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

Modeling of Statistical Variability in Nanoscale Charge-trap Flash Memories

Doctoral Dissertation of:
Salvatore Maria Amoroso

Advisor:
Prof. Alessandro Sottocornola Spinelli

Tutor:
Prof. Ivan Rech

The Chair of the Doctoral Program:
Prof. Carlo Fiorini

2012 – XXIV cycle

To My Family

The difference between false memories and true ones is the same as for jewels: it is always the false ones that look the most real, the most brilliant.
(Salvador Dalì)

Si è così profondi, ormai, che non si vede più niente. A forza di andare in profondità si è sprofondati. Soltanto l'intelligenza, intelligenza che è anche leggerezza, che sa essere leggera, può sperare di risalire alla superficialità, alla banalità.
(Leonardo Sciascia)

Anyone who says he can see through women is missing a lot.
(Groucho Marx)

Abstract

THE ever growing market of mobile communication and digital consumer electronics has stimulated the rapid development of Flash memory technology. The Flash memory market is today estimated in tens of billions of dollars, and it has a huge growth potential, some of it at the expense of volatile memories and magnetic storage media. For more than 20 years the conventional floating-gate technology has been able to meet the requirements of higher storage density, higher programming/erasing speed, higher reliability and lower power design through a continuous scaling of the cell size. However, the floating gate technology faces nowadays difficult technical challenges and some physical limitations towards further scaling[1].

The charge-trap memory cell is considered today the most practical evolution of the floating-gate Flash cell, allowing improved reliability and scaling perspectives[1]. Stress-induced leakage current immunity, strongly reduced cell-to-cell parasitic interference, and the possibility to decrease the thickness of the gate dielectric stack and, therefore, the program/erase biases appear as the main promises of the charge-trap technology[2]. However, the discrete nature of the stored charge necessarily gives rise to statistical issues related to the number and position fluctuation of the electrons in the storage-layer, determining a statistical dispersion of the threshold voltage shift after the program operation. This statistical dispersion is expected to be further worsened when considering the additional contribution of atomistic doping to non-uniform substrate inversion, enhancing percolative source-to-drain conduction[3]. Moreover the statistical nature of the process ruling the injection of charge from the substrate into the storage layer, may represent a further important variability source for the program operation of nanoscale charge-trap memory devices, compromising the tightness of the programmed threshold-voltage distribution, as already pointed out for floating-gate devices[4]. Cell scaling increases the impact of these variabil-

[1]C. H. Lee et al, IEDM Tech. Dig., 613-616, 2003; Y. Shin Kim et al., IEDM Tech. Dig., 337-340, 2005; Y. Park et al, IEDM Tech. Dig., 29-32, 2006; J. S. Sim et al., Proc. Non-Volatile Semicon. Memory Workshop, 110-111, 2007.

[2]C.-H. Lee et al, Symp. VLSI Tech. Dig., 21-22, 2006; T. Ishida et al., IEEE Electron Device Lett., 920-922,2008.

[3]N. Sano et al, Microelectron. Reliab., 189-199, 2002; A. Asenov et al., IEEE Trans. Electron Devices, 1837-1852, 2003; G. Roy et al., IEEE Trans. Electron Devices, 3063-3070, 2006; M. F. Bukhori et al., Microelectron. Reliab., 1549-1552, 2008; A. Ghetti et al., IEEE Trans. Electron Devices, 1746-1752, 2009.

[4]C. Monzio Compagnoni et al., IEDM Tech. Dig., 165-168, 2007; C. Monzio Compagnoni et al., IEEE Trans. Electron Devices, 2695-2702, 2008.; C. Monzio Compagnoni et al., IEEE

ity sources, as the number of charges (electrons and ionized dopants) decreases shrinking the cell dimensions.

This work presents part of the research on these topics author has been involved in during the cycle XXIV of the PhD Course in Information Technology.

In particular the thesis focuses the attention on the statistical variability affecting the reading and programming operations of nanoscale charge-trap memories.

The manuscript is organized as follows. **Chapter 1** briefly introduces the floating-gate Flash technology, pointing out the main scaling limits for both NAND and NOR architectures. Then the charge-trap technology is presented highlighting its potential benefits in terms of reliability, scaling perspective and technological feasibility. The end of the chapter is devoted to present the major sources of statistical variability for the charge-trap technology.

Chapter 2 gives a thorough overview of the issues related to the resolution of individual discrete charges in 3D drift-diffusion simulations. Three possible approaches to deal with this problem are outlined, the first being the use of charge smearing over a fine mesh, the second being the splitting of the Coulomb potential into short and long range components, and the third being the introduction of quantum corrections for both the electrons and the holes in the solution domain. It will be shown that even the quantum corrections are not enough to remove the artificial charge localization introduced in drift-diffusion simulations dealing with single Coulombic attractive centers and a mobility model correction will be proposed to relieve this artifact from simulation results.

Chapter 3 presents a comprehensive investigation of threshold voltae shift variability in deeply-scaled charge-trap memory cells, considering both atomistic substrate doping and the discrete and localized nature of stored charge in the nitride layer. The first part of the chapter outlines the physics-based 3D TCAD model developed for this study: the statistical dispersion of the threshold voltage shift induced by a single localized electron in the nitride is evaluated in presence of non-uniform substrate conduction. The role of 3-D electrostatics and atomistic doping on the results is highlighted, showing the latter as the major spread source. The threshold voltage shift induced by more than one electron in the nitride is then analyzed, showing that for increasing numbers of stored electrons a correlation among single-electron shifts clearly appears. The second part of the chapter is devoted to the scaling trend and the practical impact of these statistical effects on cell operation: for fixed density of trapped charge, the average threshold voltage shift decreases with scaling the cell dimensions as a consequence of fringing fields, not predictable by any 1-D simulation approach. Moreover, the distribution statistical dispersion increases with technology scaling due to a more sensitive percolative substrate conduction in presence of atomistic doping and 3-D electrostatics. The impact of the discrete electron storage in the nitride on random telegraph noise (RTN) instabilities is also investigated, showing that despite single cell behavior may be modified, negligible effects result at the statistical level.

Trans. Electron Devices, 3192-3199, 2008; C. Friederich et al., IEDM Tech. Dig., 831-834, 2008; H.-T. Lue et al., Proc. IMW, 92-95, 2010; H. Liu et al., Proc. WMED, 1-3, 2009; K. Prall et al., IEDM Tech. Dig., 102-105, 2010.

Chapter 4 addresses the study of charge-trap memory programming variability. The first part of the chapter presents the physics-based Monte-Carlo model developed to simulate the statistical electron injection process from the substrate to the storage layer: for a correct evaluation of the threshold-voltage dynamics, cell electrostatics and drain current are calculated by means of a 3D TCAD approach in presence of atomistic doping, largely contributing to percolative substrate conduction, and in presence of discrete traps in the storage layer. Results show that the low average programming efficiency commonly encountered in nanoscale charge-trap memories mainly results from the low impact of locally stored electrons on cell threshold voltage in presence of fringing fields at the cell edges. The second part of the chapter evaluates the impact of the statistical process ruling electron injection on the statistical dispersion of cell threshold voltage during the program operation: it is shown that the discrete electron injection process plays the dominant role in determining the threshold voltage spread, compared to the fluctuation in the number and position of the trapping sites and to the fluctuation of the threshold-voltage shift induced by stored electrons in presence of percolative substrate conduction.

As will be shown in chapter 3, a further burden for the programming accuracy is given by RTN instabilities, whose amplitude is enhanced by percolative substrate conduction in presence of atomistic doping[5]. In particular, the threshold-voltage shift given by single RTN traps was shown to follow an exponential distribution[6], with standard deviation proportional to the square root of the channel doping concentration when a uniform doping profile is adopted. These results reveal that channel doping is one of the most important parameters for RTN in MOS devices, opening the possibility for technology optimizations by engineered doping profiles. To this purpose **Chapter 5** presents a thorough numerical investigation of the effect of non-uniform doping on random telegraph noise in nanoscale Flash memories, considering both discrete RTN traps and discrete channel dopants. For fixed average threshold voltage, the statistical distribution of the random telegraph noise fluctuation amplitude is studied with non-constant doping concentrations in the length, width or depth direction in the channel, showing that doping increase at the active area edges and retrograde and δ-shape dopings appear as the most promising profiles for random telegraph noise suppression. To carry out this optimization study we have adopted a floating-gate device template instead of a charge-trap device template. The reasons for this choice are manifold: (i) as previously mentioned and as will be shown in chapter 3, the statistical distribution of RTN instability is not influenced by the discrete electron storage in the nitride, (ii) no comprehensive doping optimization studies are reported in literature for the

[5]H. Kurata et al, Symp. VLSI Circ. Dig., 140-141, 2006; K. Sonoda et al., IEEE Trans. Electron Devices, 1918-1925, 2007; C. Monzio Compagnoni et al., IEEE Trans. Electron Devices, 388-395, 2008; C. Monzio Compagnoni et al., IEEE Electron Dev. Lett., 984-986, 2009; J. P. Chiu et al., IEDM Tech. Dig., 843-846, 2009; H. H. Mueller et al., J. Appl. Phys., 1734-1741, 1998; A. Asenov et al., IEEE Trans. Electron Devices, 839-845, 2003; N. Sano et al., M. F. Bukhori et al., Microelectron. Reliab., 1549-1552, 2008; A. Ghetti et al., IEEE Trans. Electron Devices, 1746-1752, 2009

[6]A. Ghetti et al., Proc. IRPS, 610-615, 2008; A. Ghetti et al, IEDM Tech. Dig., 835-838, 2008; A. Ghetti et al., IEEE Trans. Electron Devices, 1746-1752, 2009.

floating-gate technology (that is the current technology in production), (iii) the simulation of floating-gate devices is computationally less costly respect to the charge-trap case.

The conclusions of this work will be summarized at the end of the manuscript, outlining what has been accomplished and proposing some future work that can extend and improve the understanding of the effects of variability on the charge-trap memory performances.

Acknowledgments

AMAZING, intense, valuable. Here are three adjectives I would give to someone asking to describe these years of work towards my PhD graduation. I'd like to thank my supervisors Alessandro Spinelli and Christian Monzio Compagnoni who have been a constant guide during this period. I would also like to thank my industrial supervisor Aurelio Mauri and the whole TCAD group from Micron -Augusto Benvenuti, Andrea Ghetti, Gianpietro Carnevale, Luca Laurin, Andrea Marmiroli- for having financed my PhD scholarship and for their availability in terms of knowledge, computational resources, and friendship. Special thanks are due to professor Asen Asenov for giving me the opportunity to work in the Glasgow Group during this last year of studies. I would also like to thank my colleagues at Politecnico di Milano: Alessandro Maconi, Niccolò Castellani, Carmine Miccoli, Federico Nardi, Davide Fugazza, Carlo Cagli, Simone Lavizzari, Mattia Boniardi, Ugo Russo, Andrea Bonfanti, Michele Ghidotti, Guido Zambra and Daniele Ielmini: I shared with them joys and efforts that this period has given us. My gratitude also goes to professor Andrea Lacaita that leads our micro- and nano-electronic group. Last but not least the most sincere thanks are for my family for all the aid and support they offered me.

Contents

Chapter 1

The Flash memory technology

This chapter briefly introduces the semiconductor non-volatile Flash memory. An overview of the floating-gate Flash technology is given in the first part of the chapter, pointing out the main scaling limits for both NAND and NOR architectures. Then the charge-trap technology is presented, highlighting its potential benefits in terms of reliability, scaling perspective and technological feasibility. The end of the chapter is devoted to present the main sources of statistical variability affecting the charge-trap device, as their impact on the cell electrical behavior is the topic of this PhD thesis.

1.1 Introduction

COMPLEMENTARY metal-oxide-semiconductor (CMOS) memories can be divided into two main categories (Fig. 1.1): random access memories (RAM), which are volatile, and read-only memories (ROM), which are nonvolatile. Nonvolatile memory market share has been continuously growing in the past twenty years. This is due to a simple virtuous circle: the costs per bit of memory decrease with the size scaling (Fig. 1.2(a) and Fig. 1.2(b)), paving the way for new applications and market segments; under the pressure of the new audio/video/phones segments, the demand of BMb/year is set to grow exponentially . In particular, a great influence is played today by the "Applications Convergence", a concept that find the most striking example in devices like the IPod or the IPhone (Fig. 1.3(a) and Fig. 1.3(b)). However, it should be pointed out that the costs of the leading edge lithography tools increase exponentially with the technology scaling (Fig. 1.4(a)). For this reason the memory

cost could undergo a reversal of the trend in the next future (Fig. 1.4(b)). A possible solution to overcome this problem is believed to be the 3D-stacking approach, presented later in this chapter.

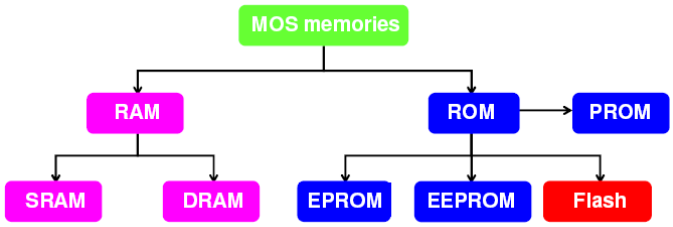
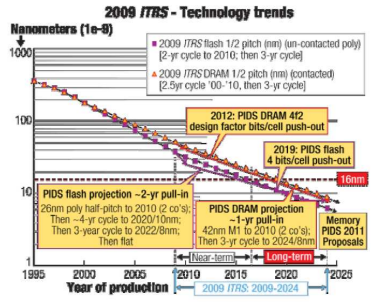
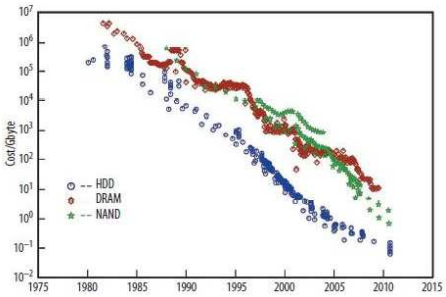


Figure 1.1: *CMOS family memory devices. Left branch represents volatile memories, right branch non-volatile memories.*



(a) *Technological scaling trend for Flash and DRAM memories.*



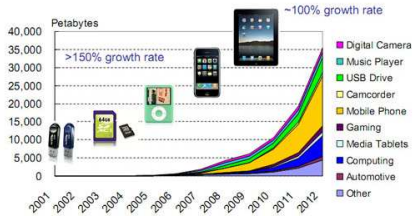
(b) *Costs scaling trend for Flash and DRAM memories.*

Figure 1.2: [ITRS and iSupply.]

Among the several semiconductor memories, DRAMs lead the volatile memories market, while Flash memories (in which a single cell can be electrically programmable and a large number of cells, called sector, are electrically erasable at the same time) dominate the non-volatile sector (Fig. 1.5) due to their enhanced flexibility against electrically programmable read-only memories (EPROM), which are electrically programmable but erasable via ultraviolet (UV) exposure. Electrically erasable and programmable read-only memories (EEPROM), which are electrically erasable and programmable per single byte, have been manufactured for specific applications only, since they use larger areas and, therefore, are more expensive.

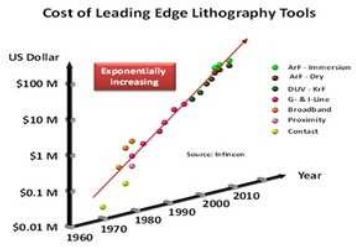
Flash memories have two major applications. One application (*code application*) is related to the nonvolatile memory integration in logic systems and microprocessors to allow software updates, store identification codes, or realize smart cards. The other application (*data application*) is to create storing elements, like memory boards or solid state hard disks. In particular, Flash

(a) *The Applications Convergence concept leads the semiconductor memory market growth.*

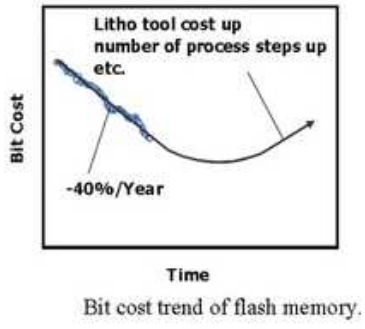


(b) *The request of semiconductor memories showed a growth rate around 100% in the last few years.*

Figure 1.3: [NVM Tutorial 2010, D. Ielmini, Politenico di Milano.]



(a) *Exponentially increase of the lithography costs with the memory technology scaling.*



(b) *Forecast of the trend reversal for the memory costs scaling curve.*

Figure 1.4: [Toshiba, H. Tanaka et al. VLSI Symposium 2007.]

memories with NOR architecture (which have faster access times and better reliability) are employed for code application, while Flash memories with NAND architecture (which have higher integration density but slower access times) are suitable for the storage application. The market evolution of the last few years is leading to an ever increasing growth of the NAND sector (Fig. 1.5), while the NOR sector shows a growth saturation.

1.2 The floating-gate device

A Flash memory is based on a MOS transistor with a threshold voltage that can change repetitively from a high to a low state, corresponding to the two states of the memory cell, i.e., the binary values (1 and 0) of the stored bit. Cells can be written into either state “1” or “0” by either programming or erasing methods, and they have to store the information independently of external conditions to meet the nonvolatile requirement.

The main idea of a floating-gate (FG) device is to store an amount of charge

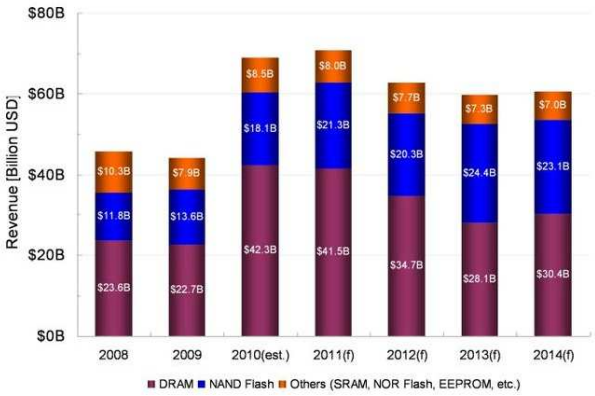
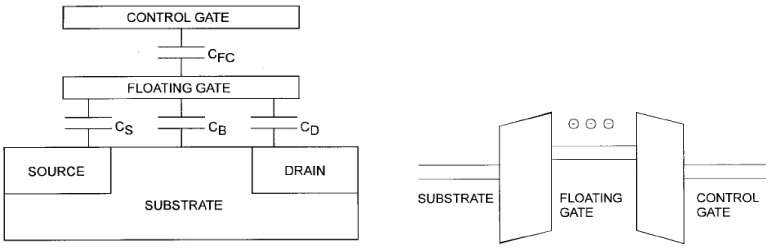


Figure 1.5: *Flash and DRAM devices dominate the semiconductor memory market [channeltimes.com].*



(a) *Schematic cross section of a generic FG device.*

(b) *Energy Band diagram of a generic floating-gate device.*

Figure 1.6: [1]

in a poly-silicon well buried in the gate oxide of a MOSFET. Indeed, the MOSFET threshold voltage can be written as [1, 2]:

$$V_T = K - \frac{\overline{Q}}{C_{ox}} \tag{1.1}$$

where K is a constant that depends on the gate and substrate material, doping, and gate oxide thickness, $\overline{Q}$ is the charge weighted with respect to its position in the gate oxide, and C_{ox} is the gate oxide capacitance. As can be seen, the threshold voltage of the memory cell can be altered by changing the amount of charge present between the gate and the channel.

Fig. 1.6(a) shows the schematic cross section of a generic FG device: the upper gate is the control gate and the lower gate, completely isolated within the gate dielectric, is the FG. The FG acts, energetically, as a potential well. If a charge is forced into the well, it cannot move from there without applying an external force: the FG stores charge, and then the information (Fig. 1.6(b)).

The simple model shown in Fig. 1.6(a) helps to understand the electrical behavior of an FG device. Here C_{FC}, C_S, C_D and C_B are the capacitances

between the FG and control gate, source, drain, and substrate regions, respectively. Consider the case when no charge is stored in the FG, i.e. $\overline{Q} = 0$:

$$\overline{Q} = 0 = C_{FC}(V_{FG} - V_{CG}) + C_S(V_{FG} - V_S) + C_D(V_{FG} - V_D) + C_B(V_{FG} - V_B) \tag{1.2}$$

where V_{FG} is the potential on the FG, V_{CG} is the potential on the control gate, V_S, V_D, V_B are potentials on source, drain, and bulk, respectively. If we name $C_T = C_{FG} + C_D + C_S + C_B$ the total capacitance of the FG, and we define $\alpha_J = C_J/C_T$ the coupling coefficient relative to the electrode J, then the potential on the FG due to capacitive coupling is given by

$$V_{FG} = \alpha_G V_{CG} + \alpha_D V_D + \alpha_S V_S + \alpha_B V_B \tag{1.3}$$

It should be noted that (1.3) shows that the FG potential does not depend only on the control gate voltage but also on the source, drain, and bulk potentials. If the source and bulk are both grounded, (1.3) can be rearranged as

$$V_{FG} = \alpha_G \left(V_{GS} + \frac{\alpha_D}{\alpha_G} V_{DS} \right) = \alpha_G (V_{GS} + f \cdot V_{DS}) \tag{1.4}$$

where

$$f = \frac{\alpha_D}{\alpha_G} = \frac{C_D}{C_{FG}} \tag{1.5}$$

Device equations for the FG MOS transistor can be obtained from the conventional MOS transistor equations by replacing MOS gate voltage with FG voltage and transforming the device parameters, such as threshold voltage V_T and conductivity factor β, to values measured with respect to the control gate. If we define for $V_{DS} = 0$

$$V_T^{FG} = V_T(floating \quad gate) = \alpha_G V_T(control \quad gate) = \alpha_G V_T^{CG} \tag{1.6}$$

and

$$\beta^{FG} = \beta(floating \quad gate) = \frac{1}{\alpha_G}\beta(control \quad gate) = \frac{1}{\alpha_G}\beta^{CG} \tag{1.7}$$

it is possible to compare the current-voltage equations of a conventional and an FG MOS transistor in the triode region (TR) and in the saturation region (SR) [3].

