has been derived by assuming:

- an isotropic collision mechanism
- a parabolic dispersion relation
- a heated Maxwellian distribution function : $n(\varepsilon) \propto \exp\left(-\varepsilon/k_B T_e\right)$

The *a priori* assumption of such a distribution allows the authors to derive simple closed-form expressions which only depend on the carrier's density and mean temperature. Both these scalar quantities need to be known in order to use the above expressions. n and T_e are intrinsically related to the distribution function, hence they are position-dependent. For a given device length and bias condition, it is not possible to universally derive an expression for both quantities. Hence an iterative scheme has been considered in this thesis, as depicted in Figure 3.35.

The first iteration is performed with the model including only the Optical Phonons, as described in section 3.2. This allows to extract the carrier density n (in cm^{-3}) along the channel using Equation 3.13, and the mean temperature using the following expression:

$$T_e(x) = \frac{2}{3k_B} \frac{\sum_0^{\varepsilon_{max}} \varepsilon_i n(\varepsilon_i, x)}{n(x)} \tag{3.31}$$

Next, the position-dependent electron-electron scattering rates can be calculated. As both energy-exchange processes (emission and absorption) are continuous functions of the exchanged energy w, a considerable number of the exchanged energies ought to be considered in the system. This would practically mean that a given node of the system (Figure 3.2) is coupled with every other node of the same column due to all the possible energy exchanges. Keeping track of all the possible exchanges in energy would significantly increase the computational burden and thus render the approach less usable. Hence, simplifications are imposed for the implementation of the electron electron scattering interaction during the second iteration of the model (Figure 3.35).

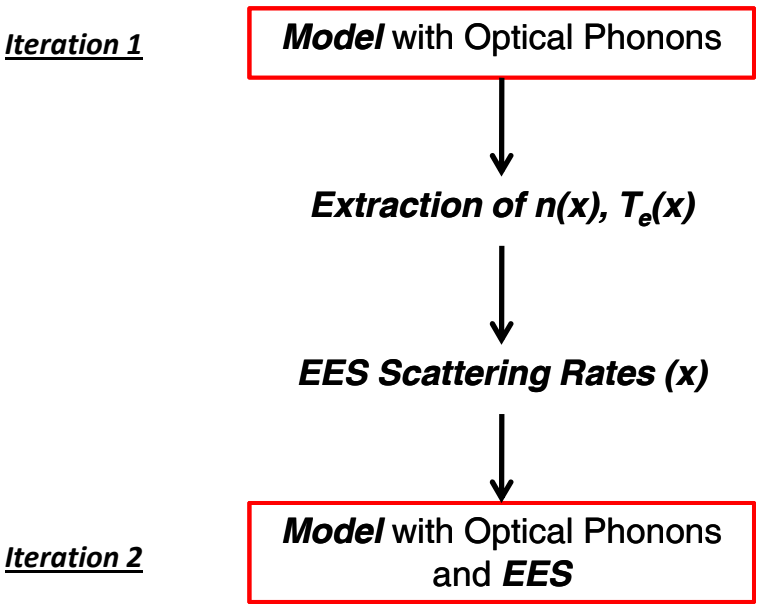


Figure 3.35: Two-step simulation process for the inclusion of the electron-electron scattering (EES) mechanism in the semi-analytic Model.

The careful examination and analysis of the previously published distribution functions including EES effect [Childs 1996], [Abramo 1996], [Ferry 1999], [Ghetti 2002], [Fixel 2008] has led to the following considerations. While EES has a clear effect on the exponential tail of the distribution functions at high energies, it has no or negligible impact on the core of the distribution in the plateau or the low energy part. This clearly means that this interaction affects only a small number of electrons. This is translated in small scattering rates with respect to phonon scattering, as it will be shown in the following paragraphs. Thus, from the modeling point of view, what seems important is to populate the exponential tail of the distribution rather than cool down those electrons which have lost a part of their energy; the latter fluxes will not be considered at all. The picture of the fluxes at each node, becomes:

In addition to the optical phonon related fluxes already described in section 3.2, the picture includes flux s which represents the carriers reaching the considered node after having gained a given energy due to EES interaction. The k-index is the level of the energy exchange due to EES. In addition, as the scattering rates are orders of magnitude smaller with respect to those of the Optical Phonons, the same probability scheme (Figure 3.4) as the one described in section 3.2 including only Optical Phonons, is maintained for the case when EES is included in the system. Thus, the P^{EES_k} probabilities directly stem from the non-local expression of Equation 3.1 without any additional weighting with Optical Phonons. In this case the local conservation equation becomes:

$$a + b + e + g + s \cong c + f + h + d \tag{3.32}$$

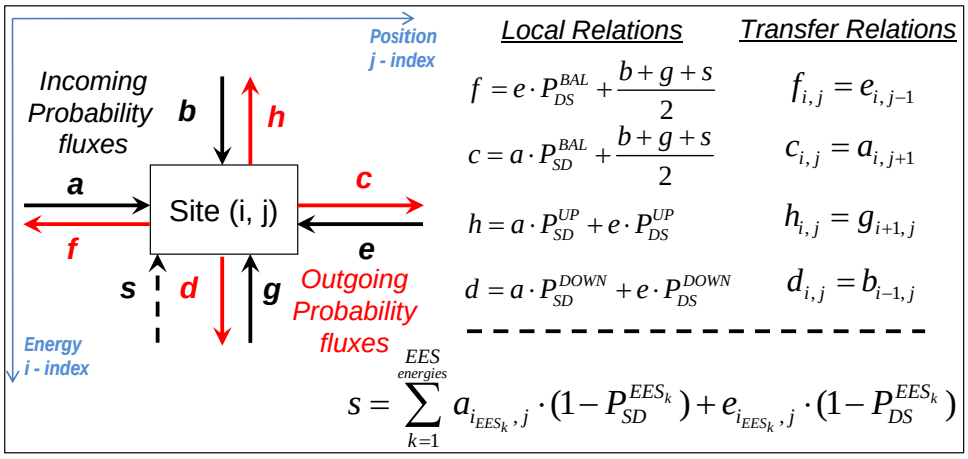


Figure 3.36: Schematic representation of the fluxes for each site and local and transfer relations when optical phonons and electron-electron scattering are included in the simulation. EES_k represents the k-th EES-exchange energy.

The above s-expression is a sum of carrier fluxes coming from different energies below the considered one. In this study, although an indefinite number of EES fluxes can be included, a compromise between accuracy and speed has been found while considering only six transition energies for EES, given in Table 3.1. Each energy exchange level is representative of an extended energetic interval whose bounding values are the w_1 and w_2 limits of Equation 3.28.

Energy (eV)	Interval(eV)
0.5	0.25 - 0.75
1	0.75 - 1.25
1.5	1.25 - 1.75
2	1.75 - 2.25
2.5	2.25 - 2.75
3	2.75 - 3.25

Table 3.1: The default exchange energies due to electron-electron scattering and their integration interval.

For instance, the scattering rate of a primary carrier gaining an energy included between 0.25 and 0.75 eV, calculated with Equation 3.28, is associated to a flux coming from 0.5 eV below the considered node. The same operation applies to the other energies. In this way an energy exchange due to EES interaction as high as 3 eV is included in the model. The energy exchanges smaller than 0.25 eV have been neglected as EES has a limited impact on the shape of the distribution when its energy exchange approaches the optical phonon energy (0.06 eV), the latter mechanism being much more efficient. The effect of the choice of these energy levels is presented and discussed in the following paragraph.

3.4.1.2 Results

The previous analysis on the effect of EES leads to many simplifications in terms of its implementation. As a consequence, several tests have been made to check the model accuracy and validity. First of all, we verified that the scattering rates are indeed well below the optical phonons' ones. Figure 3.37 reports the scattering rates as a function of the kinetic energy for the absorption of each of the six energies as defined in Table 3.1. In addition, the optical phonon scattering rates are given for comparison.

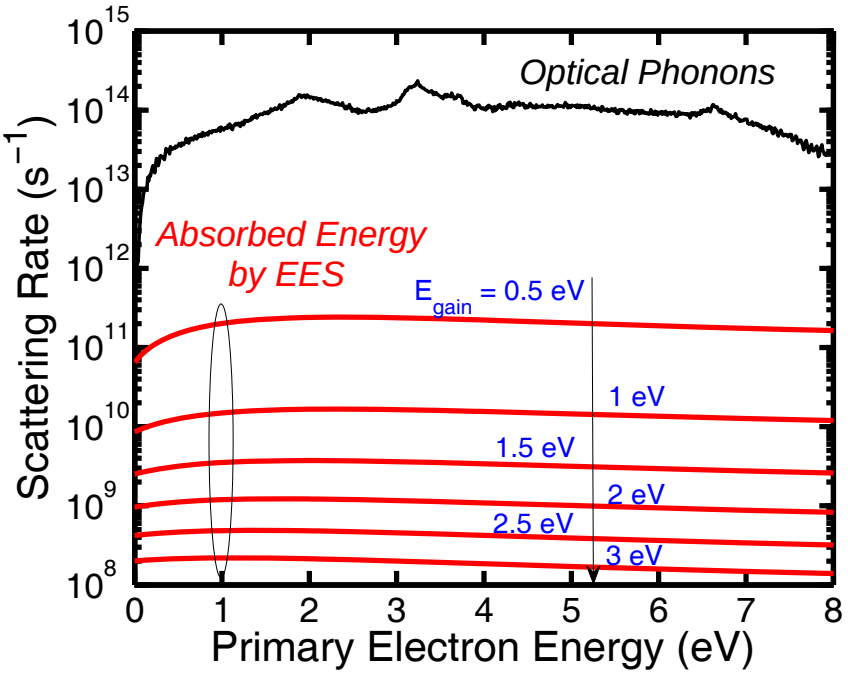


Figure 3.37: Scattering rates vs. carrier energy for Optical Phonon scattering and energy absorption by Electron-Electron Scattering (EES). The scattering rates of different EES absorbed energies from 0.5 to 3 eV are plotted from top to bottom.

It is noticeable that the scattering rates corresponding to the EES-related energy exchanges are much lower with respect to the standard optical phonon ones.. Such a big gap justifies the idea not to modify the probability scheme when considering EES interactions. Figure 3.37 also shows that the promotion of carriers to high energies (E_{gain} increasing) is globally less probable than the absorption of small energy quantities. These rates are calculated close to the channel/drain junction. Similar trends are obtained for all channel positions. However, the absolute value of the scattering rates depends on the carrier density and the mean carrier temperature at the considered position (Equation 3.28), so that slightly different results would be obtained at different positions.

The necessary inclusion of many EES exchange energies and the relevance of the choice made in this study is shown in Figure 3.38, where the distribution functions

obtained from our model and FBMC have been compared over more than fourteen orders of magnitude. The EES mechanism has been included in the FBMC simulations following the approach after [Ghetti 2002].

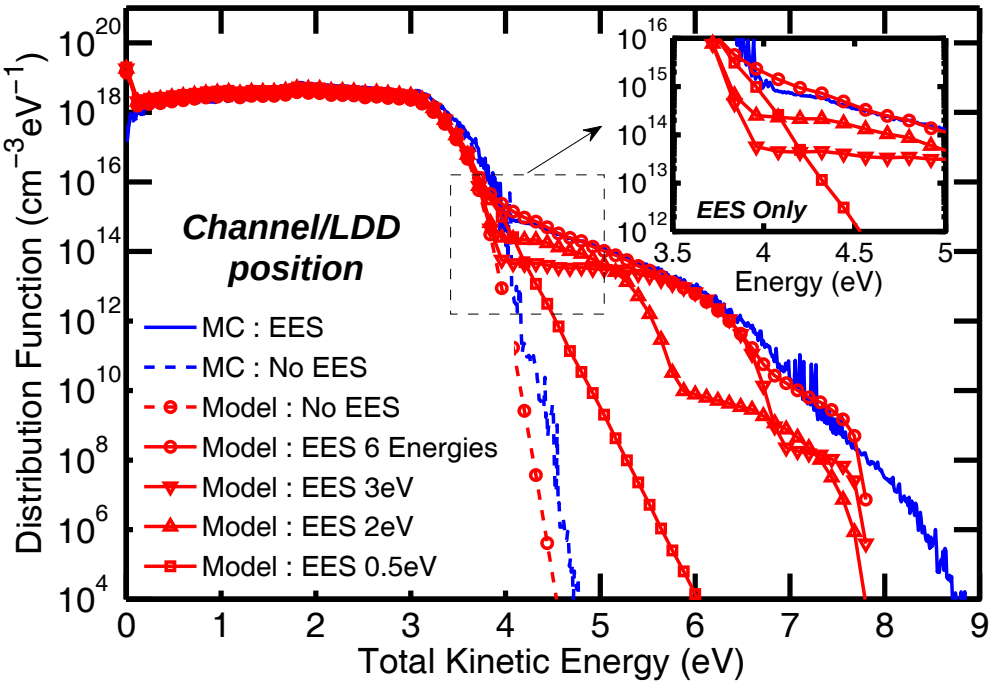
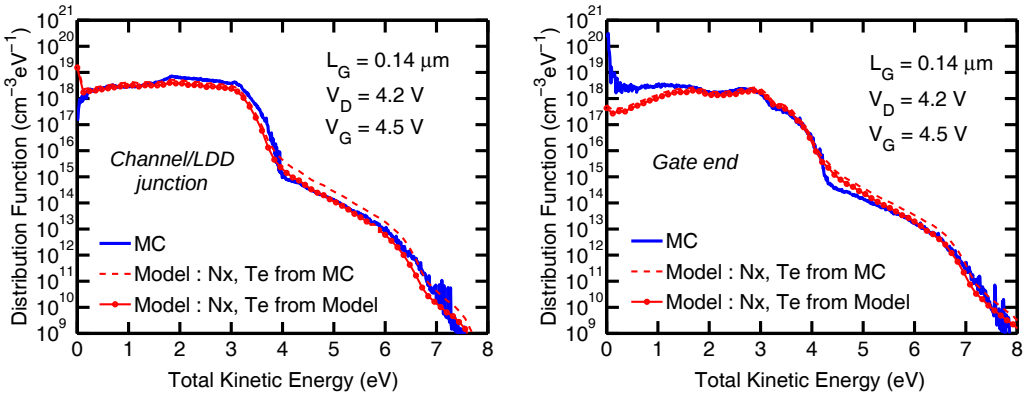


Figure 3.38: Distribution functions obtained with Monte Carlo (MC) and the Model with and without Electron-Electron Scattering (EES). Different EES configurations are represented for the Model. The distributions are taken at the channel/LDD junction for a $0.14\mu m$ gate length device biased at $V_d/V_g = 4.2/4.5$ V.

The previously discussed distribution function (section 3.2) calculated without EES has been reported as well (dashed lines) for comparison. The FBMC simulations clearly demonstrate the importance of the EES mechanism at high energy which completely modifies the usual Maxwellian tail. The results of the model featuring different EES setups are shown. The configuration which includes the six energies of Table 3.1 well captures the EES tail (circles). Three additional curves show the effect of single EES exchange energies. Small energy exchanges (0.5 eV) start to alter the tail at a lower kinetic energy but the modification is not so important at high energy (squares). Considering higher energy exchanges (3 eV), we observe that the tail is appreciably modified but the effect is appreciable only at higher kinetic energies (lower triangles). In order to correctly capture the shape and amplitude of the distribution function, a full range of energy exchanges should be included. This explains and justifies the choice made to include six different EES-related energy gain values.

In the above comparison, although the simulations apparently well reproduce

the MC results, it should be kept in mind that the proposed model involves several approximations. As already mentioned in the previous paragraph 3.4.1.1, the calculation assumes the knowledge of the total carrier density and of the mean carrier temperature along the channel, the latter being approximated as a heated Maxwellian distribution. Although section 3.2 showed the good aptitude of the model to capture the hot carrier population under very different conditions, the distribution of the cold carriers has been shown to be less accurately predicted. Unfortunately, the distributions as low energy affects both the carrier density and the mean carrier temperature, thus increasing the uncertainty related to the accuracy of the EES evaluation. In order to reduce the sources of uncertainty, the MC-predicted carrier density and mean carrier temperature have been used for EES evaluation in test simulations that thus bring the focus only on the EES calculation framework. Figure 3.39 reports the distribution functions computed according to this assumption.



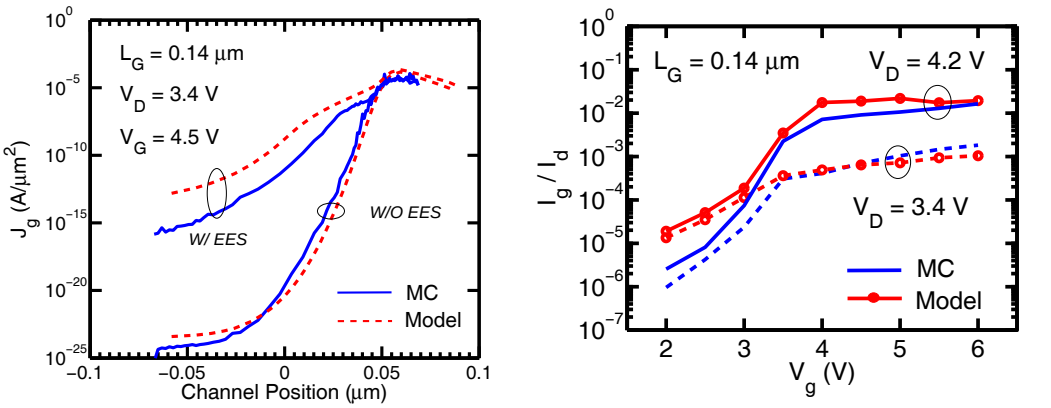
(a) (b)

Figure 3.39: Distribution functions obtained with Monte Carlo (MC) and the Model at two channel positions close to and inside the drain. The Model-predicted electron-electron effects are successively evaluated using the carrier density (Nx) and mean carrier temperature (Te) provided by the Model(circles) or the MC (dashed).

The simulations performed with the model include six energies as in Table 3.1. The figure suggests that the EES tail is further enhanced when the MC quantities are used with respect to the case where internal model quantities are used for EES calculation. The shape of the tail is however strictly preserved and the increase is limited to less than an order of magnitude. Thus, the uncertainty introduced by the internal model quantities has but a limited impact on the final result in the relevant channel positions. Furthermore, this comparison shows that the EES calculation framework which assumes a heated Maxwellian distribution, constitutes a good alternative to the Full Band EES calculation.

Finally, Figure 3.40 reports the gate current density along the channel and the injection efficiency as a function of the gate voltage for a $0.14\mu m$ gate length device,

calculated following the procedure described in section 3.1.



(a) (b)

Figure 3.40: (a) Gate current density along the channel calculated after the Model and the Monte Carlo (MC) with and without Electron Electron Scattering (EES). (b) Injection efficiencies vs. gate voltage calculated after both approaches for two drain voltages. MC results without EES are also reported.

The enhanced tail due to EES has a clear impact on the gate current density along the channel up to the junction with the drain where the tail has but a limited effect compared to the plateau of the distribution. The proposed model overestimates the gate current in the channel, although a very good qualitative agreement is achieved. In particular, the 'shoulder' of the gate current density due to EES appearing close to the junction at the channel side (peak around 0.03 μm) is nicely reproduced. The injection of "warm" carriers in the first part of the channel plays an important role especially at low voltage operation regimes. In such conditions ($V_g < 3.5V$), the simulated injection efficiencies are qualitatively well reproduced. Thus, the transition from high ($V_g > 3.5V$) to low voltage operating regimes is much smoother when EES is included. The inclusion of such effect reveals to be critical for comparison with measurements as will be shown in the next chapter.

3.4.2 Impact Ionization

The Impact Ionization (II) mechanism is another important process which involves the hot carriers and subsequently shapes their distribution function. The same aspects as for the inclusion of the EES have been considered, although different solutions are adopted.

The scattering rate of the II process as a function of the carrier energy [Bude 1992] is given in Figure 2.6 of Chapter 2. This figure shows that both Optical Phonons and II scattering processes exhibit comparable scattering rates for hot electrons. Thus, differently from the EES, the II mechanism should be considered when calculating the probabilities around each node. Figure 3.5 and Equation 3.4, which were initially introduced when only Optical Phonons were considered, are still valid when this additional process is considered. However the probabilities P^{BAL}, P^{UP}, P^{DOWN} which appear in the fluxes' calculation, have different expressions from those written in Figure 3.4. The new expressions and their derivation are given in Annex A.

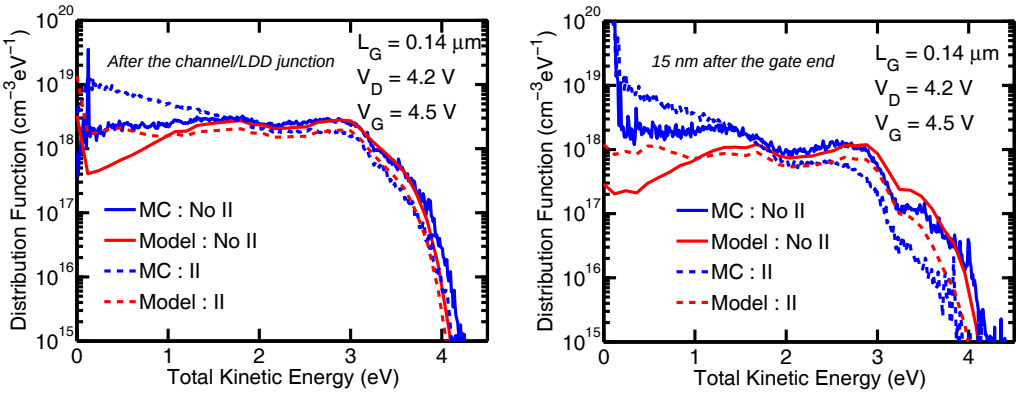
The physical process of Impact Ionization consists in extracting a Valence Band electron and promoting it to the Conduction Band. This endothermic process, which results in an electron-hole pair generation, is triggered by an electron (the primary) having an energy ε_{PRIM} higher than a certain threshold. Although this energy threshold depends on the carrier momentum [Bude 1992],[Sano 1994], thus making the process anisotropic, it has been already shown that isotropic scattering rates, averaged over all momenta having the same energy, well reproduce experimental data [Cartier 1993] with a threshold value close to the silicon band gap ($\varepsilon_{gSi} = 1.12eV$). Thus, it seems plausible that the anisotropy of the II process is hidden by electron-phonon scatterings which efficiently randomize the distribution function especially at high electron energies [Fischetti 1995].

The remaining energy of the primary electron is divided between the secondary particles (electron and hole) and the primary electron with a certain distribution [Bude 1992],[Kamakura 1994]. However, simpler approaches considering an equipartition of the remaining energy (Equation 3.33) between all particles still yield satisfactory results [Jungemann 2003]. In particular, the secondary carriers' energy can be simply set to:

$$\varepsilon_{SEC} = \frac{\varepsilon_{PRIM} - \varepsilon_{gSi}}{3} \tag{3.33}$$

This expression has been adopted in the model because it greatly simplifies the implementation and the calculations as only a limited number of additional fluxes should be considered with respect to the optical phonons case. Furthermore, the generated electrons are easily accounted for by increasing the total carrier concentration.

The contribution of the Impact Ionization process in shaping the distribution function is reported in Figure 3.41. The FBMC simulations include an isotropic impact ionization rate where secondary electrons are generated according to the distribution calculated after [Bude 1992].



(a) (b)

Figure 3.41: Distribution functions at two channel positions close to and inside the drain, obtained with the Monte Carlo (MC) and the Model, with and without the impact ionization (II) process.

Two channel positions at the junction and inside the drain featuring a large quantity of hot carriers have been considered. The inclusion of the II does not change the general shape of the curve. However, the high energy part of the curves is reduced due to this interaction (dashed lines) and the resulting electrons contribute to raise the low energy part of the curve. At both considered positions, the model qualitatively reproduces the effects of impact ionization. Note that significant ionization occurs also beyond the gate edge (right plot) as expected from previous studies [Fischetti 1995].

3.4.3 Conclusions

In this section, the introduction of two additional scattering mechanism and their impact on hot carrier effects have been presented. Preliminary physical considerations have allowed to simplify the implementation of both electron-electron scattering and impact ionization while conserving the main signatures of both processes. Electron-electron interactions have been shown to enhance the hot carrier distribution tails thus considerably impacting the injection process at low voltages. This in turn enables injection modeling at these operating conditions which are extremely important for advanced technologies. The decrease of the hot carrier distribution due to the impact ionization process has been as well discussed. The achieved agreement with full band Monte Carlo simulations consolidates the adopted approach and completes the modeling of the major mechanisms affecting the hot carrier transport.

3.5 Conclusions

A new semi-analytic approach for hot carrier modeling has been introduced in this chapter allowing to grasp the main features of the process of carrier heating and injection into the gate. The proposed quasi-ballistic modeling perspective accounts for the non-local carrier transport in which inelastic phonon scattering, impact ionization and carrier-carrier scattering have been included with a full-band description of silicon. The role and importance of these ingredients has been investigated and extensive comparisons with Monte Carlo simulations on the main figures of merit of hot carrier transport has proven the validity of this approach.

In particular it was shown that inelastic optical phonons within an isotropic approximation greatly shape the carrier distribution function and well reproduce the macroscopic bulk current. On the one hand, this confirms the predominant role of phonon scattering in the channel. On the other hand, these results show that a single inelastic phonon interaction (the one with the highest deformation potential) can reproduce most of the carrier heating behavior in silicon. In addition, the study showed that while the gate current at high gate and drain voltages can be well reproduced by accounting only for the phonons, injection at low voltage operating regimes cannot be captured but accounting for carrier-carrier scattering which enhances the hot carrier tail and contributes to a significant increase of the gate current. A computationally efficient method to account for such a complex mechanism has been suggested and implemented in this work.

Finally, in a more general perspective, this work has demonstrated that the apparent complexity of hot carrier processes can be modeled using only few variables. Interestingly enough, the channel potential and several commonly-made considerations and commonly-used silicon data much reduce the complexity of the problem in this regime where scattering still plays an important role. This approach can be used in conjunction with a compact transport model for a fast estimation of the carrier injection features for devices down to several deca-nanometers. However, care has to be taken for shorter gate lengths, where the ballistic component becomes significant, for which some of the model assumptions such as the consideration of the total energy instead of the longitudinal energy and the isotropic approximation, should be revisited.

Bibliography

[Abramo 1996] 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, page 889, 1996. (Cited on pages 18, 27, 49, 66 and 87.)

[Baranger 1984] H.U. Baranger and J.W. Wilkins. *Ballistic electrons in an inhomogeneous submicron structure: Thermal and contact effects.* Physical Review B, vol. 30, no. 12, page 7349, 1984. (Cited on page 50.)

[Brews 1978] J.R. Brews. *A charge-sheet model of the MOSFET.* Solid-State Electronics, vol. 21, no. 2, pages 345–355, 1978. (Cited on pages 57 and 118.)

[Bude 1992] J. Bude, K. Hess and G.J. Iafrate. *Impact ionization in semiconductors: Effects of high electric fields and high scattering rates.* Physical Review B, vol. 45, no. 19, page 10958, 1992. (Cited on pages 17, 18, 26, 59, 93 and 114.)

[Bufler 2005] F.M. Bufler and A. Schenk. *On the Tunneling Energy within the Full-Band Structure Approach.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2005, pages 155–158. IEEE, 2005. (Cited on pages 19, 20, 35 and 59.)

[Cappelletti 1999] P. Cappelletti. Flash memories. Springer Netherlands, 1999. (Cited on pages 2, 73 and 140.)

[Cartier 1993] E. Cartier, M.V. Fischetti, E.A. Eklund and F.R. McFeely. *Impact ionization in silicon.* Applied Physics Letters, vol. 62, no. 25, pages 3339–3341, 1993. (Cited on pages 18, 93, 113 and 114.)

[Childs 1996] P.A. Childs and C.C.C. Leung. *A one-dimensional solution of the Boltzmann transport equation including electron–electron interactions.* Journal of Applied Physics, vol. 79, page 222, 1996. (Cited on pages 18, 26, 27 and 87.)

[Eitan 1981] B. Eitan and D. Frohman-Bentchkowsky. *Hot-electron injection into the oxide in n-channel MOS devices.* IEEE Transactions on Electron Devices„ vol. 28, no. 3, pages 328 – 340, March 1981. (Cited on pages 73 and 114.)

[Ferry 1999] D.K. Ferry, S.M. Goodnick and K. Hess. *Energy exchange in single-particle electron-electron scattering.* Physica B: Condensed Matter, vol. 272, no. 1-4, pages 538–541, 1999. (Cited on pages 18, 85 and 87.)

[Fiegna 1991] C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi and B. Ricco. *Simple and efficient modeling of EPROM writing.* IEEE Transactions on Electron Devices, vol. 38, no. 3, pages 603 –610, March 1991. (Cited on pages 6, 22, 24, 49 and 66.)

[Fischetti 1995] M.V. Fischetti, S.E. Laux and E. Crabbe. *Understanding hot electron transport in silicon devices: Is there a shortcut?* Journal of Applied Physics, vol. 78, no. 2, pages 1058 –1087, July 1995. (Cited on pages 6, 18, 19, 20, 21, 22, 28, 29, 93, 94, 113 and 114.)

[Fixel 2008] D.A. Fixel and W.N.G. Hitchon. *Kinetic investigation of electron–electron scattering in nanometer-scale metal-oxide-semiconductor field-effect transistors.* Semiconductor Science and Technology, vol. 23, page 035014, 2008. (Cited on pages 18 and 87.)

[Ghetti 2002] A. Ghetti. *Explanation for the temperature dependence of the gate current in metal-oxide-semiconductor transistors.* Applied Physics Letters, vol. 80, page 1939, 2002. (Cited on pages 18, 87 and 90.)

[Gnudi 1993] 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, pages 575–581, 1993. (Cited on pages 23 and 49.)

[Hasnat 1996] K. Hasnat and C. Yeap. *A pseudo-lucky electron model for simulation of electron gate current in submicron NMOSFET's.* IEEE Transactions on Electron Devices, vol. 43, no. 8, pages 1264–1273, 1996. (Cited on pages 21, 24, 25, 28, 48, 66, 84 and 114.)