Conventional MOSFET:

$$\begin{aligned} TR \quad |V_{DS}| &< |V_{GS} - V_T| \\ I_D &= \beta \left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2 \right] \end{aligned} \tag{1.8}$$

$$\begin{aligned} SR \quad |V_{DS}| &\geq |V_{GS} - V_T| \\ I_D &= \frac{\beta}{2}(V_{GS} - V_T)^2 \end{aligned} \tag{1.9}$$

FG MOSFET:

$$\begin{aligned} TR \quad |V_{DS}| &< \alpha_G |V_{GS} + fV_{DS} - V_T| \\ I_{DS} &= \beta \left[(V_{GS} - V_T)V_{DS} - \left(f - \frac{1}{2\alpha_G} \right) V_{DS}^2 \right] \qquad (1.10) \\ SR \quad |V_{DS}| &\geq \alpha_G |V_{GS} + fV_{DS} - V_T| \\ I_{DS} &= \frac{\beta}{2} \alpha_G (V_{GS} + fV_{DS} - V_T)^2 \qquad (1.11) \end{aligned}$$

where β and V_T of (1.10)-(1.11) are measured with respect to the control gate rather than with respect to the FG of the stacked gate structure. They are to be read as $\beta(control \quad gate) = \beta_{CG}$ and $V_T(control \quad gate) = V_T^{CG}$.

It should be noted that the capacitive coupling between drain and floating gate makes different the behavior of the FG MOSFET respect the conventional MOSFET [3]. For example, the FG transistor can go into depletion-mode operation and can conduct current even when $|V_{GS}| < |V_T|$ (*drain turn-on*); moreover the FG MOSFET drain current will continue to rise as the drain voltage increases and saturation will not occur, contrary to the saturation region of a conventional MOSFET where the drain current is essentially independent of the drain voltage.

It should be noted also that the capacitive coupling ratio f depends on C_D and C_{FC} only, and its value can be verified by $f = -\frac{\partial V_{GS}}{\partial V_{DS}}_{(I_{DS}=cost)}$ in the saturation region

Many techniques have been presented to extract the capacitive coupling ratios from simple dc measurements [4–6]. The most widely used methods [7, 8] are (1)linear threshold voltage technique, (2) subthreshold slope method, and (3) transconductance technique. These methods require the measurement of the electrical parameter in both a memory cell and in a dummy cell, i.e., a device identical to the memory cell, but with floating and control gates connected. By comparing the results, the coupling coefficient can be determined. Other methods have been proposed to extract coupling coefficients directly from the memory cell without using a dummy one, but they need a more complex extraction procedure [9–11].

Let us consider the case when charge is stored in the FG, i.e. $\bar{Q} \neq 0$. All the other hypotheses made above hold true. Equations (1.4), (1.6) and (1.10) respectively, become

$$\begin{aligned} V_{FG} &= \alpha_G V_{GS} + \alpha_D V_{DS} + \frac{\bar{Q}}{C_T} \qquad (1.12) \\ V_T^{CG} &= \frac{1}{\alpha_G} V_T^{FG} - \frac{\bar{Q}}{C_T \alpha_G} = \frac{1}{\alpha_G} V_T^{FG} - \frac{\bar{Q}}{C_{FC}} \qquad (1.13) \\ I_{DS} &= \beta \left[\left(V_{GS} - V_T - \left(1 - \frac{1}{\alpha_G} \right) \frac{\bar{Q}}{C_T} \right) V_{DS} + \right. \\ &+ \left. \left(f - \frac{1}{2\alpha_G} \right) V_{DS}^2 \right] \qquad (1.14) \end{aligned}$$

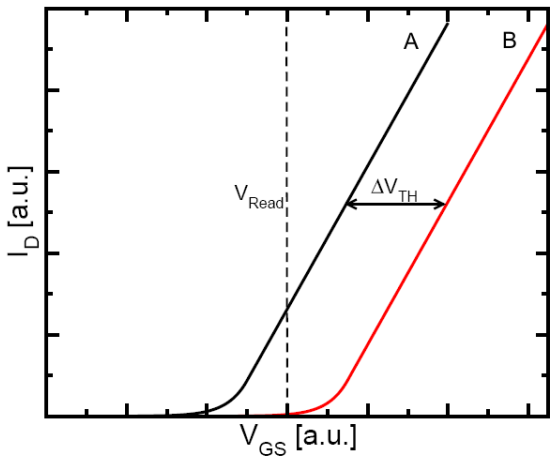


Figure 1.7: *IV curves of an FG device when there is no charge stored in the FG (curve A) and when a negative charge is stored in the FG (curve B) [12]*

Equation (1.13) shows the V_T dependence on $\bar{Q}$. In particular, the threshold voltage shift is derived as

$$\triangle V_T = V_T - V_{T0} = -\frac{\bar{Q}}{C_{FC}} \tag{1.15}$$

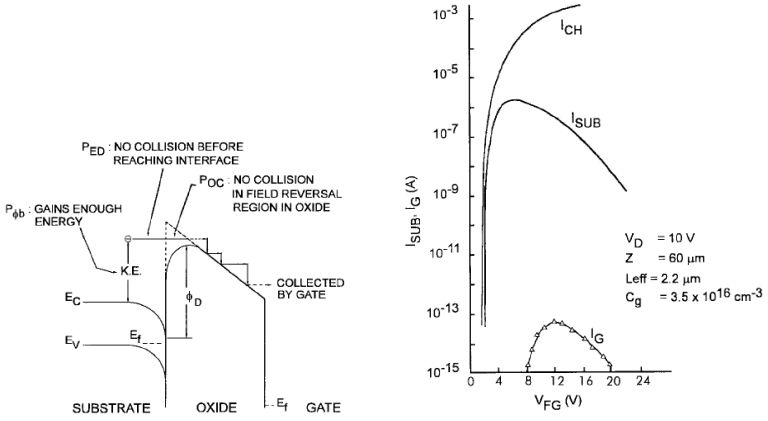
where V_{T0} is the threshold voltage when $\bar{Q} = 0$.

Equation (1.14) shows that the role of injected charge is to shift the IV curves of the cell (Fig. 1.7). In modern devices, a constant-current method is usually used to read the memory state: the threshold voltage is determined when a drain current of a fixed value (in the order of hundreds of nA to work in the sub-threshold region) is reached.

1.2.1 Charge injection mechanisms

In order to program (erase) a Flash memory cell, electric charge has to be transferred from the channel to the floating gate (and viceversa) through a dielectric material. To this aim, two physical mechanisms can be exploited: the Channel Hot Electron (CHE) injection and the Fowler-Nordheim (FN) tunneling.

Channel Hot Electron Injection: is the mechanism used for programming NOR Flash memories. An electron traveling from the source to the drain gains energy from the lateral electric field and loses energy to the lattice vibrations (acoustic and optical phonons). At low fields, this is a dynamic equilibrium condition, which holds until the field strength reaches approximately 100 kV/cm [15]. For fields exceeding this value, electrons are no longer in equilibrium with the lattice, and their energy relative to the conduction band edge begins to increase. Electrons are heated by the high lateral electric field, and a small fraction of them have enough energy to surmount the barrier between oxide and



(a) *A schematic energy band diagram describing the three processes involved in the hot electron injection*

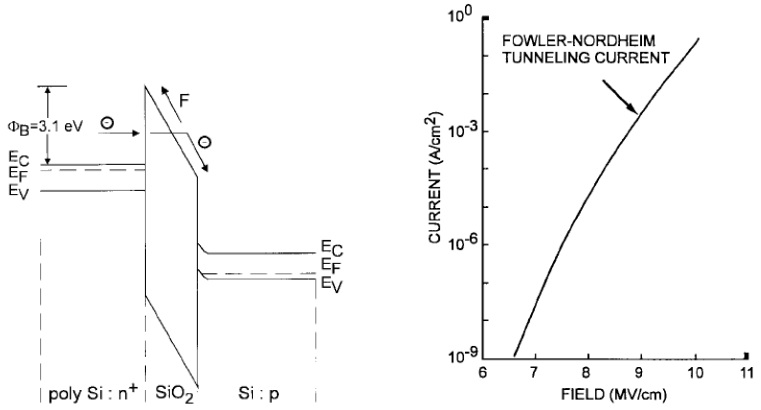
(b) *Channel, substrate and gate currents as a function of FG voltage.*

Figure 1.8: [1]

silicon conduction band edges. It is not then a tunneling mechanism since electrons do not "pass through" the energy barrier but "jump over" it. To evaluate how many electrons will actually cross the barrier, one should know the energy distribution $f_E(\varepsilon, x, y)$ as a function of lateral field ε, the momentum distribution $f_k(E, x, y)$ as a function of electron energy E (i.e., how many electrons are directed toward the oxide), the shape and height of the potential barrier, and the probability that an electron with energy E, wave vector k, and distance d from the Si/SiO interface will cross the barrier. Each of these functions needs to be specified in each point of the channel. A quantitative model, therefore, is very heavy to handle. Moreover, when the energy gained by the electron reaches a threshold, impact ionization becomes a second important energy-loss mechanism [13], which needs to be included in models. A simpler approach to obtain a first order evaluation of the hot electron gate current is based on the *lucky electron* model [14]. This model assess the probability of an electrons being lucky enough to travel ballistically in the field ε for a distance several times the mean free path without scattering, eventually acquiring enough energy to cross the potential barrier if a collision pushes it toward the Si/SiO_2 interface. Consequently, the probability of injection is the lumped probability of the following events, which are depicted in Fig. 1.8(a). Although this simple model does not fit precisely with some experiments, it allows a straightforward and quite successful evaluation of the relationship between the substrate current and the injection current:

$$I_G/I_{ch} \sim I_{sub}/I_{ch} e^{-\Phi/\Phi_i} \tag{1.16}$$

where I_{ch} is the channel current, Φ_i is the impact ionization energy and Φ the energy barrier seen from electrons (lowered by the image force). The substrate current is composed of holes generated by impact ionization in the



(a) *FN tunneling through a potential barrier in a MOS structure.*

(b) *FN tunneling current as a function of electric field.*

Figure 1.9: [1]

drain region. Holes are always generated since the energy ionization threshold (~ 1.6 eV) is lower than the injection energy barrier (~ 3.1 eV). Moreover, some holes can acquire enough energy from the lateral electric field to be injected into the oxide, thus degrading it. Although the CHE represents a very fast injection mechanism, it is very inefficient and leads to large power consumption, as understandable from Fig. 1.8(b).

Fowler-Nordheim Tunneling: is the mechanism used for programming NAND Flash memories and erasing both NAND and NOR devices. It is slower than the hot carrier injection mechanism, but more efficient because no excess substrate currents are generated. The probability of electron-tunneling depends on the distribution of occupied states in the injecting material and on the shape, height and width of the potential barrier. However, in presence of thin dielectric layers and high applied voltages, the energy barrier seen from electron for tunneling has a simple triangular shape (Fig. 1.9(a)). In this case, using a free electron gas to model the electron population in the injecting material and the Wentzel-Kramers-Brillouin (WKB) approximation to calculate the tunneling probability, the well-known analytical Fowler-Nordheim formula can be obtained

$$J = \frac{q^3 F^2}{16\pi^2 h^2 \Phi_B} exp\left[-4(2m_{ox}^*)^{1/2}\Phi_B^{3/2}/3\hbar qF\right] \tag{1.17}$$

where Φ_B is the barrier height, m_{ox}^* the effective mass in the dielectric, h the Planck's constant, q the electron charge, and F the electric field across the dielectric.

Fig. 1.9(b) shows $log(J)$ vs F . Since the field is roughly the applied voltage divided by the oxide thickness, a reduction of oxide thickness without a proportional reduction of applied voltage produces a rapid increase of the tunneling current. With a relatively thick oxide (20nm) one must apply a high voltage

(20V) to have an appreciable tunnel current. With thin oxides, the same current can be obtained by applying a much lower voltage. An optimum thickness (about 10 nm) is chosen in present devices, which use the tunneling phenomenon to trade off between performance constraints (programming speed, power consumption, etc.), which would require thin oxides, and reliability concerns, which would require thick oxides.

Although the simple and classic form of FN current density is in quite good agreement with experimental data, many features have been still undervalued: the temperature dependence of the phenomenon, the quantum effects at the silicon interface, the influence of band bending at the Si/SiO_2 interface and the voltage drop in silicon, the fact that the correct statistics for electrons are not Maxwellian but FermiDirac, and the image-force barrier lowering. Nevertheless, following a quantum approach it is possible to maintain the FN relationship as valid

$$J = A \cdot F^2 exp[-B/F] \tag{1.18}$$

on condition that, in this case, the coefficient A and B are functions of the applied electric field [15–17].

Recently the Fowler-Nordheim formula has been extended also to cylindrical geometry [18, 19] allowing the analytical computation of current densities in gate-all-around MOS devices.

1.2.2 ISPP programming algorithm

Whatever is the charge injection mechanism used for programming a memory cell, some programming scheme is required to achieve a strict control of the programmed threshold voltage. This is mandatory for the multi-level applications, where the placing of more than two bit for each cell requires very narrow threshold voltage distributions to be obtained. For a population of nominal identical cells, an accurate threshold voltage programming could be simply achieved, in theory, by an accurate control of the applied gate voltage and the programming duration. However in a realistic memory array, cells differ from each other because of the process spread, introducing for example a spread in the tunnel oxide thickness and in the initial threshold voltage value. Moreover, the threshold voltage value at a given time depends also from the cell's previous history, as different cells can loose, for example, different amounts of charge during the retention time. Thus, in practice, an adequately reproducible and narrow threshold voltage distribution cannot be achieved using a single programming pulse having a controlled amplitude and duration.

The most widely adopted solution to obtain the necessary programming precision, is that of dividing the program operation in a number of partial steps and at the end of each of them reading the cell (with the same circuitry used for the normal sensing operation) in order to determine whether or not the target threshold voltage is reached (Fig. 1.10). If it is not the case, another programming pulse of increased amplitude is applied to the gate and the whole procedure is repeated until successful completion. This algorithm is usually referred as *Incremental Step Pulse Programming* (ISPP) or also *Program and*

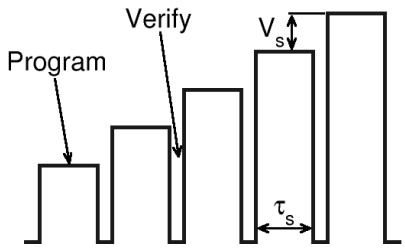


Figure 1.10: *Schematic representation of the Incremental Step Pulse Programming algorithm.*

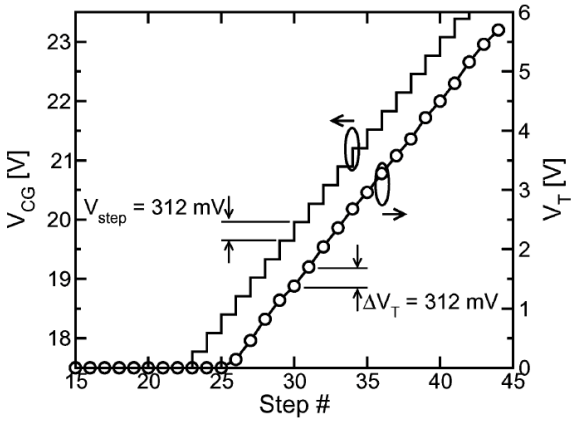


Figure 1.11: *Example for a control-gate voltage waveform used to program a FG cell and resulting V_T transient on a 60nm device.*

Verify scheme [20, 21]. Beside the programming precision, this algorithm offers also reliability advantages because only the minimum required charge flows through the tunnel oxide.

After an initial transient, a typical linear relation links the threshold voltage shift to the applied gate voltage (and in turn to the number of applied pulses) during the ISPP algorithm, as shown in Fig. 1.11. It should be noted that the slope $\partial \Delta V_T / \partial V_G$ is unitary for the floating-gate device [20–23], as expected from a straightforward analysis. Indeed, assuming a grounded substrate, we can express the electric field across the tunnel oxide as:

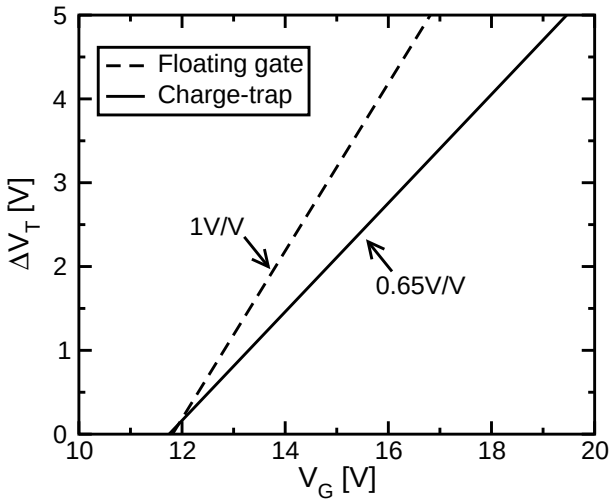


Figure 1.12: *Typical threshold voltage shift as a function of the gate voltage (i.e. as a function of the number of programming pulses) for a floating-gate device and for a charge trap device.*

$$F_{ox} = \frac{V_{FG}}{T_{top}} \tag{1.19}$$

where T_{top} is the gate oxide thickness. Considering that V_{FG} can be written as

$$V_{FG} = \alpha_G \left(V_G - \Delta V_T\right) \tag{1.20}$$

we obtain

$$F_{ox} = \frac{\alpha_G \left(V_G - \Delta V_T\right)}{T_{top}} \tag{1.21}$$

and finally

$$\frac{\partial \Delta V_T}{\partial V_G} = 1 - \frac{T_{top}}{\alpha_G} \frac{\partial F_{ox}}{\partial V_G} \tag{1.22}$$

that means that the slope $\partial \Delta V_T / \partial V_G$ is equal to 1 if the field F_{ox} is constant during the ISPP staircase. The fact that, after an initial transient, the field F_{ox} reaches a stationary condition can be easily understood: in fact an increase of ΔV_T lower than an increase of V_G would mean an increase in the field F_{ox} and in turn an increase in the injected charge and then and increase in ΔV_T. Hence, after an initial transient, the expected regime slope is 1.

On the other hand, for charge-trap memories (presented in the next section) it has been always reported a threshold variation per step lower than V_s (i.e.

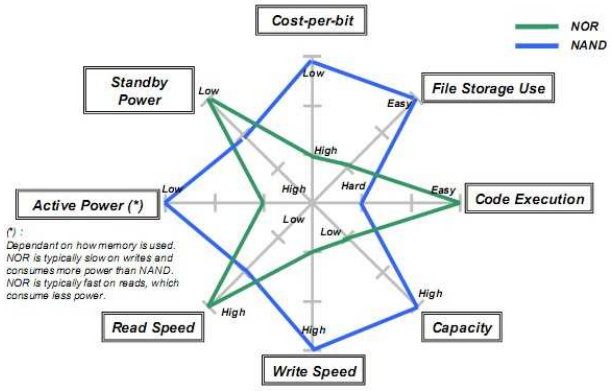


Figure 1.13: *NAND and NOR performances comparison [Toshiba].*

$\partial \Delta V_T / \partial V_G < 1$) [24, 25], as exemplified in Fig. 1.12. This behavior is anomalous because the above mentioned analysis holds also for charge-trap devices, provided that the tunnel oxide electric field is written as:

$$F_{ox} = \frac{V_G - \Delta V_T}{EOT} \tag{1.23}$$

where EOT is the *equivalent oxide thickness* of the gate stack dielectric. Hence, also for these devices the expected ISPP slope is unitary.

The explanation of this anomaly is not trivial. Some attempt to understand this behavior, in terms of electron trapping inefficiency, was made in [26]. However, in chapter 4 of this thesis we will demonstrate that the sub-unitary ISPP slope value can be explained in terms of the low impact of trapped electrons on the threshold voltage shift due to 3D electrostatics effects.