[Jin 2008] S. Jin, T.W. Tang and M.V. Fischetti. *Simulation of silicon nanowire transistors using Boltzmann transport equation under relaxation time approximation.* IEEE Transactions on Electron Devices, vol. 55, no. 3, pages 727–736, 2008. (Cited on page 50.)

[Jin 2009] S. Jin, A. Wettstein, W. Choi, F.M. Bufler and E. Lyumkis. *Gate Current Calculations Using Spherical Harmonic Expansion of Boltzmann Equation.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2009, pages 1 –4, 2009. (Cited on pages 20, 24, 26, 60 and 155.)

[Jungemann 1996] C. Jungemann, R. Thoma and W.L. Engl. *A soft threshold lucky electron model for efficient and accurate numerical device simulation.* Solid-State Electronics, vol. 39, no. 7, pages 1079–1086, 1996. (Cited on pages 20, 31, 79 and 111.)

[Jungemann 2003] C. Jungemann and B. Meinerzhagen. Hierarchical device simulation: The monte-carlo perspective. Springer Verlag, 2003. (Cited on pages 13, 19, 26, 31 and 93.)

[Kamakura 1994] Y. Kamakura, H. Mizuno, M. Yamaji, M. Morifuji, K. Taniguchi, C. Hamaguchi, T. Kunikiyo and M. Takenaka. *Impact ionization model for full band Monte Carlo simulation.* Journal of Applied Physics, vol. 75, no. 7, pages 3500–3506, 1994. (Cited on pages 18 and 93.)

[Lenzi 2008] M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi and G. Baccarani. *Investigation of the Transport Properties of Silicon Nanowires Using Deterministic and Monte Carlo Approaches to the Solution of the Boltzmann Transport Equation.* IEEE Transactions on Electron Devices, vol. 55, no. 8, pages 2086 –2096, aug. 2008. (Cited on page 50.)

[Lundstrom 2000] M. Lundstrom. Fundamentals of carrier transport. Cambridge Univ Pr, 2000. (Cited on pages 12, 15, 16, 18, 19 and 52.)

[Palestri 2006] P. Palestri, N. Akil, W. Stefanutti, M. Slotboom and L. Selmi. *Effect of the gap size on the SSI efficiency of split-gate memory cells.* IEEE Transactions on Electron Devices, vol. 53, no. 3, pages 488–493, 2006. (Cited on pages 6, 19, 24, 30 and 62.)

[Sano 1994] N. Sano and A. Yoshii. *Impact ionization rate near thresholds in Si.* Journal of Applied Physics, vol. 75, no. 10, pages 5102–5105, 1994. (Cited on pages 17, 93 and 114.)

[Sano 2004] N. Sano. *Kinetics of quasiballistic transport in nanoscale semiconductor structures: Is the ballistic limit attainable at room temperature?* Physical Review Letters, vol. 93, no. 24, page 246803, 2004. (Cited on page 50.)

[Shockley 1961] W. Shockley. *Problems related to pn junctions in silicon.* Solid-State Electronics, vol. 2, no. 1, pages 35–60, 1961. (Cited on pages 20 and 48.)

[Sze 2007] S.M. Sze and K.K. Ng. Physics of semiconductor devices. Wiley-Blackwell, 2007. (Cited on page 63.)

[Tam 1984] S. Tam, P.K. Ko and C. Hu. *Lucky-electron model of channel hot-electron injection in MOSFET's.* IEEE Transactions on Electron Devices, vol. 31, no. 9, pages 1116–1125, 1984. (Cited on pages 20, 21, 48 and 78.)

Chapter 4

Comparison between measurements and modeling results

The first chapters of this thesis were devoted to the modeling and simulation of hot carriers during Flash programming conditions. This study pointed out the main aspects and ingredients of the complex physics involved during this operation, such as the band structure, the non local carrier transport and the scattering mechanisms. Thus, a benchmarking process between several modeling approaches was proposed, whose conclusions allowed us to elaborate a new semi-analytic model including the most relevant ingredients. However, no comparison with measurements has been shown yet. This will be the guiding line of this chapter, which carefully addresses the comparison with measurements. Hence, Section 4.1 describes the adopted measurement methodology throughout this thesis. The intrinsic injection properties have been extracted from the characterization of the Flash cell and of the equivalent transistor for various device lengths and hot carrier biases. The extracted figures of merit can indeed be used either to compare different measurement conditions between them or measurement against simulations. Such a setup has been applied to two injection regimes.

The hot electron injection will be the object of the Section 4.2, where the previously described Spherical Harmonics Expansion method, the Monte Carlo simulator and the 1D non-local model will be successively compared against measurements. The necessary calibration steps (carried out prior to the comparison) will be also discussed as an important part of this procedure.

Section 4.3 instead introduces and discusses the hot hole injection phenomenon during the drain disturb regime which is an undesirable operation condition commonly found in today's Flash arrays. Such mechanism requires the attention as it may be the cause of significant charge loss in the floating gate as well as potential reliability problems. The proposed TCAD-MC coupled simulation methodology is applied for device optimization purposes with a final successful comparison with measurements.

4.1 Measurements

This section describes the characterization methodology applied for the study of the Flash cell operating regimes throughout this work. A few preliminary considerations and definitions on the employed methodology and the cell itself are first given in subsections 4.1.1 and 4.1.2, respectively. This will allow to introduce the measurement setup required to extract useful hot carrier quantities followed by the discussion of some of the experimental errors, respectively in subsection 4.1.3 and 4.1.4.

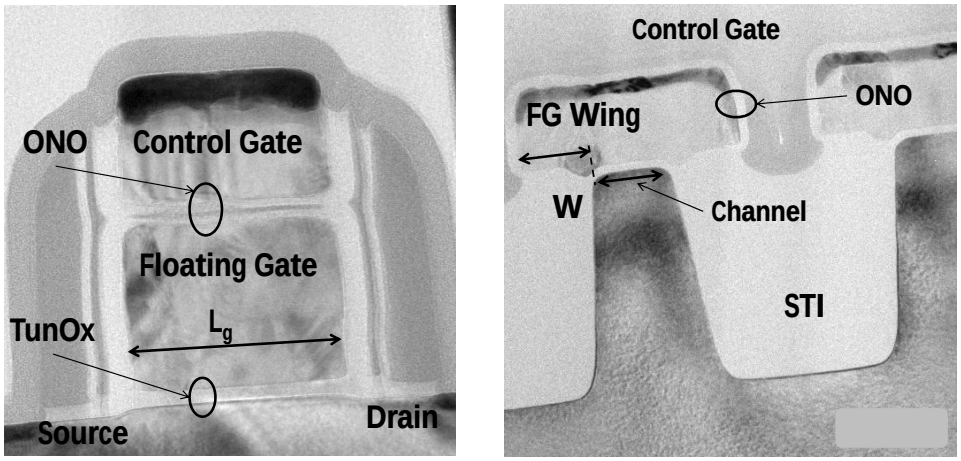
4.1.1 Motivation and methodology

In the introduction of this thesis we have shown that the Flash cell can be seen as a nMOS transistor with a modified gate stack incorporating two n-doped poly-silicon areas (gates) one of which is floating and contains the electrons, the other being electrically addressable. The electrons stored in the floating node are responsible for the threshold voltage shift giving rise to state "'0"' (programmed) and state "'1"' (erased). The objective of the characterization step is to measure as accurately as possible the current flowing *in* and *out* of the floating gate (fg). In the perspective of comparing different measurements conditions among them as well as measurements against simulations, a proper choice of relevant quantities should additionally be made. The simulations presented in Chapters 2 and 3 have introduced $I_{fg}(V_{fg})$ and $I_{fg}/I_d(V_{fg})$ characteristics which have been shown to be adequate for such comparisons. Indeed, they reflect the *intrinsic* behavior of the underlying MOS transistor whose impact on cell operation depends on the cell electrostatics.

For this reason, the following procedure aims to extract such quantities in different operating regimes. This is achieved by using the Flash cell and the *equivalent transistor*. The latter is a modified Flash cell which allows direct electrical access to the floating after shorting both gates at the layout level. Although an additional etching step with respect to the Flash cell is needed, the transistor beneath is processed identically. Hence, in principle, the equivalent transistor is able to provide all the above characteristics. However, as it will be shown in the following pages, the accuracy obtained for the gate current measurements is not satisfactory. Therefore, transient operation of the Flash cell has been used to extract the dynamic gate current. Expressing the latter current as a function of the gate voltage requires the knowledge of the Flash electrostatics (coupling coefficients) as well as the measurement of the MOS threshold voltage. The coupling coefficients will be extracted on cells using a simple geometrical model validated on 2D/3D TCAD simulations, while the MOS threshold voltage is measured on the equivalent transistor. The equivalent transistor is also used to measure the bulk currents that compared with the simulations is an important indicator of the good modeling of hot carriers at high drain voltage.

4.1.2 Cell description

The embedded NOR-type Flash devices investigated throughout this study have been fabricated in a 65nm CMOS technology. For convenience, in this subsection we report and extend the Flash cell description made in Chapter 1. Figure 4.1 shows two TEM cuts along the length (a) and width (b) direction of a typical Flash device and it defines its main components.



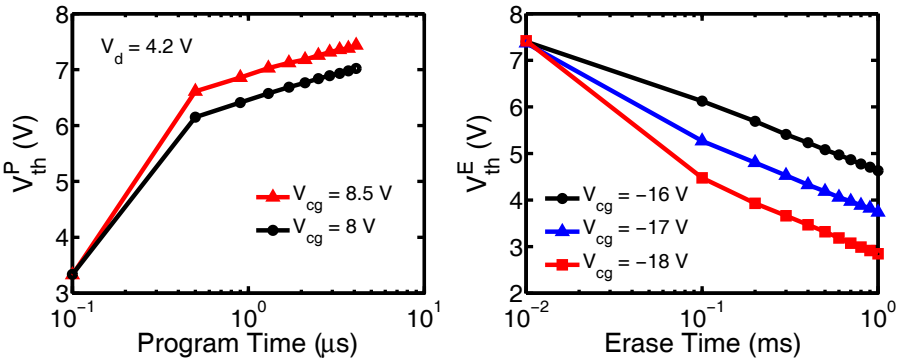
(a) (b)

Figure 4.1: TEM images of the Flash cell structure along the channel *(a)* and in the width direction *(b)*. The major components of the cell have been highlighted.

The cell has four electrically accessible terminals: the Source (s), the Drain (d), the Bulk (b) and the Control Gate (cg). The Floating Gate (fg) is isolated from the control gate and the substrate terminals (s, b, d) via the Oxide-Nitride-Oxide layer (ONO) and the Tunnel Oxide ($TunOx$), respectively. The charge exchange operates between the floating gate and the substrate through the tunnel oxide. The normal Flash operating regimes involve the *Program* and the *Erase* (P/E) phases, respectively defined as the mechanisms during which the electrons are put inside or pulled out of the floating gate. Figure 4.2 reports a typical example of the erase and program phase dynamics at different control gate biases.

The threshold voltages (V_{th}) reported in Figure 4.2, measured during the so-called *Read* phase, have been defined as the control gate voltages needed to achieve a drain current of 8 μA with a drain voltage of 0.7 V. Figure 4.3 reports an example of curves obtained during this phase. At the end of the erase and program transient regimes, V_{th}^E and V_{th}^P are respectively obtained..

Figure 4.2 shows that program/erase processes are accelerated when higher absolute values of control gate voltage are employed. For short channel devices, the larger the potential difference between the floating gate and the substrate, the more efficient these mechanisms are. However, as the floating gate cannot be electrically addressed, its potential depends on the electrostatic effect of the other terminals.



(a) (b)

Figure 4.2: Threshold voltage variation during program *(a)* and erase *(b)* phases for different control gate voltages. The measured cell features $L_g/W/FGWing = 0.14/0.08/0.115\ \mu m$.

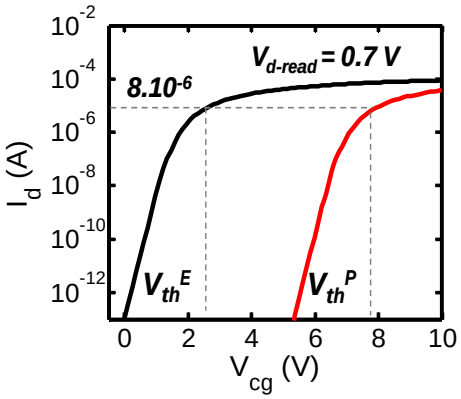


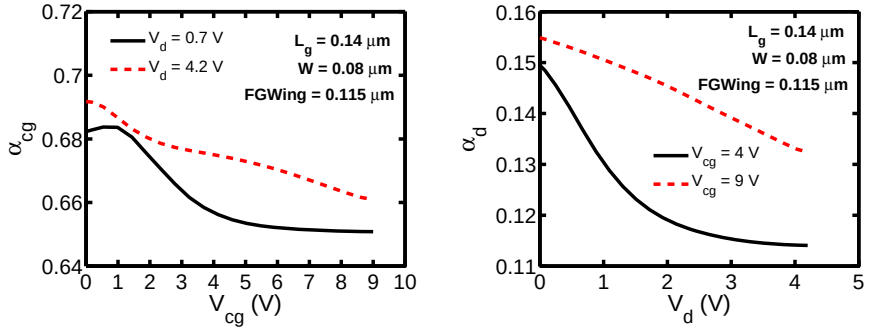
Figure 4.3: Drain current vs. control gate voltage obtained after the erase and the program phase on a $L_g/W = 0.18/0.14\ \mu m$ device.

This effect is quantified via the coupling coefficients, defined as the variation of the floating gate potential due to the potential variation of each of the other terminals:

$$\alpha_i = \frac{\partial V_{fg}}{\partial V_i} = \frac{C_i}{C_{TOT}} \quad \text{with} \quad i \in \{s, d, b, g\} \tag{4.1}$$

The coupling coefficients are equally defined as the variation of a given terminal capacitance with respect to the floating gate over the total structure capacitance (C_{TOT}). The coupling with the control gate (α_{cg}) is the most significant among them and it constitutes a key parameter for Flash optimization. A high gate coupling allows the use of lower supply voltages at equivalent times or smaller writing times for the same supply voltages, as shown in Figure 4.2. In order to increase α_{cg}, the capacitance between the floating gate and the control gate should be increased without increasing the capacitance between the floating gate and the other terminals.

This is achieved by extending the overlapping distance of both gates in the width direction beyond the active width of the device (W), as shown in Figure 4.1b. This gives rise to two symmetrical Floating Gate Wing regions (FGWing) which can substantially increase the coupling. Figure 4.4 reports the gate and drain coupling coefficients as a function of the control gate and drain voltage, respectively. The gate coupling is given for drain voltages in *Read* (0.7 V) and *Program* (4.2 V) conditions while the drain coupling has been plotted for control gate voltages roughly representing the start (9 V) and the end (4 V) of the programming phase.



(a) (b)

Figure 4.4: TCAD-simulated gate coupling (a) and drain coupling (b) coefficient as a function of the control gate and drain voltage, respectively, while the other terminal has been used as a parameter. Source and bulk terminals are grounded.

The coupling coefficients are voltage dependent with the most important variations being situated around the inversion threshold. As a matter of fact, the traditional MOS capacitances strongly vary in this region and thus directly impact the different couplings. However, at high drain and control gate voltages, the couplings are rather independent on V_{cg} and V_d.

4.1.3 Measurement setup and extraction methodology

In the following, the characterization procedure is exposed with particular focus on the programming phase. The other operating conditions, such as as the erase phase and the disturb phenomena are characterized using the same procedure. The developped test structures consist of mini matrices of Flash and equivalent transistor devices where the source and the substrate are in common. Only a single device placed in the middle of the array is characterized, whereas the surrounding ones act as dummy devices to reproduce a realistic memory environment.

A methodology combining DC and transient measurements has been set up to characterize the Flash cell as accurately as possible (Figure 4.5). The programming operation performed on Flash cells, is divided into small pulses (Figure 4.6). Each of the transient pulses of Figure 4.6 feature a rise and fall time (t_{rise}, t_{fall}) equal to 50 ns. After each pulse, the cell threshold voltage is measured in DC without

affecting the cell's state following the procedure in Figure 4.3. This allows to build the transient characteristics of Figure 4.2*a*.

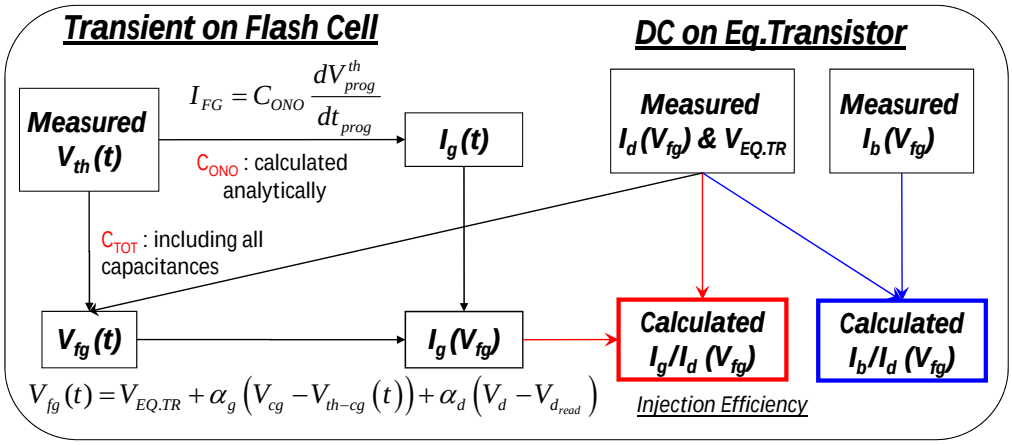


Figure 4.5: Mixed characterization scheme to extract the cell properties. DC and transient measurements have been used on equivalent transistor and Flash cells, respectively. The knowledge of the capacitances and the coupling coefficients in the structure allows to extract the injection current during programming. The unit-less ratios I_b/I_d and I_g/I_d can be additionally computed by combining both measurements.

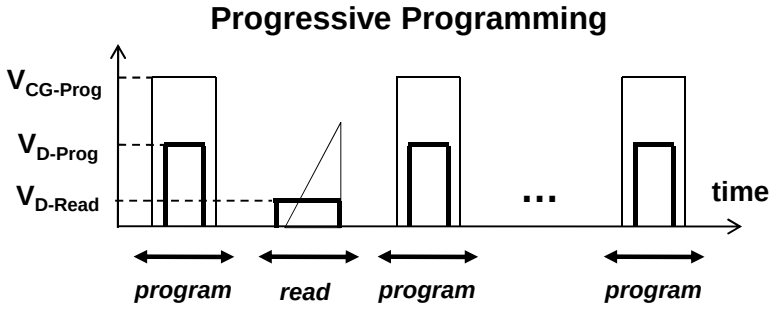
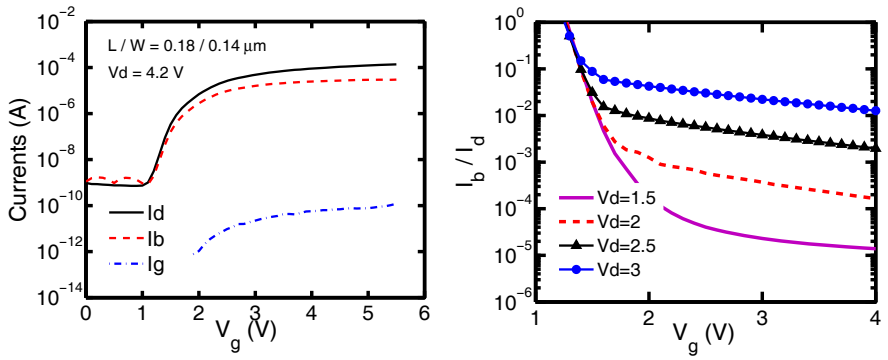


Figure 4.6: Schematic of the applied voltages during programming. The total programming time has been divided into small *program* pulses with drain and control gate terminals respectively fixed at V_{D-Prog} and $V_{CG-Prog}$. The threshold voltage is measured after each of the pulses during the *read* phase (the drain is biased at $V_{D-Read} = 0.7V$) following the description in 4.1.2. The same scheme with different voltages is applied for erase and disturb regimes.

DC measurements are additionally performed on the equivalent transistor. Typical currents as a function of the floating gate voltage (floating gate = gate, in the case of the equivalent transistor) are reported in Figure 4.7a. In the framework of the Hot Carrier Injection (HCI) regime, the hot carrier effects are also investigated by measuring the bulk current (c.f. Chapter 3). An integration time of 80 μs has been applied for all these measurements in order to limit the degradation. Starting

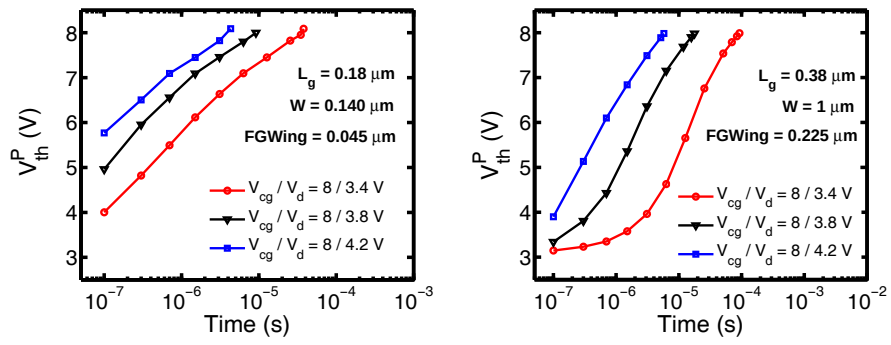
from the DC measurements the I_b/I_d ratio is reported in Figure 4.7b for different drain biases. The ratio increases with drain voltage and decreases with rising gate voltage. It is important to mention that for the shortest cells, the DC measurements are repeatable for drain biases lower than 3V, while degradation of I_d is observed for higher voltages even with the smallest integration time.



(a) (b)

Figure 4.7: Typical drain, bulk and gate currents measured in DC on an equivalent transistor (a) and the resulting I_b/I_d ratios obtained at different drain voltages (b).

DC measurements also provide the gate current which could enable us to calculate the injection efficiency. However, as it will be shown below, the gate currents measured in DC are considerably lower than the ones from transient measurements. As a consequence, they are not used in any further calculation. Instead, the threshold voltage of the cell at different programming times (transient procedure of Figure 4.6) is monitored and used to extract I_{fg}. Examples of such curves for different drain biases and gate lengths are reported in Figure 4.8.



(a) (b)

Figure 4.8: Threshold voltage evolution during programming operation obtained for a 0.18 (a) and 0.38 (b) μm gate-length device at different drain voltages.

As expected, programming is faster at higher V_d and for the shortest device. Notice the significant threshold voltage variation in the case of the shortest cell after only 100ns. This time duration has been considered as the first reliable measurement point which is not affected by the ramp up of the drain and control gate terminals (rise time of 50ns). From the threshold voltage shift it is possible to extract the dynamic floating gate current using the following expression:

$$I_{fg} = C_{ONO}\frac{\Delta V_{th}}{\Delta t} \tag{4.2}$$

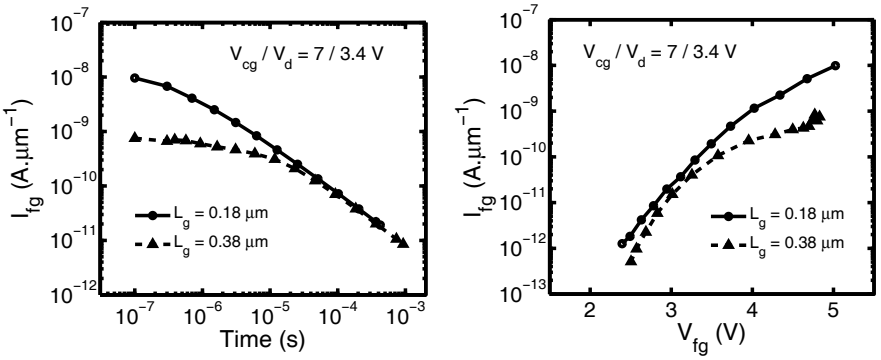
C_{ONO} is the capacitance between both gates as defined in subsection 4.1.2. Its value depends on the physical dimensions of the overlapping of both gates as well as on electrical permittivity of the layers. Throughout this chapter, C_{ONO} has been analytically calculated after the model published in [Garetto 2009a], which includes parallel plate, corner and fringe contributions. This expression has been validated against 3D TCAD simulations. The result of such an extraction is shown on Figure 4.9a for two different cells at a given bias condition. At the beginning of the programming, the shorter cell features a higher injection current, while near the end of the programming phase both cells display the same injection current dynamics. At this stage, no precise comparison between cells is possible as no information on the floating gate voltage is available. The dynamics of the latter is extracted from a standard capacitive network approach with constant coupling ratios [Kolodny 1986]:

$$V_{fg}(t) = V_{EQ.TR} + \alpha_g\left(V_{cg} - V_{th}(t)\right) + \alpha_d\left(V_d - V_{d_{read}}\right) \tag{4.3}$$

$V_{EQ.TR}$ is the threshold voltage of the equivalent transistor extracted during DC measurements at the drain voltage $V_{d_{read}}$=0.7 V, which has been equally used to measure the cell threshold voltage during the programming phase $V_{th}(t)$. During programming, the control gate and the drain are respectively biased at V_{cg} and V_d. α_g and α_d are respectively the gate and drain coupling ratios. Putting together the results of Equation 4.2 and 4.3, we are able to build the $I_{fg}(V_{fg})$ curve, as shown on Figure 4.9b. At this point, the comparison between cells and programming conditions is possible as the gate current obtained in transient conditions is expressed as a function of floating gate voltage. Hence, both structures show similar gate currents when $V_{fg} < V_d$. However, at higher floating gate voltage, the shortest cell injects more current.

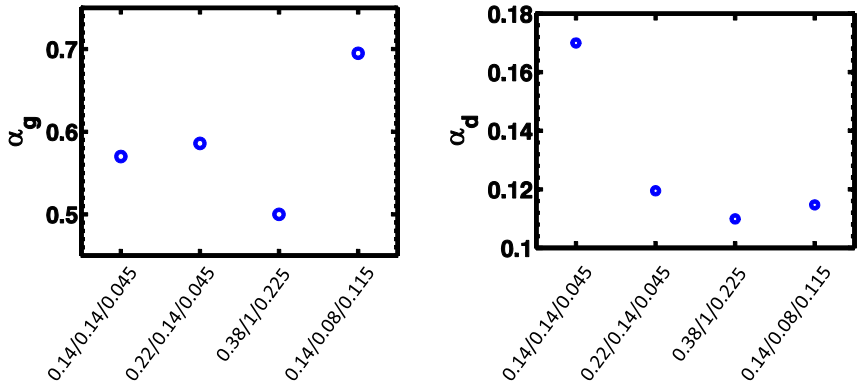
The results of Figure 4.9*b* have been obtained by assuming constant drain and gate coupling coefficients. An overview of the constant coupling coefficients used for different devices is reported in Figure 4.10.

Finally, the injection efficiency is calculated by dividing the extracted gate current on the Flash cell by the measured drain current on the equivalent transistor as a function of the floating gate voltage. Figure 4.11*a* reports the injection efficiencies obtained from the approach described for a 0.18 μm gate-length device at different drain voltages. Similarly to the threshold voltage variation and to the gate current, the injection efficiency increases with drain voltage. For comparison,



(a) (b)

Figure 4.9: Floating gate current as a function of the programming time (a) and the floating gate voltage (b) for two devices with different gate-lengths under a particular programming condition. The graphs are built upon transient measurements.

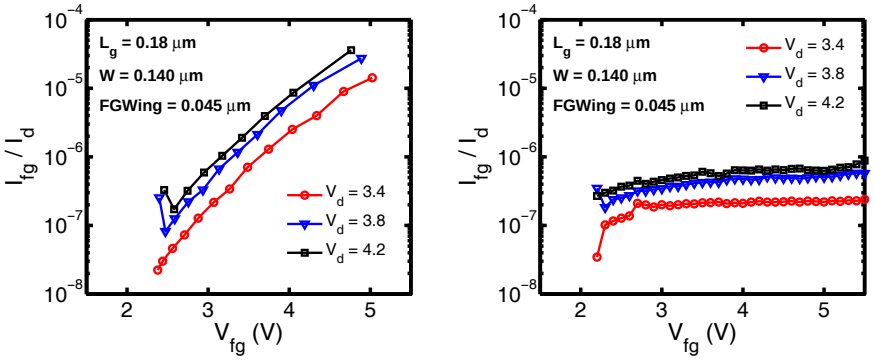


(a) (b)

Figure 4.10: Examples of gate *(a)* and drain *(b)* coupling coefficients used during the extraction. The ordered triple in the x-axis stand for the device geometrical dimensions L_g/W/FGWing.

the result of the gate-to-drain current ratio obtained from the DC measurements previously shown in Figure 4.7a for the equivalent transistor is also given, showing a significant discrepancy between the two methods.

In particular, the DC-only method shows a rather flat injection efficiency with increasing floating gate voltage for small gate lengths, with maximum values around 10^{-6}. Instead, the efficiency extracted from the mixed transient-DC approach rises with gate voltage with maximum measurable values greater than 10^{-5}. To test the soundness of the DC-values, we calculate the threshold voltage variation of the cell during the programming phase (using Equations 4.2 and 4.3) assuming the gate current of the equivalent transistor from DC measurements is also valid for the



(a) (b)

Figure 4.11: Injection efficiency as a function of the floating gate voltage extracted from mixed transient-DC (a) or DC-only (b) measurements at different drain voltages for a 0.18 μm gate-length device.

Flash cell. Figure 4.12 compares the DC-only resulting threshold voltage with the one measured during transient operation for different drain voltages. The couplings and the other voltages are kept constant. The observed discrepancy proves that the current levels measured in DC on the equivalent transistor are much lower than the one needed to program the Flash cell according to transient measurements.