Finally we want to point out that the theoretical programming accuracy of the ISPP algorithm is given by the voltage increase per step (V_s in Fig. 1.10). Indeed, neglecting errors due to sense amplifier accuracy or voltage statistical fluctuations, the last program pulse applied to a cell will cause its threshold voltage to be shifted above the decision verify level by an amount at most as large as V_s (or even less than V_s for the charge-trap device).

1.2.3 NOR and NAND architectures

In order to achieve high memory capacity per chip, several memory devices have to be organized in array structures. There are two dominant array organizations used today for Flash memories: NOR and NAND architectures. In the internal circuit configuration of NOR Flash, the individual memory cells are connected in parallel, which enables the device to achieve random access. This configuration enables the short read times required for the random access of microprocessor instructions. NOR Flash is ideal for lower-density, high-speed read applications, which are mostly read only, often referred to as *code-storage*

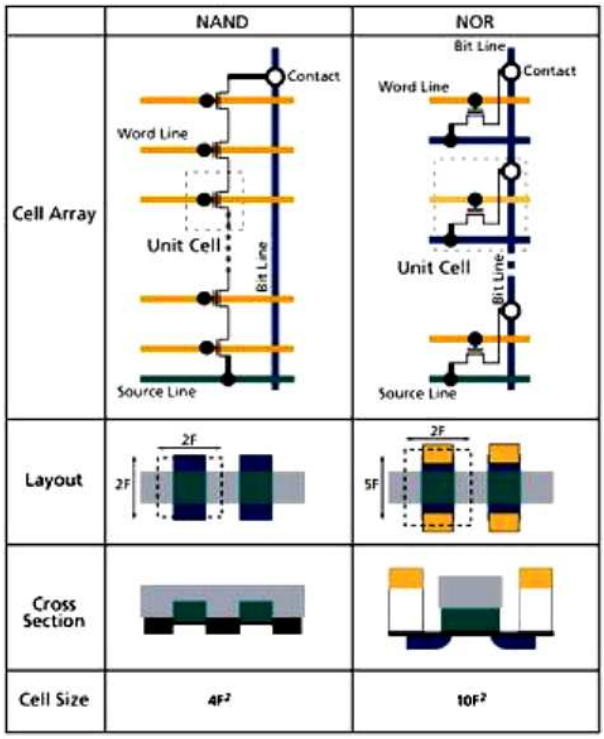


Figure 1.14: *Schematic, layout and cross-section of NAND and NOR array organization.*

applications. NAND Flash was developed as an alternative optimized for high-density data storage, giving up random access capability in a tradeoff to achieve a smaller cell size, which translates to a smaller chip size and lower cost-per-bit. This was achieved by creating an array of several memory transistors connected in a series. Utilizing the NAND Flash architectures high storage density and smaller cell size, NAND Flash systems enable faster write and erase by programming blocks of data. NAND Flash is ideal for low-cost, high-density, high-speed program/erase applications, often referred to as *data-storage* applications. Fig. 1.13 provides a qualitative summary of how NAND and NOR Flash vary for a number of important design characteristics: capacity, read speed, write speed, active and standby power consumption, cost-per-bit, and ease of use for file storage and code storage applications.

NOR architecture was proposed for the first time by Dr. Masuoka (Toshiba) in 1980. The schematic, the layout and the cross-section of this array organization are shown in Fig. 1.14 (right), while Fig. 1.15 shows the typical threshold voltage distributions for erased and programmed cells. It is peculiar of NOR architecture that the erase cells have always a positive threshold voltage to avoid the presence of cells in the ON state (also when non selected) precluding the correct reading operation of a selected cell in the same bit-line. To read a cell, the selected word-line voltage is raised and the selected bit-line is connected

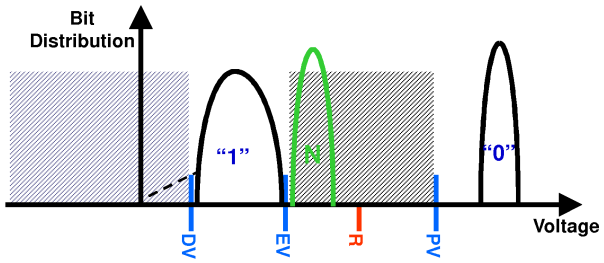


Figure 1.15: *Qualitative threshold voltage distributions for erased and programmed cells in a NOR architecture.*

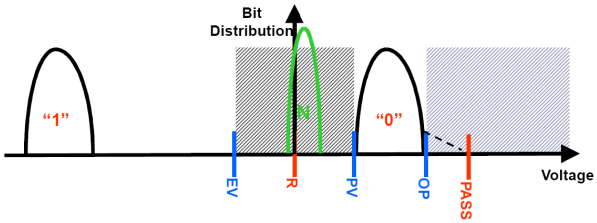


Figure 1.16: *Qualitative threshold voltage distributions for erased and programmed cells in a NAND architecture.*

to the sense amplifier: only if the cell is in the erased status, a current flow is measured at the sense amplifier. To program a cell (by means of the channel hot electron injection), both bit-line voltage and word-line voltage are raised in order to create hot electrons with the electric field between source and drain, and inject them through the tunnel oxide with the electric field between channel and gate. The erase operation is not applied to a single cell but to a whole sector: in this case the sector is first programmed to bring the cells to the same high threshold voltage value and then the substrate voltage (common to all the cells in the sector) is raised to remove the electrons from the floating gates by means of the FN tunneling mechanism. A issue for NOR architecture is represented by the *over-erasing* phenomenon: faster cells (e.g. that with a thinner tunnel oxide) risk to reach a negative erased threshold. To avoid this problem, all the sector is checked after the erase operation and the cells with negative threshold voltage are *soft-programmed* to correct their threshold value. The presence of the source and drain contact for each cell and the poor scalability of their junction (which have to avoid punch-through phenomena during the CHE injection) are the reasons that make the NOR device unable to meet the *data-storage* market requirements. On the other hand, the high reading speed and the random accessibility make the NOR device suitable for the *code-storage*.

NAND architecture was proposed again by Dr. Masuoka (Toshiba) in 1987. The schematic, the layout and the cross-section of this array organization are shown in Fig. 1.14 (left), while Fig. 1.16 shows the typical threshold

voltage distributions for erased and programmed cells. In this case, unlike NOR devices, the threshold voltage of erased cells is negative. The serial organization complicate the reading operation (Fig. 1.17). There are three phase to read a cell: *pre-charging*,*evaluation* and *sensing*. The cell to be read out is first selected, setting the corresponding word-line voltage to zero (V_{read} = 0 V). The selected bit-line is then *pre-charged* (due to the parasitic capacitance) turning on the source and drain selectors and then is left floating to *evaluate* the selected cell status: only if the selected cell is in an erased status, a current flows toward ground discharging the pre-charged bit-line. In this way the cell status can be read during the *sensing* phase, when the source and drain selectors are turned off, reading the voltage present on the bit-line. In order to make the bit-line discharging dependent only on the selected cell, a voltage V_{pass} (higher than the maximum threshold voltage) is applied to the other word lines. The programming operation is realized by means of the FN tunneling, and require high voltages ($\sim 20V$) to be applied across the tunnel oxide layer. To this aim, the channel of the selected cell is grounded, grounding the corresponding bit-line with the drain selector ON and the source selector OFF. The word-lines of unselected cells are raised to a voltage high enough to allow the inversion of each channel (to have a uniformly grounded bit-line), but low enough to avoid a FN programming of these cells (Fig. 1.18). On the other hand, to avoid the programming of unselected cells placed on the same word-line of the selected cell, a *self-boosting* technique is used (Fig. 1.19): this technique provides the necessary program inhibit voltage by electrically isolating the unselected bit-lines (after pre-charge) and applying a pass voltage (e.g. 10 V) to the unselected word-lines during programming. The unselected word-lines couple to the channel of the NAND strings corresponding to the unselected bit lines, causing a voltage (e.g. 8 V) to be impressed in the channel of the unselected bit lines, thereby preventing program disturbs. It should be mentioned that usually an incremental step pulse programming (ISPP) algorithm is used to program the selected cell: pulses of increases amplitude are applied to the gate, followed by verify operations to determine whether a target threshold voltage is reached or not; if the target threshold is reached then the algorithm stops, otherwise another pulse is applied. Finally, the erasing operation is carried on by sectors, similarly to the NOR case. Since in the NAND architecture only one source/drain contact is present for each string, the cells have smaller dimensions than NOR cells: this allows the NAND devices to be better scaled, making this architecture suitable for the *data-storage*. On the other hand, the complexity of the reading operation and the necessity to read all the cells on the same word-line ($\sim$ tens of thousands) make the NAND device unable to meet the *code-storage* market requirements. It should be noted that, due to the array organization, several disturbs can affect the reading, programming and erasing operation: these disturbs are briefly reported in Figs. 1.20(a)-1.20(d).

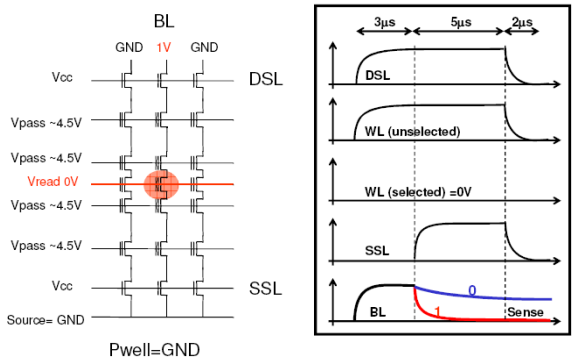


Figure 1.17: *NAND read out operations: pre-charge, evaluation, sense.*

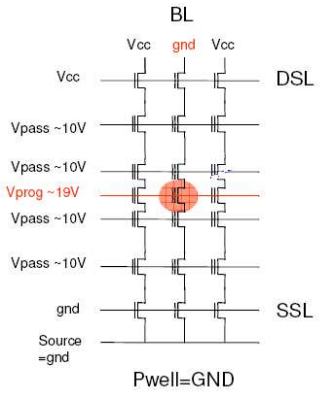


Figure 1.18: *NAND program operation.*

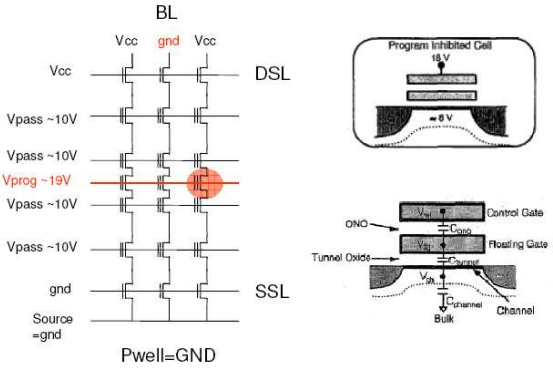
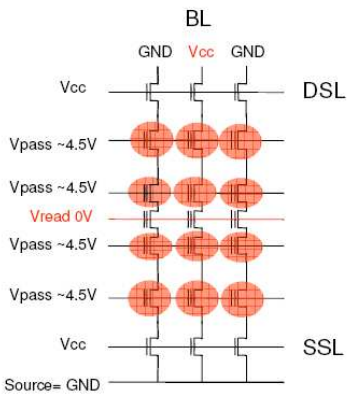
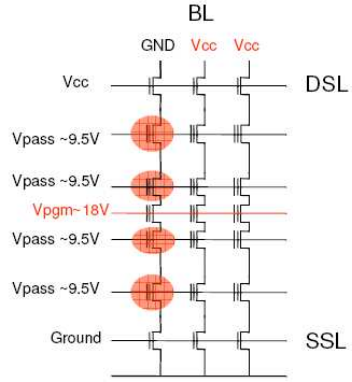


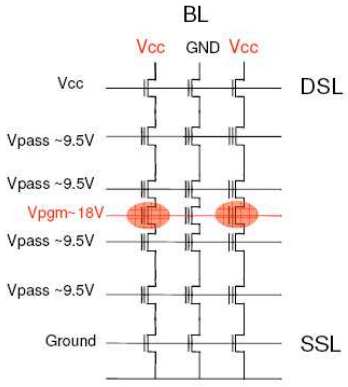
Figure 1.19: *Self-boosting technique used to reduce the program disturb along the selected word-line.*



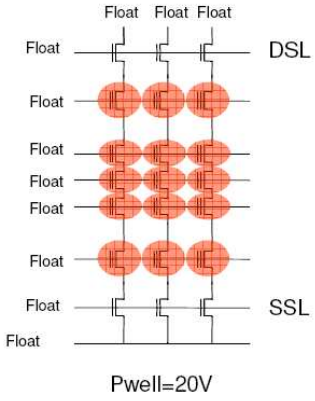
(a) *Read disturb affecting the unselected word-lines and all bit-lines.*



(b) *Pass disturb affecting the unselected cells of the selected bit-line.*

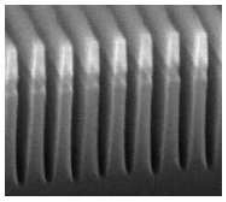


(c) *Program disturb affecting the unselected cells on the selected word-line.*

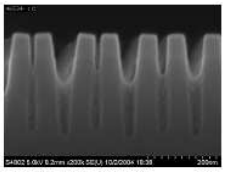


(d) *Erase disturb affecting the unselected neighbor sectors.*

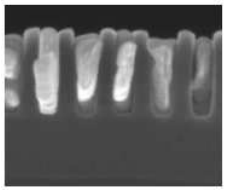
Figure 1.20:



(a) *Active area structure for an optimized 25nm FG memory cell.*

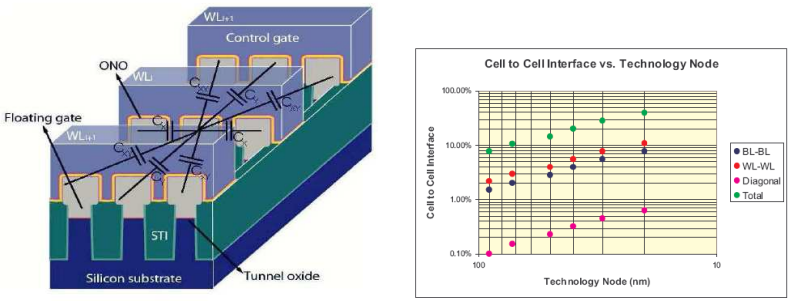


(b) *STI structure bending after the SOD (Spin On Dielectric) application.*



(c) *Deformed DMI structure for a 40nm cell after metallization.*

Figure 1.21: [27]



(a) *The capacitive coupling between adjacent cells is a major limit for the FG device scaling.*

(b) *Relative threshold voltage shift due to the interference between adjacent floating-gates as a function of technology node.*

Figure 1.22: [27]

1.2.4 Scaling issues

The size scaling of nonvolatile floating-gate Flash memories below 30nm presents several kind of issues. It is necessary, in fact, to take into account (a) the mechanical integrity of the device structures, (b) the problem related to the floating-gate scaling, (c) the coupling noise between closer cells in the same array.

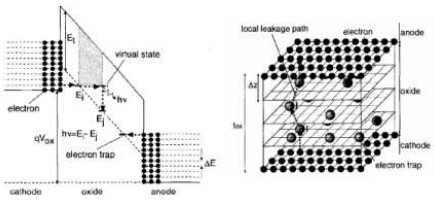
Mechanical and structural integrity. As reported in Figs. 1.21(a)-1.21(c), shared structures between more cells -such as STI (*Shallow Trench Isolation*) o DMI (*Damascene Metal Patterning*)- can be subject to collapse or to bending under mechanical/thermal/electrostatic stress.

Capacitive interference between adjacent cells. This interference phenomenon related to the capacitive coupling between adjacent cells is a major limiting factor for the conventional floating-gate Flash memory scaling (Fig. 1.22(a)). The programming operation of one cell induces a disturb in the eight neighbors cells, causing an unwant shift of their threshold voltage. As shown in Fig. 1.22(a), this threshold voltage shift can reach over the 50% of the actual threshold voltage value for the 20nm technology node.

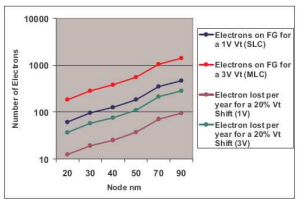
Charge losses through the tunnel oxide defects. The tunnel oxide thickness scaling is limited by the Stress Induced Leakage Current (SILC) phenomenon. The continuous program/erase cycling induces, in fact, defects on the tunnel oxide layer. These defects promote the charge loss from the floating-gate to the substrate by Trap Assisted Tunneling (TAT) (Fig. 1.23(a))[1]. This worsens the device retention performances, but also enhances the read/program disturbs. The tunnel oxide thickness is usually limited to 7nm. This limitation is more strict for scaled devices since the number of stored electrons in the floating-gate is exiguous (Fig. 1.23(b)).

Variability. A number of variability sources acquires more importance as the floating-gate cell dimensions are scaled down. Several of these sources in-

[1]Multiple traps aligned in space and energy can form direct charge loss paths, called percolative paths.



(a) *Energy band diagram and schematic representation of the TAT mechanism.*



(b) *Number of electrons in the floating-gate as a function of the technology node.*

Figure 1.23: [27]

creases proportional to $1/WL$, showing a rising of 30-50% for each technological node. Among them we can enumerate: the random dopant fluctuation, the random telegraph noise, the poly-silicon gate granularity, the line-edge roughness. The variability sources will be introduced in the next section where we will address the statistical variability of charge-trap memories.

Scaling Flash vs. Scaling CMOS. It should be noticed that the Flash device requires high programming/erasing voltages. If the scaling of conventional MOS transistor goes hand in hand with the scaling of applied voltages, this cannot be true for the floating-gate MOS. Indeed, the charge injection mechanisms require voltages depending on non-scalable parameters, such as the silicon-oxide barrier (3.1eV) and the oxide thickness ($>$ 7nm).

1.3 The charge-trap device

The floating-gate Flash technology scaling issues are particularly severe for the NAND architecture. This architecture allows a high density of devices per chip and it is indeed used for *data-storage* applications. However, for the same reason, this architecture suffers more than the other from the capacitive interferences problem.

As visible from the ITRS (*International Technology Roadmap for Semiconductors*) projections reported in Fig. 1.24, the charge-trap memory cell is considered today the most practical evolution of the floating-gate Flash cell for NAND architectures, allowing improved reliability and scaling perspectives. Stress-induced leakage current immunity, strongly reduced cell-to-cell parasitic interference, and the possibility to decrease the thickness of the gate dielectric stack and, therefore, the program/erase (P/E) biases appear as the main promises of the charge-trap technology [28]. In a charge trap memory device

- the charge is stored in the electronic defects of an high defects-density material (e.g. Si_3N_4);
- it is not required the lithographic definition of a *floating gate* during the process flow (Fig. 1.25): the memory stack consist of a dielectric trilayer between substrate and gate. This stack is then more scalable in comparison to a structure with a floating-gate;

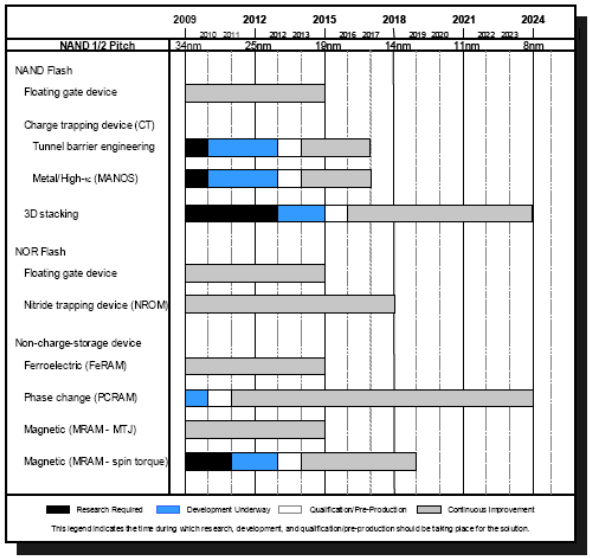


Figure 1.24: *2009 ITRS forecasting for the post-FG memory scenario.*

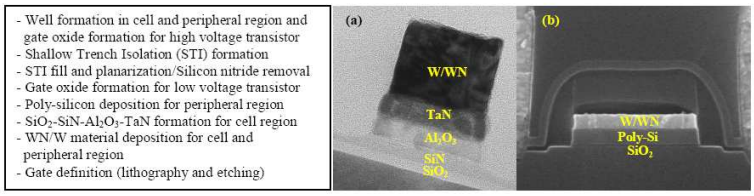


Figure 1.25: *Process flow for a TANOS memory cell; (a) TANOS cell TEM; (b) peripheral transistor TEM. [[29]]*

- cell-to-cell capacitive interferences are sensibly reduced, because the trapping layer is a dielectric material (instead of a semiconductor) and because its thickness ($\sim$ 5nm) is by far less than a floating-gate thickness ($\sim$ 70nm)

A thin ($\sim$ 4nm) silicon dioxide layer is always used as *tunnel oxide*, while silicon dioxide or high-k materials (e.g. Al_2O_3) are used as top dielectric (*blocking oxide*). The trapping layer is usually silicon nitride, though also high-k materials (e.g. HfO_2) are employed. Polysilicon or metal (e.g. TaN) are adopted as gate electrode. Depending on the materials employed for the blocking oxide and for the gate electrode, several names are adopted in literature for the charge-trap device: SONOS (*PolySi-Oxide-Nitride-Oxide-Si*), SANOS (*PolySi-AlO-Nitride-Oxide-Si*), TANOS (*TaN-AlO-Nitride-Oxide-Si*) (Fig. 1.26). Both programming and erasing are realized by FN tunneling, this kind of device belonging to the NAND type memories.