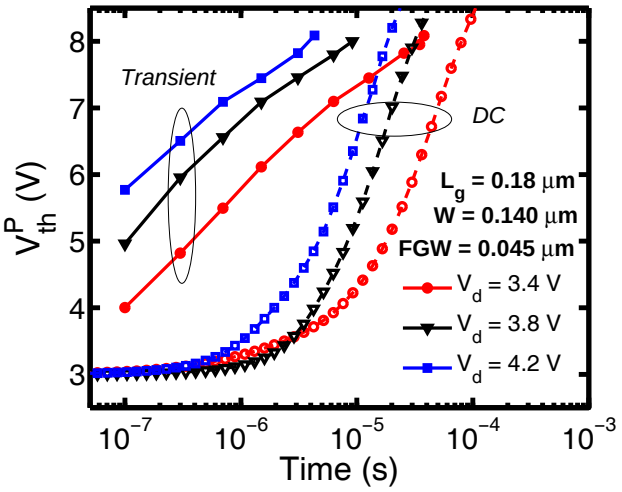


Figure 4.12: Threshold voltage variation as a function of the programming time for different drain voltages, extracted from transient measurements (solid curves) or from DC measurements (dashed curves) using Equations 4.2 and 4.3.

However, the discrepancy between the two methods is considerably reduced for longer cells, as shown in Figure 4.13. These comparisons indicate that the estimation

of the injection current (thus of the injection efficiency) from a DC-only approach can be quite misleading especially for short channel devices of interest here ($L_g < 0.22\mu m$).

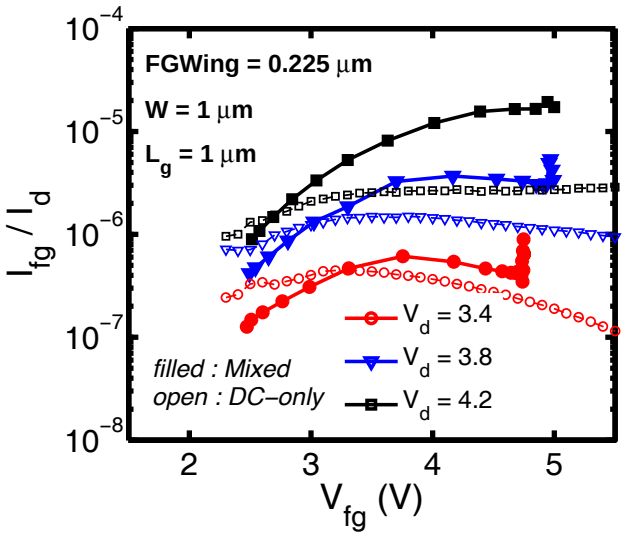


Figure 4.13: Injection efficiency as a function of the floating gate voltage extracted from mixed transient-DC (filled symbols) or DC-only (open symbols) measurements at different drain voltages for a 1 μm gate-length device.

4.1.4 Experimental errors

Although we have shown that a reasonable approach to extract the cell properties should rely on mixed transient-DC measurements, this method could suffer form other sources of uncertainty. Most of the measurements are performed on several dies which are used to extract the mean values of the characteristics. However, if measurements are performed on a full-wafer scale, considerable variations may be seen, as shown in Figure 4.14. The characteristics have been extracted using the corresponding equivalent transistor threshold voltage of each die. Additionally, the transient operation is affected by the drain voltage rise time, therefore to avoid overshoot effects while establishing the desired drain voltage, a rise time of 50ns has been found to be adequate for all the devices and biases under consideration.

The most important source of error may come from the use of constant (over bias, but different for different devices as shown in Figure 4.10) coupling coefficients, especially for the gate coupling ratio as it will impact the accurate evaluation of the floating gate voltage. Although the gate coupling coefficient is not strictly constant, TCAD simulations in Figure 4.4 show that such ratio is rather flat in the programming working regime. The same remark can be made for the drain coupling coefficient. Hence, what reveals to be important is the gate (respectively, drain) coupling value used which is calculated as the ratio of the gate (respectively, drain)

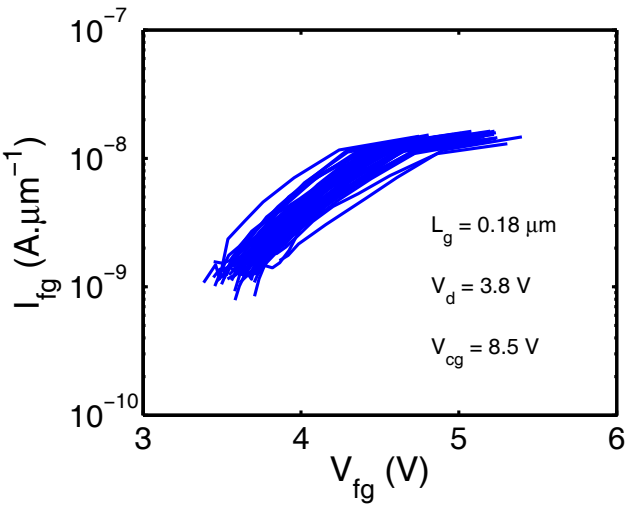
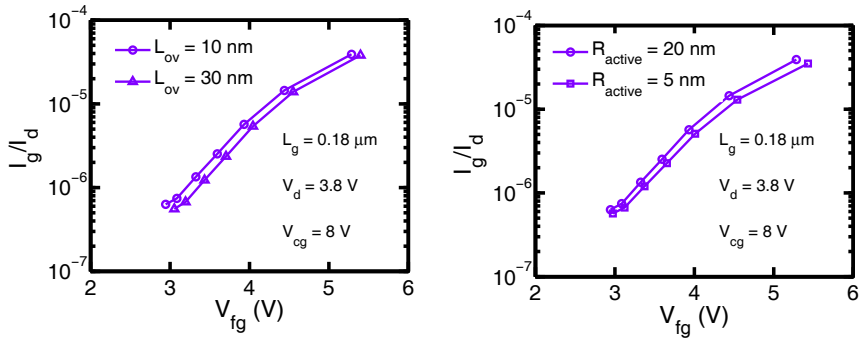


Figure 4.14: Floating gate current as a function of the floating gate voltage extracted after transient measurements over 50 dies on the same wafer.

capacitance over the total structure capacitance (Equation 4.1). It is clear that predicting different values of the capacitances will affect the extracted floating gate value.

Figure 4.15 reports the effect on the extracted injection efficiency characteristics of the overlap extension L_{ov} and of the radius of the channel R_{active} in the width direction (c.f. Figure 4.1). The same measurements and the same methodology described in section 4.1.3 have been used.



(a) (b)

Figure 4.15: Injection efficiencies as a function of the floating gate voltage resulting from different drain overlap extensions (a) and channel radius (b) used for the extraction. The same experiments are used for all the cases.

As a matter of fact, L_{ov} and R_{active} dimensions are hardly measured compared to oxide thicknesses and geometrical lengths where the precision is higher. Figure

4.15 shows that by changing the relative weight on the capacitances, that is: when L_{ov} increases, α_d increases and α_{cg} decreases or when R_{active} decreases, α_b increases and α_{cg} decreases, the injection efficiency is almost rigidly shifted towards higher V_{fg} values.

4.2 Hot carrier injection regime

The previous section described the experimental extraction of the intrinsic cell properties. In this section, hot carrier injection measurements are compared to simulation results obtained with the methods presented in Chapters 2 and 3. TCAD and Monte Carlo simulations are presented in subsection 4.2.1 while the use of the 1D non-local model is presented in subsection 4.2.2.

4.2.1 TCAD and Monte Carlo simulations

Before carrying out hot carrier simulations using the Spherical Harmonics Expansion method (SHE) in the TCAD environment or using the Monte Carlo (MC) simulator, preliminary steps are required. On the one hand, we should verify that the simulated structure obtained from TCAD process simulations is close to the measured one (4.2.1.1). On the other hand, we should also verify that the simulators well reproduce standard low and high energy figures of merit, such as the mean carrier velocity, the ionization coefficient and the quantum yield (4.2.1.2). Finally, the hot carrier effects can be studied and simulations can be compared to measurements (4.2.1.3).

4.2.1.1 Structure calibration

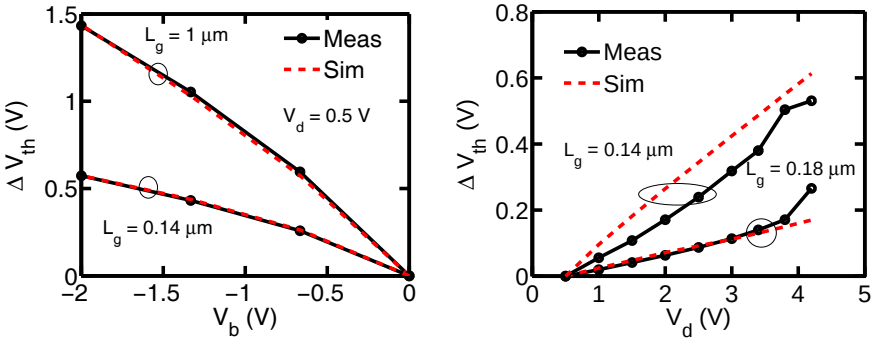
Before studying the hot carrier effects we should first make sure that the structure we aim to simulate reproduces the basic electrostatic effects of the real one. The complete structure is obtained following a realistic 2D process simulation; in particular, care has been taken regarding the implant recipes, the thermal budget and the dopant diffusion process [Synopsys 2010b]. However, when comparing to measurements a careful check on the validity of the doping profiles and of the geometrical dimensions is always necessary [Jungemann 1996]. Hence, the following measurements have been used to extract useful information about the structure:

- The gate capacitance
- The threshold voltage variation with bulk voltage
- The threshold voltage variation with drain voltage

All the measurements were performed on the equivalent transistor as it undergoes exactly the same process as the Flash cell. Thus, the gate capacitance allows to tune the tunnel oxide properties which are considered to be reasonably the same as for the Flash cell. In order to avoid any parasitic effects the capacitance is measured

on a long and wide structure and a tunnel oxide of 98 Åand an oxide dielectric permittivity of 4.1 ε_0 has been found to match the measurements. These values are within expected experimental values and are kept constant for the rest of the structure calibration.

The next step is to validate the vertical doping profile. By reverse biasing the bulk terminal, the depletion depth increases with respect to the Si/SiO_2 interface. Thus, the subsequent increase of the threshold voltage with the bulk bias provides a useful indication on the Boron (P-type well) profile evolution in the first 50 nm close the interface (the Body effect). In particular, the Boron segregation parameter at the Si/SiO_2 interface in the process simulation has been adjusted in order to match the $V_{th}(V_b)$ curves, as shown in Figure 4.16*a*. During the diffusion process, the interface acts as a large defect which may trap more or less dopant atoms. Thus increasing the segregation of atoms at the interface changes the fluxes at this boundary and finally results in an increased dose loss in the channel. The study has been first performed on long transistors which are adequate due to the absence of short channel effects. The same setup has then been successfully applied to shorter devices.



(a) (b)

Figure 4.16: Threshold voltage variation as a function of the bulk voltage *(a)* and drain voltage *(b)* obtained after measurements and 2D TCAD simulations.

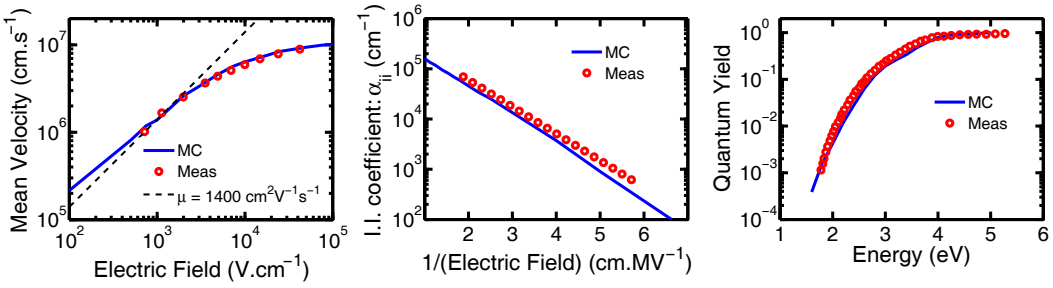
The final step of the proposed procedure concerns the validation of the lateral doping profile. The latter has been adjusted in order to reproduce the threshold voltage variations as a function of the drain voltage (Drain Induced Barrier Lowering - DIBL effect) for two target cells (Figure 4.16*b*). Slightly smoother junctions compared to the original simulations were necessary. This has been achieved by increasing the number of silicon interstitials created by the heavy Arsenic atoms during the LDD implant phase in the process simulation. The presence of the interstitials increases the Arsenic diffusion towards the channel during the annealing phase. Furthermore, the diffusion of Phosphorous, used during Source/Drain implants, has been slightly increased due to the presence of high concentration of Arsenic. Structures with different gate length obtained upon the calibrated process will be used for both SHE and MC simulations.

4.2.1.2 Calibration of the MC model

The second step of the calibration procedure concerns the MC simulator ingredients. As shown in Chapter 2, the band structure, the phonon scattering parameters and the impact ionization in silicon have a great influence on the hot carrier simulation. As a crucial ingredient for hot carrier transport modeling [Fischetti 1995], the full-band structure is not under discussion. It directly determines the density of states which in turn influences the electron-phonon scattering rate. What is discussed in the following, is the adjustment one needs to make between phonons and impact ionization parameters in the full-band approach. Traditionally, the validity of the latter is tested using the following relations [Jacoboni 1983], [Fischetti 1988], [Cartier 1993], [Kamakura 1999], [Ghetti 2003]:

- Mean carrier velocity at low fields
- Ionization coefficient at moderate and high fields
- Ionization quantum yield at high energies

Figure 4.17 reports the comparison of the MC results with the published measurements for these figures. The mean carrier velocity as a function of the electric field is correctly reproduced. At 300 K, for fields lower than 50 kV, the slope of the curve closely follows the phonon-limited mobility of electrons. The obtained agreement shows that the deformation potentials and phonon energies responsible for low energy transport have been chosen correctly. The phonon parameters in the MC are close to the ones given in [Jacoboni 1983].



(a) (b) (c)

Figure 4.17: *a* Mean electron velocity as a function of the electric field in homogeneous silicon along <100> direction; measurements are from [Canali 1975]. *b* Electron impact ionization coefficient vs. inverse of the electric field at 300 K; measurements are from [Van Overstraeten 1970]. *c* Quantum yield as a function of the electron energy; measurements are from [DiMaria 1985].

The impact ionization coefficient and the quantum yield simulations, reported in Figure 4.17*b*,*c*, also well compare to experiments. The impact ionization coefficient is defined as the number of electron-hole pairs generated by an electron per unit

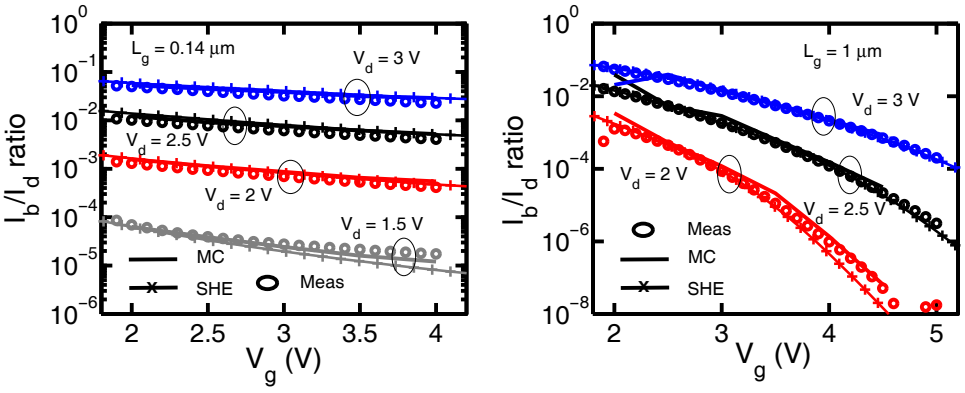
length while the quantum yield is the ratio between the bulk and the channel current following the injection of an electron of a given energy in the silicon. Both quantities scan the electronic distribution function at high energies. Due to this fact, it is difficult to accurately separate the strength of each mechanism at high fields. Thus, no unique set of phonons and impact ionization exists in literature, but rather multiple sets which are able to reproduce these data with almost the same accuracy due to compensation effects between these physical mechanisms [Fischetti 1995]. The MC results are obtained with the impact ionization rates calculated by Bude [Bude 1992b] and adjusted with a multiplication factor of 0.25. Such correction procedures have been already presented and discussed in [Cartier 1993], [Sano 1994], [Bude 1995]. This correction ensures a good agreement with measured bulk currents as shown in the next paragraph. Finally, as the scattering rates of the MC and of the SHE method are quite close (c.f. Figure 2.6)), this calibration procedure is thus valid for both approaches.

4.2.1.3 Comparison with measurements for Flash cells

Once the device structure and the simulators have been calibrated independently, the simulated hot carrier currents can be compared to measurements resulting from the methodology exposed in 4.1.3. Figure 4.18 reports the substrate to drain current ratio for two gate length devices under different bias conditions. A very good matching between MC, SHE and the measurements is obtained over the whole investigated length and bias range throughout this work. Such an agreement validates the calibration procedures previously presented. In particular, the reduction of the impact ionization rates by a factor of 4 with respect to the ab-initio calculations in [Bude 1992a] is here further justified. In addition, the close agreement between MC and SHE comes as a consequence of almost identical distribution functions obtained throughout the channel and discussed in Chapter 2.

Next, the measured injection efficiency as a function of the floating gate voltage at a constant drain voltage for different gate lengths is reported in Figure 4.19 and compared with SHE and MC simulations. Observing the measurements, it can be noticed that the shape of the injection efficiency changes considerably as a function of the gate length. For the shortest devices, the injection efficiency increases as a function of the gate voltage, while a bell-shape curve is obtained in the case of the longest ones with the maximum of the injection roughly situated at $V_g = V_d$. Such behavior is in agreement with [Eitan 1981], [Goldsman 1988], [Hasnat 1996], and also with the considerations in Chapter 3. Furthermore, the injection efficiency at high gate voltage, where most of the programming occurs, increases with decreasing length.

The reported simulation results have been obtained with a conduction band offset at the Si/SiO_2 interface equal to 3.25 eV for all the investigated conditions. It ought to be mentioned that the tunnel oxides are subject to high doses of Nitrogen which contribute to increase the barrier height. Furthermore, no barrier lowering and no parallel momentum conservation has been included in these simulations.



(a) (b)

Figure 4.18: Comparison between the normalized bulk current I_b/I_d vs. gate voltage obtained after Monte Carlo (MC), the Spherical Harmonics Expansion (SHE) and Measurements for a 0.14 µm *a* and 1 µm *b* gate-length device at different drain biases.

As a matter of fact, Figure 2.18 of Chapter 2 showed that including or excluding simultaneously both effects yields similar results in terms of I_g/I_d. Figure 4.19 shows that MC simulations with electron-electron scattering provide a very good setup for the shortest devices for the whole gate voltage range. In particular, for channel lengths in the 0.14 - 0.22 μm interval, the injection efficiency at low gate voltages is correctly captured. This matching cannot be achieved otherwise than by including electron-electron interactions, as shown by the MC and SHE without including such a mechanism. However, a discrepancy with respect to measurements is observed for the longest devices for both the MC and the SHE which predict a lower gate current by 1-2 decades. Although the core of the hot carriers is correctly reproduced (c.f. bulk currents in Figure 4.18), the modeling of the carrier heating after carrier scattering at energies close to the barrier (exponentially decaying tails) reveals to be more difficult and requires further investigation.

In addition to gate length variation, the models are compared to measurements for different drain voltages as well. Figure 4.20 compares the measured injection efficiencies with MC simulations for a 0.18 μm gate length device. Notice the decrease of the injection efficiency with the reduction of the drain voltage. This is a direct consequence of the finite energy that a carrier can gain in the channel which depends on V_d. MC with electron-electron scattering well captures the measurements for all the investigates biases. In addition, Figure 4.20 also reports the MC simulations without electron electron scattering, which significantly differ from measurements especially at low drain voltages. This stresses again the necessity to account for such a mechanism whenever low voltage operation is considered.

Finally, Figure 4.21 reports the same comparison as in Figure 4.20 but considering the SHE method. For the highest V_d (3.4 V) at high V_g, a good matching is

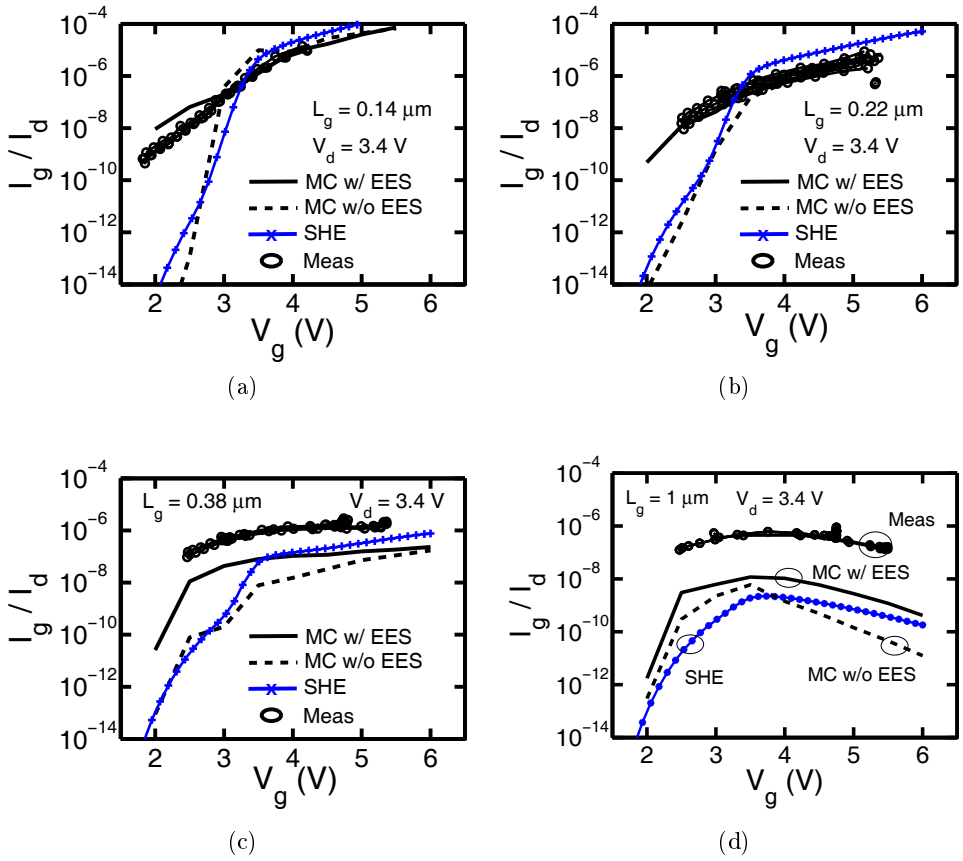


Figure 4.19: Comparison of the injection efficiencies (I_g/I_d) vs. gate voltage obtained from measurements (Meas), the Monte Carlo (MC) and the Spherical Harmonics (SHE) at constant drain voltage for different gate lengths: 0.14 *(a)*, 0.22 *(b)*, 0.38 *(c)* and 1 μm *(d)*. MC simulations are shown with or without Electron-Electron Scattering (EES).

obtained similarly to the case of the 0.14 and 0.22 μm gate length devices of Figure 4.19. However, the low voltage operation regime is not well reproduced. This discrepancy is especially visible at $V_d < 3V$ for all the range of V_g values. The behavior is similar to the case of the MC simulations without electron-electron scattering in Figure 4.20. Thus, while the SHE method can be comfortably used to predict gate currents at high drain and gate voltages, at low bias bias operating regimes it reveals to be inaccurate for gate current prediction.

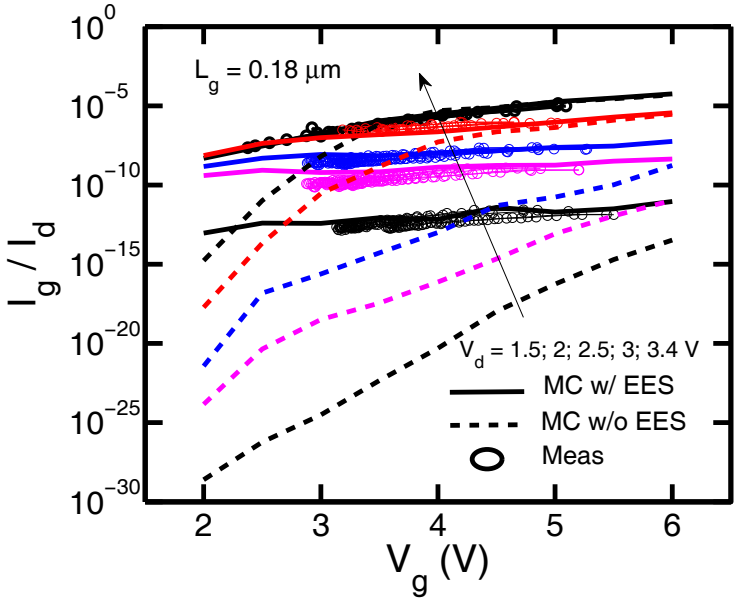


Figure 4.20: Comparison of the injection efficiencies (I_g/I_d) vs. gate voltage obtained from measurements (Meas) and the Monte Carlo (MC) on a 0.18 μm gate length device for various drain voltages. MC simulations are shown with or without Electron-Electron Scattering (EES).

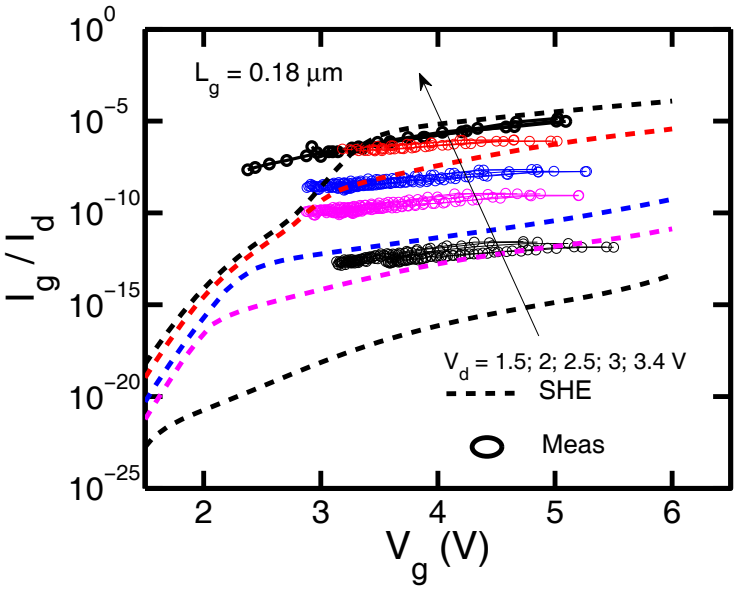


Figure 4.21: Comparison of the injection efficiencies (I_g/I_d) vs. gate voltage obtained from measurements (Meas) and the Spherical Harmonics Expansion (SHE) method on a 0.18 μm gate length device for various drain voltages.

4.2.2 Simulations using the 1D semi-analytic approach

In this paragraph, the HCI regime is simulated using the 1D non-local injection model presented in Chapter 3. This model will be used in conjunction with a Charge Sheet Model (CSM), as shown in Figure 4.22. CSM is based on Brews' equation [Brews 1978], [Gilibert 2004] which additionally includes overlap [Rideau 2010], fringe capacitance effects [Garetto 2009b] and charge sharing correction [Quenette 2009]. The calculation of the channel potential enables the use of the non local injection model which has been previously described and benchmarked against Monte Carlo simulations. The combined CSM - injection model methodology has been exposed in [Zaka 2011]. Subsection 4.2.2.1 recalls the necessary ingredients for the understanding of this work; a complete review can be found in [Tsividis 1987]. Subsection 4.2.2.2 proposes a potential correction which will be used in the third subsection 4.2.2.3 where the model is compared against measurements.

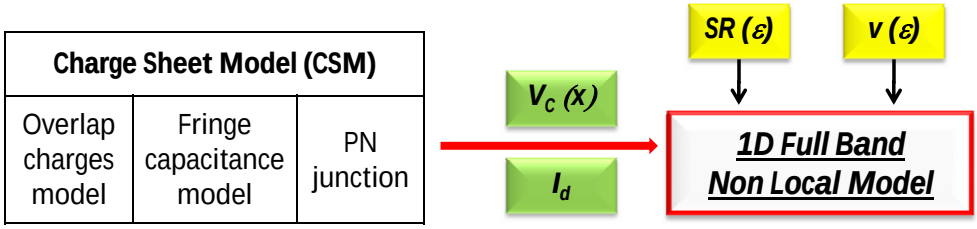


Figure 4.22: Overview of the compact modeling oriented approach including a Charge Sheet Model (CSM) capable to calculate the channel potential ($V(x)$) and the drain current (I_d) used as inputs by the non-local injection model with the scattering rate (SR) and the group velocity (v(ε)) (c.f. Chapter 2).