Contrary to what was previously said for the floating-gate device, in charge-

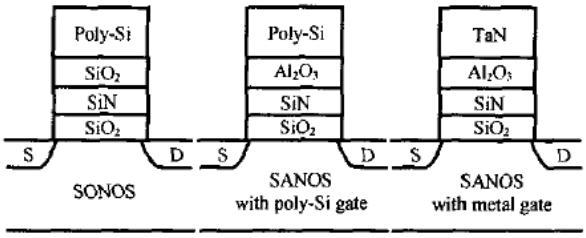


Figure 1.26: *Schematic representation of SONOS, SANOS and TANOS stacks.*

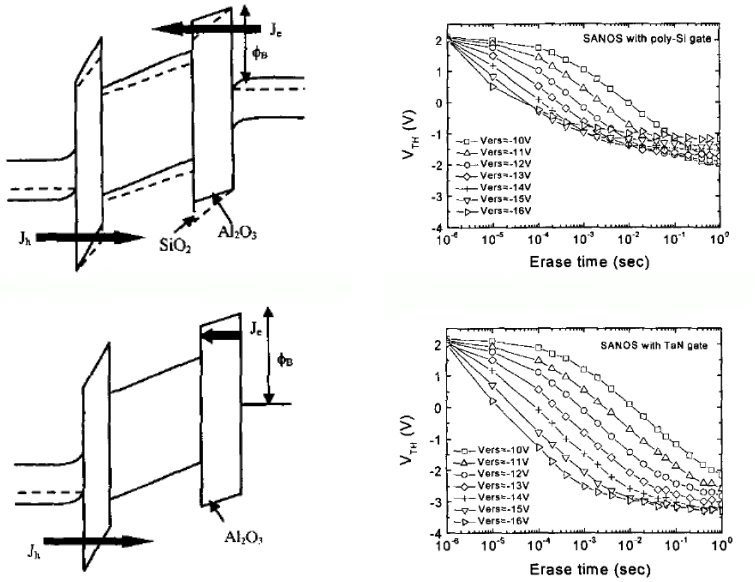
trap devices the tunnel oxide is scalable below 7nm. Indeed, electrons are trapped in non-communicating traps in the nitride layer: this assure the SILC immunity because, if some defect is present in the tunnel oxide, only the traps spatially and energetically near to the defect can be discharged. However, it is impossible to use thicknesses below 4nm if the requirement of 10years of data retention has to be meet, because *direct-tunneling* start to play an important role below 4nm [30]. On the other hand, a tunnel oxide thicker than 4nm make the program/erase operations slow: this problem can be solved using an high-k material (Al_2O_3) as top dielectric. Figs. 1.27(a) and 1.27(b) show the enhanced erasing performances of a TANOS stack compared to SONOS and SANOS stacks [30]. In these figures it is also shown the *erase saturation* phenomenon, appearing when the fraction of electrons extracted from the nitride equals the fraction of electrons injected from the gate during the erase operation. The adoption of an high work-function metal gate solves this problem. The erase saturation level decreases increasing the work-function of the metal (Fig. 1.28) [30]. Finally, Fig. 1.29 shows the increased performance for programming obtained with the employment of an high-k top oxide.

BE-SONOS. Another interesting solution proposed in literature to enhance the program/erase/retention performances of a charge-trap device is based on the so called *Band-Gap Engineering* (BE) [32]. In a BE-SONOS the tunnel oxide is substituted by an ultrathin ($\sim 5nm$) oxide-nitride-oxide tri-layer, as shown in Fig. 1.30.

The structure O1/N1/O2 represents a *crested tunnel barrier*: it suppresses the direct tunneling at low fields (e.g. retention conditions), but enhances an highly efficient electrons injection during program and and highly efficient holes injection during erase (Fig. 1.31(a) e Fig. 1.31(b)). In [32] it was demonstrated a wide threshold voltage window (>6V) enabling the employment of this device for multi-level applications. It should be noted that the N1 nitride layer does not contribute to the charge trapping since nitride layers thinner than 2nm do not exhibit trapping properties [32].

Moreover, the endurance tests and the retention tests show promising results for the BE devices (Fig. 1.32(a) and Fig. 1.32(b)).

3D-stacking approach. As the ITRS roadmap shows (Fig. 1.24), three-dimensional architectures appear today as the most viable solutions for the integration of non-volatile memory cells in Terabit arrays [33–38] (Fig. 1.33).



(a) *Energy band diagram during the erase operation for a SANOS (solid) and a SONOS (dashed) device (top) and for a TANOS device (bottom).*

(b) *Threshold voltage shift during erase operation for a SANOS (top) and a TANOS (bottom) device.*

Figure 1.27: [30]

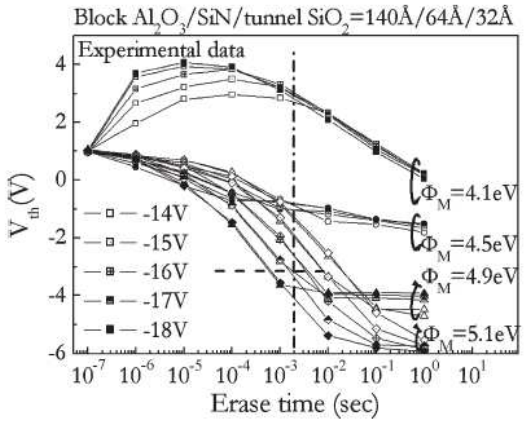


Figure 1.28: *The erase saturation level as a function of the metal gate work-function. Metals employed for the gate are: Al, TiAl, Pd and Au with work function 4.1, 4.5, 4.9, 5.1 respectively [31]*

In these architecture, several layers of memory cells, featuring a vertical channel and a gate-all-around geometry, are realized in the same wafer. Exploiting

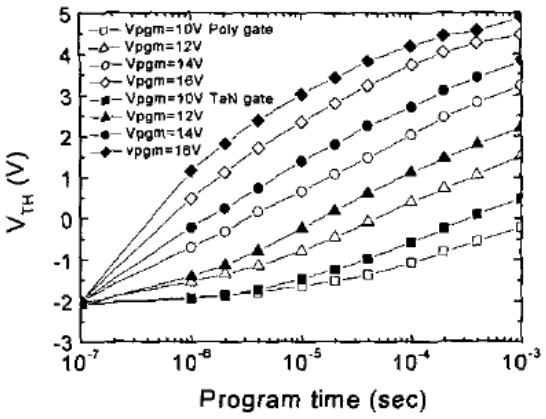


Figure 1.29: *Program performances enhancement for the TANOS stack when compared to the SANOS stack. [30]*

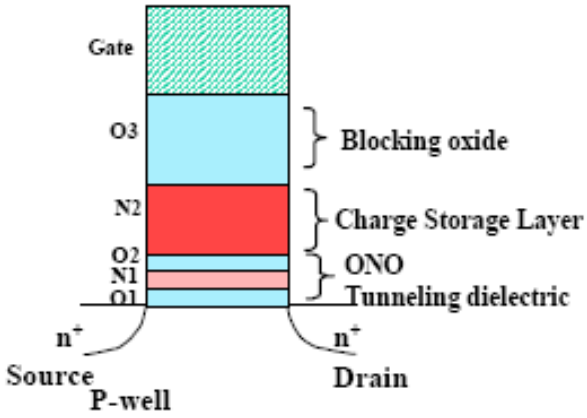
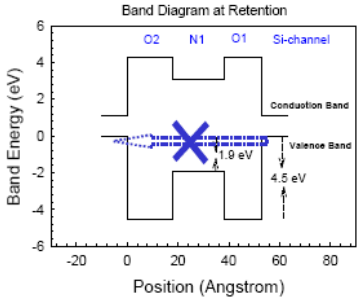


Figure 1.30: *BE-SONOS memory stack. [32]*

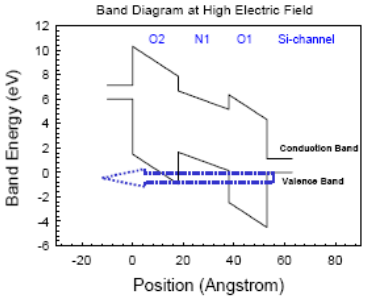
the vertical direction, the planar scaling constraints can be relaxed obtaining, nevertheless, a high-density integration.

The increasing emphasis on 3D NAND comes as the cost of advancing planar 2D NAND to the next technology node lithography in particular may be prohibitive (Fig. 1.4(a)). 3D NAND could come in with a 55nm half pitch, possible using a dry lithography tool-set. And the basic steps are fairly straightforward. Indeed, once several layers are deposited, just one lithographic/etching step is necessary to form a *vertical channel* common to several cells (Fig. 1.34(a)).

In particular, the gate-all-around (GAA) cell (Fig. 1.34(b)) with vertical channel is considered one of the most promising structures for future NAND Flash technologies, showing improved program/erase and retention performance

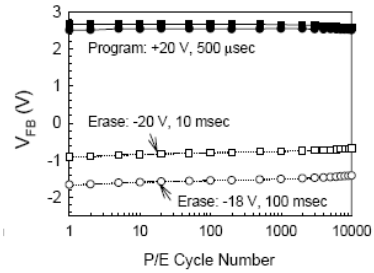


(a) *Energy band diagram of the crested barrier tunneling tri-layer employed in BE-SONOS memory stack during retention.*

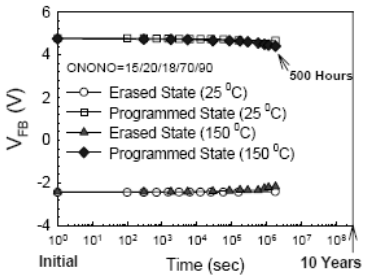


(b) *Energy band diagram of the crested barrier tunneling tri-layer employed in BE-SONOS memory stack during erase.*

Figure 1.31: [32]



(a) *BE-SONOS cells endurance performance.*



(b) *BE-SONOS cells retention performance.*

Figure 1.32: [32]

with respect to planar devices [39–42]. Moreover, thanks to the reduction of corner and fringing field effects during both program/erase and read, GAA-CT cells allow more uniform trapped charge distributions in the storage layer and provide, in turn, steeper incremental step pulse programming (ISPP) transients than planar cells [24, 25].

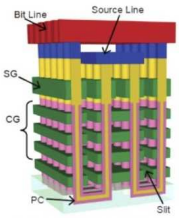
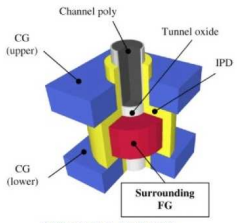
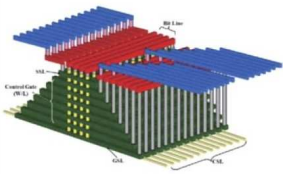
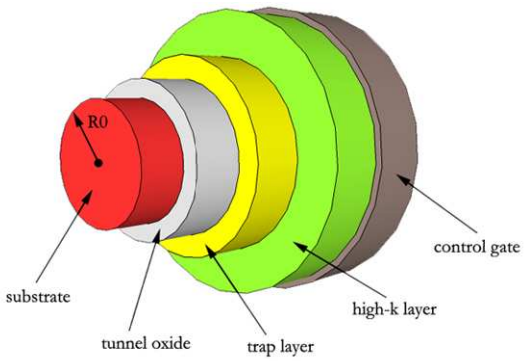
	Gate First		Gate Last
	Toshiba/P-BICS	Hynix DC-SF	Samsung/TCAT
Type of 3D NAND	 R. Katsumata et al, VLSI 2009	 S Whang et al, IEDM 2010	 J. Jang et al, VLSI 2009
Transistor	Gate all around; Salicided Poly Si gate	Gate all around; Salicided Poly Si gate	Gate all around; Damascene metal gate
Storage	Charge trap	Floating gate	Charge trap

Figure 1.33: *Schematic representation of different approaches to the 3D memory stacking. (Source: Applied Materials)*

(a) *TEM image of three vertical channels connecting several layers of memories in a 3D-stacked architecture.*



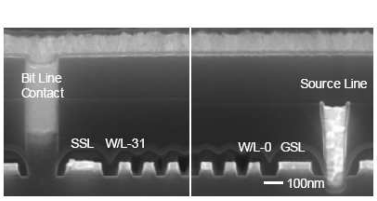
(b) *Schematic representation of a gate-all-around charge-trap memory cell.*

Figure 1.34:

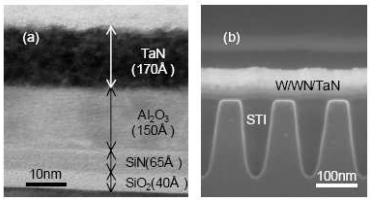
1.3.1 Technological feasibility

Several studies about the technological feasibility on real chips were reported during these years for the charge-trap memory devices. All these studies show that the charge-trap technology is feasible and controllable.

Fig. 1.35(a) and 1.35(b) show respectively a SEM and a TEM image of a section of a 4Gb charge-trap NAND array realized for the 63nm technology node (ArF lithography + phase change masks + *optical proximity correction*)(pattern 126nm on word-line e 130nm on bit-line) [43]. Fig. 1.36(a) shows the chip prototype, while Fig. 1.36(b) shows the threshold voltage distribution for programmed and erased cells.

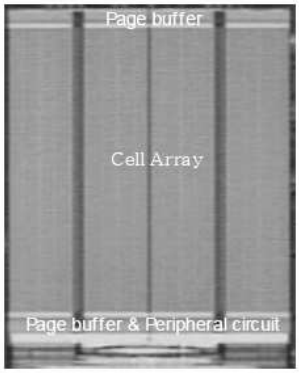


(a) *SEM cross-section of a TANOS NAND string, including ground and source selectors.*

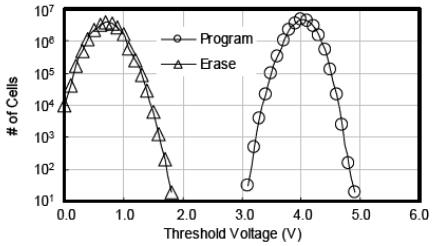


(b) *(a) TEM cross-section of TANOS stack; (b) SEM cross-section of TANOS cells, including STI.*

Figure 1.35: [43]



(a) *4Gb NAND TANOS prototype.*



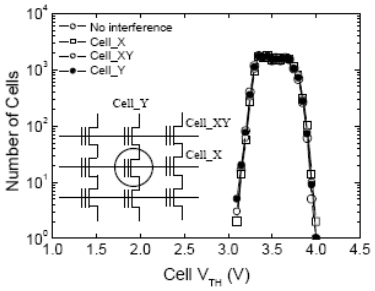
(b) *Threshold voltage distributions for erased and programmed cells of a 2Mb TANOS array.*

Figure 1.36: [43]

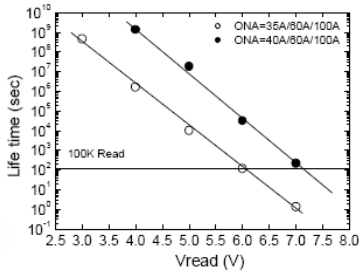
Multi-level memories. In order to have a multi-level memory device (i.e. a cell containing more than two logic states), a wide threshold voltage window and narrow threshold voltage distribution are required. Moreover a strict control of program/read/pass disturbs has to be applied.

Again for the 63 technology node, the feasibility of a multi-level TANOS NAND array was demonstrated in [29]. Fig. 1.37(a) shows the immunity of the

charge-trap cell to the capacitive interference during program/erase operation. Fig. 1.37(b) shows the retention performance for this multi-level device. Fig. 1.38(a) shows the typical two-phases multi-bit program operation. Finally, Fig. 1.38(b) shows the threshold voltage distribution on the different logic values: the width of each distribution is less than the separation between two logic values.

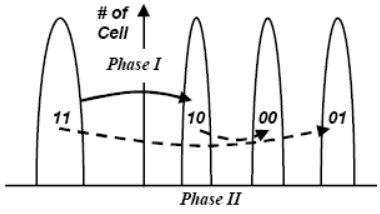


(a) *TANOS cell-to-cell interference immunity during program and erase.*

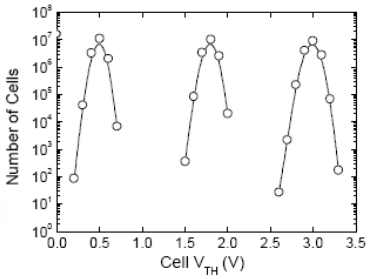


(b) *Retention performance for a TANOS device as a function of tunnel oxide thickenss.*

Figure 1.37: [29]



(a) *Typical two-phases multi-bit program operation.*



(b) *Threshold voltage distributions for a 64Mb multi-bit TANOS array.*

Figure 1.38: [29]

3D-stacking approach. The feasibility of a 3D-stacking approach has been demonstrate by several company and research groups [33–38]. As explicative example, Fig. 1.39 shows a spectacular cross-sectional SEM image of the 60nm 3D-stacked Toshiba P-BiCS memory (the first structure schematized in Fig. 1.33). This architecture realizes a 32Gb flash memory with outstanding performances [33].

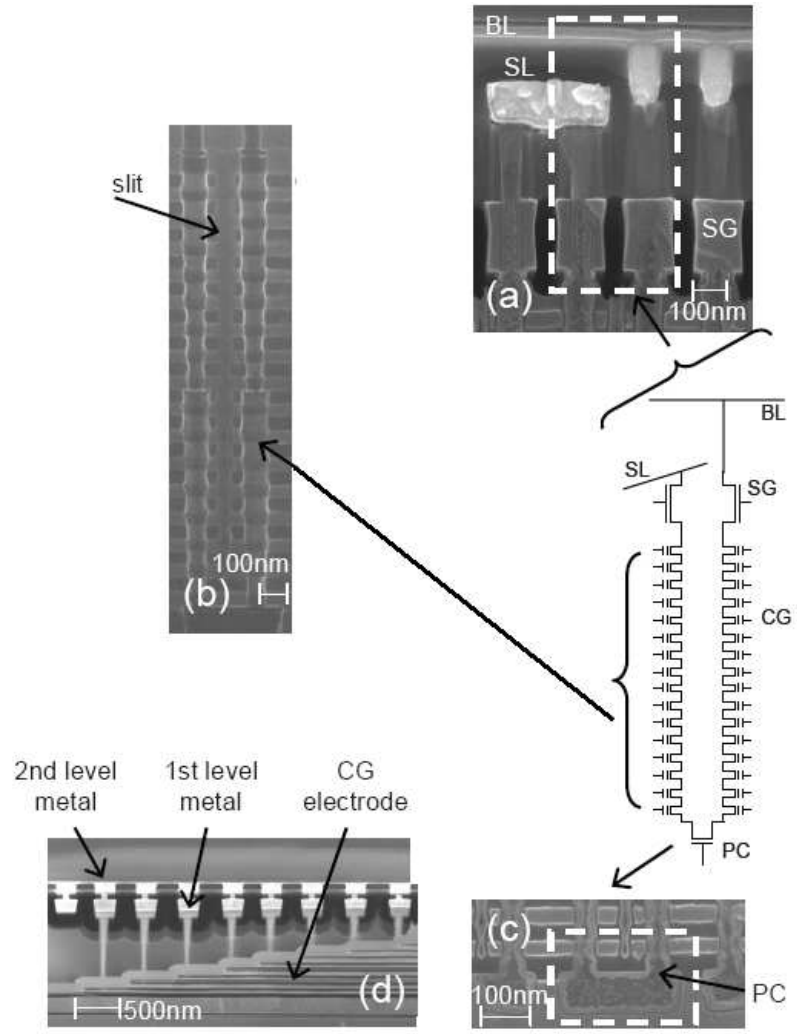


Figure 1.39: *Cross-sectional SEM image of 60nm 3D-stacked Toshiba P-BiCS flash memory and equivalent circuit. (a) Source line, bit line and select-gate (b) Memory hole after the removal of sacrificial film. (c) Pipeconnection. (d) Contact via for access to control-gate electrode. [33]*

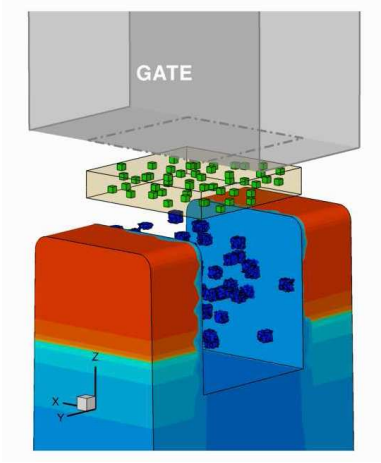


Figure 1.40: *Schematic SONOS cell structure adopted for the 3D TCAD simulation of the variability effects in nanoscale charge-trap memories.*

1.3.2 Statistical variability sources

Referring to the program and read operation of charge-trap planar device, three main source of *intrinsic* (i.e. related to the granular nature of matter and charge) variability can be identify:

- random dopant fluctuations, in number and position, in the MOSFET channel;
- random trap fluctuations, in number and position, in the storage layer;
- statistical charge injection process, during program, from the channel to the storage layer.