4.2.2.1 Description of the Charge Sheet Model

Starting from the Pao et Sah equation [Pao 1966], Brewer has proposed a CSM approach integrating the *Gradual Channel Approximation* (GCA) and valid for the subthreshold, weak and strong inversion regimes. GCA translates the fact that the lateral electric field in the channel is slowly varying compared to the vertical field. In this approach the surface potential $\Psi_S(x)$ is obtained from the iterative solution of the following implicit equation:

$$[V_{fg} - V_{fb} - \Psi_S(x)]^2 = \gamma^2 \Psi_T \left[\exp\left(-\frac{\Psi_S(x)}{\Psi_T}\right) + \frac{\Psi_S(x)}{\Psi_T} - 1\right] + \gamma^2 \Psi_T \exp\left(-2\frac{\Psi_F}{\Psi_T}\right)\left[\exp\left(\frac{\Psi_S(x) - V_C(x)}{\Psi_T}\right) - \frac{\Psi_S(x)}{\Psi_T} - 1\right] \quad (4.4)$$

This equation is solved at the source (yielding $\Psi_S(0)$) and at the drain (yielding $\Psi_S(L)$) where the channel potential $V_C(x)$ is set to 0 and V_{ds}, respectively. However, due to the model assumptions (GCA) a discussion will follow concerning the

boundary condition at the drain. V_{fg} and V_{fb} are the floating gate and flat-band voltages, respectively, while $\Psi_T = k_bT/q$ is the thermal voltage, $\Psi_F = -(E_F - E_i)/q = k_bT/q \log(N_A/n_i)$ is the Fermi level potential defined with respect to the intrinsic silicon level. The body factor γ is defined as:

$$\gamma = \frac{\sqrt{2qN_A\varepsilon_0\varepsilon_{Si}}}{C_{ox}} \tag{4.5}$$

N_A is the channel doping while $C_{ox} = \varepsilon_0\varepsilon_{ox}/t_{ox}$ is the tunnel oxide capacitance per unit surface. The drain current is calculated by:

$$I_d = \mu\frac{W}{L}C_{ox}\left[F(L) - F(0)\right] \tag{4.6}$$

where F is a function of the surface potential defined as:

$$F(x) = \left[V_{fg} - V_{fb} - \Psi_S(x)\right]\Psi_S(x) - \Psi_T\Psi_S(x) - \Psi_T\gamma\sqrt{\Psi_S(x) - \Psi_T} - \frac{1}{2}\Psi_S(x)^2 - \frac{2}{3}\gamma\left[\Psi_S(x) - \Psi_T\right]^{3/2} \tag{4.7}$$

Thus, the drain current is given after the calculation of the surface potential at only two points, i.e. at the source and at the drain. From drain current conservation, the surface potential $\Psi_S(x)$ along the channel can be calculated using [Tsividis 1987]:

$$\frac{x}{L} = \frac{F(x) - F(0)}{F(L) - F(0)} \tag{4.8}$$

The easiest way to evaluate this expression is to give $\Psi_S(x)$ values contained in the $[\Psi_S(0);\ \Psi_S(L)]$ interval and retrieve x using Equation 4.8. The electrostatic channel potential ($V_C(x)$) can be finally calculated by inverting Equation 4.4. Together with the drain current, it constitutes one of the inputs of the non-local injection model (c.f. Figure 4.22). Chapter 3 has shown that the calculation of the distribution function, leading to the gate current, is based on the discretization of the potential profile along the channel. Hence, the accuracy of the profile directly impacts the model results. Below we report an extension proposed to increase the calculation accuracy.

4.2.2.2 Potential correction

When calculating the surface potential Ψ_S from the electrostatic channel potential (V_C) (Equation 4.4) at constant V_g, it can be seen that the relation $\Psi_S(V_C)$ shows different regimes [Tsividis 1987]. In particular, Ψ_S varies linearly with V_C in strong inversion and then it starts to saturate in moderate and weak inversion before attaining a constant value Ψ_{Sa} depending on V_{gb}, calculated as :

$$\Psi_{Sa} = \left(\sqrt{\gamma^2/4 + V_{gb} - V_{fb}} - \gamma/2\right)^2 \tag{4.9}$$

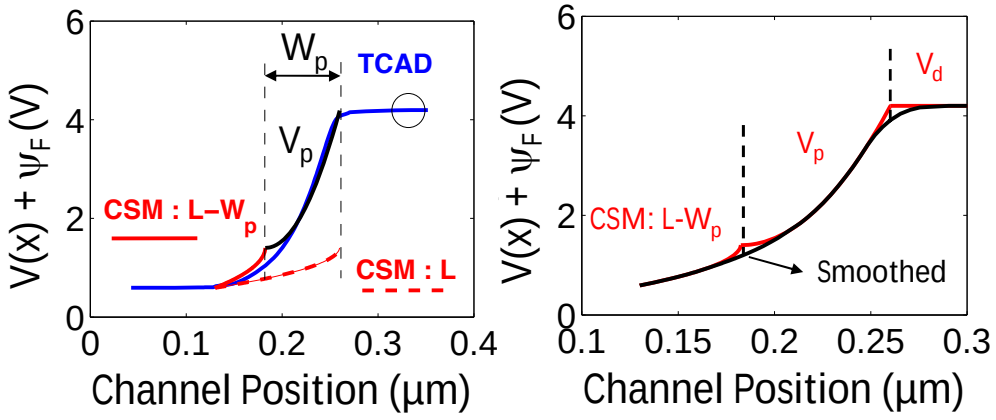
Any further increase of V_C does not change Ψ_{Sa}. However, in the HCI regime, the drain and the gate bias ranges ($V_d \sim 3-4V$ and $V_g \sim 2.5-5.5V$, respectively) are such that the portion of the channel close to the drain junction is often in the weak inversion/depletion regime. This means that in this portion, the surface potential is pinned at Ψ_{Sa} starting from a given channel position. The part of the $V_C(L)$ greater than the electrostatic potential corresponding to Ψ_{Sa} is not accounted for any spatial variation. The discrepancy between $(\Psi_S(L) - \Psi_S(0))$ and V_{ds} reflects the failure of the GCA. This in an intrinsic limitation of the model which does not account for this *transition* region where the lateral electric field is significant. In order to use the CSM in the framework of the Flash memory modeling, it is thus important to be able to describe the electrostatic potential in this region. To meet this requirement, an analytic PN-junction model has been employed at the channel/LDD junction in a two-step CSM iteration scheme.

The first iteration is performed with a channel length equal to the effective length $L = L_{eff} = L_g - L_{ov}$, with L_{ov} being the overlap extension. The channel potential is calculated by inverting Equation 4.4 with $\Psi_S(L) = \Psi_{Sa}$. Figure 4.23*a* reports the obtained channel potential for a 0.14 μm gate-length device at a given V_d/V_g bias configuration (*CSM : L* curve). Notice that the new value of the electrostatic potential at the drain $V_C(L)^{CSM}$ is smaller than V_d for the reasons explained above. In addition, this value is given at $x = L$ and not at the beginning of the depletion regime as no information on the latter exists in CSM. The latter region is determined by the reverse-biased PN junction model where the P-side (channel) is biased at $V_C(L)^{CSM}$ while the N-side (LDD) is biased at the externally applied bias $V_C(L)^{EXT} = V_d$. Under such conditions, the extension of the space charge region inside the channel W_p and the potential variation therein V_p, can be respectively calculated as:

$$W_p = \sqrt{\frac{2\varepsilon_{Si}}{qN_A}\left(V_d + V_{bi} - V_C(L)^{CSM}\right)} \tag{4.10}$$

$$V_p = V_C(L)^{CSM} + \frac{qN_A}{2\varepsilon_0\varepsilon_{Si}}(x_1)^2 \tag{4.11}$$

In these equations, $V_{bi} = k_bT/q\log(N_AN_{drain}/n_i^2)$ refers to the build-in potential for the PN junction, while the x_1 variable is defined in the $[L - W_p; L]$ interval. In this procedure, the extension of the space charge region inside the LDD has been neglected due to its much higher doping concentration. Then, CSM is rerun with an effective length of $L = L_g - L_{ov} - W_p$ which represents the inversion layer length in the channel where the GCA holds. Keeping in mind that the other parameters (doping, bias, ...) have been kept constant, the same potential $V_C(L)^{CSM}$ is obtained at $x = L$ ($CSM : L - W_p$ curve in Figure 4.23*a*). The V_p profile is also reported in Figure 4.23*a*. At this point, the channel potential consists of three parts: the potential issued from CSM with $L = L_g - L_{ov} - W_p$, V_p potential in the W_p*extension*, the constant drain voltage region. The continuous piecewise channel potential requires a final smoothing in order to be safely used for injection purposes (Figure 4.23*b*).



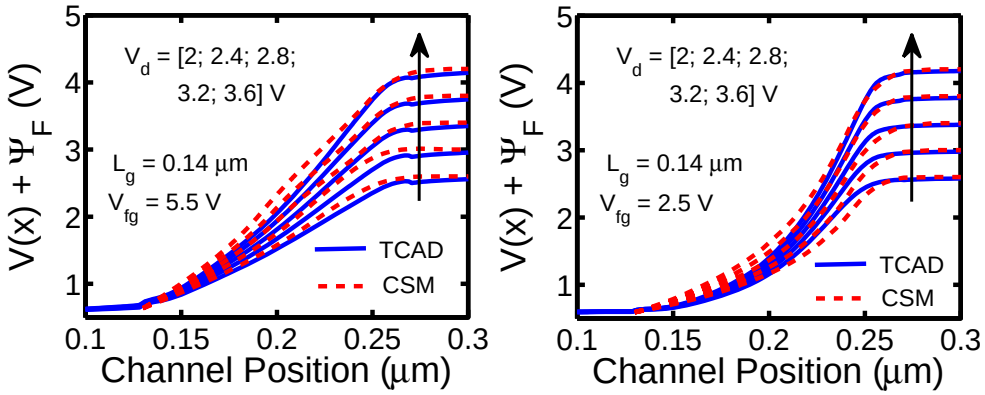
(a) (b)

Figure 4.23: *(a)* Channel potential V(x) calculated with the Charge Sheet Model (CSM) for a gate length of $L = L_g - L_{ov}$ (dashed) and for a gate length of $L = L_g - L_{ov} - W_p$ (plain). The figure is completed by the potential in the PN junction region V_p of length W_p (delimited by the two vertical lines) and the potential obtained at the interface by a 2D TCAD simulation. *(b)* Piecewise channel potential composed of $CSM : L - W_p$, V_p and *constant* V_d curves alongside the smoothed curve ready to be used in the non-local injection model. For both figures: $L_g = 0.14\ \mu m$, V_d=3.6 V and V_{fg}=2.5V.

The results of the above procedure are compared with TCAD results in Figure Figure 4.24, reporting the potential profiles for various bias conditions on a 0.14 μm device. Note that TCAD simulations, here taken as a reference, are performed including real process simulations in agreement with the calibration procedure described in 4.2.1.1. The non-constant doping in the 2D structure has been at best approximated by an average channel doping value in CSM simulations. Nevertheless, a fairly good matching has been achieved in the programming bias range. In particular, this method is able to reproduce the channel potential when the latter goes from negligible ($V_g = 5.5$, $V_d = 2$) to strong ($V_g = 2.5$, $V_d = 3.6$) depletion region.

4.2.2.3 Comparison with I_g/I_d measurements

Following the procedure described in Chapter 3 the 1D non-local injection model, which includes full-band elements of silicon band structure, has been applied to calculate the injection efficiency. Figure 4.25 reports such ratio as a function of the gate length for a constant drain bias and two floating gate voltages. The condensed view of Figure 4.25, with respect to the previous figures showed in 4.20 and 4.21, has the merit to immediately provide the evolution of the efficiency with gate length, which is often a major request during device optimization. Thus, at high gate voltage regime ($V_g > V_d$) where most of the injection occurs, the reduction of the



(a) (b)

Figure 4.24: Channel potential V(x) calculated with 2D TCAD simulations at the Si/SiO_2 interface and with the Charge Sheet Model including the potential correction. The comparison is performed for a 0.14 μm gate-length device at different bias conditions relevant for Flash programming.

gate length increases the device efficiency. Good agreement between the model and the measurements has been achieved, thus demonstrating its use in the context of compact approaches.

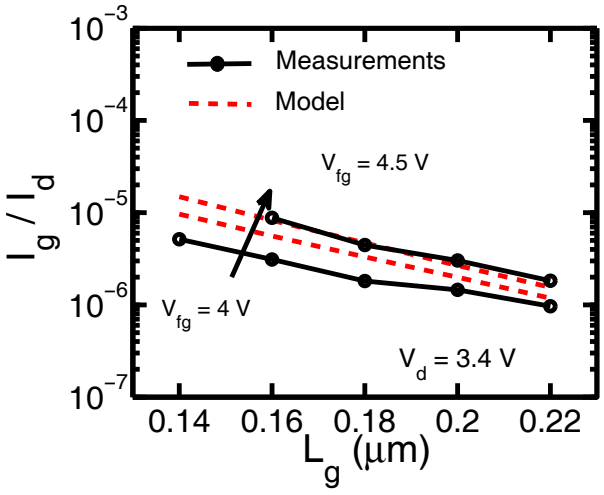


Figure 4.25: Injection efficiency (I_g/I_d) as a function of the device gate length obtained from measurements and the 1D non-local model at a constant drain voltage (V_d=3.4 V) for two different floating gate voltages.

4.3 Drain disturb regime

The hot carrier injection during the programming phase certainly constitutes the most important aspect of hot carrier mechanisms in a Flash cell. However, injection of hot carriers may also occur in unwanted circumstances, such as during the drain disturb phases. This encountered phenomenon is briefly introduced in subsection 4.3.1, while the simulation methodology is described in subsection 4.3.2. Finally, application to realistic cells and an optimization case-study will be presented in subsection 4.3.3.

4.3.1 The drain disturb phenomenon

The memory cells are organized in an array where each row and column is specifically addressable. In order to program a cell, the line connecting the drains of all cells in a given column (called *bitline*) and the line connecting the control gates of all cells in a given row (called *wordline*) should be biased (source and substrate contacts are grounded), as shown in Figure 4.26. The cell located at the intersection of the selected lines is subject to programming conditions. However, the other cells situated along the bitline and the wordline are subject to unwanted electrical stress, respectively called *drain disturb* and *gate disturb*. Hence, these unwanted operating regimes should first be evaluated and then the cell should be optimized to be as immune as possible.

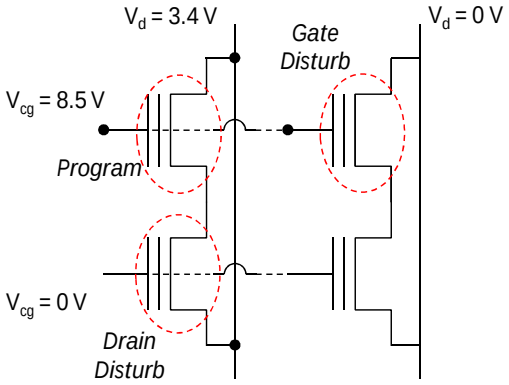
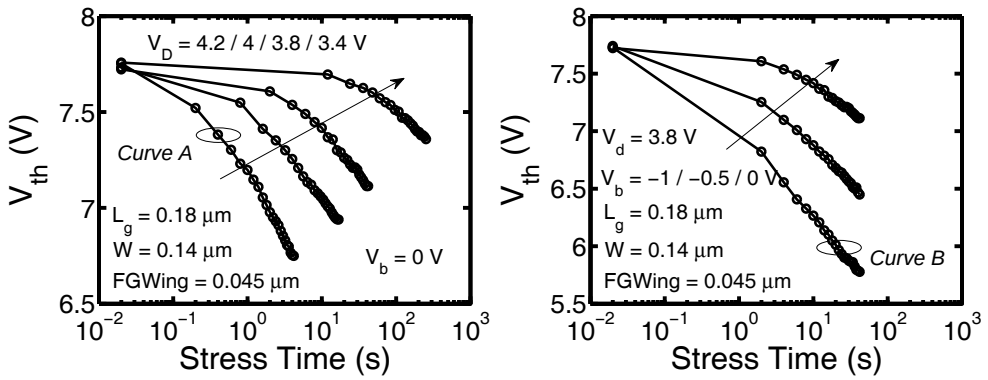


Figure 4.26: Schematic illustration of the cells under programming, drain disturb and gate disturb conditions.

The effect of these regimes on the cell depends on whether the latter is in an erased or a programmed state, due to different floating gate potentials. Figure 4.27 reports the charge loss (V_{th} decrease) in the floating gate as a function of the drain disturb time duration for a programmed cell. Notice that the charge loss is enhanced by increasing drain and bulk (in absolute values) voltage. The same measurements have been performed for cells which are initially in the erased state (Figure 4.28*a*). In this case, charge gain occurs yielding a V_{th} increase. Traditionally, the charge loss for a programmed cell is often compared to charge gain for an erased cell in order to

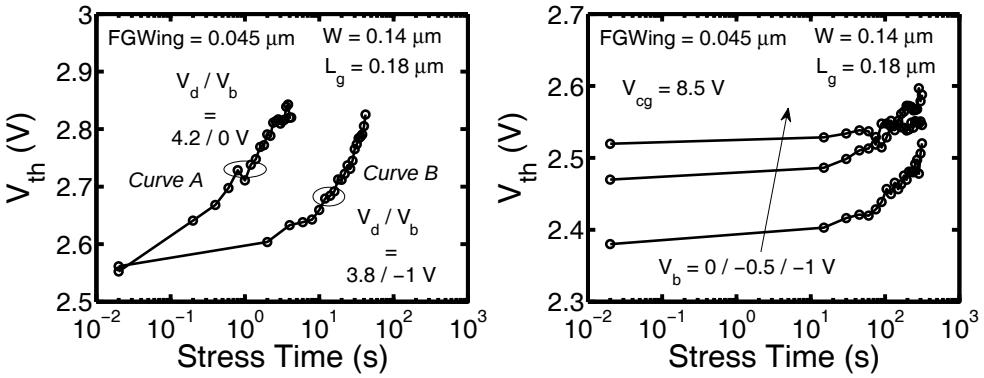
determine the worst case scenario and consequently determine the underlying dominating physical mechanism [Nair 2004], [Kumar 2006]. The comparison between *Curves A* of Figures 4.27*a* and 4.28*a* as well as the comparison between *Curves B* of Figures 4.27*b* and 4.28*a*, representative of two different programming conditions, shows that the threshold voltage shift is higher in the case of charge loss compared to the case of charge gain. Furthermore, the threshold voltage shift under gate disturb conditions (only charge gain has been reported in Figure 4.28*b*) is less stronger than the one occurring in drain disturb conditions. Therefore, charge loss for a programmed cell submitted to drain disturb constitutes the most critical aspect for the investigated cells and will be the object of this section. It ought to be mentioned that this observation depends on the cell architecture. For example, the same worst case scenario was found and discussed in [Chimenton 2002], [Ielmini 2006] while the charge gain effect during drain disturb was found the most critical and analyzed in [Nair 2004], [Kumar 2006].



(a) (b)

Figure 4.27: Measured threshold voltage variation vs. stress time duration for a programmed cell subject to different drain *(a)* and bulk *(b)* biases. The first point of the curves is the threshold voltage after programming.

The charge loss process observed in Figure 4.27 may possibly have two different origins: either electrons tunnel out of the floating gate or holes are injected into the floating gate. Considering the applied biases under such conditions, the previous studies in the field [Rakkhit 1990], [Chimenton 2002], [Ielmini 2006] have indicated hot hole injection during the drain disturb phase as the main mechanism. The V_b dependence shown in Figure 4.27*(b)* supports this argument for the investigated cells. The origin of hot holes is ascribed to pair generation mechanisms such as band-to-band tunneling (BTBT) and impact ionization (II). It is well known that for high electric fields ($\sim$ 1MV/cm) electrons tunneling from the conduction band to the valence band of silicon become significant [Schenk 1993]. Such fields may exist all around the channel/drain junction and close to the oxide. The following paragraph describes the approach used to simulate this phenomenon.



(a) (b)

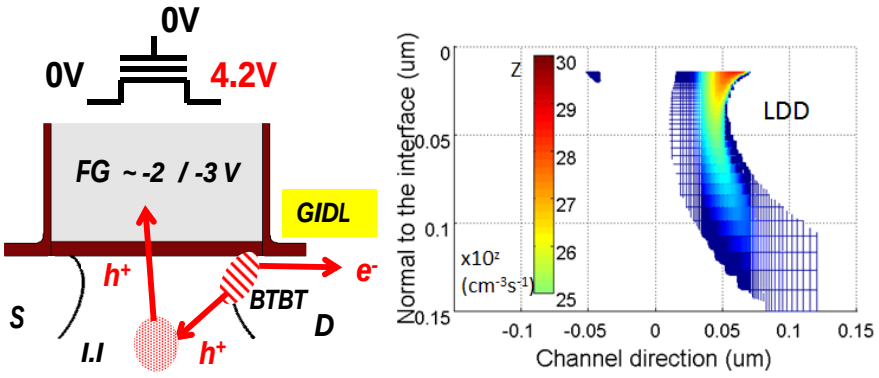
Figure 4.28: Measured threshold voltage variation vs. stress time duration for an erased state submitted to drain disturb *(a)* and gate disturb conditions *(b)*. The first point of the curves is the threshold voltage after erase.

4.3.2 Simulation methodology

A coupled 2D TCAD - MC simulation approach has been employed to estimate the hot hole injection into the floating gate. An accurate description of the electric field around the substrate/drain junction is necessary for the calculation of the space dependent BTBT rate. The latter is calculated within SDevice [Synopsys 2010a] using Schenk's model [Schenk 1993] and is reported in Figure 4.29*f* or bias conditions representative of drain disturb regime. The electrons generated by BTBT are collected at the drain and constitute the Gate Induced Drain Leakage (GIDL) [Rideau 2010]. The drain current is then used as an indicator of the strength of the BTBT mechanism. On the other hand the generated holes are pushed away from the drain and the field lines redirect them towards the source, the interface or the substrate [Ielmini 2006]. During the travelled distance, the holes are accelerated and a fraction of them generate other electron-hole pairs by impact ionization (Figure 4.29*(a)*). The secondary carriers are again accelerated giving birth to other carriers via impact ionization. During this scheme, a small fraction of the holes gain enough energy to overcome the Si/SiO_2 barrier ($\sim$4.7 eV) and de-program the cell.

While the BTBT rate can be accurately calculated within the TCAD tools, the generated carriers and their complex trajectories are not well accounted for within this framework. For this reason, a combined TCAD - MC methodology has been used to simulate this phenomenon [Ingrosso 2002], [Ielmini 2006]. Holes are generated within the MC proportionally to the 2D-TCAD BTBT cartography which is used as an input. The MC simulations include phonon (acoustic and optical) scattering, impact ionization and hole-hole interactions.

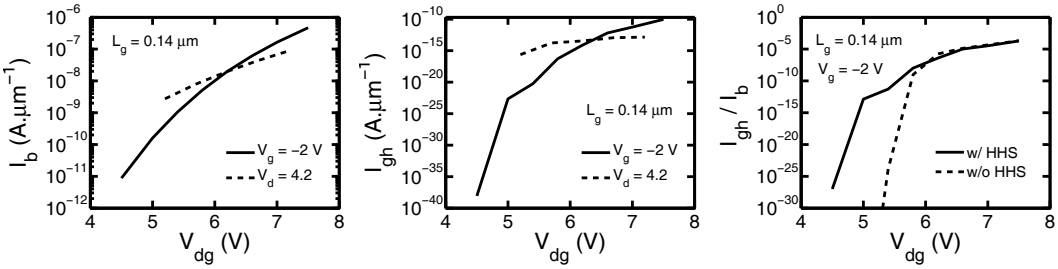
Similarly to the hot electron injection during the programming phase, several indicators can be employed to analyze this regime. Figure 4.30 reports the substrate current, the gate current and gate-to-substrate current ratio, henceforth called the



(a) (b)

Figure 4.29: *(a)* Schematics of the structure in the drain disturb regime and its main mechanisms. *(b)* Zoom of the 2D cartography of Band to Band tunneling around the channel/LDD junction, calculated with Schenk's model [Schenk 1993].

drain disturb efficiency, as a function of $V_{dg} = V_d - V_g$ for a 0.14 μm device. When the floating gate voltage varies in the interval [-1; -3]V, representative of a programmed cell state, both the substrate and the gate current show a rather stable behavior, whereas a strong reduction of the both current is observed when the drain voltage decreases. A similar trend is obtained for the drain disturb efficiency which has been reported for a constant floating gate voltage with or without hole-hole scattering. Similar to the case of electron injection during programming conditions, this interaction plays an important role at low drain voltage. Notice that only the gate current is affected by this mechanism, as identical substrate currents are obtained with or without the inclusion of this mechanisms.

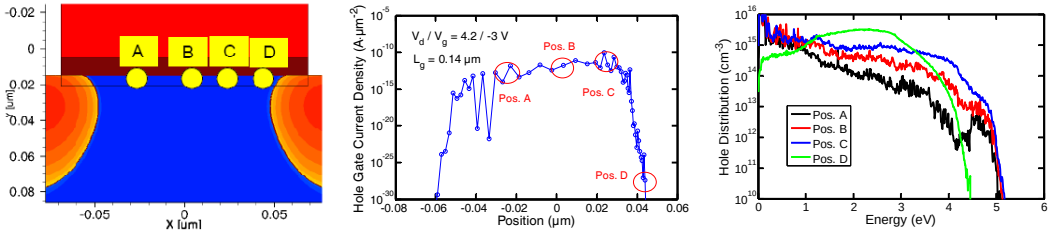


(a) (b) (c)

Figure 4.30: Substrate *(a)* and gate *(b)* currents as a function of the $V_{dg} = V_d - V_g$ voltage obtained with the MC for constant V_g (solid curves) or V_d (dashed curves) voltages. The drain disturb efficiency I_{gh}/I_b is given in *(c)* for a constant V_g with or without Hole-Hole Scattering (HHS).

In terms of microscopic quantities, Figure 4.31 reports the gate current density

and the hole distribution functions at four positions along the channel (A, B, C, D) highlighted in the 2D device structure image. The gate current density shows a maximum at around 20-30 nm from the junction with a rapid decrease for positions located towards the drain. However, a smoother reduction of the gate current is observed towards the source.



(a) (b) (c)

Figure 4.31: *(a)* Image of the simulated structure, highlighting four channel positions. *b* Gate current density along the channel obtained with the MC. *(c)* Hole distribution function vs. energy at the four positions highlighted in *(a)*.

The exposed methodology is applied in the next paragraph to optimize the previously calibrated structure (4.2.1.1) in terms of immunity to the drain disturb. In this process, the injection efficiency during programming conditions will be considered as well.

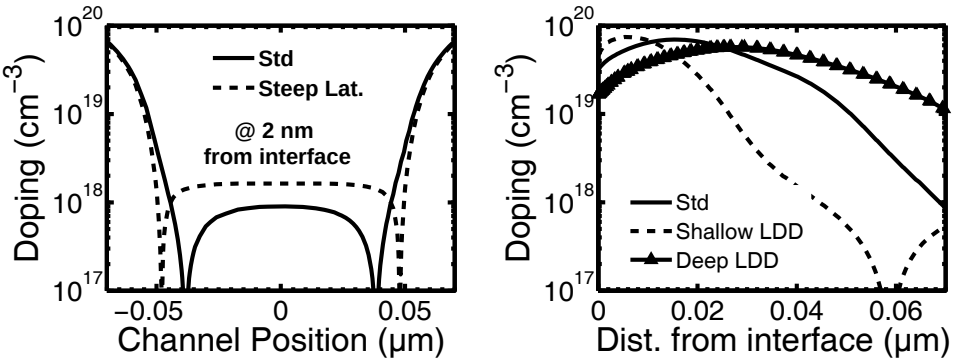
4.3.3 Device optimization and comparison with measurements

Table 4.1 shows the investigated structures and the abbreviations used in the following graphs. The standard structure is the one described in 4.2.1.1 and constitutes the starting point of this optimization procedure. Figure 4.32 reports lateral and vertical cuts of the doping profiles close to the interface for the considered structures. It can be seen that particular attention has been paid about the gradient of the doping profiles. The lateral steepness effect has been studied by increasing the channel doping and the lateral doping gradient with respect to the standard structure (Figure 4.32*a*). On the contrary, the source/drain metallurgic junction depth has been varied in order to study the vertical doping steepness effect (Figure 4.32*b*). For both cases, care has been taken not to strongly modify the doping in the other direction : vertical and lateral direction for the former and latter case, respectively. Channel, LDD dose and energy implants are tuned to obtain the desired profiles, starting from the standard one.

Figure 4.33 reports the injection efficiencies as a function of the gate voltage for the considered structures. Channel doping and lateral junction steepness appear to be critical for the injection efficiency variation: a two-fold increase is observed in Figure 4.33*a* for the steep junction despite the fact that the channel-length increases. On the contrary, Figure 4.33*b* shows that, for the considered structure, the injection efficiency is much less sensitive to the vertical LDD profile (i.e. the junction

Structure Description	Abbreviation
Standard matching exp.	Std
Steeper Lateral Junctions	Steep Lat.
Shallower LDD implant	Shallow LDD
Deeper LDD implant	Deep LDD

Table 4.1: The investigated structures and their abbreviations. The standard structure is the one described in 4.2.1.1.



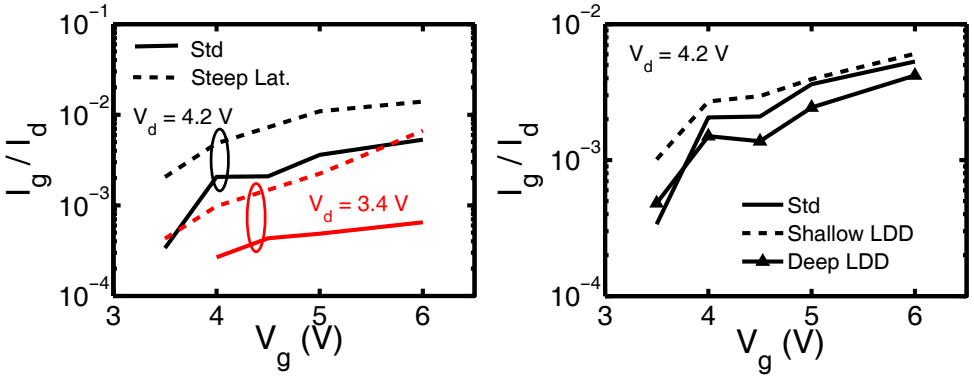
(a) (b)

Figure 4.32: *(a)* Doping profiles of the Std and Steep Lat. structures in the channel direction. *(b)* Vertical doping profiles of Std, Shallow and Deep LDD structures in the overlap region.

depth variation). The qualitative trends of Figure 4.33 could have also been separately predicted with a rapid estimation of the maximum lateral electric field in the channel, assuming a lucky electron model approach [Hu 1983]. However, structure optimization also requires a quantitative balance between many concurrent effects : channel doping, channel length, junction steepness and junction depth.

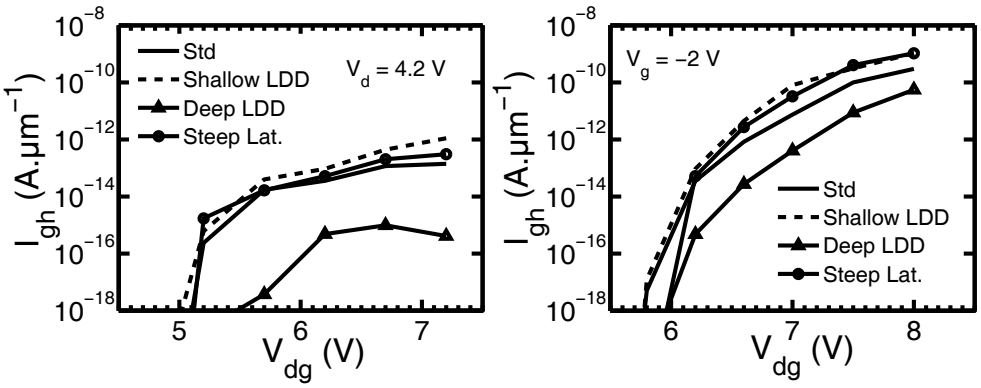
Figure 4.34 reports the drain disturb current as a function of V_{dg} for constant V_d = 4.2V and constant V_g = -2V, respectively. We see that structures having steeper junctions, either horizontally or vertically, show a constant or increased current compared to the standard structure. This effect is particularly enhanced when V_g varies within [-1; -3V] interval which corresponds to the floating gate voltage for a programmed cell. On the contrary, the drain disturb current is reduced for all bias conditions if a smoother LDD vertical doping (corresponding to increased source/drain junction depths) is adopted. Such a trend has been previously reported [Nair 2004] and shown to be consistent with experimental data.

Hence, from Figures 4.33 and 4.34 it seems that the Deep LDD structure shows an improvement in terms of drain disturb (reduced hole gate current) without too much sacrificing the injection efficiency during programming conditions. To un-



(a) (b)

Figure 4.33: Injection efficiency for CHE programming (I_g/I_d) vs. gate voltage (V_g) reported for the Std and the Steep Lat. structures at two drain voltages *(a)* and for Std, Shallow and Deep LDD structures for V_d=4.2V *(b)*.



(a) (b)

Figure 4.34: Hole gate current (I_{gh}) during drain disturb vs. V_{dg} at constant V_d=4.2V *(a)* and constant V_g=-2V *(b)* for all the investigated structures.

derstand such a behaviour, Figure 4.35 compares the Standard and Deep LDD structures in terms of BTBT current and Drain Disturb Efficiency (DDE). BTBT current is reduced for the Deep LDD structure by approximately one decade. Moreover, the DDE is also reduced especially at low V_{dg}. These observations imply that the disturb reduction observed for the Deep LDD structure is not only due to a BTBT reduction, but also to a different injection efficiency of the generated holes.

Finally, Figure 4.36 reports the comparison between the above MC simulations and the measurements performed on 0.14 μm cells. The increase of the floating gate potential due to charge loss as a function of disturb time for a constant V_d is reported for the Standard and the Deep LDD structures. The latter is indeed less subject to hole injection and therefore constitutes an improvement in terms of drain

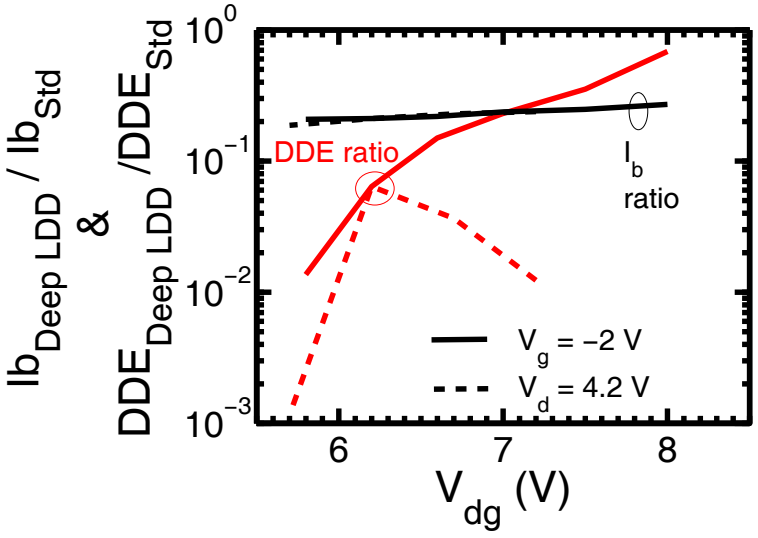


Figure 4.35: Relative variation of the substrate current (I_b) and of the Drain Disturb Efficiency (DDE = I_{gh}/I_b) generated in the Deep LDD structure with respect to Std structure. Ratios are plotted vs. V_{dg}: V_g=-2V and V_d varying (solid curves), V_d=4.2V and V_g varying (dashed curves).

disturb. Good agreement between measurements and simulations has been achieved for both structures, demonstrating the soundness of the simulation approach.

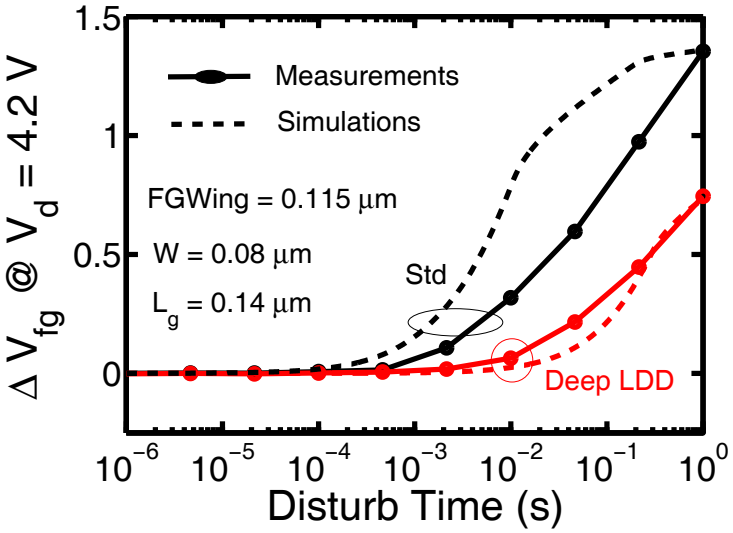


Figure 4.36: Floating gate potential variation (ΔV_{fg}) vs. disturb time at constant V_d=4.2V for Std and Deep LDD structures obtained with measurements (symbols), from the procedure described in section 4.1.3, and TCAD - MC simulations (dashed).

4.4 Conclusions

In this chapter, simulation and experimental results on hot carrier injection have been compared on a broad range of gate-lengths and bias and process configurations. The first section describes the applied characterization methodology for the purpose of extracting the intrinsic cell injection characteristics. Measurements on Flash cells and equivalent transistors are required in order to obtain accurate results. In this perspective, several sources of uncertainty which allow to have an estimation of the error bar in the considered measurements, are also discussed. The developed methodology has been adopted in two hot carrier injection configurations.

The injection of hot electrons into the floating gate during the programming phase has been the object of the second section where various Flash device gate lengths and bias conditions have been considered for comparison with the simulators described in the previous chapters. On one hand, 2D simulations have been performed using the MC simulator and the SHE method. Prior to the comparisons, the simulated cell has been calibrated to best match the measured electrostatic parameters and the simulator ingredients have been slightly adjusted to match the mean velocity at low fields, the II coefficient and the quantum yield. In the subsequent comparisons, the MC was shown to correctly reproduce experiments in almost all the investigated cases and in particular in the low voltage regime ($V_d < 3V$) where the Electron-Electron Scattering mechanism plays a crucial role. This regime also demonstrates the limits of the present SHE method which predicts a smaller injection current due to the absence of EES. On the other hand, 1D simulations have also been performed after combining the developed 1D non-local injection model and a Charge Sheet Model. A correction has been included in the latter in order to obtain a realistic potential profile close to the channel/drain junction. It was shown that this methodology compares well with the experimental results as a function of the gate length, in particular in the high voltage injection regime ($V_g > V_d > 3$) where most of the injection occurs in standard bias configuration.

Finally, the drain disturb regime has been presented in the third section. First, it was experimentally shown that the charge loss (V_{th} decrease) in a programmed cell due to hot hole injection is the strongest effect of this unwanted regime. As a result, a combined TCAD-MC simulation methodology was set up in order to grasp the intrinsic properties of this regime and to provide guidelines for the reduction of hole injection. Simulations results, confirmed by experimental data, showed that deeper LDD implants, which result in smoother profiles, make Flash devices more immune to drain disturb as a consequence of both a lower band-to-band tunneling current and a reduced injection efficiency for the generated holes.

Bibliography

[Brews 1978] J.R. Brews. *A charge-sheet model of the MOSFET.* Solid-State Electronics, vol. 21, no. 2, pages 345–355, 1978. (Cited on pages 57 and 118.)

[Bude 1992a] J. Bude, K. Hess and G.J. Iafrate. *Impact ionization: beyond the Golden Rule.* Semiconductor Science and Technology, vol. 7, page B506, 1992. (Cited on page 114.)

[Bude 1992b] J. Bude, K. Hess and G.J. Iafrate. *Impact ionization in semiconductors: Effects of high electric fields and high scattering rates.* Physical Review B, vol. 45, no. 19, page 10958, 1992. (Cited on pages 17, 18, 26, 59, 93 and 114.)

[Bude 1995] J.D. Bude and M. Mastrapasqua. *Impact ionization and distribution functions in sub-micron nMOSFET technologies.* IEEE Electron Device Letters, vol. 16, no. 10, pages 439–441, 1995. (Cited on page 114.)

[Canali 1975] C. Canali, C. Jacoboni, F. Nava, G. Ottaviani and A. Alberigi-Quaranta. *Electron drift velocity in silicon.* Physical Review B, vol. 12, no. 6, page 2265, 1975. (Cited on page 113.)

[Cartier 1993] E. Cartier, M.V. Fischetti, E.A. Eklund and F.R. McFeely. *Impact ionization in silicon.* Applied Physics Letters, vol. 62, no. 25, pages 3339–3341, 1993. (Cited on pages 18, 93, 113 and 114.)

[Chimenton 2002] A. Chimenton, A.S. Spinelli, D. Ielmini, A.L. Lacaita, A. Visconti and P. Olivo. *Drain-accelerated degradation of tunnel oxides in flash memories.* In International Electron Devices Meeting (IEDM) 2002, pages 167–170. IEEE, 2002. (Cited on page 124.)

[DiMaria 1985] D.J. DiMaria, T.N. Theis, J.R. Kirtley, F.L. Pesavento, D.W. Dong and S.D. Brorson. *Electron heating in silicon dioxide and off-stoichiometric silicon dioxide films.* Journal of Applied Physics, vol. 57, no. 4, pages 1214–1238, 1985. (Cited on page 113.)

[Eitan 1981] B. Eitan and D. Frohman-Bentchkowsky. *Hot-electron injection into the oxide in n-channel MOS devices.* IEEE Transactions on Electron Devices,, vol. 28, no. 3, pages 328 – 340, March 1981. (Cited on pages 73 and 114.)

[Fischetti 1988] M.V. Fischetti and S.E. Laux. *Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects.* Physical Review B, vol. 38, no. 14, pages 9721–9745, 1988. (Cited on pages 19 and 113.)

[Fischetti 1995] M.V. Fischetti, S.E. Laux and E. Crabbe. *Understanding hot electron transport in silicon devices: Is there a shortcut?* Journal of Applied Physics, vol. 78, no. 2, pages 1058 –1087, July 1995. (Cited on pages 6, 18, 19, 20, 21, 22, 28, 29, 93, 94, 113 and 114.)

[Garetto 2009a] 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 European

Solid State Device Research Conference (ESSDERC) 2009, 2009. (Cited on page 106.)

[Garetto 2009b] 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. (Cited on page 118.)

[Ghetti 2003] A. Ghetti. *Hot-electron induced MOSFET gate current simulation by coupled silicon/oxide Monte Carlo device simulation.* Solid-State Electronics, vol. 47, no. 9, pages 1507–1514, 2003. (Cited on pages 6 and 113.)

[Gilibert 2004] F. Gilibert, D. Rideau, S. Bernardini, P. Scheer, M. Minondo, D. Roy, G. Gouget and A. Juge. *DC and AC MOS transistor modelling in presence of high gate leakage and experimental validation.* Solid-State Electronics, vol. 48, no. 4, pages 597 – 608, 2004. <ce:title>ULIS 2003 conference</ce:title>. (Cited on page 118.)

[Goldsman 1988] N. Goldsman and J. Frey. *Electron energy distribution for calculation of gate leakage current in MOSFETs.* Solid-State Electronics, vol. 31, no. 6, pages 1089–1092, 1988. (Cited on pages 23 and 114.)

[Hasnat 1996] K. Hasnat and C. Yeap. *A pseudo-lucky electron model for simulation of electron gate current in submicron NMOSFET's.* IEEE Transactions on Electron Devices, vol. 43, no. 8, pages 1264–1273, 1996. (Cited on pages 21, 24, 25, 28, 48, 66, 84 and 114.)

[Hu 1983] C. Hu. *Hot-electron effects in MOSFETs.* In International Electron Devices Meeting (IEDM) 1983, volume 29, pages 176–181. IEEE, 1983. (Cited on page 128.)

[Ielmini 2006] D. Ielmini, A. Ghetti, A.S. Spinelli and A. Visconti. *A study of hot-hole injection during programming drain disturb in flash memories.* IEEE Transactions on Electron Devices, vol. 53, no. 4, pages 668–676, 2006. (Cited on pages 6, 124 and 125.)

[Ingrosso 2002] G. Ingrosso, L. Selmi and E. Sangiorgi. *Monte Carlo Simulation of Program and Erase Charge Distributions in NROM (TM) Devices.* In European Solid-State Device Research Conference (ESSDERC) 2002, pages 187–190. IEEE, 2002. (Cited on page 125.)

[Jacoboni 1983] C. Jacoboni and L. Reggiani. *The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials.* Review of Modern Physics, vol. 55, no. 3, pages 645–705, Jul 1983. (Cited on pages 15, 16, 17, 19 and 113.)

[Jungemann 1996] C. Jungemann, R. Thoma and W.L. Engl. *A soft threshold lucky electron model for efficient and accurate numerical device simulation.* Solid-State Electronics, vol. 39, no. 7, pages 1079–1086, 1996. (Cited on pages 20, 31, 79 and 111.)

[Kamakura 1999] Y. Kamakura, I. Kawashima, K. Deguchi and K. Taniguchi. *Monte Carlo simulation of quantum yields exceeding unity as a probe of high-energy hole scattering rates in Si.* In International Electron Devices Meeting (IEDM) 1999, pages 727–730. IEEE, 1999. (Cited on page 113.)

[Kolodny 1986] A. Kolodny, S.T.K. Nieh, B. Eitan and J. Shappir. *Analysis and modeling of floating-gate EEPROM cells.* IEEE Transactions on Electron Devices, vol. 33, no. 6, page 835, 1986. (Cited on pages 106 and 141.)

[Kumar 2006] P.B. Kumar, D.R. Nair and S. Mahapatra. *Using soft secondary electron programming to reduce drain disturb in floating-gate NOR flash EEPROMs.* IEEE Transactions on Device and Materials Reliability, vol. 6, no. 1, pages 81–86, 2006. (Cited on page 124.)

[Nair 2004] D.R. Nair, S. Mahapatra, S. Shukuri and J.D. Bude. *Drain disturb during CHISEL programming of NOR flash EEPROMs-physical mechanisms and impact of technological parameters.* IEEE Transactions on Electron Devices, vol. 51, no. 5, pages 701–707, 2004. (Cited on pages 124 and 128.)

[Pao 1966] H.C. Pao and C.T. Sah. *Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors.* Solid-State Electronics, vol. 9, no. 10, pages 927–937, 1966. (Cited on page 118.)

[Quenette 2009] V. Quenette, D. Rideau, R. Clerc, S. Retailleau, C. Tavernier and H. Jaouen. *Dynamic Charge Sharing modeling for surface potential based models.* 2009. (Cited on page 118.)

[Rakkhit 1990] R. Rakkhit, S. Haddad, C. Chang and J. Yue. *Drain-avalanche induced hole injection and generation of interface traps in thin oxide MOS devices.* In International Reliability Physics Symposium (IRPS) 1990, pages 150 –153, march 1990. (Cited on page 124.)

[Rideau 2010] D. Rideau, V. Quenette, D. Garetto, E. Dornel, M. Weybright, J.-P. Manceau, O. Saxod, C. Tavernier and H. Jaouen. *Characterization and modeling of gate-induced-drain-leakage with complete overlap and fringing model.* In International Conference on Microelectronic Test Structures (ICMTS) 2010, pages 210 –213, 2010. (Cited on pages 118 and 125.)

[Sano 1994] N. Sano and A. Yoshii. *Impact ionization rate near thresholds in Si.* Journal of Applied Physics, vol. 75, no. 10, pages 5102–5105, 1994. (Cited on pages 17, 93 and 114.)

[Schenk 1993] A. Schenk. *Rigorous theory and simplified model of the band-to-band tunneling in silicon.* Solid-State Electronics, vol. 36, no. 1, pages 19–34, 1993. (Cited on pages 124, 125 and 126.)

[Synopsys 2010a] Synopsys. *Synopsys Sentaurus, release D-2010.12, SDevice simulators*, 2010. (Cited on pages 21, 23, 24, 30, 125 and 156.)

[Synopsys 2010b] Synopsys. *Synopsys Sentaurus, release D-2010.12, SProcess simulators*, 2010. (Cited on pages 111 and 155.)

[Tsividis 1987] Y. Tsividis. Operation and modeling of the mos transistor. McGraw-Hill, Inc., 1987. (Cited on pages 118 and 119.)

[Van Overstraeten 1970] R. Van Overstraeten and H. De Man. *Measurement of the ionization rates in diffused silicon pn junctions.* Solid-State Electronics, vol. 13, no. 5, pages 583–608, 1970. (Cited on page 113.)

[Zaka 2011] A. Zaka, D. Garetto, D. Rideau, P. Palestri, J.P. Manceau, E. Dornel, Q. Rafhay, R. Clerc, Y. Leblebici, C. Tavernier and H. Jaouen. *Characterization and modelling of gate current injection in embedded non-volatile flash memory.* In International Conference on Microelectronic Test Structures (ICMTS) 2011, pages 130–135. IEEE, 2011. (Cited on page 118.)

Chapter 5

Modeling the cell degradation

In the previous chapters, a modeling and characterization methodology was established for the study of the Flash program operation involving hot electron injection in the floating gate. A large range of device lengths and bias configurations were analyzed by simulation and compared with measurements, demonstrating that they grasp the main features of this regime. However, the operation of the Flash cell includes many physical mechanisms, some of which constitute unwanted operating phases. A typical example is the drain disturb phenomenon discussed in the previous chapter, in which hot holes are injected into the floating gate resulting in a sizable charge loss. Oxide degradation is another typical and worrisome phenomenon affecting Flash performances during its lifetime. This chapter is devoted to this topic. Considering the vastness of this subject, only some specific aspects of the phenomenon will be presented and discussed.

The aptitude of the cell to conserve its performance over time strongly depends on the robustness against defect creation during the program, erase and read phases. Section 5.1 focuses on the endurance characteristics of the Flash cell. In particular, the observed reduction of the programming window during cycling, clearly pointing to oxide degradation processes, has been experimentally investigated with the purpose of separating the impact of the defects on each of the program, erase and read phases.

The observed degradation results from the traps created in the bulk oxide as well as near the Si/SiO_2 interface. These latter traps will be the subject of Section 5.2 which presents a microscopic modeling and simulation approach for the generation of such traps during the hot carrier injection regime. Such an approach allows for an estimation of the interface traps along the channel for various stress times. Comparisons with measurements on spatial trap profiles and electrical macroscopic parameters for various stress conditions demonstrate the soundness of this approach for the estimation of hot carrier induced interface traps.

5.1 Cell endurance

The endurance characteristics constitute an important indicator for Flash cells. This characteristics will be introduced in subsection 5.1.1. Our attention will then be focused on the analysis of the programming window degradation with cycling. In particular, subsection 5.1.2 will present an experimental method aiming to split the observed degradation into three components, each corresponding to the program, the erase and the read operation regime.

5.1.1 Endurance characteristics

In its lifetime, a Flash cell is meant to work over a great number of cycles. Each cycle is composed by a complete program and erase operation, respectively characterized by their threshold voltages V_{th}^P and V_{th}^E. Typically, the cell should endure $10^5 - 10^6$ cycles during which the erased and the programmed states should still be easily distinguishable. This is quantified by introducing a new indicator called *Programming Window*: $W = V_{th}^P - V_{th}^E$, defined as the difference between the cell states. The term *endurance characteristics* of a cell reflects the reduction of the window with increasing number of cycles [Aritome 1993]. The causes of the endurance degradation fall into two categories: *usual* oxide wear-out and *erratic* single-cell failures [Modelli 2004]. Among these two causes, only the former is addressed in the following section.

Figure 5.1 reports the experimental endurance characteristics of a 0.14 μm cell which is subject to two different cycling regimes. In one case, the cell is programmed by Channel Hot Electrons (CHE) and erased by Fowler-Nordheim (FN) which are the usual cycling conditions. In the other case, for comparison purposes, full Fowler-Nordheim cycling (program: V_{cg} at 18.9V for 10ms; erase: V_{cg} at -17.65V for 1ms) is considered. Such an operation is employed in NAND Flash memories. The threshold voltages throughout the cycles have been normalized with respect to the threshold voltage in the erased state (0%) and to the threshold voltage in the programmed state (100%) at the beginning of the endurance characteristics (after 1 cycle). The threshold voltage is defined as the control gate voltage needed to achieve a drain current of 8 μA with a drain voltage of 0.7V.

Figure 5.1 shows that the threshold voltage in the erased state increases during cycling for both operations. Instead, the program threshold voltage increases for FN-FN operation while it is nearly stable for CHE-FN operation. However, a closer look at the latter operation reveals that the threshold voltage first decreases until 5.10^3 cycles and then starts to increase. The programming window W is also shown in Figure 5.1. Similar behaviors have been recently observed in [Lee 2006], [Tao 2007b] and [Fayrushin 2009]. In order to reduce the endurance degradation, both process - and bias - based optimization solutions can be considered. However, prior to applying any solution, a good understanding of the degradation mechanism is necessary. In the following subsection, a characterization methodology is introduced to quantify the degradation during each of the main operating regimes as a

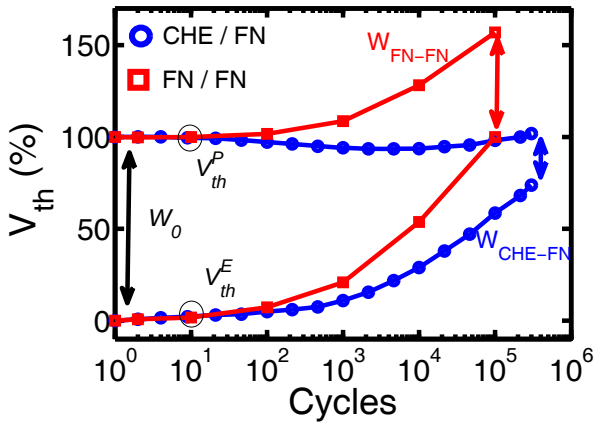


Figure 5.1: Normalized Flash cell threshold voltage after Erase (V_{th}^E) and Program (V_{th}^P) operation during cell cycling performed by Channel Hot Electron (CHE) - Fowler Nordheim (FN) or FN-FN configurations. The evolution of the *programming window* W, defined as the difference between program and erase threshold voltages, constitutes the endurance characteristics.

first step towards degradation reduction.

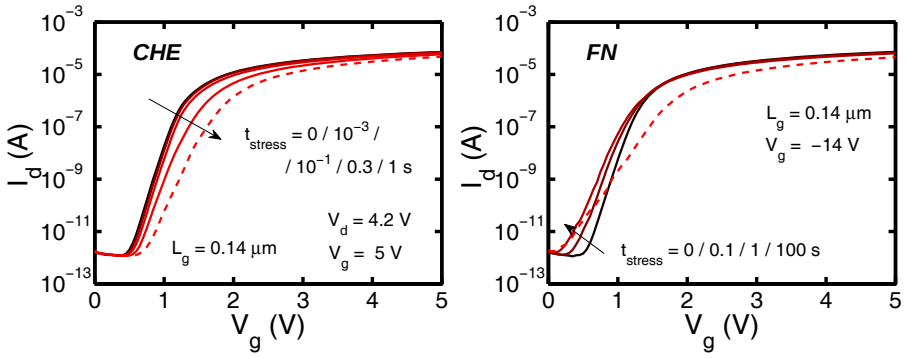
5.1.2 Experimental analysis

Both the CHE injection and the FN tunneling are known to degrade the tunnel oxide properties. In this subsection, the effects of traps on the window are analyzed from the experimental point of view. The threshold voltage after Program/Erase (P/E) operation reported in Figure 5.1 includes both the *intrinsic* behavior of program/erase operation and the read operation. It is important to separate these operations in order to acknowledge any variation of program/erase efficiency during cycling. As both operations strongly depend on the floating gate voltage (c.f. Chapter 4), the program/erase phases should be evaluated at a constant floating gate charge during cycling, so that only the effect of the traps created during stress be investigated. However, the separation of the *negative charges* coming from the electrons in the floating gate and the ones trapped in the oxide is not a trivial task.

5.1.2.1 Characterization of the equivalent transistor

The most direct way to tackle this problem is to perform Constant Voltage/Current Stress (CVS and CCS, respectively) experiments on equivalent transistors (floating and control gates are shorted). In these experiments, where no charge is stored in the floating gate (that is directly connected to the control gate), a constant bias or current close to the desired condition is applied at the drain/gate terminals or imposed at the gate terminal, respectively, for a given time duration (stress time). Figure 5.2 reports $I_d(V_g)$ characteristics after constant voltage stress CHE and FN

conditions at different stress times. Notice that in the case of CHE stress, the curves gradually shift towards higher threshold voltage while for FN-like stress first the sub-threshold slope is degraded and then the threshold voltage increases as well. It has been observed that at the beginning of the electron injection from the gate in the FN condition, a positive charge builds up in the oxide [Itsumi 1981]. At the same time, traps created during this phase and their subsequent filling during I_dV_g measurements increase the V_{th}. Both effects are thus globally counterbalanced at the beginning of the FN stress. Similar results have been reported for CCS setups in [Cappelletti 1999] and references therein.



(a) (b)

Figure 5.2: Drain current vs. gate voltage after Channel Hot Electron (CHE) *(a)* and Fowler-Nordheim gate injection (FN) *(b)* constant voltage stress conditions at different stress times for a 0.14 μm equivalent transistor device. All the curves are obtained at V_d=0.7V; the most degraded curve is dashed.

In order to study the evolution of the stress with time, commonly employed macroscopic electrical parameters are extracted from I_dV_g curves and reported in Figure 5.3 as a function of the stress time. All results are given relatively to the fresh value of the parameter (threshold voltage, sub-threshold slope, maximum of the transconductance). In the case of CHE, the drain voltage has been kept constant at 4.2 V and the gate voltage has been varied from 2 to 6 V in order to cover the whole V_g variation during the programming conditions. The same approach has been followed for FN stress with V_g varying from -15 to -11 V. Higher stress times are used for FN in agreement with the higher erase time which is 100-1000 times longer than the program duration in the cell operation. For all CHE and FN stress conditions, the parameters are degraded faster when higher V_g values (absolute value) are used. It can be noticed that while V_{th} is degraded faster for CHE conditions, for the reasons explained above, the sub-threshold slope and the transconductance are more sensitive to the FN stress condition. However, there is a strong effect of V_g on the degradation level and the analysis of the curves may lead to different conclusions if it is performed at a lower (absolute value) V_g. Considering that the floating gate voltage in a cell, by definition, varies in the program or erase

phases, the real cell degradation would be a mixture of all the investigated equivalent transistor DC conditions.

Although the evaluation on the equivalent transistor clearly indicates that both CHE and FN processes degrade the device, such information cannot be directly applied to the cell. Hence, an evaluation on cells can lead to very useful complementary information to investigate the effect of traps on transient operation.

5.1.2.2 Characterization of the Flash cell

An interesting idea was proposed in [Tseng 2001] where the authors designed special test structures which integrate a switch to access the floating gate. When the switch is not connected, the structure is used in a cell configuration with the charges in the floating gate being injected by CHE and extracted by FN. When the switch is connected, the charges in the floating gate are removed and a transistor-like configuration is recovered. This method thus allows to study the effect of traps (degradation) independently for program and erase phases. However, the layout modifications needed for the test structures result in a giant cell where program/erase dynamics are performed thousands of times more slowly compared to the case of a standard Flash cell. Thus, the insight gained in this approach is not directly applicable to realistic cells.

In a recent publication, Tao et al. [Tao 2007a] have come up with an interesting method to isolate the trap effect using only the Flash cell. This work, inspired by [Kolodny 1986], allows the authors to determine the neutral threshold voltage corresponding to the cell state with no charges in the floating gate. This voltage is determined during the cycling as a fitting parameter from the combination of the I_g current expression in the FN regime and of the electrostatic relations ruling the cell (c.f. Chapter 4). Hence, the traps' effect on endurance can be quantified. This methodology was applied to cells in which program and erase are performed by FN tunneling where the classical analytical expression [Lenzlinger 1969] of the current has been fitted to reproduce the measurements. However, a reliable extraction of the neutral threshold voltage in NOR Flash memories involving CHE injection, for which an accurate analytical expression is difficult to obtain (c.f. Chapter 2 and 3), has not yet been proposed.

In the following, a step-by-step experimental approach is proposed, aiming to separate the contributions of each operating regime on the endurance characteristics [Garetto 2011]. This is achieved by analyzing a set of measurements which include program and erase transients at different number of cycles obtained with the same bias conditions as in Figure 5.1. The proposed analysis is carried out on the CHE/FN condition and will eventually allow to build the endurance characteristics at the end of this section.