Fig. 1.40 shows a schematic SONOS cell with a random dopants distribution in the channel and a random traps distribution in the nitride.

Random dopant fluctuations. Random dopant fluctuations are caused by variation in the position and number of the dopant atoms within the channel which result in potential fluctuations which allow percolation paths to form as the device turns on [44–52]. Different devices have different microscopic doping distribution resulting in different conduction characteristics from device to device. This can be understood from Fig. 1.41 comparing the channel conduction simulation for the case of uniform doping and for the case of atomistic doping.

Using 3D simulation it has been shown that the random discrete dopants introduce fluctuations in the threshold voltage of a MOSFET and a lowering of the average threshold voltage of an ensemble of devices [49]. Moreover these effects become more prominent as the scaling process goes on. Indeed, the the overall number of discrete random dopants within the channel decreases due to the physical size reduction. Because of this reduction in the average number of

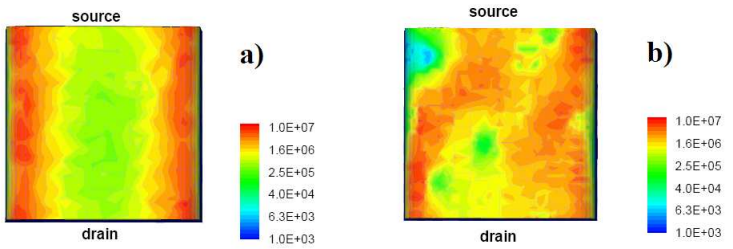


Figure 1.41: *Simulation results for the source-to-drain conduction in a SONOS cell at threshold condition in case of (a) uniformly doped substrate and (b) atomistic doped substrate.*

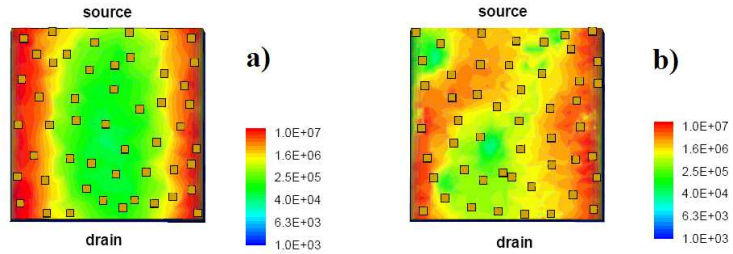


Figure 1.42: *Simulation results for the source-to-drain conduction in a SONOS cell at threshold condition in case of (a) uniformly doped substrate and (b) atomistic doped substrate, in presence of discrete charge trapped in the nitride. Traps located at different positions have a different impact on the threshold voltage shift.*

dopant atoms, the magnitude of the fluctuations increases. Numerical simulation has also shown that the doping concentration dependence of random dopant induced fluctuations is stronger than that obtained from analytical models [53]. This is due to the fact the the analytical models only take into account variations in the number of dopants and not the variation in their relative positions. This positional dependence is of great importance as it has been found that the dopants closest to the interface are responsible for a large fraction of the intrinsic fluctuation.

Random trap fluctuations. Random trap fluctuations are caused by variation in the position and number of the trap sites within the volume of the silicon nitride layer [54, 55]. We can understand that the percolative source-to-drain conduction, due to atomistic substrate doping, makes not only the fluctuation of the trap number in the cell but also of the trap position over the channel a major variability source for nanoscale cells. Indeed, traps located over a percolative conduction path will have a stronger impact on the device threshold voltage shift respect to traps located over a channel region with poor conductivity. This is sketched in Fig. 1.42.

Charge injection variability. In addition to the previous sources of variability, the program operation of charge-trap devices is also affected by the charge injection variability [21,22,56–58]. Being the granular injection a stochas-

tic process, the number of electrons injected from the channel to the nitride at a given time is statistically dispersed, resulting in a dispersion of the the devices threshold voltage. Indeed, let's consider that the *average* time required for the injection of the first electron into the nitride after the beginning of the program operation (time $t = 0$, tunneling current I_i) is τ_1, which is related to I_i by the relation $\tau_1 = q/I_i$. Then, the time required for the first electron injection (ΔT_1) is exponentially distributed (as confirmed by shot-noise measurements of the tunneling current in MOS capacitors [13]) as

$$P_{\Delta T_1} = \frac{1}{\tau_1} e^{-\frac{\Delta T_1}{\tau_1}} \tag{1.24}$$

where $P_{\Delta T_1}$ is the probability density function of ΔT_1. When the first electron is actually injected into the nitride, the tunneling current is reduced by a factor f, because the field across the tunnel oxide is reduced after the potential increasing in the nitride consequent to the electron trapping. Thus the *average* injection time for the second electron becomes $\tau_1 f$. Therefore, the time required for the second electron injection (ΔT_2) becomes exponentially distributed with a probability density function given by

$$P_{\Delta T_2} = \frac{1}{\tau_1 f} e^{-\frac{\Delta T_2}{\tau_1 f}} \tag{1.25}$$

and so forth for the subsequent injection times. The analytical modeling of the factor f has been reported reported in [22] for the floating-gate devices, but it represents a prohibitive problem for the charge-trap devices because the injection process is complicated by the presence of the atomistic doping, making non uniform the inversion charge in the channel, and by the presence of discrete traps in the nitride layer, making dependent on the trap position the electrostatic feedback f of a trapped electron on the subsequent injection event.

Random telegraph noise fluctuations. Random Telegraph Noise (RTN) is a statistical effect caused by traps at the Si/SiO_2 interface on MOS transistors operation. This noise originates from the alternate capture and emission of charge carriers by traps (defects) at the Si/SiO_2 interface, causing discrete drain current fluctuations. This component is a random signal that oscillates between discrete levels with a random period, naming the effect (Fig. 1.43). The traps can become filled with charge carriers which should be in the channel contributing to current conduction. Once the carrier is trapped it causes a change in channel current. Evidence for random telegraph noise (RTN) in MOSFET conduction due to single-electron trapping/detrapping events near the substrate/oxide interface has been reported since the mid-1980s [59]. The RTN amplitude has been shown not only to increase with the reduction of device dimensions but also to depend on bias conditions, trap position over the channel area, and substrate doping due to the nonuniform inversion caused by the atomistic nature of the dopants [60–62]. Recently, the miniaturization of cell dimensions following the development of sub-90-nm Flash technologies made the RTN instabilities clearly observable in the operation of memory arrays [63–65]. In this case, the modulation of cell conduction results into threshold-voltage fluctuations that could eventually affect those applications requiring very severe V_T

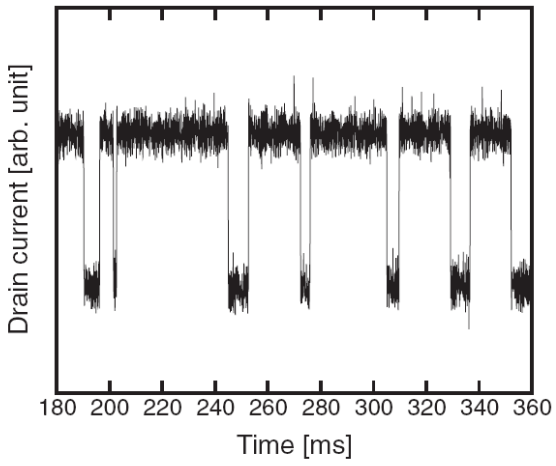


Figure 1.43: *Example of a two-level RTN waveform measured on a selected 65nm Flash cell. [66]*

control [63]. It should be noted that, in charge-trap memories, this instability could be also influenced by the discrete electron storage in the nitride.

What presented so far makes understandable that the use of 3D simulation is mandatory to account for the complex electrostatics and conduction profiles. For this reason in all the author's works presented in this manuscript a 3D simulation approach has been used for studying the several statical variability effects in nanoscale charge-trap memories.

Finally, it is worth to be pointed out that in all the works presented in the next chapters we do not take into account other secondary variability sources such as the line edge roughness, the poly-silicon gate granularity or the metal-gate granularity [51, 53, 67–72]. The impact of these sources on the threshold voltage variability of nanoscale MOSFET devices is, in fact, by far less than the impact of the random dopant variability [51].

1.4 Conclusions

This chapter presented the semiconductor non-volatile Flash memory. The floating-gate technology, representing the current dominant technology, was analyzed in the first part of the chapter. The NOR and NAND architectures were then introduced, showing their application in the *code storage* and *data storage* sector, respectively. We have shown the main scaling limitations afflicting especially the NAND architecture, namely the cell-to-cell capacitively coupling and the SILC. One of the most promising solution to overcome the floating-gate scaling issues is represented by the charge-trap technology, which store charge in electronic defects inside an insulating film. We have shown different kinds of charge-trap memories, including the 3D-stacking approach recently pro-

posed to achive the Terabit level of integration. Finally we have presented the main statistical variability sources affecting the electrical behavior of charge-trap memories: this is the major topic of this thesis and it will be developed in the next chapters.

Chapter 2

Resolving discrete charges in a TCAD framework

In order to study the intrinsic statistical variability of charge-trapping memory devices, charges (e.g. ionized dopants in channel or trapped electrons in the nitride) have to be treated as discrete entities. This represents a non-trivial issue in the framework of TCAD (Technology Computer Aided Design) simulation, where a drift-diffusion (DD) formalism is usually employed to study the carrier conduction. Indeed, the Coulomb potential associated with each discrete charge becomes physically inconsistent with the concepts of electrostatic potential presumed in DD device simulations. The first part of this chapter introduces the problems involved with the resolution of discrete charges in drift-diffusion simulation, showing the main solutions proposed up to date in literature. The second part of the chapter will show an improved study on this topic presented by the author at the 2011 SISPAD (Simulation of Semiconductor Processes and Devices) Conference.

2.1 Introduction

THE fact that a statistical configuration of discrete dopants could lead to a variability of the threshold voltage in MOSFET device was pointed out by Hoeneisen and Mead [73] several decades ago, and has now became a real problem in deca-nanometer MOSFETs [74]. It has been shown that, the threshold voltage fluctuations result from the variations of both dopants number and dopants arrangement in the device substrate [75].

The presence of discrete dopants give rise to complicated 3-dimensional (3-D) potential configurations, making compulsory the use of 3-D numerical drift-diffusion (DD) simulations for the study of the threshold voltage fluctuation

problem [49,52,76]. The key question is how to introduce the microscopic non-uniformity of localized dopant distributions inside the device to be simulated. The pioneering work of Nishinohara [44] first proposed to assign the dopant density at each mesh node in accordance with the number of dopants generated in each mesh region from a Poisson distribution. This approach has been extended to the extreme atomistic regime, where most mesh regions contain no dopant or, at most, one dopant, in order to represent the granularity nature of the doping in real deca-nanometers MOSFETS [45,47–50,77]. However, several works [51,52,78] have pointed out that such a naive extension of the conventional dopant model to the atomistic regime can be not consistent with the physics presumed in the DD simulation approach. In fact, Sano first highlighted this inconsistency [79,80], showing that subthreshold characteristics in sub-100 nm MOSFETs could be drastically changed when all dopants inside the entire device regions are treated as being atomistic. This implies that the naive atomistic approach may lead to erroneous results in threshold voltage evaluations.

The purpose of the first part of this chapter is to show in details the problems involved with the resolution of atomistic charges within the drift-diffusion framework, reviewing the three main solutions proposed in literature to relieve these issues: (1) the charge smearing method [45,49,50,76], (2) the Sano model [52] and (3) the quantum correction approach [78,81]. In the second part we will present an original extended study of the quantum correction approach, proposing a modified mobility model able to remove the residual artifacts of drift-diffusion atomistic simulation.

2.2 Implications of a discrete doping

In order to study the threshold voltage variations caused by random dopants in deca-nanometers devices, a physics based method to introduce the microscopic non-uniformity of localized dopant distributions inside the device has to be found. Conventional atomistic simulation, used in the first pioneering works on this topic [44], employed the following scheme: considering a completely random dopant arrangement inside the device (uniform continuous doping), the number of dopants included in a certain region should fluctuate in accordance with the Poisson distribution. The number of dopants in each mesh region is thereby determined from a Poisson distribution with mean given by the continuous (macroscopic) dopant density, and it is then translated into the local dopant density and assigned to the corresponding mesh node. In the case of large size devices, the number of dopants included in each mesh element exceeds unity and the dopant density has smooth variations between adjacent mesh nodes. On the other hand, in the case of deca-nanometers devices, most meshes contain no dopant or, at most, one dopant, and, thus, the dopant density changes abruptly from one mesh point to another, behaving like a δ-function (Fig. 2.1).

The electrostatics laws tell us that the electrostatic potential is a bare Coulomb potential well with a singularity at the dopant position, when the dopant density is given by the δ-function. Instead, no singularities are present in the potential solution when the dopant density is a smooth function as in the case of large devices. This is because the short-range variation of the Coulomb

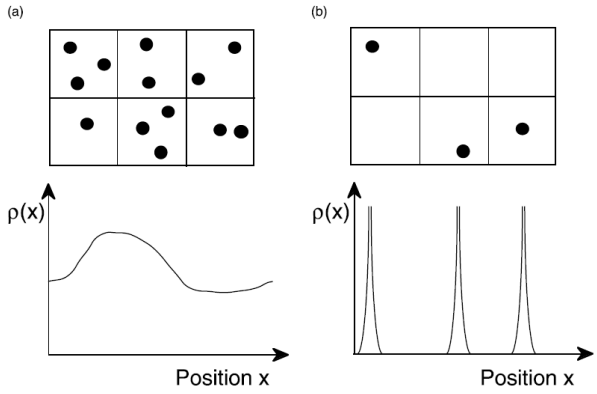


Figure 2.1: *Schematic representation of the dopant arrangements and the mesh configurations employed in the (a) classical and (b) atomistic DD simulations. The lower drawings sketch schematically the corresponding dopant densities as a function of position for the dopant arrangements shown above. [52]*

potential, whose wavelength is smaller than the mesh spacing, is implicitly eliminated since the mesh spacing Δ is always greater than the mean separation between the dopants $N_D^{-1/3}$, as shown in Fig. 2.1, where N_D is the continuous doping density. On the other hand, the mesh spacing Δ is always smaller than the mean separation of dopants in the atomistic cases, then both the short-range part and the long-range part of the Coulomb potential are explicitly included. As a result, the electric potential for an atomistic dopant is given by the full Coulomb potential. Fig. 2.2 shows schematically the electrostatic potential arising in the case of atomistic ionized acceptors.

It remains to understand which potential, whether the full Coulomb potential or the long-range part of the potential, is appropriate for classical DD device simulations. When the full potential is employed in the DD simulations, the majority carriers near the dopants are strongly localized by the sharply resolved attractive Coulomb well because the charge concentration follows exponentially (through the Boltzmann or Fermi-Dirac distribution) the electrostatic potential, obtained from the solution of the Poisson equation. Such charge "trapping" is physically impossible since, in quantum mechanical terms, the confinement in space keeps the ground electron energy state high in the well. This is shown in fig. 2.3 where is reported the 1D Poisson-Schrodinger solution for a Coulomb potential well corresponding to a charge plane and the bounded energy states [82]. It is clear that a large fraction of the full potential cannot be followed by the electron concentration due to such quantization.

This problem is obviously associated with the length-scale involved in the DD simulations. The current continuity equation in the DD approximation is derived under the continuous limit, taking into account only the first two moments of the Boltzamm transport equation. This means that this equation is valid over a length scale beyond l_c which is approximately equal to the mean distance of carriers and/or dopants. In other words, the spatial frequency of the

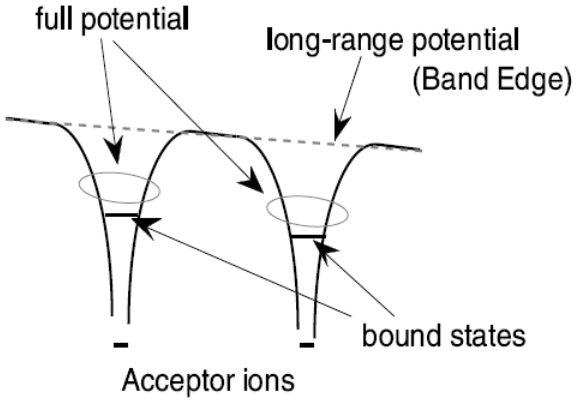


Figure 2.2: *Schematic drawing of the electric potential caused by ionized acceptors. The solid curves represent the full Coulomb potential and its long-range part. Notice that the spatially smoothed band edge is represented by the long-range part of the Coulomb potential. [52]*

electric potential has to be smaller than $1/l_c$.

It should be noted that the short-range part of the Coulomb potential of ionized dopants, that is the screened Coulomb potential, is responsible for the short-ranged impurity scattering. In the framework of the DD simulations, local quasi-equilibrium, which is attained through the scattering processes such as the phonon scattering and the impurity scattering, is assumed in the derivation of the current continuity equations. Therefore, the short-range part of the Coulomb potential is implicitly included in DD simulations, by means of the mobility models. It is important to point out that in the case of large devices simulation, this short-range part of the potential is explicitly eliminated in the Poisson equation because the mesh spacing Δ is usually greater than the mean distance of the dopants. Therefore, the double counting of the short-range part of the Coulomb potential is avoided.

In conclusion, only the long-range part of the Coulomb potential of ionized dopants can be consistent with the DD approach. Three are the methods proposed in literature to deal with the high spatial frequencies of the electrostatic potential (i.e. short-range components): (1) the charge smearing method [45, 49, 50, 76], (2) the Sano model [52] and (3) the quantum correction approach [78, 81]. We are going to analyze each of these methods in the next section.

2.3 Discrete dopant models

2.3.1 Charge smearing

The first and easiest method we present is the *charge smearing*. The aim of this approach is to smear the charge assignment over several mesh nodes, in order to

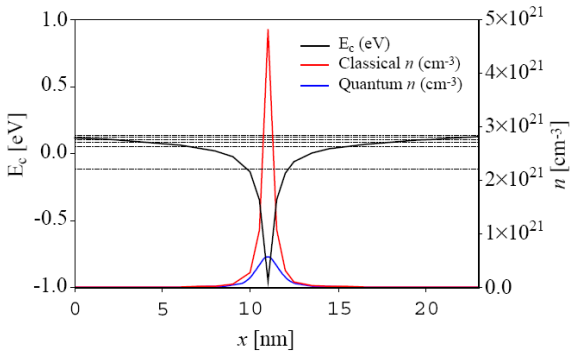


Figure 2.3: *1D Poisson-Schrodinger solution, showing the Coulomb potential well and bound energy states. It is shown the classical electron concentration and the quantum electron concentration associated with this potential well [82]*

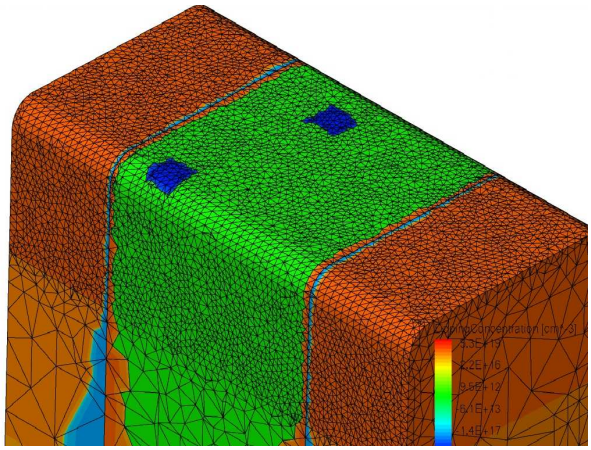
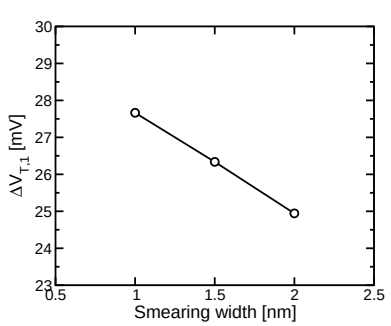


Figure 2.4: *3D mesh for MOSFET atomistic simulation, representing the charge smearing approach for the discrete dopants assignment.*

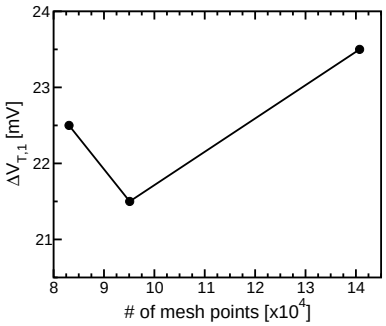
avoid the presence of a δ-function for the charge density (Fig. 2.4). Constant-smearing and Gaussian-smearing have been adopted in literature [49, 54, 76, 82]. Considering the Gaussian assignment [82], a discrete dopant is represented as a normalized Gaussian distribution, which has a total associated charge of one. The doping concentration at each mesh node is then calculated from this distribution as:

$$\rho(x) = exp\left(\frac{(x_i - x_m)^2}{\sigma^2}\right) \tag{2.1}$$

where $\rho(x)$ is the charge density at position x, x_i is the x coordinate of the dopant in real space, x_m is the x coordinate of the mesh node in real space, and



(a) *Threshold voltage shift given by one stored electron in a 25nm-SONOS memory cell as function of the smearing width.*



(b) *Threshold voltage shift given by one stored electron in a 25nm-SONOS memory cell as function of the mesh points.*

Figure 2.5:

σ is the distribution width.