Figure 5.4 reports the current-voltage measurements during the read operation after cell erase (a) and the threshold voltage evolution during the erase phase after various number of cycles (b). The increase of the sub-threshold slope in Figure 5.4a with cycling indicates the presence of interface traps (also called interface

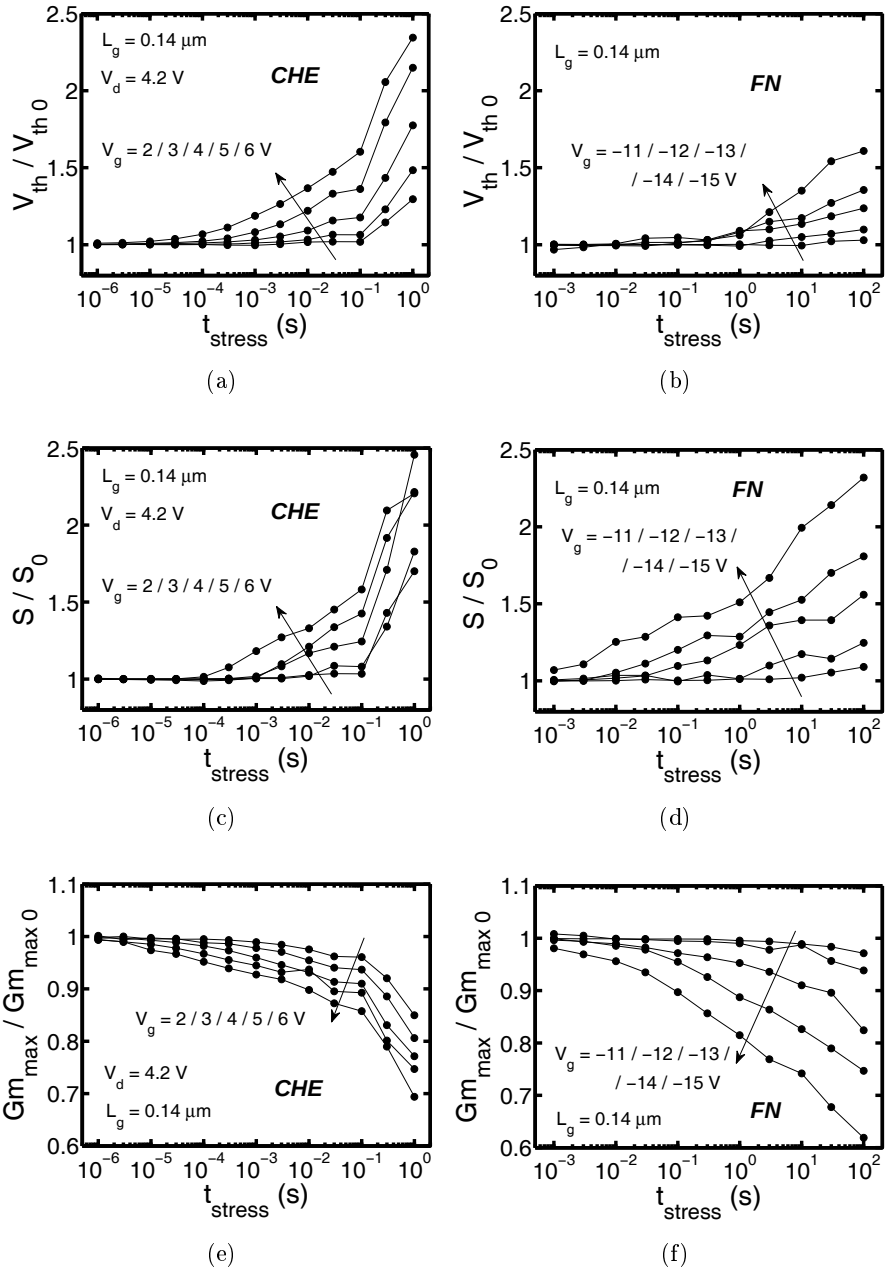


Figure 5.3: Threshold voltage *(a-b)*, sub-threshold slope *(c-d)* and maximum of transconductance *(e-f)* variation with stress time after different CHE and FN stress conditions performed on a 0.14 μm gate length equivalent transistor device.

states) created during the cycling and filled during the read operation. The most probable cause of this behavior are amphoteric P_b-centers located within the first angstroms of the oxide from the interface. They behave as acceptors in an inverted nMOS structure [Lenahan 1984]. In addition, the rigid shift of the threshold

voltage can suggest the presence of trapped electrons inside the oxide (E'-like defects) [Lenahan 1984], since both interface and oxide traps contribute to increase the threshold voltage shown in Figure 5.4*a*. In what follows we will not explicitly distinguish between P_b-centers and E' defects. In order to extract the intrinsic properties of the program/erase regimes during cycling, the trap contribution in the read operation should be first separated and then the P/E phases be evaluated at the same floating gate charge.

This can be achieved using the erase transient characteristics (Figure 5.4*b*) based on the following considerations. Figure 4.2 of Chapter 4 showed that by applying higher control gate voltages (in absolute values) the final threshold voltage is lower. However, the erase dynamics, i.e. the pace at which the threshold voltage decreases with erase time, was not affected by the control gate voltage change during most of the erase phase. This result is further confirmed in the cell calibration procedure described in [Zaka 2011], where the presence of parasitic capacitances which lower the gate coupling coefficient did not affect the erase dynamics either. In fact, the FN tunneling dynamics during the erase phase depends on the oxide field, which in turn depends on the initial floating gate condition, the charged traps in the oxide, the oxide thickness and the total capacitance [Zous 2004]. For a given structure with constant geometrical boundaries, the oxide field will only depend on the value of the floating gate potential at the beginning of the erase phase and on the quantity of trapped charges inside the oxide. As the number of traps increases with cycling, the floating gate potential evolutes as well at the beginning of a given erase cycle. However, we cannot know the exact floating gate potential as we only know the threshold voltage affected by the filling of an unknown oxide trap distribution during read operation. From the above considerations, we neglect the first part of the erase transient characteristics up to the $t_{E-V_{th}} = 0.2ms$, position where all the curves (i.e. transient for different erase cycles) will approximately show the same oxide field. In order to additionally guarantee the same *apparent* floating gate voltage, all the curves of Figure 5.4*b* are, rigidly shifted in order to join together at the same threshold voltage at $t = t_{E-V_{th}}$ equal to the one at the beginning of cycling, as shown in Figure 5.5(a). The shift during this procedure ΔV_{th}^R, reflects the effect of the traps during the read operation and has been calculated as:

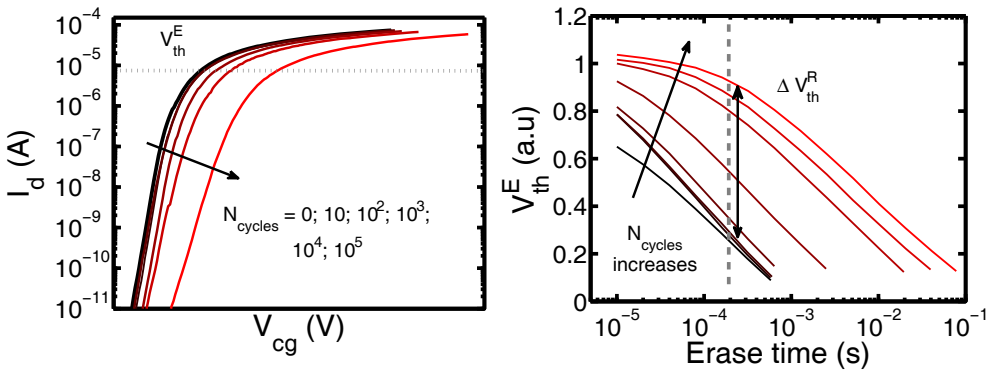
$$\Delta V_{th}^R(n_{cycles}) = V_{th}(t_{E-V_{th}}, n_{cycles}) - V_{th}(t_{E-V_{th}}, 0) \tag{5.1}$$

Figure 5.5*b* plots the extracted threshold voltage shift ΔV_{th}^R curve vs. number of cycles, whose evolution reflects the increase of the trap concentration in the oxide with cycling. Thus, the ΔV_{th}^R quantity is subtracted to the raw curves of Figure 5.4*a* to obtain the curves in 5.5*b*. All the curves are put at the same *apparent* floating gate voltage at $t_{E-V_{th}} = 0.2ms$. The degradation of the erase efficiency due to the presence of traps can be quantified at the end of the erase dynamics as:

$$\Delta V_{th}^E(n_{cycles}) = V_{th}^E(t_E, n_{cycles}) - \Delta V_{th}^R(n_{cycles}) - V_{th}^E(t_E, 0) \tag{5.2}$$

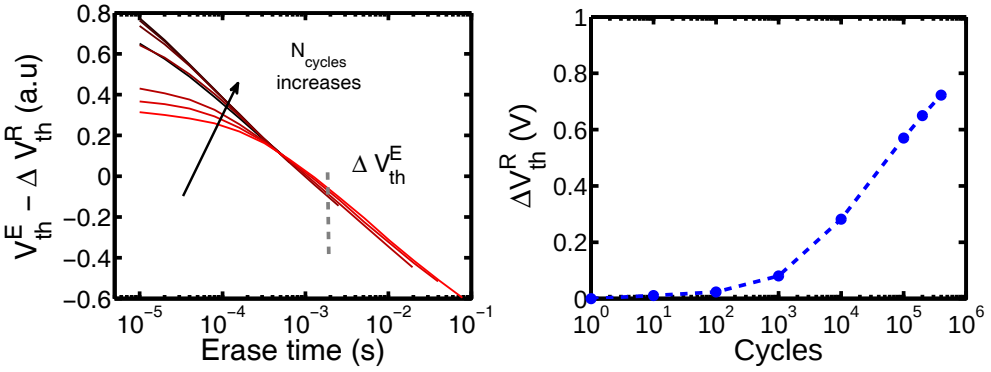
t_E is the time where the erase operation is considered finished (in this case it is

taken equal to 2 ms, c.f. Figure 5.5*(b)*). For the considered devices and conditions (CHE-FN and FN-FN), the ΔV_{th}^{E} is very small. Hence, the degradation of the intrinsic erase efficiency during cycling can be considered negligible.



(a) (b)

Figure 5.4: *(a)* Drain current vs. control gate voltage (read operation at V_d=0.7V) measured after the erase phase for various program/erase cycles. V_{th}^{E} is extracted at a drain current of 8 μA. *(b)* Threshold voltage during erase phase for an increasing number of program/erase cycles. The vertical dashed line (at $t = 0.2ms$) shows the threshold voltage shift (ΔV_{th}^{R}) between fresh and stressed devices reported in Figure 5.5.

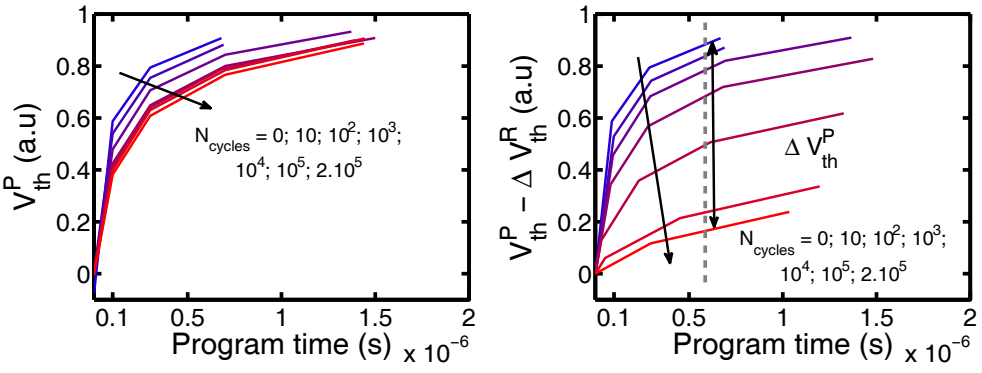


(a) (b)

Figure 5.5: *(a)* Threshold voltage during erase phase after different number of cycles corrected by ΔV_{th}^{R} to obtain the same floating gate charge for comparison purposes. The erase efficiency degradation (ΔV_{th}^{E}) is evaluated at 2 ms. *(b)* Threshold voltage shift (ΔV_{th}^{R}) vs. number of cycles, extracted after the data of Figure 5.4.

The effect of the traps during the read operation $\Delta V_{th}^{R}(n_{cycles})$ is equally present during the measurements of the program dynamics. Hence, the same quantity is

subtracted to the raw programming curves of Figure 5.6*a* to obtain the shifted curves plot *b* which allow to quantify the effect of the traps during the program operation.



(a) (b)

Figure 5.6: Threshold voltage during program phase after different number of cycles reported as-measured *(a)* and corrected by the threshold voltage shift *(b)*. The degradation of the program efficiency (ΔV_{th}^{P}) is evaluated after 0.7 μs.

The degradation of the program efficiency (ΔV_{th}^{P}) is defined as the negative threshold voltage shift at a given t_P programming time:

$$\Delta V_{th}^{P}(n_{cycles}) = V_{th}^{P}(t_P, n_{cycles}) - \Delta V_{th}^{R}(n_{cycles}) - V_{th}^{P}(t_P, 0) \tag{5.3}$$

t_P is here taken equal to 0.7 μs. Figure 5.7 reports the ΔV_{th}^{P} during cycling. Differently from the erase case of Figure 5.5*b*, the program efficiency is strongly impacted by the traps even in the very first cycles. The V_{th} decrease becomes significant after 10^3 cycles.

Finally, in order to validate this approach, Figure 5.8 compares the endurance obtained from classical read operation after the program/erase phases (as in Figure 5.1) with the endurance resulting from the analysis of the above program/erase dynamics (c.f. Figures 5.4 to 5.7). The program/erase threshold voltages during the cycling in the second approach are calculated as:

$$V_{th}^{E}(n_{cycles}) = V_{th}^{E}(0) + \Delta V_{th}^{R}(n_{cycles}) + \Delta V_{th}^{E}(n_{cycles}) \tag{5.4}$$

$$V_{th}^{P}(n_{cycles}) = W_0 + \Delta V_{th}^{R}(n_{cycles}) + \Delta V_{th}^{P}(n_{cycles}) \tag{5.5}$$

W_0 and $V_{th}^{E}(0)$ are respectively, the window and erase threshold voltage at a *fresh* state. As the degradation of the intrinsic erase efficiency (ΔV_{th}^{E}) is negligible, the erase threshold voltage during cycling (V_{th}^{E}) bears the signature of trap filling during the read operation (ΔV_{th}^{R}). In the meanwhile, the program threshold voltage (V_{th}^{P}) is the sum of two antagonistic affects: the trap filling during read operation and the degradation of the intrinsic program efficiency (ΔV_{th}^{P}), which respectively,

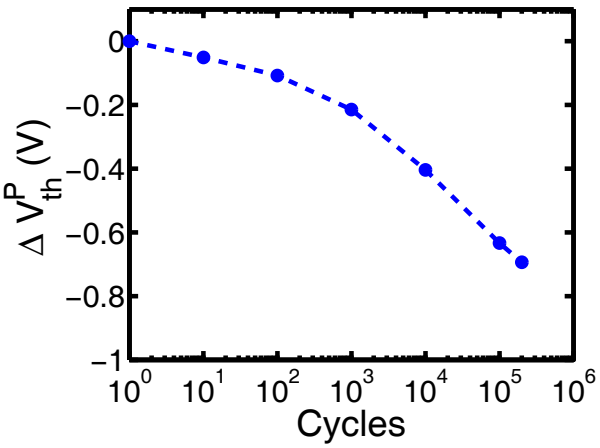


Figure 5.7: Decrease of the program efficiency vs. number of cycles, extracted after the data of Figure 5.6.

increase and decrease the threshold voltage. For the chosen CHE/FN conditions, both effects almost compensate resulting in a stable V_{th}^P. Notice that $\Delta V_{th}^R(n_{cycles})$ in Figure 5.5*a* as well as $\Delta V_{th}^P(n_{cycles})$ in Figure 5.7 can be directly compared since they are given in the same arbitrary units.

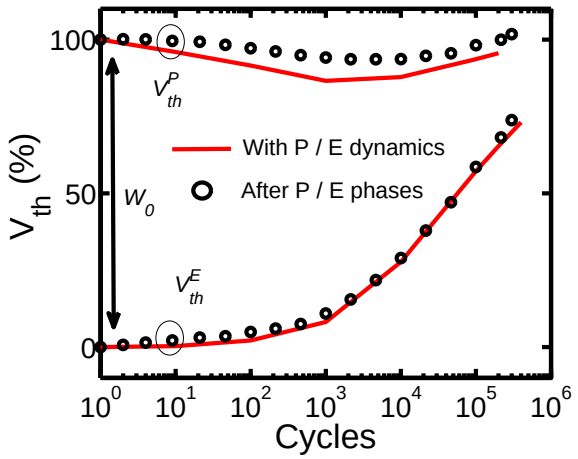


Figure 5.8: Programming window evolution for the CHE-FN operation, obtained from the measurements after program / erase operation (symbols, c.f. Figure 5.1) and the proposed extraction technique from program/erase dynamics (lines) where the threshold voltages are calculated with Equations 5.4 and 5.5.

The agreement obtained in Figure 5.8 between the two measurement approaches validates the proposed methodology to extract ΔV_{th}^E and ΔV_{th}^P. Furthermore, the use of program/erase dynamics has allowed to separate the effect of traps in terms

of threshold voltage during the three main operating conditions, i.e. *read*, *program* and *erase*. We found out that both FN and CHE are responsible, although the microscopic mechanisms involved in FN and CHE degradation are certainly very different. While an exhaustive treatment of the degradation mechanisms is beyond the scope of this work, we will focus our attention on the interface traps generation due to hot carrier injection. In the following section, a general methodology for the simulation of such traps and an application to a realistic case will be presented.

5.2 Interface state modeling

The presence of hot carriers inside the channel gives birth to phenomena described in the previous chapters, in particular to the carrier injection into the gate. As a side-effect of this phenomenon, the resulting oxide degradation should be considered carefully. In particular, the following work focuses on the interface states generation as they become a constant issue for thin oxides. Subsection 5.2.1 gives a brief historical background of such modeling built around [Hu 1985] approach, whose limitations give rise to another modeling framework [Guerin 2009] presented in subsection 5.2.2. The experience of the previous chapters on hot carrier simulations has been associated to this latter framework in subsection 5.2.3, in the purpose of simulating the degradation of a realistic device under different hot-carrier stress conditions.

5.2.1 Historical background

The hot carrier stress degradation of MOS transistor has been a constant issue in the modeling and characterization communities in the last thirty years [Takeda 1983], [Doyle 1990], [Hess 1998], [Rauch 2005]. The degradation of electrical parameters, such as the threshold voltage (V_{th}), the transconductance (g_m), the linear and saturation currents (I_{dlin} and I_{dsat}), has since accompanied each technology node and drained the efforts towards the understanding of the degradation mechanisms as well as toward accurate device modeling and structure optimization. Many studies focused on measuring the variation of the above parameters with stress time under different stress conditions as well as the correlations between these parameters. This has allowed Takeda et al. [Takeda 1983] to propose an empirical device lifetime (or time-to-failure) projection based on the bulk current. The device lifetime, corresponding to a stress time for which the nominal value of a given parameter varies by 5 or 10 %, is now considered as an important indicator to be integrated during the early design phase [Huard 2011].

Several characterization techniques have allowed to probe and point out different types of defects which are held responsible for the observed parameter variation. The defects are often classified based on their spatial location with respect to the Si/SiO_2 interface: interface charge traps (N_{it}), bulk oxide traps (N_{ot}) as well as border traps located inside the oxide but very close to the interface [Fleetwood 2008]. Both electrons and holes can create and populate such traps depending on the bias

configurations [Mistry 1993], [Doyle 1997]. However, contrary to the thick oxides where charge trapping in the oxide was frequently observed [Doyle 1990], interface states have been found to be the predominant defect for $T_{ox} \leq 4nm$ [Momose 1997], [Li 2001]. Given the rising interest in such defects, the following results and discussion are focused only on interface defects. For this reason, in the following comparisons, devices featuring thin oxides have been used in order to minimize the oxide trapping effects.

The generation of interface traps has been traditionally correlated to the bulk current or to the bulk-over-drain current ratio (I_b/I_d) [Hu 1985], [Weber 1995]. The maximum electron interface trap generation was observed at the maximum bulk current (I_b). For thick oxides, this corresponded also to the so-called *worst case* scenario, which provided the engineers the shortest device lifetime. For the technologies investigated at that time, this condition was achieved at approximately $V_g \sim V_d/2$ [Hu 1985]. Considering that hot carriers are the origin of both N_{it} and I_b, the authors modeled them using the Lucky Electron Model approach (LEM, c.f. Chapter 2) and proposed to use the I_b/I_d ratio as a direct indicator of the degradation level as it is proportional to the device lifetime (τ) in log scales (Equation 5.6). The proportionality factor is given by the ratio of the energy thresholds to create a defect ($\varepsilon_{th} \sim 3.7$ eV) and to impact ionize ($\varepsilon_{II} \sim 1.1$ eV).

$$\tau \propto \frac{I_b}{I_d}^{\varepsilon_{th}/\varepsilon_{II}} \tag{5.6}$$

This elegant approach has since been widely used as it correlates between easy-to-measure macroscopic currents (I_b, I_d) and hard-to-measure microscopic quantities (N_{it}). In this approach the carrier heating is calculated based on the maximal electric field value which has been reported to be in rather a poor agreement with rigorous Monte Carlo results (c.f. Chapter 2). The latter results have shown that the maximal energy a carrier can acquire is limited by the drain voltage (apart from electron-electron scattering, c.f. Chapter 3). Thus, according to the original theory [Hu 1985], limited degradation should have been observed for drain voltages below the generation threshold of 3.7 eV [Hu 1985]. However, continuous degradation of macroscopic parameters is still observed for drain voltages as low as 1V [Guerin 2009]. Moreover, with the scaling of the gate length and oxide thickness, the worst case reliability scenario has shifted towards higher gate voltages ($V_g \geq V_d$), requiring the investigation of more than one bias condition. In fact, a whole range of V_d/V_g biases should be investigated to establish a reliable margin. The analysis of such biases in advanced technologies [Rauch 2005], [Guerin 2009] has revealed that the variation of the device lifetime follows the trend predicted by [Hu 1985] only for a limited range of the bias conditions. As a matter of fact, faster lifetime degradation has been observed at higher gate voltages. All these elements have shown the necessity to revisit the original approach in order to propose a more accurate modeling of the interface state creation, able to predict the lifetime degradation in a full range of hot carrier bias conditions.

5.2.2 Microscopic modeling framework

Over the last decade, a new modeling framework in the hot carrier degradation modeling has emerged [Bude 1998], [Ghetti 2001], [Rauch 2005] based on an *energy-driven* approach instead of the classical *field-driven* LEM approach. This paradigm shift reflects the passage from macroscopic (electric field) to microscopic (carrier energy) modeling. It is interesting to notice that the same evolution has occurred in the gate current injection modeling context (c.f. Chapter 2). In order to predict the degradation, two ingredients are necessary. First, the knowledge of the carrier distribution as a function of length and bias conditions is needed. This can be obtained by the Monte Carlo (MC) simulations, the Spherical Harmonics Expansion (SHE) method or the 1D Semi-analytic approach presented in Chapters 2 and 3. In addition, the probability to create an interface defect (S_{it}) as a function of the carrier energy (ε) is also needed. In order to establish S_{it} a preliminary knowledge of the microscopic defect at the interface is necessary.

Many studies have confirmed that the interface traps are created after the dissociation of the $Si-H$ bond at the Si/SiO_2 interface [Hess 1998], [Chen 2000], since hydrogen is commonly introduced in CMOS technologies to passivate the dangling bonds at the interface after the gate oxide growth. Advanced simulation ([Tuttle 1999b], [Tuttle 1999a], [Kaneta 2003]) and experimental ([Avouris 1996]) studies have shown that the desorption of hydrogen from the interface can follow different paths. The bond presents *stretching* and *bending* dissociation modes and each may have different dissociation energies with the minimum situated around 2.5eV and 1.5eV, respectively. Thus, the bending mode is the most favorable dissociation path and is used in the subsequent modeling. Furthermore, it has been shown that dissociation can be achieved either by a Single Electron (SE) having a sufficiently high energy or by Multiple Electrons (ME) which collaboratively excite the bond until its dissociation [Persson 1997]. In the context of the microscopic approach, both mechanisms are characterized by different effective cross-sections, leading to S_{it-SE} and S_{it-ME} (in units of m^2) plotted in Figure 5.9, and mathematically written in the same form:

$$S_{it-m}(\varepsilon) = A_m \left(\varepsilon - \varepsilon_{th-m}\right)^{p_m} \tag{5.7}$$

m stands either for SE or ME; A_m are multiplication coefficients. These expressions have been established by analogy with carrier scattering mechanisms in the channel. Hence, the SE process is similar to impact ionization ($\varepsilon_{th-SE} = \varepsilon_{bond} = 1.5eV$ and $p_{SE} = 11$) [Rauch 2005], while the ME process similar to the emission of inelastic phonons ($\varepsilon_{th-SE} = \hbar\omega = 0.075eV$ and $p_{SE} = 0.5$) [Persson 1997]. The energy thresholds are respectively the bond dissociation and the phonon energy, while the exponents are either fitted on measurements (SE) or from analogy with phonons scattering rate expression (ME).

In the following paragraphs, two possible approaches using the above S_{it} expressions will be presented. First, a rigorous microscopic approach involving the exact knowledge of the carrier distribution function in energy is discussed in 5.2.2.1. Then,

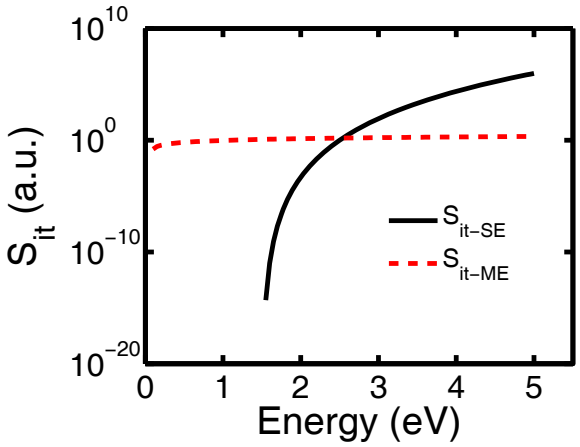


Figure 5.9: Interface state generation probability for Single Electron (SE) and Multiple Electron (ME) processes, plotted after Equation 5.7.

a possible approximation of the rigorous approach will be discussed in 5.2.2.2, and finally several comparisons between both will be reported in 5.2.2.3.

5.2.2.1 Rigorous approach

Similarly to the calculation of the hot carrier effects in the channel (Chapter 3), the bond dissociation rate (R_{it}, in units of s^{-1}) from the SE and ME mechanisms can be calculated as:

$$R_{it-m} = \int f(\varepsilon)g(\varepsilon)v(\varepsilon)S_{it-m}(\varepsilon)\mathrm{d}\varepsilon \tag{5.8}$$

f, g and v are respectively the occupation function, the density of states and the electron group velocity. m stands for SE or ME. The interface state concentration generated by each mechanism is given by [Guerin 2009]:

$$N_{it-SE} = N_0\left[1 - \exp\left[-\left(R_{it-SE}\cdot t\right)^{\alpha_{SE}}\right]\right] \tag{5.9}$$

$$N_{it-ME} = N_0\left[\left(\frac{P_{up}}{P_{down}}\right)^{N_{levels}}\cdot\left(1-\exp(\zeta_{emi}\cdot t)\right)\right]^{\alpha_{ME}} \tag{5.10}$$

The N_{it-SE} expression is obtained after solving a first-order differential equation typically employed for bond dissociation reactions. In both expressions, N_0, t and α_m are respectively the concentration of the $Si-H$ bonds at the interface which can be potentially broken, the stress time and an empirical factor reflecting the time-evolution of the interface state generation. A detailed discussion on α_m is carried out in the next subsection. Instead, the N_{it-ME} expression has been obtained from different considerations. In this case, the Si-H bond is modeled as an harmonic oscillator (Figure 5.10) having N_{levels} (Equation 5.11) equi distant vibrational states

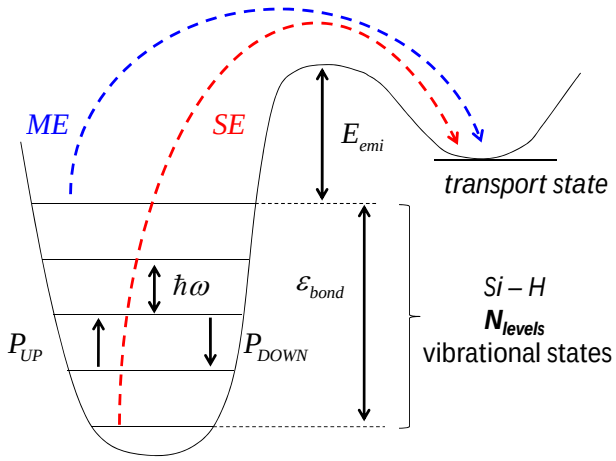


Figure 5.10: Schematic view of Si-H bond dissociation processes and the related quantities: the Si-H bond energy ε_{bond}, the energy between vibrational states $\hbar\omega$ (assumed to be constant), the total number of vibrational levels N_{levels}, the probability for the bond to be excited or to decay to the upper (P_{UP}) or lower level (P_{DOWN}), respectively. Single Electron (SE) and Multiple Electron (ME) processes dissociate the bond from the lowest and the highest vibrational state, respectively. The figure is inspired from [Persson 1997], [Guerin 2009] and [Starkov 2011].

[Persson 1997]. The bond is excited to a higher or lower vibrational state with probability P_{up} and P_{down}, respectively (Equations 5.12 and 5.13). However, an excited vibrational state should have a lifetime, defined as $1/\omega_E$ and taken equal to 0.1 ps and constant for each of the equidistant vibrational levels, comparable or longer than the average time between subsequent carrier scattering events. Finally, an energy barrier E_{emi} separates the highest vibrational state from the transport state which eventually marks the bond dissociation. The emission rate over this barrier is given by Equation 5.14.

$$N_{levels} = \frac{\varepsilon_{th-SE}}{\varepsilon_{th-ME}} = \frac{\varepsilon_{bond}}{\hbar\omega} \tag{5.11}$$

$$P_{up} = R_{it-ME} + \omega_E \exp\left(-\frac{\hbar\omega}{k_b T}\right) \tag{5.12}$$

$$P_{down} = R_{it-ME} + \omega_E \tag{5.13}$$

$$\zeta_{emi} = \nu \exp\left(-\frac{E_{emi}}{k_b T}\right) \tag{5.14}$$

5.2.2.2 Possible approximation

The application of the above model requires the knowledge of the whole distribution function which may sometime constitute a blocking point for the N_{it} evaluation. For this reason, another proposal was made in [Rauch 2005], [La Rosa 2007],

where macroscopic quantities, such as the drain current, are used instead to describe the main features of the distribution function. The degradation is theneval-uated based on the S_{it} values calculated on carefully chosen *dominant* energies, instead of the whole energy range. These energies, are defined as the energies for which the integrand of Equation 5.8 shows a local or global maximum [Rauch 2005], [La Rosa 2007] (c.f. Figure 5.11).

The interface state generation rate may then be calculated as:

$$R_{it-SE} = \int f(\varepsilon)g(\varepsilon)v(\varepsilon)S_{it-SE}(\varepsilon)\mathrm{d}\varepsilon = \sum_i F(\varepsilon_{dom-i}) \cdot S_{it-SE}(\varepsilon_{dom-i}) \qquad (5.15)$$

F stands for the electron flux equivalent to the product $f \cdot g \cdot v$. When no EES is considered, there is only one dominant energy (ε_{dom-1}). If EES is included, one additional dominant energy is present (ε_{dom-2}), hence the sum over i ($i \in \{1, 2\}$) in Equation 5.15. The second dominant energy has been found to be situated at around $\varepsilon_{dom-2} = 1.8\varepsilon_{dom-1}$ [Rauch 2005]. Such approach has been only applied to SE-processes, and the interface state density is calculated as:

$$N_{it-SE} \propto \sum_i [G(I_d)_i S_{it-SE}(\varepsilon_{dom-i})t]^{\alpha_{SE}} \qquad (5.16)$$

In this expression $G(I_d)_1 = I_d$ and $G(I_d)_2 = I_d^2$ [La Rosa 2007]. Equation 5.16 can be deduced from Equation 5.9 following several steps. First, if we replace Equation 5.15 into Equation 5.9, we find:

$$N_{it-SE} = N_0 \left[1 - \exp\left[-\left(F(\varepsilon_{dom-i}) \cdot S_{it-SE}(\varepsilon_{dom-i}) \cdot t\right)^{\alpha_{SE}}\right]\right] \qquad (5.17)$$

Then, a 1^{st} order Taylor expansion of Equation 5.17 leads to the following equation:

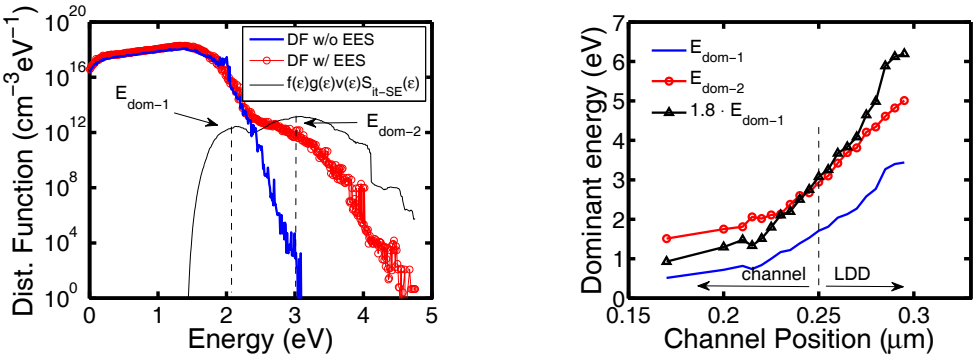
$$N_{it-SE} = N_0 \left[\left(F(\varepsilon_{dom-i}) \cdot S_{it-SE}(\varepsilon_{dom-i}) \cdot t\right)^{\alpha_{SE}}\right] \qquad (5.18)$$

The electron flux $F(\varepsilon_{dom-i})$ constitutes the drain current coming from the ε_{dom-i} energy. As the distribution function is not known, the above information is not available. Thus, it has been assumed that $F(\varepsilon_{dom-i})$ is proportional to either I_d ($i = 1$) or I_d^2 ($i = 2$) [Rauch 2005], respectively, thus leading to Equation 5.16. Unfortunately, the information on the position-dependent interface states generation rate is lost. Hence, only a global value of N_{it} can be extracted, which has been proven to reproduce device lifetime in a satisfactory way [Rauch 2005], [La Rosa 2007].

However, this methodology has not yet been applied to the ME-processes which have been reported to play an important role especially in the channel [Starkov 2011]. In the following paragraph, the results of the dominant-energy approach will be compared against the full integration method previously discussed.

5.2.2.3 Discussion

To investigate the above assumptions, full band Monte Carlo (MC) simulations have been performed with and without EES. Figure 5.11*a* reports the distribution functions obtained for a 0.28 μm device featuring a gate oxide thickness of 5 nm. Notice the enhanced hot carrier tail due to the EES, as already discussed in Chapter 3. Figure 5.11*a* also reports the product $f \cdot g \cdot v \cdot S_{it-SE}$ constituting the integrand of Equation 5.8 which shows two peaks previously defined as the dominant energies E_{dom-1} and E_{dom-2}. These energies can be extracted as a function of the channel position using the distribution functions. However, it should be mentioned that the authors of [Rauch 2005] and [La Rosa 2007] defined only global dominant energies which were functions of the drain voltage and not of the channel position, due to the lack of the distribution functions. Figure 5.11*b* reports the dominant energies extracted along the channel using the above method, thus extending the original concept of the dominant energy. Distribution functions obtained with and without EES have been used for this purpose. In addition, this figure reports the second dominant energy calculated after a multiplication of the first one with a factor of 1.8, as suggested by [Rauch 2005]. This compares rather well with the MC results, in particular in the region close to the channel/LDD junction

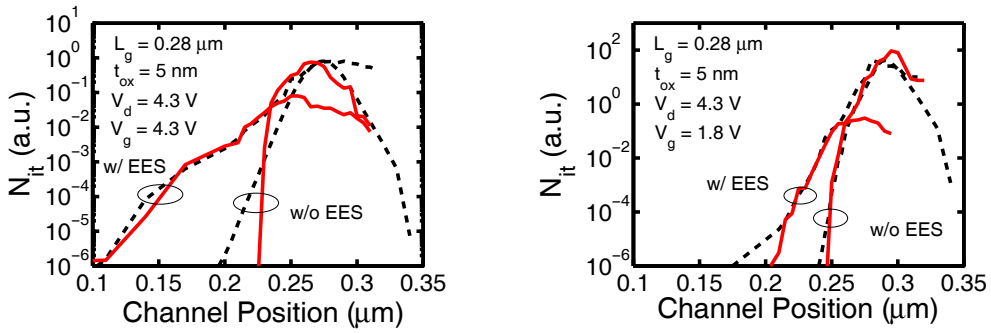


(a) (b)

Figure 5.11: *(a)* Example of electron distribution function (DF) simulated by the Monte Carlo with and without Electron-Electron Scattering (EES) at V_g/V_d = 1.8/4.3V on a device featuring a gate length of 0.28 μm and tunnel oxide of 5 nm. The extraction of the dominant energies (E_{dom-1} and E_{dom-2}) based on the maximum position of the integrand of Equation 5.8 is also illustrated. The y-axis of the integrand has been scaled for the sake of clarity. *(b)* First and second dominant energies along the channel extracted from MC distributions and the second dominant energy extracted using the approach in [Rauch 2005].

Then, the interface states calculated after the dominant energy picture approach (Equation 5.16) and the full integral solution (Equation 5.9), are compared in Figure 5.12. When EES is included in the model, a penetration of the interface traps

towards the channel is obtained, due to the hot carrier tail enhancement, for both the investigated biases. The reasonable qualitative matching obtained in Figure 5.12 allows to use the dominant energy picture approach for the modeling of SE-processes. In particular, in the following section, the second dominant energy will be used to model the EES effect starting from a distribution function calculated without including such a scattering mechanism.



(a) (b)

Figure 5.12: Single Electron-induced interface state density along the channel calculated after dominant energy picture (right side of Equation 5.15, dashed lines) and after the full integration of f(E) (left side of Equation 5.15, solid lines), on a device featuring a T_{ox}=5nm and L=0.28 μm, under V_g/V_d = 4.3/4.3V hot carrier stress condition.

5.2.3 Application to 65nm technology

In this subsection, the microscopic modeling framework of section 5.2.2 is applied to investigate the generation of interface traps along the channel for various bias conditions on a transistor featuring a 0.28 μm gate length and a 2.8 nm oxide thickness, fabricated and simulated after a 65nm process. For integration purposes in the TCAD environment, the SHE method has been used. Since SHE does not account for EES (c.f. Chapter 2), the dominant energy picture approach previously described has been used to model the effect of such mechanism. Figure 5.13 shows the applied methodology which is split up into three steps. A 2D TCAD structure obtained with realistic process simulation has been used [Synopsys 2010b]. First, the distribution function is calculated with the SHE [Jin 2009] on the device featuring no interface traps (unstressed). The terminals are ramped up to the considered stress configuration and the distributions along the channel at the interface ordinate (y_{int}) are collected. Note that due to the source/drain reoxidation process, the interface ordinate is not constant along the channel and has to be taken into consideration in the extraction procedure. The calculation of the distribution function in such non-planar channels is particularly facilitated by using the SHE method. A self-consistent approach would require the evaluation of the distribution function during the stress time at periodic time intervals. However, in the proposed approach, a non self consistent approach has been adopted, similarly to [Tyaginov 2010] and [Starkov 2011]. It has thus been assumed that the carrier distribution is not significantly affected by the traps created during the stress.

Next, the interface trap concentration is calculated at each point of the channel using the Equation 5.9, for the SE-processes, and Equation 5.10 for the ME-processes. SE-processes are slightly modified by the inclusion of the EES using the Equation 5.16 by taking into account only the $i = 2$ term since $i = 1$ is naturally included in Equation 5.9 with the distribution functions calculated with SHE. The EES contribution is thus proportional to the S_{it-SE} evaluated at $\varepsilon_{dom-2} = 1.8\varepsilon_{dom-1}$ which depends on the channel position (c.f. Figure 5.11*b*) and is given by:

$$N_{it-SE_{EES}}(x,t) = \frac{Max\left(N_{it-SE_{SHE}}\right) \cdot \left(S_{it-SE}(1.8\varepsilon_{dom-1}(x)) \cdot t\right)^{\alpha_{SE}}}{Max\left(\left[S_{it-SE}(1.8\varepsilon_{dom-1}(x)) \cdot t\right]^{\alpha_{SE}}\right)} \tag{5.19}$$

Here, $S_{it-SE}[\varepsilon_{dom-2}(x)]$ is multiplied and divided by constant factors ($Max(...)/Max(...)$ ratio) which indirectly contain the I_d^2 term appearing in Equation 5.16. This ratio is bias (but not energy) dependent (since the current itself is bias dependent and not space-dependent) and it has been chosen in a way such that the peak value of N_{it-SE} inside the drain calculated with or without EES is the same, i.e. small contribution of EES-induced degradation inside the drain. This contribution depends only on the adjustment parameters of $N_{it-SE_{SHE}}$. This scenario corresponds to the one presented in Figure 5.12 (dashed lines). The total trap concentration is given by the sum of SE and ME processes. Note that a $N_{it}(x)$ curve is obtained for each stress time t used as a parameter in the formulae. Finally, for each stress time the

obtained interface density is integrated at the Si/SiO_2 interface of the 2D structure and electrically simulated [Synopsys 2010a]. The degradation of macroscopic quantities (V_{th}, G_m, S) can thus be studied.

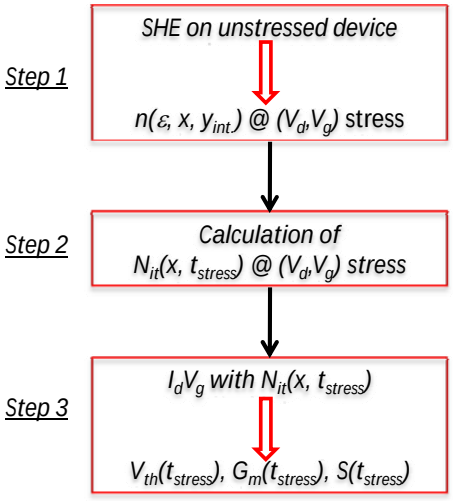


Figure 5.13: The three-step procedure proposed in this modeling approach. The distribution function $n(\varepsilon,\ x,\ y_{int.})$ along the channel is calculated for each (V_d, V_g) stress condition (step 1) which is then used to calculate the interface state density N_{it} along the channel (step 2). Electrical macroscopic parameters of interest are then extracted from the I_dV_g characteristics simulated including the defects profile (step 3).

Each of the steps is systematically evaluated as it bears uncertainties and requires adjustment parameters. The distribution functions obtained with the SHE method have already been successfully compared to reference MC results at different channel positions, channel lengths and bias configurations (c.f. Chapter 2). This validates the accuracy of the first step in the absence of EES. Whenever EES should be considered, the dominant energy picture can be reasonably applied.

The results of the second step have been compared to Lateral interface trap Profile (LP) measurements yielding interface state densities along the channel. Charge pumping measurements have been performed for this purpose [Randriamihaja 2012]. Figure 5.14 reports the simulated and measured interface state densities at three different biases. Two different regions can be observed for each of the curves: a peak at the junction or inside the LDD and a rather flat distribution in the channel. As already suggested in [Tyaginov 2010] and [Starkov 2011], the peak at the drain side is due to SE-processes as this mechanism favors the high energy electrons (due to a high exponent: $p_{it-SE} = 11$) which are situated exactly in this part of the device (c.f. Chapter 3). The interface states in the channel are instead due to ME-processes for which the most important parameter is the carrier density instead of the carrier energy (due to a small exponent: $p_{it-ME} = 0.5$). The energy thresholds are taken equal to 1.5 and 0.075 eV, respectively for SE- and ME-processes, in

agreement with [Guerin 2009]. One last important point in the model setup, concerns the time-dependent interface state dynamics α_m, with $m \in \{SE, ME\}$. The experimental data of Figure 5.14 suggest an α_{SE} in the range of 0.35 - 0.54 and an α_{ME} in the range of 0.45 - 0.60 for the considered bias conditions. The same values have been used in the corresponding modeling cases throughout the time stress duration. As a matter of fact, such dispersion of α_m has long been observed and reported (e.g [Doyle 1990], [Mahapatra 2006]). This effect may presumably be linked to the dispersion of the dissociation bond energy at the surface. In fact, the latter can be caused by an interface disorder at a microscopic level [Hess 1999] or to the effect of an external electric field [Guerin 2009]. It ought to be mentioned that whenever α_m is changed, a readjustment of A_{SE} and A_{ME}, which are devoid of any physical meaning, is necessary. Notice that for all the considered conditions, the ME process is rather significant and thus the effect of EES is not much visible as it only contributes to smooth the transition between SE and ME and regions.

Finally, Figure 5.15 reports the measured and simulated macroscopic parameters' degradation due to the interface states. All the parameters (V_{th}, G_m, S) have been extracted after measured and simulated I_dV_g characteristics at a drain voltage of 0.1 V. Mobility degradation due to remote coulomb scattering has been included in the simulations to account for the mobility reduction due to the presence of charged traps. Acceptor traps with an uniformly distributed energy in the silicon bandgap [Lenahan 1984] and a cross-section of $2 \cdot 10^{-16} cm^{-2}$ have been included and simulated within the framework of the classical SRH approach. A good agreement between simulations and measurements has been obtained for all the investigated biases.

These comparisons confirm that the developed modeling framework is likely to capture most of the observed hot electron induced degradation. Although the global trends are well reproduced, a continuous investigation is still required. In particular, an in-depth understanding of the evolution of the trap creation dynamics (α) with applied bias (V_d, V_g) would be helpful to better grasp the physics of these processes. Furthermore, a theoretical evaluation of the prefactors used in this study and in the above references would certainly shed more light into the observed variations. Finally, this framework can also include a hot hole degradation component, as already suggested in [Tyaginov 2011].

Overall, in order to further establish the approach, additional stress and electrical characterization should be performed, such as: 1) reverse I_dV_g measurements (source and drain inverted during characteristics extraction) to enhance the non-uniform distribution of the traps, 2) HCI stress including bulk voltage in order to test the effect of holes (CHISEL conditions), 3) characterization of devices with even shorter gate length.

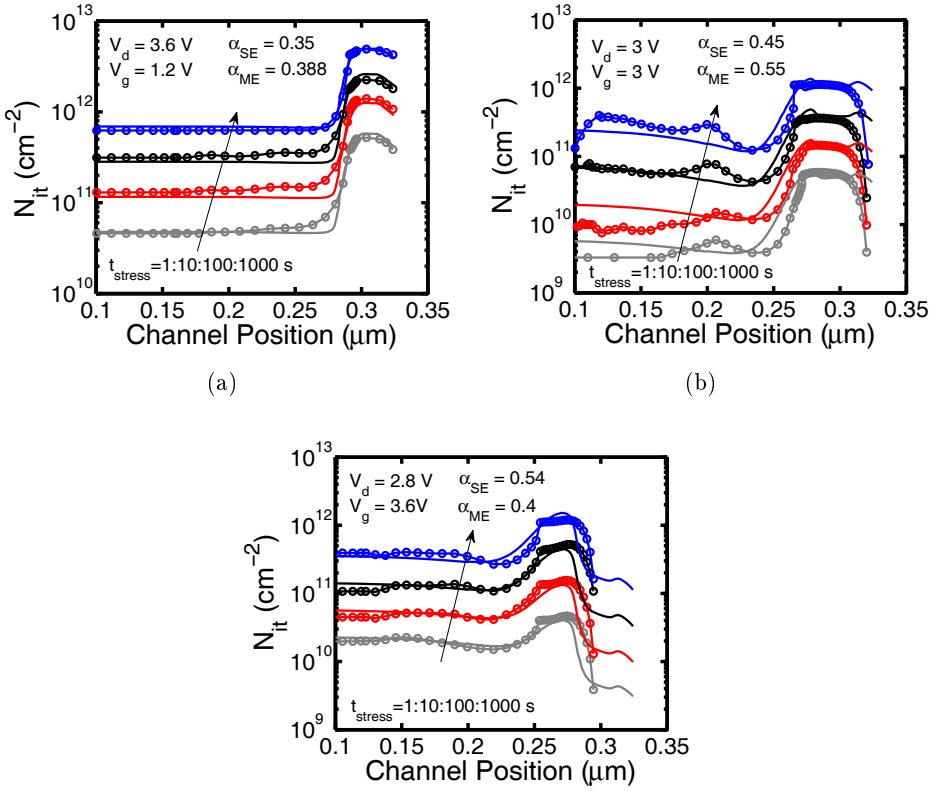


Figure 5.14: Interface state density along the channel obtained after lateral profile measurements and with the proposed modeling framework for a device featuring L=0.28 μm and T_{ox}=2.8nm subject to V_g/V_d = 1.2/3.6V *(a)*, 3/3V *(b)* and 3.6/2.8V *(c)* hot carrier stress conditions for t_{stress} = 1/10/100/1000 s. The following parameters are used in the simulations: ε_{th-SE} = 1.5eV, ε_{th-ME} = 0.075eV, p_{it-SE} = 11, p_{it-ME} = 0.5, E_{emi} = 0.26eV.

5.3 Conclusions

In this chapter we dealt with several specific aspects of oxide degradation due to carrier injection which has been historically considered as a harmful process for the gate oxide. The first part of the study discussed the programming window closure with increasing number of cycles as one of the most critical manifestations of oxide degradation in Flash memories. An experimental methodology was set up to separate the impact of traps created during cycling on each of the program, erase and read phases. It was found that traps highly affect the reading of the threshold voltage and the intrinsic program efficiency, while negligibly affect the erase efficiency. The combination of the three explains the observed endurance results. Furthermore, the experimental study on equivalent transistors has shown that both programming and erase operations analyzed separately significantly degrade the device.

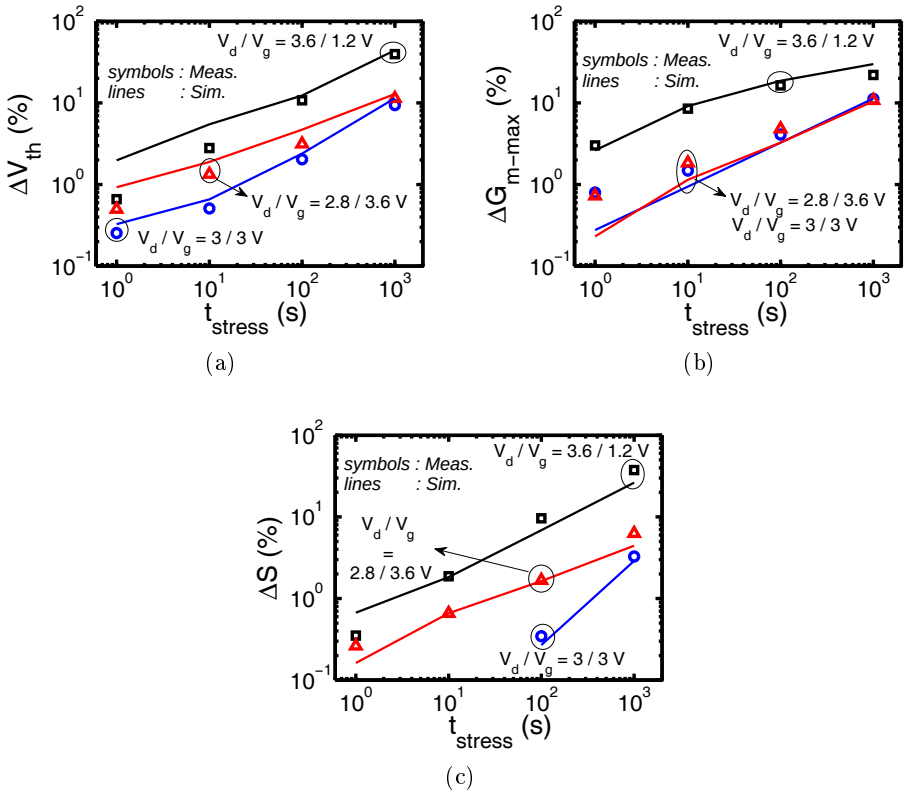


Figure 5.15: Relative electrical parameters variation vs. stress time obtained after measurements and TCAD device simulations integrating the defect profiles of Figure 5.14. Two electrostatic-related parameters, the threshold voltage V_{th} *(a)* and subthreshold slope S *(b)*, as well as the maximum of the transconductance G_{m-max} *(c)*, have been reported.

The second part of the study dealt with the modeling of the hot carrier injection induced degradation. A microscopic perspective, involving the calculation of the carrier distribution function in the channel and including the electron-electron scattering effect, was adopted in conjunction with an energy-driven model for the generation of interface states, the latter being the only defects treated in this investigation. The results of this study showed that the generation of interface states at the drain side and in the channel (due to hot and cold carriers, respectively) is consistent with lateral interface trap profiling measurements which allows to probe the traps concentration along the channel. Furthermore, a good correlation was found between simulated and measured macroscopic electrical parameters, such as threshold voltage, sub-threshold slope and transconductance, after different stress durations and hot carrier injection bias configurations. Thus, the energy-driven approach used in the framework of a TCAD environment is a promising approach to gain more insight on the complex degradation phenomena. Its application to different gate lengths and stress regimes would provide more details on the involved

physical phenomena and the required model ingredients.

Bibliography

[Aritome 1993] S. Aritome, R. Shirota, G. Hemink, T. Endoh and F. Masuoka. *Reliability issues of flash memory cells.* Proceedings of the IEEE, vol. 81, no. 5, pages 776–788, 1993. (Cited on page 138.)

[Avouris 1996] P. Avouris, R.E. Walkup, A.R. Rossi, T.-C. Shen, G.C. Abeln, J.R. Tucker and J.W. Lyding. *STM-induced H atom desorption from Si(100): isotope effects and site selectivity.* Chemical Physics Letters, vol. 257, no. 1-2, pages 148 – 154, 1996. (Cited on page 149.)

[Bude 1998] D.J. Bude, B.E. Weir and P.J. Silverman. *Explanation of stress-induced damage in thin oxides.* In International Electron Devices Meeting (IEDM) 1998, pages 179 –182, dec 1998. (Cited on page 149.)

[Cappelletti 1999] P. Cappelletti. Flash memories. Springer Netherlands, 1999. (Cited on pages 2, 73 and 140.)

[Chen 2000] Z. Chen, K. Hess, J. Lee, J.W. Lyding, E. Rosenbaum, I. Kizilyalli, S. Chetlur and R. Huang. *On the mechanism for interface trap generation in MOS transistors due to channel hot carrier stressing.* IEEE Electron Device Letters, vol. 21, no. 1, pages 24 –26, jan 2000. (Cited on page 149.)

[Doyle 1990] B. Doyle, M. Bourcerie, J.C. Marchetaux and A. Boudou. *Interface state creation and charge trapping in the medium-to-high gate voltage range during hot-carrier stressing of n-MOS transistors.* IEEE Transactions on Electron Devices, vol. 37, no. 3, pages 744–754, 1990. (Cited on pages 147, 148 and 157.)

[Doyle 1997] B.S. Doyle, K.R. Mistry and J. Faricelli. *Examination of the time power law dependencies in hot carrier stressing of n-MOS transistors.* IEEE Electron Device Letters, vol. 18, no. 2, pages 51–53, 1997. (Cited on pages 36 and 148.)

[Fayrushin 2009] A. Fayrushin, K.S. Seol, J.H. Na, S.H. Hur, J.D. Choi and K. Kim. *The new program/erase cycling degradation mechanism of NAND Flash memory devices.* In International Electron Devices Meeting (IEDM) 2009, pages 1–4. IEEE, 2009. (Cited on page 138.)

[Fleetwood 2008] D.M. Fleetwood, S.T. Pantelides and R.D. Schrimpf. Defects in microelectronic materials and devices. CRC, 2008. (Cited on page 147.)

[Garetto 2011] 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

International Memory Workshop (IMW) 2011, pages 1 –4, may 2011. (Cited on page 141.)

[Ghetti 2001] A. Ghetti. *Characterization and modeling of the tunneling current in Si-SiO2-Si structures with ultra-thin oxide layer.* Microelectronic Engineering, vol. 59, no. 1-4, pages 127–136, 2001. (Cited on page 149.)

[Guerin 2009] C. Guerin, V. Huard and A. Bravaix. *General framework about defect creation at the Si/SiO2 interface.* Journal of Applied Physics, vol. 105, no. 11, pages 114513–114513, 2009. (Cited on pages 147, 148, 150, 151 and 157.)

[Hess 1998] K. Hess, I.C. Kizilyalli and J.W. Lyding. *Giant isotope effect in hot electron degradation of metal oxide silicon devices.* IEEE Transactions on Electron Devices, vol. 45, no. 2, pages 406 –416, feb 1998. (Cited on pages 147 and 149.)

[Hess 1999] K. Hess, L.F. Register, W. McMahon, B. Tuttle, O. Aktas, U. Ravaioli, J.W. Lyding and I.C. Kizilyalli. *Theory of channel hot-carrier degradation in MOSFETs.* Physica B: Condensed Matter, vol. 272, no. 1-4, pages 527 – 531, 1999. (Cited on page 157.)

[Hu 1985] C. Hu, S.C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan and K.W. Terrill. *Hot-Electron-Induced MOSFET Degradation – Model, Monitor, and Improvement.* IEEE Journal of Solid-State Circuits, vol. 20, no. 1, pages 295 – 305, February 1985. (Cited on pages 20, 147 and 148.)

[Huard 2011] V. Huard, F. Cacho, Y.M. Randriamihaja and A. Bravaix. *From defects creation to circuit reliability: A bottom-up approach (invited).* Microelectronic Engineering, vol. 88, no. 7, pages 1396 – 1407, 2011. (Cited on page 147.)

[Itsumi 1981] M. Itsumi. *Positive and negative charging of thermally grown SiO2 induced by Fowler-Nordheim emission.* Journal of Applied Physics, vol. 52, no. 5, pages 3491 –3497, may 1981. (Cited on page 140.)

[Jin 2009] S. Jin, A. Wettstein, W. Choi, F.M. Bufler and E. Lyumkis. *Gate Current Calculations Using Spherical Harmonic Expansion of Boltzmann Equation.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2009, pages 1 –4, 2009. (Cited on pages 20, 24, 26, 60 and 155.)

[Kaneta 2003] C. Kaneta, T. Yamasaki and Y. Kosaka. *Nano-scale simulation for advanced gate dielectrics.* Fujitsu Scientific and Technical Journal, vol. 39, no. 1, pages 106–118, 2003. (Cited on page 149.)

[Kolodny 1986] A. Kolodny, S.T.K. Nieh, B. Eitan and J. Shappir. *Analysis and modeling of floating-gate EEPROM cells.* IEEE Transactions on Electron Devices, vol. 33, no. 6, page 835, 1986. (Cited on pages 106 and 141.)

[La Rosa 2007] G. La Rosa and S.E. Rauch. *Channel hot carrier effects in n-MOSFET devices of advanced submicron CMOS technologies.* Microelectronics Reliability, vol. 47, no. 4-5, pages 552 – 558, 2007. 14th Workshop on Dielectrics in Microelectronics (WoDiM 2006). (Cited on pages 18, 36, 151, 152 and 153.)

[Lee 2006] W.H. Lee, C.K. Park and K. Kim. *Temperature dependence of endurance characteristics in NOR flash memory cells.* In International Reliability Physics Symposium (IRPS) 2006, pages 701–702. IEEE, 2006. (Cited on page 138.)