The width of this distribution is usually set around 2nm [51,54] which corresponds to the effective diameter of a dopant using the Bohr model for hydrogen adapted to Silicon.

It should be noted that the effect of this charge assignment is somewhat similar to remove the Coulomb potential short range components. However, in this case the cut-off frequency is arbitrary and there is a possibility of double counting the screening effects since screening is also included in the Drift Diffusion equations. Another problem of the charge spreading approach is represented by the loss of resolution with respect to actual atomistic effects as they become averaged out as a result of the dopants spreading over such a wide area. This can be a huge issue especially for the simulation of very small devices. Moreover, simulation results can be mesh-size dependent [51]: indeed a fine mesh resolve better the potential well compared to a coarse mesh, resulting in a stronger charge trapping effect. Nevertheless, this method is widely used in literature [49, 54, 76, 82] because it is computational efficient and gives reliable results, in terms of threshold voltage shift, provided that the charge distribution width and the mesh resolution are accurately chosen. An example is showed in Fig. 2.5(a) and 2.5(b), where it is reported the threshold voltage shift given by one stored electron in a 25nm-SONOS memory cell: though the charge assignment method may have an impact of the threshold voltage value, 3D simulation results show negligible variations for the threshold voltage shift in a quite large range of smearing width and mesh size.

2.3.2 Sano's model

The Sano approach [52,79,80] is a refined variant of the charge smearing method. It essentially split, in a more formal and elegant way, the charge density of a discrete dopant between the long-range and short-range parts for properly taking into account of localized dopants in ultra-small devices. Considering

ionized acceptors as discrete dopants, the charge density $\rho_{ac}(r)$ is expressed as

$$\rho_{ac}(x) = -q \sum_{i=1}^{N} \delta(r - r_i) \tag{2.2}$$

where q is the magnitude of the electronic charge, N is the number of acceptors contained in the entire device regions, and r_i is the position of the ith acceptor. The core of an acceptor ion is ignored and it is treated as a point charge. It should be emphasized that this expression exactly corresponds to the acceptor density for the conventional atomistic dopant approach, where, in fact, the charge density at the jth mesh node is given by

$$\rho_j^{atm} = -q \begin{cases} 1/\nu_j & \text{if an acceptor is inside the jth mesh element} \\ 0 & \text{otherwise} \end{cases}$$

where ν_j is the volume of the jth mesh element. As the mesh-size shrinks, $1/v_j$ increases reaching the δ-function limit. This explain better why the physical quantities such as the dopant density and electric potential depend on the size of the mesh employed in conventional atomistic simulations. The smaller the mesh is, the deeper the electric potential becomes. As a result, the amount of trapped carriers strongly depend upon the mesh-size employed in DD simulations. This is of course impermissible in any numerical simulations and another indication that atomistic DD simulations are inappropriate.

Rewriting the δ-function in Eq. 2.2 in terms of the wave vector k, the charge density can be easily separated into the long-range and short-range parts

$$\begin{aligned} \rho_{ac}(x) &= -q \sum_{i=1}^{N} \frac{1}{V} \sum_{k<k_c} e^{ik(r-r_i)} - q \sum_{i=1}^{N} \frac{1}{V} \sum_{k>k_c} e^{ik(r-r_i)} \\ &= -q \sum_{i=1}^{N} n_{ac}^{long}(r - r_i) - q \sum_{i=1}^{N} n_{ac}^{short}(r - r_i) \end{aligned} \tag{2.3}$$

where V is the volume of the device and assumed to be very large compared with the characteristic length scale. The first (second) term in the second line represents the long (short) range part of the acceptor charge density and n_{ac} is the corresponding number density. The cut-off spatial frequency k_c is the parameter that determines the split between the long-range and short-range components. So far this approach is very similar to the charge smearing approach. What makes the Sano's approach more refined than the charge smearing approach is that the cut-off parameter is chosen in a physics-based manner. Physically, the cut-off parameter should be related to the screening length: indeed more the doping density increases more the charge acceptors are screened by the mobile charge. This is the well known concept of *Debye length* or the *Thomas−Fermi screening length*, sometimes also referred as *Brooks−Herring screening model*. However Sano [52] pointed out that this screening model is inadequate for our aim because it does not reproduce the correct mobility in

high-doped regions and this kind of screening concept breaks down in the depletion regions where majority free carrier density becomes very small so that the screening length becomes extremely large. For these reasons Sano suggested to use the *Conwell– Weisskopf model* [83], in which the screening is mainly due to the overlap of electrostatic potentials of ionized dopants. In this case, the magnitude of the cut-off parameter k_c is then given by the inverse of the mean separation of dopants:

$$k_c = 2N_{ac}^{1/3} \tag{2.4}$$

where N_{ac} is the macroscopic acceptor density.

From Eq. 2.3, the long-range part of the number density $n^{long}(r)$ of a single acceptor located at the origin ($r = 0$) is given by

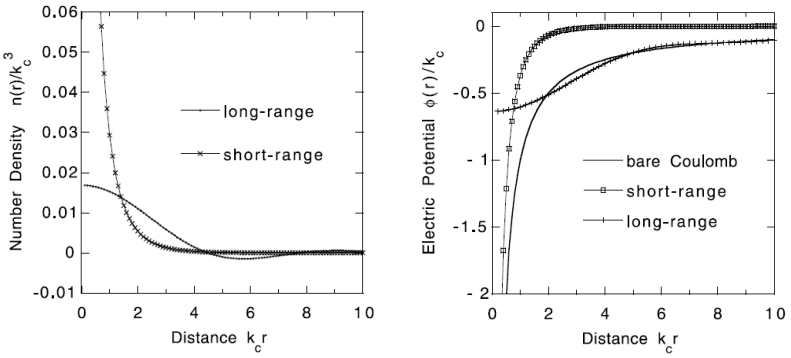
$$n_{ac}^{long}(r) = \frac{1}{V}\sum_{k<k_c} e^{ikr} \rightarrow \frac{k_c^3}{2\pi^2}\frac{\sin(k_c r) - (k_c r)\cos(k_c r)}{k_c r}^3 \tag{2.5}$$

where the coefficient in the last expression is normalized such that the integral of the number density $n_{ac}^{long}(r)$ over the entire space becomes unity.

$$\int n_{ac}^{long}(r) d^3 r = 1 \tag{2.6}$$

Since the short-range part of the charge density is eliminated in the Poisson equation of the DD simulations, the number density of each acceptor has to be re-normalized so that the total acceptor charge in the substrate is conserved. It is very important to note that Eq. 2.5 is oscillatory at large r, then it gives no contribution when the integral region is sufficiently large. This means that the exact expression of the long-range part of the charge density is not essential. The fact that the charge density is broadened is of most importance. *This observation makes the Sano approach really close, even identical from a practical point of view, to the charge-smearing approach.* As well as the charge smearing approach, Sano's method's weakness is represented by the loss of resolution with respect to actual atomistic effects as they become averaged out as a result of the dopants spreading. Another problem of this method is that, as also pointed out by Sano [52], the cut-off parameter k_c is never calculated from Eq. 2.4 in practical simulations, but usually it is obtained as a fitting parameter. This results in a sort of arbitrariness on the choice of k_c and makes vain the elegance that this method owns respect to the charge smearing method.

Fig. 2.6(a) shows the long-range part (with the normalization coefficient properly modified) and the short-range part of the number density due to a single acceptor located at the origin. Notice that the singularity of the number density at the origin is properly removed. The long-range part of the number density spreads over the inverse of the cut-off parameter $1/k_c$, whose magnitude corresponds to the mean separation of the acceptors. This is a sharp contrast to the conventional atomistic dopant approach which shows a δ-function-like behavior. From Eq. 2.5 the long-range part of the electronic potential $\phi_{ac}^{long}(r)$ by a single acceptor with unit charge is readily obtained



(a) *Long-range and short-range parts of the number density caused by a single dopant located at the origin.*

(b) *Long-range and short-range parts of the electronic potential caused by a single dopant with unit charge located at the origin. The bare Coulomb potential is also plotted with the solid line.*

Figure 2.6: [52]

$$\phi_{ac}^{long}(r) = \frac{-2}{\pi r}\int_0^{k_c r}\frac{\sin(t)}{t}dt = \frac{-2k_c}{\pi}\frac{Si(k_c r)}{k_c r} \tag{2.7}$$

where $Si(r)$ is the sine integral. The short range part of the potential is the difference between the bare Coulomb potential and the long-range part given by Eq. 2.7, and this is well approximated with the screened potential

$$\phi_{ac}^{short}(r) \approx \frac{-e^{-k_c r}}{r} \tag{2.8}$$

Fig. 2.6(b) shows the long-range and short-range parts of the electronic potential by a single acceptor with unit charge located at the origin. The long-range part of the potential changes very smoothly and does not diverge at the origin, then preventing the charge trapping at the dopant position. It is interesting to note that the hollow around the origin becomes deeper as the macroscopic acceptor density increases (k_c decreases). This potential hollow corresponds to the *band-edge fluctuations* with which the carriers are locally accelerated or decelerated during their drift motion [84].

2.3.3 Quantum corrected model

As we have previously seen, the charge "trapping" phenomenon affecting conventional atomistic DD simulations is physically impossible since, in quantum mechanical terms, the confinement in space keeps the ground electron energy state high in the well (fig. 2.3). Therefore one should expect that the introduction of quantum mechanical correction in DD simulations solves the charge "trapping" problem. Indeed this is the idea followed by Asenov and his group [51, 78, 81], that have applied quantum corrections to the conventional atomistic simulations

by means of the Density Gradient (DG) approximation. If these quantum corrections are applied to the region that contains discrete random dopants a marked reduction in the charge localization at impurity sites is actually observed. In the case of a sharply resolved potential well electrons are unable to follow the real electrostatic potential because their concentration is determined by the effective quantum potential which introduces a force pushing the electrons out of the well. The DG formalism introduces the most important (first-order) quantum effects are required for the purpose of practical Computer-Aided Design of future generations of devices, preserving a good computational efficiency.

The DG formalism resembles the Bohm interpretation of quantum mechanics [85, 86], adding into the classical equation of state for electrons a quantum correction term ψ_{qm}, which is proportional to the second-derivative of the square root of the carrier density:

$$\psi_{qm} = 2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = \phi_n - \psi + \frac{k_B T}{q} \ln\left(\frac{n}{n_i}\right) \tag{2.9}$$

where $b_n = \hbar^2/4qm_n^* r$, with r being a dimensionless parameter.

This approximation can be derived from the Wigner transport equation [87]:

$$\frac{\partial f(p,r,t)}{\partial t} + v \cdot \nabla_r f(p,r,t) - \frac{2}{\hbar} V(r) \sin\left[\frac{\hbar \overleftarrow{\nabla}_r \overrightarrow{\nabla}_p}{2}\right] f(p,r,t) = 0 \tag{2.10}$$

The quantum effects are included through the non-local driving potential in the third term on the left-hand side, noting that ∇_r acts only on V and ∇_p acts only on the distribution function f . The operator within the sin function may be written in terms of a Taylor series, so that the transport equation for the Wigner distribution function can be written in the form of a modified Boltzmann Transport Equation as:

$$\frac{\partial f}{\partial t} + v \cdot \nabla_r f - \frac{1}{\hbar} \nabla_r V \cdot \nabla_k f + \sum_{\alpha}^{\infty} \frac{\hbar^{2\alpha-1}(-1)^{\alpha+1}}{4^{\alpha}(2\alpha+1)!} \left(\nabla_r V \cdot \nabla_k f\right)^{2\alpha+1} = \left(\frac{\partial f}{\partial t}\right)_{coll} \tag{2.11}$$

where V represents the electrostatic potential. Expanding to first order in $\hbar$, so that only the first non-local quantum term is considered, has been shown to be sufficiently accurate to model non-equilibrium quantum transport [88].

The introduction of the DG correction in the DD system of equations can be realized using a modified Gummel iteration scheme, where the Poisson equation (2.12) and then the Density Gradient equation (2.13) are solved for the electrostatic potential and the quantum-corrected electron density respectively:

$$\nabla \cdot (\epsilon \nabla \psi) = -q\left(p - n + N_D - N_A\right) \tag{2.12}$$

$$2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = \phi_n - \psi + \frac{k_B T}{q} \ln\left(\frac{n}{n_i}\right) \tag{2.13}$$

In this way, the *effective quantum potential* can be computed as:

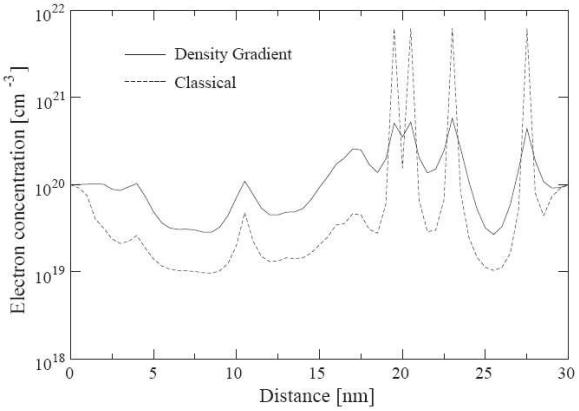


Figure 2.7: *Electron concentration along the x-direction of a 30x20x20nm resistor, showing both the classical electron concentration and the quantum electron concentration generated by the inclusion of DG quantum corrections. [82]*

$$\psi_{eff} = \psi + 2b_n \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} = \phi_n + \frac{k_B T}{q} \ln\left(\frac{n}{n_i}\right) \tag{2.14}$$

This effective potential is used as the driving force for the electron current continuity equation:

$$\nabla \cdot J_n = R \tag{2.15}$$

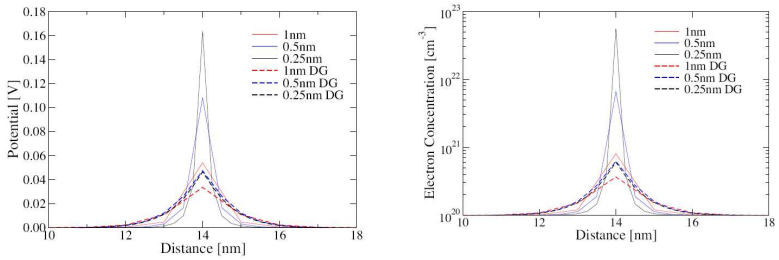
with R the net recombination-generation rate, and J_n is given by

$$J_n = -qn\mu_n \nabla \psi_{eff} + qD_n \nabla n \tag{2.16}$$

Analogous corresponding equations hold for holes.

The impact of the inclusion of DG quantum effects in simulations by is illustrated in Figure 2.7. Quantum confinement in the sharp potential wells smoothes the overall electron concentration by reducing the sharper peaks eliminating the charge localization. It can also be seen that this effect raises the concentration in the surrounding area allowing more charge to participate to the conduction.

Mesh sensitivity. In conventional atomistic simulations the amount of charge that becomes localized is strongly dependant on the mesh size used in the simulation. Indeed the doping charge density at a given mesh point is proportional to $1/\nu$, where ν is volume of the mesh element surrounding that node. Therefore a reduction of the mesh size causes a subsequent increase of the charge density which, in combination with the better resolved potential well, turns in a greater charge localization. As shown in [51], the quantum approach is much less sensitive to the mesh size. Figure 2.8(a) illustrates the mesh size dependence of the conventional electrostatic potential and the effective quantum potential associated with the DG formalism obtained from the mesh-based



(a) *Electrostatic potential and the effective quantum potential around a single point charge for mesh sizes of 1 nm, 0.5 nm and 0.25 nm*

(b) *Classical and quantum electron concentrations around a single point charge for mesh sizes of 1 nm, 0.5 nm and 0.25 nm*

Figure 2.8: [82]

solution of the Poisson's equation around a single discrete donor showing the reduced mesh sensitivity of the DG approach respect to the classical approach. It is interesting to note that the effective potential in the middle of the Coulomb well is approximately 40 mV, corresponding roughly to the energy level of the electron ground state, in good agreement with the ground state of a hydrogenic model of an impurity in Si.

Figure 2.8(b) shows the corresponding electron concentration associated with the single point charge. In the classical case the electron concentration follows the solution of the Poisson equation which in turn is strongly coupled to the mesh size, while the inclusion of quantum corrections makes the solution mesh independent below 0.5nm mesh spacing value.

DG correction for holes. The reported discussion about quantum correction has been focalized on electron carriers. Considering a conventional MOSFET device, the electron trapping phenomenon occurs only in the n-doped source and drain regions. The charge localization in this regions gives an increase in the MOSFET access resistance, reducing the MOSFET I_{ON} current. However for the accurate atomistic simulation of a MOSFET it is also important to consider the trapping of holes in the channel region when discrete acceptors are considered. This trapping leads to a change in both the size and shape of the depletion region. This in turn alters the characteristic of the simulated device by reducing, for example, the threshold voltage. This is particularly important for the case of memory devices simulations. Fig. 2.9 show the impact of DG corrections on both electrons and holes on the trans-characteristic of 50nm-MOSFET [82].

DG vs. Sano model. The strength of DG approach lies in the low mesh sensitivity of simulation results. Another very important feature of the DG approach is its high resolution. To explain the latter point we schematically compare in Fig. 2.10(a) the DG method and the Sano's method. It is evident that Sano's approach, similar to charge smearing approach from a practical point of view, loose resolution when two dopants are close enough to have an overlap of their long-range distribution functions. On the other hand, DG approach

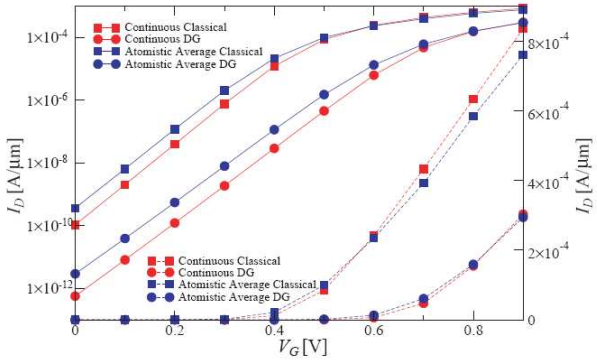
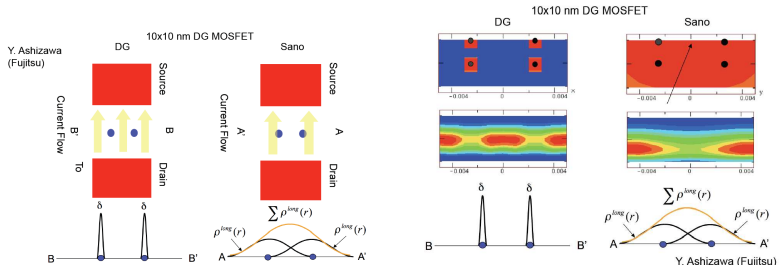


Figure 2.9: *Trans-characteristics of a 50x50 nm transistor for both continuous doping and atomistic doping, with and without the inclusion of DG quantum corrections. [82]*



(a) *Schematic representation of DG and Sano methods doping distribution functions in case of two close discrete dopants in a double gate 10nm MOSFET.*

(b) *Simulation results for the electron conduction between source and drain obtained with DG and Sano approach in a double gate 10nm MOSFET.*

Figure 2.10: [89]

can maintain the highest level of resolution because it can deal with δ-shape doping distribution function, being the quantum corrections (and not the charge smearing) to relieve the charge trapping problem. This lack of resolution in the Sano's method can lead to wrong results, as shown in Fig. 2.10(b): while the DG approach gives three conduction maxima between source and drain, the Sano's approach gives only two maxima due to the overlap of the long-range doping distribution functions.

2.4 A modified mobility model for atomistic simulation

As we have seen in the previous section, resolving individual charges within classical drift diffusion simulation using a fine mesh is problematic. We have also seen that the most promising solution is to use a quantum mechanically

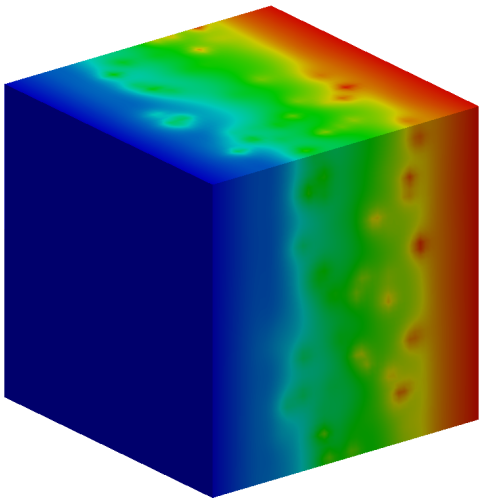


Figure 2.11: *3D plot of the effective potential in a 50x50x50nm silicon resistor doped at* 10^{19} cm^{-3}. *Note that the region with the discrete doping is long 30nm.*

consistent treatment of the electron concentration via the DG approximation. Though DG approach removes the mesh-sensitivity problem, Roy *et. al* [51,82] have pointed out that a residual non-negligible error remains in the simulation of the electron transport due to a residual trapping. In order to understand the nature of this residual trapping error we have extended the study of Roy *et. al* [51,82], proposing a modified mobility model to remove the residual error. In this section we report the main results of this study.