[Lenahan 1984] 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, pages 3495–3499, 1984. (Cited on pages 142, 143 and 157.)

[Lenzlinger 1969] M. Lenzlinger and E.H. Snow. *Fowler-Nordheim Tunneling into Thermally Grown SiO2.* Journal of Applied Physics, vol. 40, no. 1, pages 278–283, 1969. (Cited on page 141.)

[Li 2001] E. Li, E. Rosenbaum, J. Tao and P. Fang. *Projecting lifetime of deep submicron MOSFETs.* IEEE Transactions on Electron Devices, vol. 48, no. 4, pages 671 –678, apr 2001. (Cited on page 148.)

[Mahapatra 2006] S. Mahapatra, D. Saha, D. Varghese and P.B. Kumar. *On the generation and recovery of interface traps in MOSFETs subjected to NBTI, FN, and HCI stress.* IEEE Transactions on Electron Devices, vol. 53, no. 7, pages 1583–1592, 2006. (Cited on page 157.)

[Mistry 1993] K.R. Mistry and B. Doyle. *AC versus DC hot-carrier degradation in n-channel MOSFETs.* IEEE Transactions on Electron Devices, vol. 40, no. 1, pages 96 –104, jan 1993. (Cited on page 148.)

[Modelli 2004] A. Modelli, A. Visconti and R. Bez. *Advanced flash memory reliability.* In nternational Conference on Integrated Circuit Design and Technology (ICICDT) 2004, pages 211 – 218, 2004. (Cited on page 138.)

[Momose 1997] H.S. Momose, S.-I. Nakamura, T. Ohguro, T. Yoshitomi, E. Morifuji, T. Morimoto, Y. Katsumata and H. Iwai. *A study of hot-carrier degradation in n- and p-MOSFETs with ultra-thin gate oxides in the direct-tunneling regime.* In Electron Devices Meeting, 1997. IEDM '97. Technical Digest., International, pages 453 –456, dec 1997. (Cited on page 148.)

[Persson 1997] B.N.J. Persson and P. Avouris. *Local bond breaking via STM-induced excitations: the role of temperature.* Surface Science, vol. 390, no. 1-3, pages 45 – 54, 1997. <ce:title>Desorption Induced by Electronic Transitions</ce:title>. (Cited on pages 149 and 151.)

[Randriamihaja 2012] Y.M. Randriamihaja, A. Zaka, V. Huard, P. Palestri, D. Rideau, D. Roy, A. Bravaix and M. Rafik. *Hot carrier degradatino: from defect creation modeling to their impact on MOS parameters (submitted).* In Reliability Physics Symposium (IRPS), 2012 IEEE International. IEEE, 2012. (Cited on page 156.)

[Rauch 2005] S.E. Rauch and G. La Rosa. *The energy-driven paradigm of NMOSFET hot-carrier effects.* IEEE Transactions on Device and Materials Reliability, vol. 5, no. 4, pages 701 – 705, dec. 2005. (Cited on pages 147, 148, 149, 151, 152 and 153.)

[Starkov 2011] I. Starkov, S. Tyaginov, H. Enichlmair, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Ceric and T. Grasser. *Hot-carrier degradation caused interface state profile - Simulation versus experiment.* vol. 29, no. 1, page 01AB09, 2011. (Cited on pages 151, 152, 155 and 156.)

[Synopsys 2010a] Synopsys. *Synopsys Sentaurus, release D-2010.12, SDevice simulators*, 2010. (Cited on pages 21, 23, 24, 30, 125 and 156.)

[Synopsys 2010b] Synopsys. *Synopsys Sentaurus, release D-2010.12, SProcess simulators*, 2010. (Cited on pages 111 and 155.)

[Takeda 1983] E. Takeda and N. Suzuki. *An empirical model for device degradation due to hot-carrier injection.* Electron Device Letters, IEEE, vol. 4, no. 4, pages 111 – 113, apr 1983. (Cited on page 147.)

[Tao 2007a] G. Tao, H. Chauveau, D. Boter, D. Dormans and R. Verhaar. *A simple and accurate method to extract neutral threshold voltage of floating gate flash devices and its application to flash reliability characterization.* In International Integrated Reliability Workshop Final Report (IRW) 2007, pages 52 –56, oct. 2007. (Cited on page 141.)

[Tao 2007b] G. Tao, H. Chauveau and R. Verhaar. *A Quantitative Study of Endurance Characteristics and Its Temperature Dependance of Embedded Flash Memories With 2T-FNFN NOR Device Architecture.* IEEE Transactions on Device and Materials Reliability, vol. 7, no. 2, pages 304–309, 2007. (Cited on page 138.)

[Tseng 2001] J.M.Z. Tseng, B.J. Larsen, Y. Xiao and D.A. Erickson. *A novel method to separately investigate program and erase degradation mechanisms in flash memory cells.* IEEE Transactions on Electron Devices, vol. 48, no. 12, pages 2947–2951, 2001. (Cited on page 141.)

[Tuttle 1999a] B. Tuttle. *Hydrogen and P_b defects at the* $(111)\mathrm{Si} - \mathrm{SiO_2}$ *interface: An ab initio cluster study.* Physical Review B, vol. 60, pages 2631–2637, Jul 1999. (Cited on page 149.)

[Tuttle 1999b] B. Tuttle and C.G. Van de Walle. *Structure, energetics, and vibrational properties of Si-H bond dissociation in silicon.* Physical Review B, vol. 59, pages 12884–12889, May 1999. (Cited on page 149.)

[Tyaginov 2010] S.E. Tyaginov, I.A. Starkov, O. Triebl, J. Cervenka, C. Jungemann, S. Carniello, J.M. Park, H. Enichlmair, M. Karner, C. Kernstock, E. Seebacher, R. Minixhofer, H. Ceric and T. Grasser. *Interface traps density-of-states as a vital component for hot-carrier degradation modeling.* Microelectronics Reliability, vol. 50, no. 9-11, pages 1267 – 1272, 2010. <ce:title>21st European Symposium on the Reliability of Electron Devices, Failure Physics and Analysis</ce:title>. (Cited on pages 155 and 156.)

[Tyaginov 2011] S. Tyaginov, I. Starkov, O. Triebl, H. Ceric, T. Grasser, H. Enichlmair, J.M. Park and C. Jungemann. *Secondary generated holes as a crucial component for modeling of HC degradation in high-voltage n-MOSFET.* In International Conference on Simulation of Semiconductor Processes and Devices (SISPAD) 2011, pages 123–126. IEEE, 2011. (Cited on page 157.)

[Weber 1995] W. Weber, M. Brox, R. Thewes and N.S. Saks. *Hot-hole-induced negative oxide charges in n-MOSFETs.* Electron Devices, IEEE Transactions on, vol. 42, no. 8, pages 1473 – 1480, aug 1995. (Cited on page 148.)

[Zaka 2011] A. Zaka, J. Singer, E. Dornel, D. Garetto, D. Rideau, Q. Rafhay, R. Clerc, J.-P. Manceau, N. Degors, C. Boccaccio, C. Tavernier and H. Jaouen. *Characterization and 3D TCAD simulation of NOR-type flash non-volatile memories with emphasis on corner effects.* Solid-State Electronics, vol. 63, no. 1, pages 158 – 162, 2011. (Cited on pages 22 and 143.)

[Zous 2004] N.K. Zous, Y.J. Chen, C.Y. Chin, W.J. Tsai, T.C. Lu, M.S. Chen, W.P. Lu, T. Wang, S.C. Pan and C.Y. Lu. *An endurance evaluation method for flash EEPROM.* IEEE Transactions on Electron Devices, vol. 51, no. 5, pages 720–725, 2004. (Cited on page 143.)

General conclusions

Hot carrier injection on standard floating gate NOR memory cells and some of the associated degradation mechanisms have been investigated in this thesis through simulation and experimental analysis.

Chapter 1 introduced the general context of NOR memories which face a rising demand for code storage and embedded applications. Indeed, new industry-oriented and consumer products have driven the development of such memories which which suffer from many scaling issues. The optimization of the cell electrostatics, of the carrier injection during program or drain disturb as well as the resulting degradation constitute major elements which are jointly considered with the help of suitable advanced simulators, TCAD tools and alternative approaches.

In this context, Chapter 2 investigated the hot electron injection models which are commonly used to model gate current in such memories: the Fiegna Model (FM), the Lucky Electron Model (LEM), the Spherical Harmonics Expansion (SHE) method and the Monte Carlo (MC). On one hand, when calibrated, this study confirmed that FM and LEM can reproduce the injection efficiency for a particular technology. However, the assumptions in these models make them inadequate for reliable predictive simulations. Indeed, a closer look at microscopic quantities reveal their weaknesses (local models, over-simplified scattering, non-parabolic bands), confirming the earlier findings on this subject. On the other hand, comparisons with reference MC simulations showed that SHE captures relatively well electron injection at high drain and gate voltages, despite its assumptions (isotropic scattering and band structure). Furthermore, this accuracy comes along only with a small increase of the computational burden, contrarily to the MC approach. However, SHE does not include at present electron-electron scatterings (useful in case of low-voltage operation) which is however implemented in some MC tools..

Taking advantage of the previous chapter's results, Chapter 3 presented a new 1D semi-analytic model for hot carrier transport. The most relevant scattering mechanisms (inelastic phonon scattering, impact ionization and electron-electron scattering) have been implemented. Furthermore, full-band aspects of silicon have been used in agreement with previous studies. Accounting for the strongest phonon interaction with full-band scattering rates, a simple equipartition scheme for impact ionization and a simplification of electron-electron energy exchanges make this model very computationally efficient. Moreover, the inclusion of the carrier's history throughout the channel confers the non-local nature to this model. Hence, comparisons with MC simulations showed that the model well captures the distribution functions, the impact ionization generation rates and the gate current density along the channel as well as the bulk-to-drain and gate-to-drain current ratios over a wide range of gate lengths (0.14 to 1 μm) and biases ($V_d \in [2.5, 5]$ and $V_g \in [2.5, 6]$), representative of state of the art NOR memories. This approach constitutes an original alternative to SHE as it additionally integrates electron-electron scattering in the

framework of device simulation.

In Chapter 4, the hot electron injection regime was analyzed from an experimental perspective. The characterization methodology combines transient measurements on flash cells ($V_{th}(t)$) and static measurements on equivalent transistors ($I_d(V_g)$) needed for a precise device electrostatics calibration. The combination of these measurements allows to extract the intrinsic cell injection characteristics were extracted. First, these measurements were compared to 2D simulations where it was shown that MC correctly reproduces the injection for most of the investigates cases, extending previous studies by considering broader ranges of gate lengths and biases. In particular, results from drain voltages as low as 1.5 V were correctly reproduced by MC, assessing the importance of electron-electron scattering for the low-voltage regimes. Confirming the results of chapter 2, SHE correctly captures injection at high voltages but fails at low drain voltages as it does not include electron-electron scattering. On the other hand, comparisons with 1D simulations combining the developed 1D non-local injection model of chapter 3 and a Charge Sheet Model, were performed. In the high voltage injection regime ($V_g > V_d > 3$) where most of the injection occurs, the developed methodology compares well with the experiments as a function of the gate length. The last part of this chapter was devoted to the study of the unwanted disturb operation. In the case of the investigated cells, experiments showed that charge loss (V_{th} reduction) due to hot hole injection into the floating gate during drain disturb is the main issue. A combined TCAD-MC simulation methodology was developed to investigate the intrinsic properties of this regime and optimize the cell to improve the immunity to the disturb.

Finally, Chapter 5 dealt with specific aspects of oxide degradation resulting from carrier injection through the tunnel oxide. First, the programming window closure of the cell with increasing number of cycles (endurance degradation) was experimentally investigated. This allowed to separate the impact of oxide traps created during cycling on each of the program, erase and read phases. Threshold voltage reading and program efficiency were found to be highly impacted. The experimental study on equivalent transistors also showed that both programming and erase operations significantly degrade the device. Hence, the second part of the study was devoted to the hot carrier injection induced degradation and more particularly to the generation of interface defects. In the spirit of the previous chapters and in agreement with the latest works on this topic, a microscopic perspective is adopted to model the generation of such traps. Degradation resulting from hot carriers at the drain and cold carriers in the channel was included and the resulting degraded current-voltage curves were simulated. The good agreement obtained after comparisons with measurements in terms of defects profile along the channel (obtained with charge pumping) and macroscopic characteristics (V_{th}, G_m, S) at different hot carrier stress conditions, support the adopted approach.

List of publications

- **A. Zaka**, Q. Rafhay, P. Palestri, R. Clerc, D. Rideau, L. Selmi, C. Tavernier and H. Jaouen, *On the accuracy of current TCAD hot carrier injection models for the simulation of degradation phenomena in nanoscale devices*, In International Semiconductor Device Research Symposium (ISDRS) 2009, pages 1-2, dec. 2009.

- **A. Zaka**, P. Palestri, D. Rideau, M. Iellina, E. Dornel, Q. Rafhay, C. Tavernierand H. Jaouen, *Programming efficiency and drain disturb trade-off in embedded non-volatile emories*, In International Workshop on Computational Electronics (IWCE) 2010, pages 1-3, oct. 2010.

- **A. Zaka**, D. Garetto, D. Rideau, P. Palestri, J.P. Manceau, E. Dornel, Q. Rafhay, R. Clerc, Y. Leblebici, C. Tavernier and H. Jaouen, *Characterization and modelling of gate current injection in embedded non-volatile Flash memory*, In International Conference on Microelectronic Test Structures (ICMTS) 2011, pages 130-135, 2011.

- **A. Zaka**, Q. Rafhay, M. Iellina, P. Palestri, R. Clerc, D. Rideau, D. Garetto, E. Dornel, J. Singer, G. Pananakakis, C. Tavernier and H. Jaouen, *On the accuracy of current TCAD hot carrier injection models in nanoscale devices*, Solid-State Electronics, vol. 54, no. 12, pages 1669-1674, 2010.

- **A. Zaka**, J. Singer, E. Dornel, D. Garetto, D. Rideau, Q. Rafhay, R. Clerc, J.-P. Manceau, N. Degors, C. Boccaccio, C. Tavernier and H. Jaouen, *Characterization and 3D TCAD simulation of NOR-type Flash non-volatile memories with emphasis on corner effects*, Solid-State Electronics, vol. 63, no. 1, pages 158-162, 2011.

- **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*, IEEE Transactions on Electron Devices, 2011 (submitted).

- 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.

- 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 International Memory Workshop (IMW) 2011, pages 1-4, may 2011.

- D. Garetto, Y.M. 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 Conference on Ultimate Integration on Silicon (ULIS) 2011, pages 1-4, march 2011.

- D. Garetto, Y.M. Randriamihaja, **A. Zaka**, D. Rideau, A. Schmid, H. Jaouen, Y. Leblebici, *Analysis of defect capture cross sections using non-radiative multiphonon-assisted trapping model*, Solid-State Electronics, Available online 25 November 2011, ISSN 0038-1101, 10.1016/j.sse.2011.10.024, 2011.

- Y.M. Randriamihaja, **A. Zaka**, V. Huard, M. Rafik, D. Rideau, D. Roy and A. Bravaix, *Mosfet's hot carrier degradation characterization and modeling at a microscopic scale*, In International Reliability Physics Symposium (IRPS) 2011, pages XT.5.1-XT.5.3, april 2011.

- Y.M. Randriamihaja, **A. Zaka**, V. Huard, P. Palestri, D. Rideau, D. Roy, A. Bravaix and M. Rafik, *Hot carrier degradatino: from defect creation modeling to their impact on MOS parameters*, In International Reliability Physics Symposium (IRPS) 2012 (submitted).

- Y.M. Randriamihaja, V. Huard, **A. Zaka**, S. Haendler, X. Federspiel, M. Rafif, D. Rideau, D. Roy, A. Bravaix, *Oxide defects generation modeling and impact on BD understanding*, In International Reliability Physics Symposium (IRPS) 2011, april 2011.

Appendix A

Probability Scheme

This appendix details the probabilities of each scattering event when Optical Phonons and Impact Ionization scatterings are simultaneously included in the Model. Figure A.1 shows how the probabilistic events are split with their respective expressions.

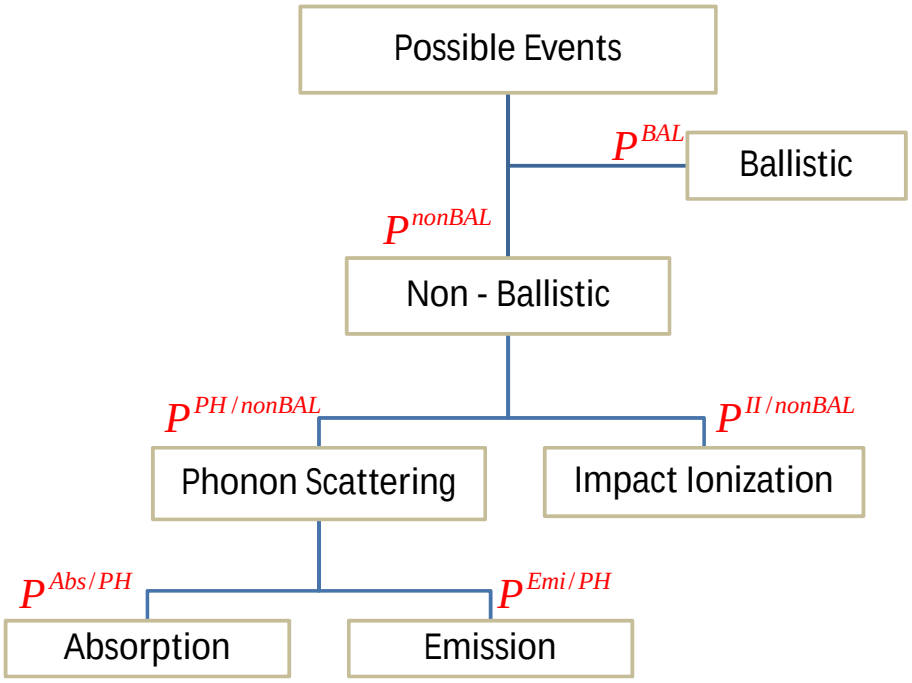


Figure A.1

The notations P^i refer to the probability for a carrier to be *ballistic* considering the i-th mechanism, while the notations $P^{i/k}$ refer to the probability to have i knowing k. The elementary scattering mechanisms are: *Phonon Absorption* - *Abs*, *Phonon Emission* - *Emi*, *Impact Ionization* - *II* calculated after Equation 3.1. For instance, P^{Emi} refers to the probability for a carrier not to emit any phonon within a given distance. Hence, the calculation of the probability for a carrier to be ballistic is quite straightforward considering that is should not interact by any elementary mechanism:

$$P^{BAL} = P^{PH} \cdot P^{II} = P^{Abs} \cdot P^{Emi} \cdot P^{II} \tag{A.1}$$

In this expression, P^{PH} is the probability not to scatter by *any* phonon mechanism (neither absorption nor emission).

The remaining non-ballistic carriers either interact with a phonon or impact ionize if they have sufficient energy. Both events are respectively expressed by the probabilities $P^{PH/nonBAL}$ and $P^{II/nonBAL}$. As all the elementary scattering probabilities are *independently* calculated using their respective scattering rates, the sum of the probabilities of all the possible events should be weighted according to the available choices. Furthermore, throughout this work it has been assumed that non-ballistic carriers undergo *exactly* one scattering at a time.

Hence, taking into account the above considerations, the phonon/impact-ionization separation is described in details. The available choices are:

1. the carrier *does not* impact ionize **and** *does* scatter with a phonon:

$$P^{II}(1-P^{PH})$$

2. the carrier *does* impact ionize **and** *does not* scatter with a phonon:

$$P^{PH}(1-P^{II})$$

The ensemble of the possible choices is thus given by their sum: $P^{II}(1-P^{PH})+P^{PH}(1-P^{II})$. At this point, the probability for a carrier to scatter with a phonon (to impact ionize) knowing that it is non-ballistic, is given by the ratio of event 1 (2) over the total possible choices:

$$P^{PH/nonBAL} = \frac{P^{II}(1-P^{PH})}{P^{II}(1-P^{PH})+P^{PH}(1-P^{II})} \tag{A.2}$$

$$P^{II/nonBAL} = \frac{P^{PH}(1-P^{II})}{P^{II}(1-P^{PH})+P^{PH}(1-P^{II})} \tag{A.3}$$

Hence, let P_1 be the *probability for a carrier to scatter by phonon* and P_2 be the *probability for a carrier to impact ionize*. The above results naturally lead to:

$$P_1 = (1-P^{BAL})\cdot P^{PH/nonBAL} \tag{A.4}$$

$$P_2 = (1-P^{BAL})\cdot P^{II/nonBAL} \tag{A.5}$$

This calculation consistently splits the remaining non-ballistic carriers between both processes. Indeed, the total probability is conserved:

$$P_1 + P_2 + P^{BAL} = 1 \tag{A.6}$$

Following the same procedure as above, P_1 is further split into phonon absorption (P_3) and emission (P_4) processes (Figure A.1), implying:

$$P^{Abs/PH} = \frac{P^{Emi}(1-P^{Abs})}{P^{Abs}(1-P^{Emi}) + P^{Emi}(1-P^{Abs})} \tag{A.7}$$

$$P^{Emi/PH} = \frac{P^{Abs}(1-P^{Emi})}{P^{Abs}(1-P^{Emi}) + P^{Emi}(1-P^{Abs})} \tag{A.8}$$

The probability to emit or absorb a phonon is thus given by:

$$P_3 = (1-P^{BAL}) \cdot P^{PH/nonBAL} \cdot P^{Abs/PH} \tag{A.9}$$

$$P_4 = (1-P^{BAL}) \cdot P^{PH/nonBAL} \cdot P^{Emi/PH} \tag{A.10}$$

The sum of the probabilities of all the possible events (the termination of all branches of Figure A.1) is thus equal to 1:

$$P_2 + P_3 + P_4 + P^{BAL} = 1 \tag{A.11}$$

Hence, when Impact Ionization is included and keeping the same notations as in Figure 3.4, the upward and downward fluxes are given by:

$$P^{UP} = P_3 \tag{A.12}$$

$$P_{DOWN} = P_2 + P_4 \tag{A.13}$$

In the case when Impact Ionization is not included in the simulations (Optical Phonons only), the Equations A.1, A.9, A.10 are readily modified, thus obtaining:

$$P^{BAL} = P^{Abs} \cdot P^{Emi} \tag{A.14}$$

$$P_3 = (1-P^{BAL}) \cdot P^{Abs/PH} \tag{A.15}$$

$$P_4 = (1-P^{BAL}) \cdot P^{Emi/PH} \tag{A.16}$$

These are indeed the expressions used in section 3.2.

Appendix B

Perpendicular Flux Calculation

This appendix shows the details of the calculation of the normal flux as a function of the normal energy starting from the carrier distribution function expressed in total energy. This projection involves the discretization of the distribution function by a Dirac comb. Thus, the calculation of the impulse response of the system is first performed before generalizing to the comb.

Let the carrier distribution function in total energy $n(\varepsilon)$ (in units of: $cm^{-3}eV^{-1}$) be centred at ε_0. The total carrier density n (in units of: cm^{-3}) can be calculated as:

$$n = \int_0^\infty n(\varepsilon)\delta\left(\varepsilon - \varepsilon_0\right) d\varepsilon = \int_0^\infty f(\varepsilon)g(\varepsilon)d\varepsilon \tag{B.1}$$

In this equation, $f(\varepsilon)$ is the probability function (in units of: ϕ) and $g(\varepsilon)$ is the density of states (in units of: $cm^{-3}eV^{-1}$). Identifying both sides of the equation, it is possible to write:

$$f(\varepsilon)g(\varepsilon) = n(\varepsilon)\delta\left(\varepsilon - \varepsilon_0\right) \Longrightarrow f(\varepsilon) = \frac{n(\varepsilon)}{g(\varepsilon)}\delta\left(\varepsilon - \varepsilon_0\right) = \frac{n(\varepsilon_0)}{g(\varepsilon_0)}\delta\left(\varepsilon - \varepsilon_0\right) \tag{B.2}$$

The value of the carriers' concentration at ε_0 ($n(\varepsilon_0)$, in units of: cm^{-3}) is already known. $g(\varepsilon)$ is instead calculated as:

$$g(\varepsilon) = \frac{2}{(2\pi)^3}\int_{-\infty}^{\infty} \delta\left(\varepsilon(\vec{k}) - \varepsilon(\vec{k'})\right) d\vec{k'} \tag{B.3}$$

Assuming parabolic bands and isotropic distributions in $\vec{k}$, the projection of Cartesian to spherical coordinates yields:

$$g(\varepsilon) = \frac{2}{(2\pi)^3}\int_0^\infty\int_0^\pi\int_0^{2\pi} \delta\left(\frac{\hbar^2k^2}{2m} - \frac{\hbar^2k'^2}{2m}\right) k'^2 \sin\theta \; dkd\theta d\phi \tag{B.4}$$

k is the amplitude of the momentum vector, θ is the polar angle with the k_z direction and ϕ is the angle between the momentum vector with the k_x direction in the xy-plane projection. After both angle integrations and exploiting Dirac-delta properties, we obtain the density of states:

$$g(\varepsilon) = \frac{\sqrt{2m^3\varepsilon}}{\pi^2\hbar^3} \tag{B.5}$$

Evaluating Equation B.5 at ε_0 ($g(\varepsilon_0)$, in units of: cm^{-3}) and combining it with Equation B.2, the probability function in total energy is finally written as:

$$\boxed{f(\varepsilon) = \frac{n(\varepsilon_0)\pi^2\hbar^3}{\sqrt{2m^3\varepsilon_0}}\delta\left(\varepsilon - \varepsilon_0\right) = A\delta\left(\varepsilon - \varepsilon_0\right)} \tag{B.6}$$

The constant A concentrates the information concerning the distribution function and the density of states at a given total energy. The next step is to calculate the flux perpendicular to the Si/SiO_2 interface as a function of the normal energy. Let k_z be the perpendicular direction which, under the working assumptions, is equivalent to all the other directions. Using the same variable change as previously, the perpendicular flux in spherical coordinates can thus be written as:

$$J_\perp = \frac{2}{(2\pi)^3}\int_0^\infty\int_0^{\frac{\pi}{2}}\int_0^{2\pi} f\left(\varepsilon(k)\right) v_\perp(k) k^2 \sin\theta \; dkd\theta d\phi \tag{B.7}$$

In the parabolic band approximation, the normal velocity is given by:

$$v_\perp(k) = \frac{\hbar k\cos\theta}{m} \tag{B.8}$$

Replacing Equation B.8 in Equation B.7 and integrating over ϕ, yields:

$$J_\perp = \frac{4\pi\hbar}{(2\pi)^3 m}\int_0^\infty f\left(\varepsilon(k)\right) k^3 \; dk \int_0^{\frac{\pi}{2}} \cos\theta\sin\theta d\theta \tag{B.9}$$

As the perpendicular energy is defined as:

$$\varepsilon_\perp = \frac{\hbar^2 k^2\cos^2\theta}{2m} \tag{B.10}$$

, the integration on θ in B.9 is not immediately calculated in order to keep the $\varepsilon_\perp$ dependence till the end. The first integral is computed by first making a variable change from k to ε and then using the Dirac delta properties:

$$J_\perp = \frac{Am}{\pi^2\hbar^3}\varepsilon_0\int_0^{\frac{\pi}{2}} \cos\theta\sin\theta d\theta \tag{B.11}$$

As fixed carrier energy has been assumed (ε_0), the perpendicular energy only depends on the θ value and becomes a single variable function. It is therefore possible to make a variable change such as:

$$d\varepsilon_\perp = -\frac{\hbar^2 k_0^2}{m}\cos\theta\sin\theta d\theta = -2\varepsilon_0\cos\theta\sin\theta d\theta \tag{B.12}$$

k_0 is the momentum corresponding the the fixed ε_0 carrier energy. Inserting B.12 into B.11 yields:

$$J_\perp = \frac{Am}{2\pi^2\hbar^3}\int_0^{\varepsilon_0} d\varepsilon_\perp \tag{B.13}$$

The perpendicular energy is a positive value smaller than the total carrier energy. Hence the normal flux as a function of the normal energy can be finally written as:

$$J_\perp(\varepsilon_\perp) = \frac{Am}{2\pi^2\hbar^3}\Theta\left(\varepsilon_0 - \varepsilon_\perp\right) \tag{B.14}$$

Θ is the Heaviside function limiting the validity of this expression in the concerned energy range. Replacing the value of A from Equation B.6, the normal flux (in units of: $Acm^{-2}eV^{-1}$) is expressed as a function of the carrier density at ε_0:

$$\boxed{J_\perp(\varepsilon_\perp) = \frac{n(\varepsilon_0)}{2\sqrt{2m\varepsilon_0}}\Theta\left(\varepsilon_0 - \varepsilon_\perp\right)} \tag{B.15}$$

This constitutes the impulse response of the system and can be generalized by considering a discretization of the distribution function with a Dirac comb with a spacing of $\Delta\varepsilon_0$ (in units of: eV) and an index p:

$$n(\varepsilon) = \sum_{p=0}^{n} n(p\Delta\varepsilon_0)\delta\left(\varepsilon - p\Delta\varepsilon_0\right) \tag{B.16}$$

From the Dirac-delta to the full distribution case, we notice that:

$$\varepsilon_0 = p\Delta\varepsilon_0 \tag{B.17}$$

$$n(\varepsilon_0) = \Delta\varepsilon_0 \cdot n(p\Delta\varepsilon_0) \tag{B.18}$$

, the normal flux at a given normal energy $\varepsilon_{\perp 0i}$ is thus given by:

$$\boxed{J_\perp\left(\varepsilon_{\perp 0i}\right) = q\sum_{p=i}^{n}\frac{\Delta\varepsilon_0 \cdot n\left(p\Delta\varepsilon_0\right)}{2\sqrt{2mp\Delta\varepsilon_0}}\Theta(p\Delta\varepsilon_0 - \varepsilon_{\perp 0i})} \tag{B.19}$$