To study the charge trapping phenomena associated with atomistic simulation, a silicon resistor with n-type doping and dimensions 50x50x50 nm was chosen as test structure and an ensemble of a thousand microscopically different devices were simulated for each doping density and each mesh spacing using the GSS[1] atomistic simulator GARAND. A Cloud-in-Cell (CIC) [90] method is adopted for the charge assignment in the simulations, combined with quantum DG corrections. The continuously doped case is taken as the reference as it matches the analytical conductivity of the resistor. The electron mobility in atomistic simulations is calculated, via the Masetti model [91], using the continuous doping profile. Hence, for each doping density, all atomistic devices use the same calculated mobility. Additionally, when specified, an electric field dependent mobility is also adopted in simulations. In this case the mobility depends on the local electric field generated by each discrete dopant. It should be noted (Fig. 1) that the atomistic region length is 30 nm. A continuously doped region is interposed between the contacts and the discrete zone to avoid any influence related to boundary conditions. The potential distribution for one

[1]Gold Standard Simulations Ltd, Rankine Building, Oakfield Avenue, Glasgow, G12 8LT, UK

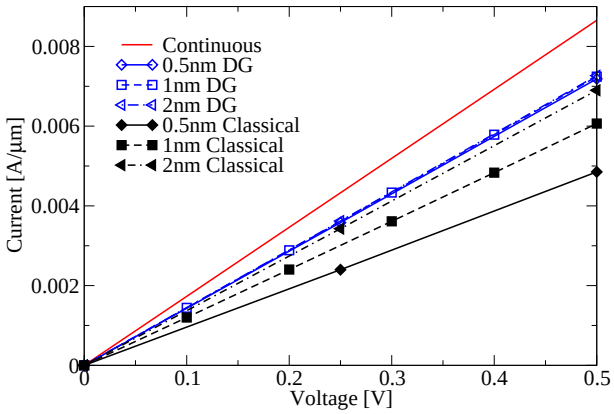


Figure 2.12: *Comparison of the IV characteristics produced from classical and DG simulation on ensembles of 1000 atomistically different devices with mesh spacings of 0.5, 1 and 2nm.*

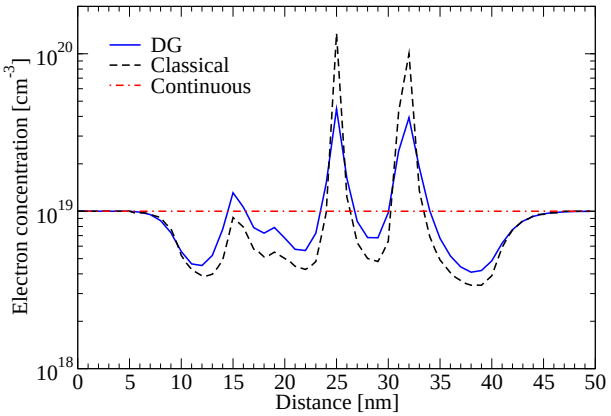


Figure 2.13: *1D plot of the electron concentration through a 50x50x50nm silicon resistor doped at* 10^{19} cm^{-3}*, comparing the profile obtained from classical ad DG simulations.*

of the simulations is shown in Fig. 2.11 for a doping value of $N_D = 10^{19} cm^{-3}$.

Fig. 2.12 depicts the average I-V characteristics obtained from classical and DG simulations of the resistor, for a doping value of $N_D = 10^{19} cm^{-3}$ and a mesh spacing of 0.5nm, 1 nm and 2nm. As can be seen from Fig. 2.12 the resistance of the device increases dramatically in the atomistic classical cases, showing a strong mesh dependence. DG results are mesh-independent and closer to the analytical result, but a discrepancy remains between the atomistic average resistance and the analytical resistance. This difference is related to the electron concentration profile through the device (Fig. 2.13). It should be noted that the integral of charge density over the device volume is the same for both the average atomistic resistor and the continuously doped resistor. Fig. 2.13 shows that the

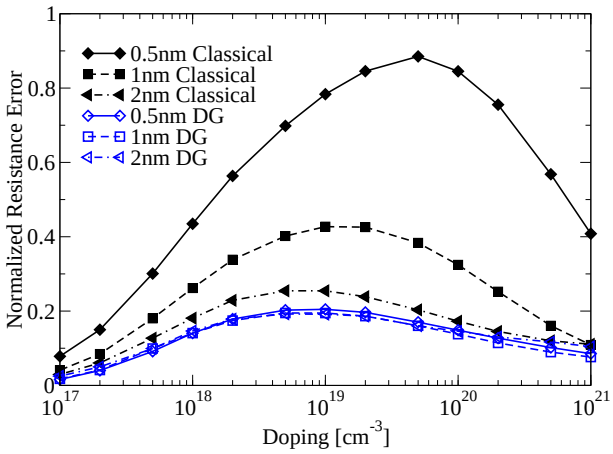


Figure 2.14: *The relative error of the resistance calculated in a 50x50x50nm silicon resistor using both classical and DG formalisms for various mesh spacing and doping densities.*

DG approach smoothes the electron density profile with respect to the classical approach, reducing the effect of charge trapping. Fig. 2.14 shows the relative error between the average atomistic resistance and the analytical resistance for different doping densities and different mesh spacing values for both classical and DG simulations. As before it can be seen that classical simulations have a larger error than DG simulations. It can also be seen that all simulations exhibit a similar shape. Starting at low doping densities the error increases until it reaches a maximum of around $10^{19} cm^{-3}$ and then decreases for higher doping densities.

This shape can be explained if we first consider high doping densities. Fig. 2.15 shows that the height of the potential peaks causing electron trapping decreases when the doping concentration increases. This can be due to the reduction of the Debye length (λ_D), causing an increase in the electrostatic screening, or the reduction of the average distance between dopants (d_{avg}), causing a smoothing of the interactive potential wells. The inset in Fig. 2.15 shows that λ_D decreases more quickly than d_{avg} when the doping concentration increases, suggesting that the electrostatic screening is the major factor responsible for the reduction in peak height. This is confirmed in Fig. 2.16 where the 1D electrostatic potential along a resistor is shown when a single discrete dopant is placed at the center of the device with varying continuous background doping densities (D_{bg}). Our previous analysis implies that the error in the resistance should increase as the doping density decreases. However Fig. 2.14 shows that the error in the resistance decreases at low doping densities. In order to understand this behavior the electron conduction profile is studied in more detail. Contrary to the situation in a MOSFET channel, the discrete dopants in a resistor are attractive potentials for electrons. As a result, the current density maxima are found corresponding to the dopants positions. This is clearly shown in Fig. 2.17(a) which shows a plot of the electron current density

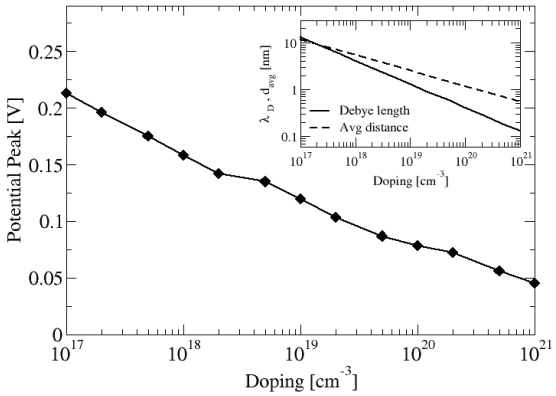


Figure 2.15: *Potential peaks height in 3D resistor simulation as a function of the average doping density. The Inset shows the trend of the Debye length (λ_D) and of the average distance between dopants (d_{avg}) with the doping density.*

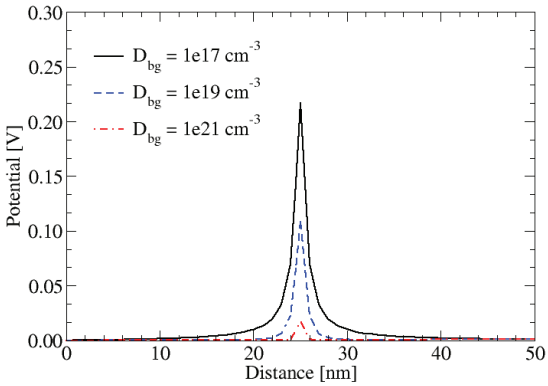


Figure 2.16: *1D electrostatic potential in the case of one discrete dopant at the center of the resistor with different continuous background doping densities.*

taken through the center of a resistor with a single discrete dopant placed at the center. However, if we integrate the current density inside the dashed circle (region B, representing the region where the potential peak of Fig. 2.17(b) has the largest impact) then we obtain a current one order of magnitude smaller than the current flowing outside the dashed circle (region A). This implies that, the carrier current still follows a percolative path between the discrete dopants, when few dopants are present in the device. This current flow between dopant atoms becomes more important as the number of dopants in the device becomes

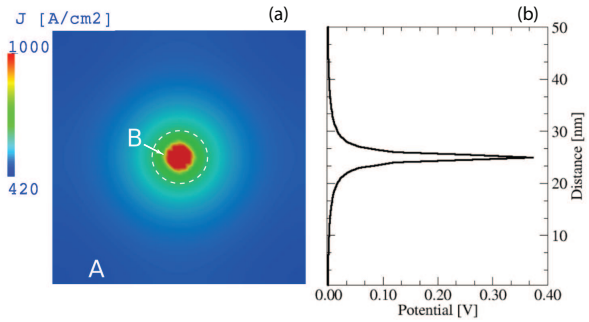


Figure 2.17: *(a) Current density in the median plane between source and drain in the case of one discrete dopant placed at the center of the resistor, and (b) corresponding 1D electrostatic potential.*

small. For this reason the error curve of Fig. 2.14 start to decrease towards low doping densities. Even though the error is reduced at low dopant numbers there is a secondary effect associated with the position of dopant atoms. Fig. 2.18(a) shows the normalized error in the resistance for different numbers of dopants extracted from the different doping densities. It can be seen that even if the dopant numbers are the same, at low numbers there is a spread in the error coming directly from the positional effect. This spread in error values is large at low doping densities and strongly decreases at high doping densities. Indeed, in the presence of percolative conduction the relative position of dopants has a strong impact in determining the resistance of atomistic devices. This is also evident in Fig. 2.18(b) where the slope of the normal probability plot of the resistance error increases with the number of discrete dopants. Moreover it should be noted that the error in the resistance (Fig. 2.14) tends to zero when the number of dopants is very small, regardless of whether DG or a classical approach is used. This again confirms that, in presence of very few dopants, the current mainly flows outside the sphere of influence of the potential wells.

The previous results have been obtained using only the Masetti [91] mobility model. If field dependent mobility models such as Caughey-Thomas [92] or Lombardi [93] are included we obtain the results shown in Fig. 2.19. Here it can be observed that there is an increase in the error due to the electrostatic potential wells associated with each discrete dopant in atomistic simulations. However, this increase in the resistance error is negligible at high doping densities, due to the Debye screening, and remains small at low doping densities, due to the percolative regime.

The previous analysis suggests that (i) in attempting to eliminate the error from the simulations the effects of the electric field can be ignored and (ii) the error in the resistance can be removed by modifying the mobility in atomistic simulations according to:

$$\mu_{atomisitc}\left(N_D\right) = \left[1 + err\left(N_D\right)\right] \mu_{continuous}\left(N_D\right) \tag{2.17}$$

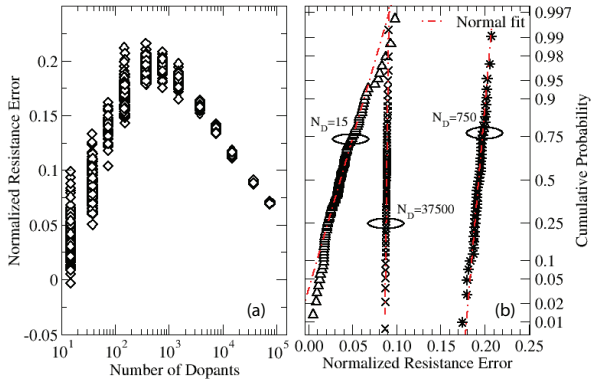


Figure 2.18: *(a) Resistance error for each atomistic device for fixed values of number of dopants in the resistor, and (b) corresponding normal probability plot for three values of number of dopants.*

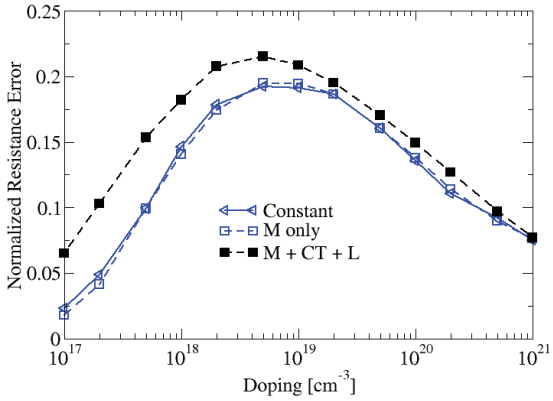


Figure 2.19: *Resistance error curve for different mobility models: Constant, Masetti [91] (M), Caughey-Thomas [92] (CT) and Lombardi [93] (L).*

where $\mu_{continuous}(N_D)$ is the mobility used in simulations of continuously doped resistors and $err(N_D)$ is the DG resistance error of Fig. 2.14. It is clear that $err(N_D)$ is defined only for a finite number of discrete values so an interpolation strategy is required, especially when this correction is applied to the simulation of MOSFETs with non-uniform doping profiles. A better strategy is to adopt an analytical expression for the corrected atomistic mobility.

Fig. 2.20 shows the electron mobility that should be adopted in the atomistic simulations to match the resistance of the continuously doped devices, compar-

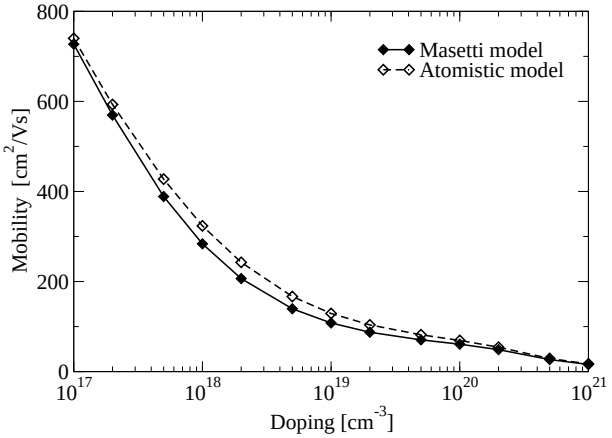


Figure 2.20: *Electron mobility used to remove the resistance relative error (dashed line) compared to the Masetti model mobility (solid line) used for the continuously doped devices.*

Parameter	Continuous	Atomistic
μ_0	1417	1429
μ_{min1}	52.2	52.2
μ_{min2}	52.2	52.2
μ_1	43.4	45
P_c	0	0
C_r	9.68×10^{16}	9.81×10^{16}
C_s	3.43×10^{20}	3.48×10^{20}
α	0.68	0.61
β	2	2

Table 2.1: *Masetti model parameters values extracted to obtain a match, within 0.5% of error, between the continuously doped device resistance and the average atomistic device resistance.*

ing it with the conventional Masetti model mobility [6]. It is important to note that the atomistic mobility only slightly differs from the Masetti model. This suggest that an analytical expression for the atomistic mobility model can be found by modifying the parameters of the Masetti model, without introducing any other complex relationship. Table 2.1 reports the conventional and the modified values for the Masetti model parameters obtained to have an error in the resistance of less then 0.5% over the whole range of doping densities.

Fig. 2.21 shows the scatter plot of ensembles of 1000 atomistically doped silicon resistors, with currents normalized to the continuously doped case for doping densities of $1 \times 10^{17} cm^{-3}$, $1 \times 10^{19} cm^{-3}$ and $1 \times 10^{21} cm^{-3}$. It is evident that the mobility correction remove the resistance error without introducing distortions in the original statistical distribution.

Finally, Fig. 2.22 points out that the percolative nature of the conduction at low doping densities makes the error dependent on the aspect ratio of the

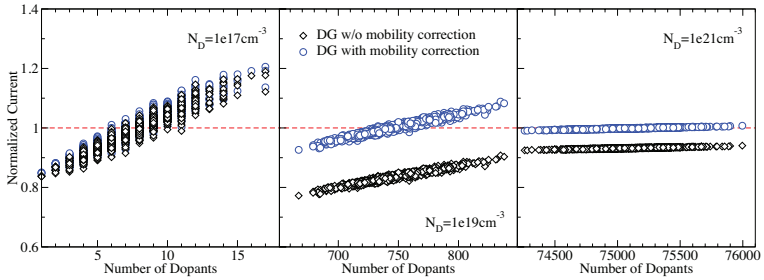


Figure 2.21: *Scatter plot of ensembles of 1000 atomistically doped 50x50x50nm silicon resistors, with currents normalized to the analytical average case for doping densities of* 10^{17} cm^{-3}, 10^{19} cm^{-3} *and* 10^{21} cm^{-3}.

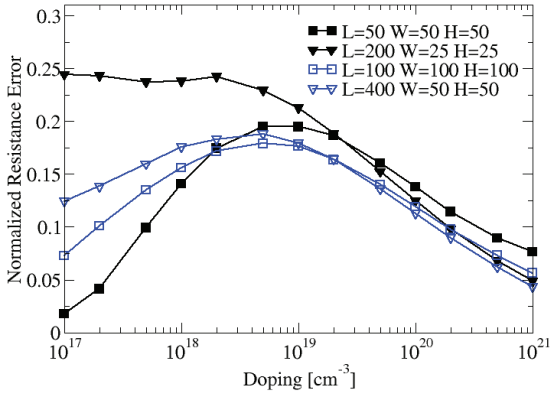


Figure 2.22: *Resistance error curve for a 50x50x50nm (filled symbols) and a 100x100x100nm (open symbols) resistors with two different aspect ratio.*

device. As expected, a shrinking in the resistor cross-section (maintaining a constant volume) results in an increased error at low doping. This effect becomes more prominent as the volume of the resistor is reduced. However it affects only the low doping region, that is usually irrelevant in the simulation of realistic MOSFETs. Moreover Fig. 2.22 shows that the error curve remains almost unchanged when increased to nearly ten times the volume of the resistor (50x50x50 nm^3 to 100x100x100 nm^3). This confirms that the charge trapping is an artifact of atomistic simulations and that the analytical resistance represents a good choice of reference with which to compute the resistance error.

2.5 Conclusions

This chapter presented the physics behind the atomistic dopant model widely used for the study of statistical threshold voltage variations with DD simulators. It has been shown that the conventional dopant model, when extended to the extreme atomistic regime, becomes physically inconsistent with the concepts presumed in drift-diffusion simulations. It has been pointed out that cutting off the short-range part of the Coulomb potential associated with each discrete dopant is essential in correctly simulating the device properties under the atomistic regime. To this aim, three dopant models, namely (i) the smearing method, (ii) the Sano model and (iii) the quantum correction, have been introduced showing their range of validity and their limitations. Finally, a mobility model correction for atomistic simulations was proposed in order to improve the accuracy and reliability of the quantum correction method.

Chapter 3

ΔV_T variability in charge-trap memories

This chapter presents a comprehensive investigation of statistical effects in deeply-scaled charge-trap memory cells, considering both atomistic substrate doping and the discrete and localized nature of stored charge in the nitride layer. By means of 3-D TCAD simulations, the statistical dispersion of the threshold voltage shift induced by a single localized electron in the nitride is evaluated in the first part of the chapter. The role of 3-D electrostatics and atomistic doping on the results is highlighted, showing the latter as the major spread source. The threshold voltage shift induced by more than one electron in the nitride is then analyzed, showing that a correlation among single-electron shifts clearly appears. The scaling trends are discussed in the second part of the chapter, showing that for fixed density of trapped charge, the average threshold voltage shift decreases with the cell scaling as a consequence of fringing fields. Moreover, the distribution statistical dispersion increases with technology scaling due to a more sensitive percolative substrate conduction. The impact of the discrete electron storage in the nitride on random telegraph noise instabilities is also investigated, showing that despite single cell behavior may be modified, negligible effects results at the statistical level.

3.1 Introduction

SONOS and TANOS memories are considered today the most practical evolution of the floating-gate (FG) NAND Flash technology, allowing improved reliability thanks to discrete charge storage in thin silicon nitride layers [29, 94–98]. In order to investigate their performance, many 1-D models

have been reported to describe the charging/discharging dynamics of relatively large area cells and capacitors [26, 99–102]. However, all these models suffer from two main limitations that question their validity for the investigation of deca-nanometer memory cells, i.e., the lack (1) of the real 3-D cell electrostatics during program/erase (P/E) and read conditions and (2) of the discrete and localized nature of stored electrons. Considering 3-D electrostatics is mandatory to account for fringing fields at the active area edges, determining both a non-uniform charge injection to/from the nitride layer during P/E and non-uniform substrate inversion during read [25, 103]. In this context, the correct evaluation of the threshold voltage shift (ΔV_T) determined by the discrete and localized electrons stored in the nitride after program requires a careful numerical simulation of source/drain conduction during read. Moreover, the discrete nature of the stored charge necessarily gives rise to statistical issues related to the number and position fluctuation of the electrons in the nitride, determining a statistical dispersion of ΔV_T after program which cannot be investigated by any 1-D model. This statistical dispersion is further worsened when considering the additional contribution of atomistic doping to non-uniform substrate inversion, enhancing percolative source-to-drain conduction [51–53, 104, 105].

3.2 Physics-based Modeling

In this section we present a comprehensive 3-D numerical investigation of deca-nanometer nitride memory cells, considering both atomistic substrate doping and discrete and localized electron storage in the nitride. In order to correctly capture the stored charge effect, not only on substrate inversion but also on source-to-drain conduction in presence of non-uniform substrate inversion, ΔV_T is evaluated from cell drain current–gate voltage ($I_D - V_G$) transcharacteristics, obtained solving the transport equations in the active area. By means of Monte Carlo simulations, the statistical distribution of the threshold voltage shift induced by a single localized electron randomly placed in the nitride volume ($\Delta V_{T,1}$) is evaluated accounting for dopant number and position randomness in the substrate. The role of 3-D electrostatics and atomistic doping on the $\Delta V_{T,1}$ distribution is highlighted, showing the latter as the major spread source. Then, the threshold voltage shift induced by N electrons stored in the nitride ($\Delta V_{T,N}$) is analyzed, showing that for large N a correlation among single-electron shifts clearly appears, reducing the spread of the $\Delta V_{T,N}$ distribution.

3.2.1 Numerical model implementation

We performed 3-D TCAD simulations on the template device structure reported in Fig. 5.1, featuring: STI trenches at the cell sides, atomistic doping in the substrate and discrete electrons in the nitride layer. A bottom oxide/nitride/top oxide (ONO) stack with thicknesses equal to 4/4.5/5 nm was assumed for the gate dielectric, with the nitride layer patterned over cell channel ($\epsilon_{ox} = 3.9$ and $\epsilon_N = 7.5$ were used for the relative dielectric constants of silicon oxide and nitride, respectively). Cell width (W) and length (L) were set to 25 nm, with a

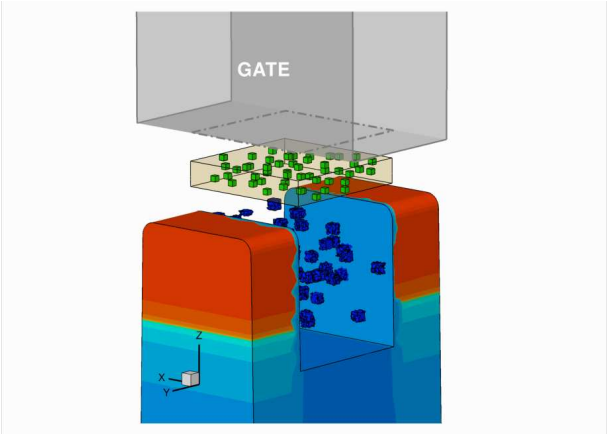


Figure 3.1: Schematics for the cell structure investigated in this work, highlighting the atomistic doping region in the substrate and the discrete localized electrons stored in the nitride. The gate dielectric comprises an oxide/nitride/oxide (ONO) stack. Red regions: n implants; blue regions: p substrate; green region: nitride; oxide regions are not highlighted.

constant substrate doping $N_a = 3 \times 10^{18}$ cm^{-3}. A planar metal word-line was assumed for the gate [103].

The simulation flow is schematically depicted in Fig. 3.2: after the definition of the template device structure, the atomistic doping region in the substrate was set with a depth from the substrate/oxide interface (25 nm) larger than the average depletion layer width at threshold. Outside this region, a continuous doping profile was used. A Monte Carlo loop was employed to collect statistical information on the cells. The loop includes the extraction of the actual number of dopants in the discretization region from a Poisson statistics with average value $N_a^d = 46$ (as resulting from the product of N_a and the volume of doping discretization) and their placement according to a uniform distribution. Poisson and drift-diffusion equations were then solved for fixed drain bias $V_D = 0.7$ V (source and bulk grounded) and increasing gate bias to obtain the $I_D - V_G$ transcharacteristics. This was then used to extract neutral cell threshold voltage ($V_{T,0}$) as the gate bias allowing 200 nA to flow from source to drain. $I_D - V_G$ and V_T were then calculated again after a fixed number N of electrons were randomly placed in the nitride volume to investigate the programmed cell state, extracting $\Delta V_{T,N}$. More than 100 Monte Carlo runs were used to obtain the $V_{T,0}$ and $\Delta V_{T,N}$ statistics. Note that the maximum number of Monte Carlo runs has to be considered when defining the thickness of the atomistic doping region, in order for the cell with the lower number of dopants out of the Poisson statistics to have a depletion layer completely included in this region.

All the simulations were performed by means of a commercial software [106], implementing in its framework the tools to deal with atomistic substrate doping and discrete electron placement in the nitride. To this aim, we followed a simulation approach similar to what reported in [107], using a constant mobility value for the drift-diffusion simulations [51, 107] and spreading the dopant

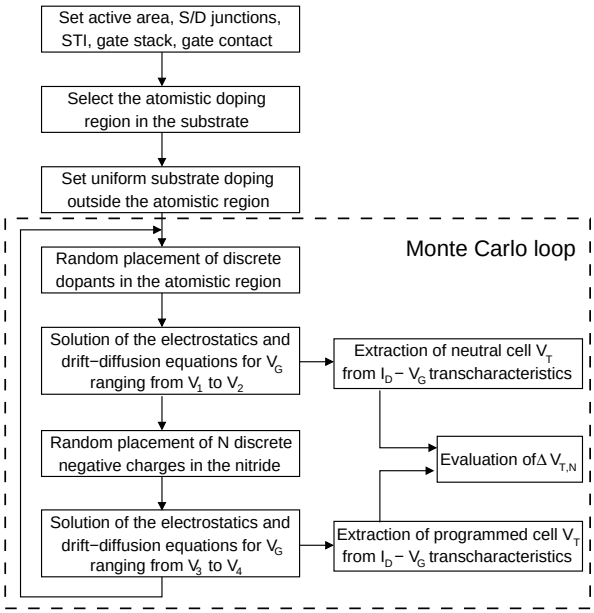


Figure 3.2: Block diagram for the simulation procedure used to collect statistical information on $V_{T,0}$ and $\Delta V_{T,N}$. $V_{G,1}$, $V_{G,2}$, $V_{G,3}$ and $V_{G,4}$ define the explored V_G ranges for the neutral and programmed cell state.

charge in a cube of side equal to 2 nm centered in the chosen atomistic dopant position. The cube side was selected from the compromise between a better resolution of percolative substrate conduction and the necessity to avoid artificial charge localizations when solving the poisson and drift-diffusion equations in presence of coulomb potential wells, as discussed in detail in the previous chapter. Similarly, localized electron storage in the nitride was reproduced by placing the electronic charge in a cube of 1 nm side centered in each selected storage position.

3.2.2 One electron ΔV_T statistical distribution

Fig. 3.3 shows the statistical distribution of $V_{T,0}$ obtained from the Monte Carlo simulation analysis previously presented. Although current crowding at the cell corners determined by 3-D electrostatics impacts the average value $\overline{V_{T,0}} = 2$ V of the distribution, its broadening is the result of atomistic doping [53,105]. In fact, substrate percolative conduction is determined by the number and position of dopants placed in the discretization region, with different cells showing different $V_{T,0}$ as a result of a more or less favorable configuration of doping atoms from the source-to-drain conduction standpoint. Note that the resulting $V_{T,0}$ statistics in Fig. 3.3 displays a good gaussian behavior, with a standard deviation $\sigma =$ 195 mV.

We begun our investigation of the programmed cell state by considering a sin-

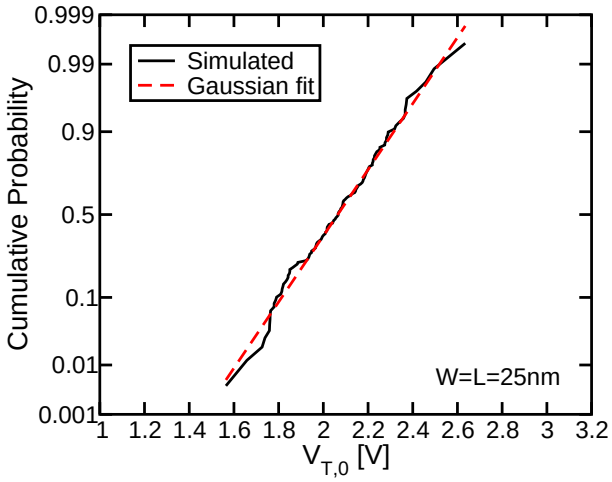


Figure 3.3: Simulated statistical distribution of V_T for neutral (no charge in the nitride) cells.

gle trapped electron (*i.e.*, $N = 1$) randomly placed in a localized position inside the nitride volume. The statistical distribution of the V_T shift obtained with respect to the neutral cell state ($\Delta V_{T,1}$) is shown in Fig. 3.4 (solid line), displaying an average value of $\overline{\Delta V_{T,1}} = 22$ mV and a standard deviation $\sigma_{\Delta V_{T,1}} = 8.5$ mV. Moreover, a clear exponential tail departs in the positive $\Delta V_{T,1}$ direction, meaning that single electrons can result into very large V_T shifts, though with low probabilities. Note that this behavior cannot be predicted by any 1-D model. In fact, in a 1-D treatment, the only spread source for $\Delta V_{T,1}$ is the vertical position of the stored electron in the nitride layer. This affects the V_T shift according to $\Delta V_{T,1}^{1D} = -q'(t_2/\epsilon_{ox} + t_x/\epsilon_N)$, where $q' = q/WL$ is the electron charge normalized to cell area and t_2 is the top oxide thickness. Assuming a random distance t_x of the electron from the nitride/top oxide interface, the resulting 1-D distribution of $\Delta V_{T,1}$ is reported in Fig. 3.4 (dots). The distribution has been calculated including a scaling factor equal to 2.09 to the $\Delta V_{T,1}^{1D}$ values in order to obtain the same average result given by 3-D simulations, as will be explain in the next section. Note that this distribution is much tighter than what predicted by the accurate 3-D analysis, revealing a dominant contribution of 3-D electrostatics and random dopant effects on the $\Delta V_{T,1}$ statistics.

In order to separate the contribution of atomistic doping and 3-D electrostatics to the statistical dispersion of $\Delta V_{T,1}$, Fig. 3.4 also shows results obtained for continuous substrate doping (dashed-dotted line). This distribution highlights the effect of a single, localized electron in the nitride on source-to-drain conduction, correctly taking into account 3-D electrostatics but neglecting the contribution of atomistic doping to non-uniform substrate inversion. The lower statistical dispersion of $\Delta V_{T,1}$ with respect to the case when atomistic doping is accounted for reveals that percolative source-to-drain conduction has a major role in determining the impact of a single trapped electron on V_T, and that

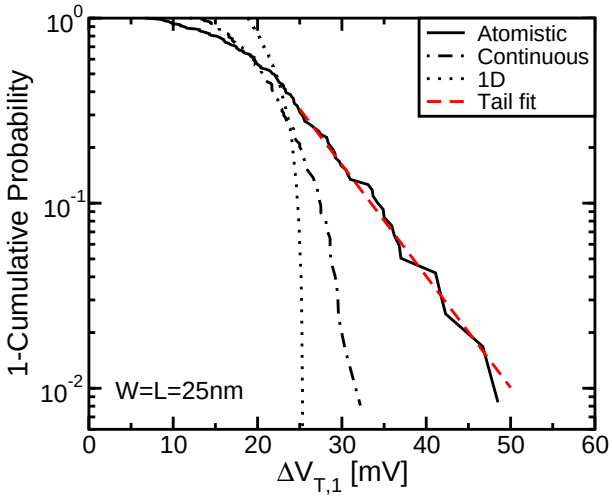


Figure 3.4: Simulated statistical distribution of the threshold voltage shift determined by one localized electron randomly placed in the nitride volume.

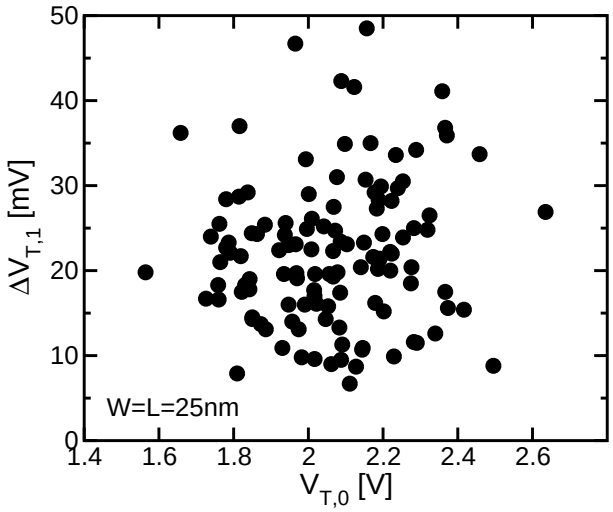


Figure 3.5: Simulated $\Delta V_{T,1}$ as a function of cell $V_{T,0}$, considering 3-D electrostatics and atomistic substrate doping.

this is strongly affected by the discrete nature of substrate dopants. However, Fig. 3.5 makes clear that there is no correlation between $V_{T,0}$ and $\Delta V_{T,1}$.

To better understand the results of Fig. 3.4, we reported the $\Delta V_{T,1}$ values as a function of the electron position along L, W and nitride thickness t_N in Figs. 3.6-3.8, for the case of continuous (left) and atomistic (right) substrate doping. Results of Figs. 3.6-3.7 are similar to what was already obtained for the

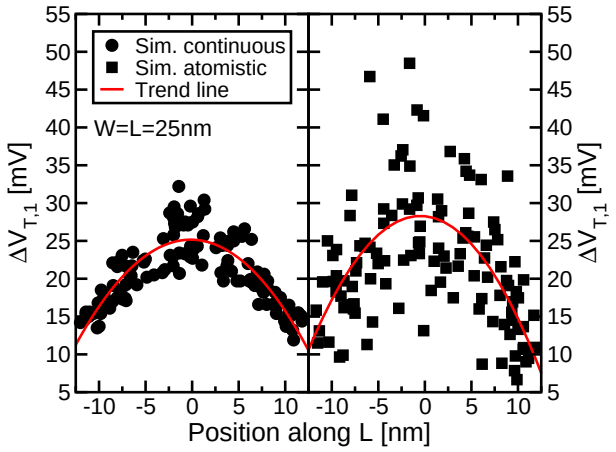


Figure 3.6: $\Delta V_{T,1}$ as a function of the electron position along L for the case of continuous (left) and atomistic (right) substrate doping. 0 is the channel center.

case of random telegraph noise (RTN) in Flash devices [76, 105]: the V_T shift is larger when the electron is placed half-way between source and drain, due to a more effective electrostatic control of channel inversion, or close to the edges along W, due to the possibility for the electron to stop a larger part of drain current, as field intensification at the edges locally increases the inversion charge and the drain current density. Fig. 3.8 then reveals the average increase of $\Delta V_{T,1}$ when the electron is moved closer to the nitride/bottom oxide interface, as also predicted by 1-D electrostatics.

Superimposed on the average trend is the statistical dispersion of the results, due to fluctuations in the electron position (*e.g.*, in Fig. 3.6 the spread is due to fluctuations along W and t_N). In particular, in Fig. 3.8 the $\Delta V_{T,1}$ spread reduces when the electron is placed closer to the top oxide. This is ascribed to a less-local electrostatic effect of the stored electron on the substrate when the distance between electron and substrate grows, reducing the impact of non-uniformities in the current density profile. In the case of atomistic doping, a larger dispersion of $\Delta V_{T,1}$ can be seen from all the Figs. 3.6-3.8, that was already evident in Fig. 3.4. For our simulation set, this increased spread totally overrides the weak average W dependence of $\Delta V_{T,1}$ that was shown by the continuous doping results (see the different trend lines). The origin of this additional spread is obviously the enhancement of percolative conduction determined by atomistic doping, resulting into a statistical dispersion of $\Delta V_{T,1}$ that is larger than that given by 3-D electrostatics alone. In fact, atomistic doping enhances the possibility to have cells where source-to-drain current takes place along few strong percolation paths, allowing a single electron to largely block cell conduction when this is exactly placed over a path where current crowding occurs [105, 108–110]. In this case, a quite large V_T shift results, contributing to the enlargement of the $\Delta V_{T,1}$ statistics.

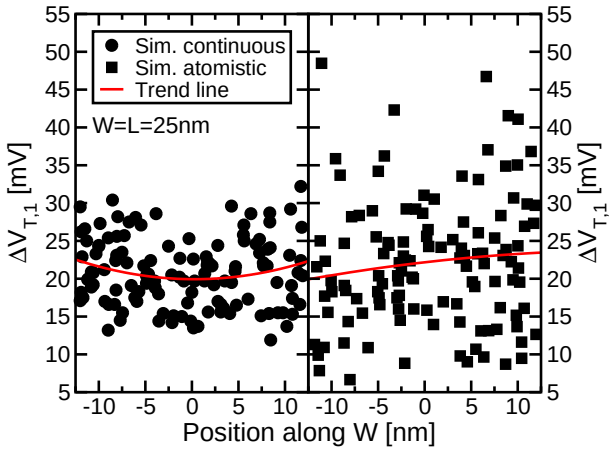


Figure 3.7: Same as Fig. 3.6 but as a function of the electron position along W. 0 is the channel center.

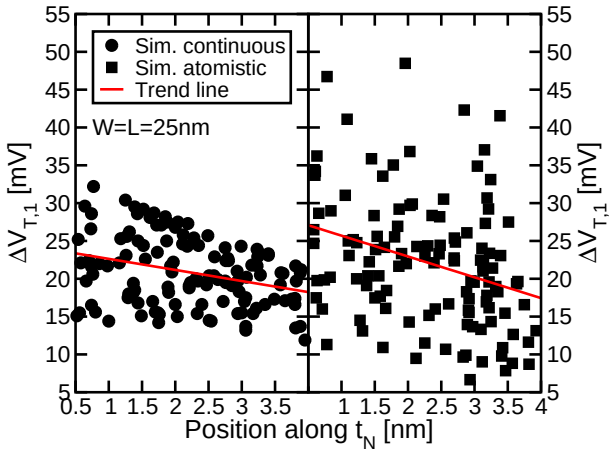


Figure 3.8: Same as Fig. 3.6 but as a function of the electron position along t_N. 0 is the nitride/bottom oxide interface.

3.2.3 Many electrons ΔV_T statistical distribution

To investigate how multiple electrons combine their effect to determine the V_T shift, we ran Monte Carlo simulations with $N > 1$. Electrons were placed randomly in the nitride volume according to a uniform distribution, therefore neglecting any disuniformity in the stored charge profile after program that may arise from a non-uniform injection field in the bottom oxide. Fig. 3.3.1 shows the statistical distributions of $\Delta V_{T,N}$ for the case of $N = 14$ and 50. A larger statistical dispersion of $\Delta V_{T,N}$ appears with respect to the case of $N = 1$, contradicting the 1-D prediction of a spread reduction due to a more stable charge

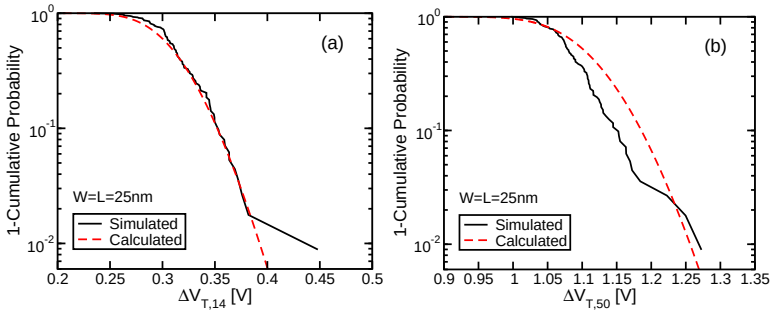


Figure 3.9: Simulated $\Delta V_{T,N}$ distribution in the case of $N = 14$ (a) and 50 (b) electrons stored in the nitride layer (black lines). Red dashed-lines represent the calculated distributions assuming independent trap superposition and the $\Delta V_{T,1}$ distribution of Fig. 3.4.

centroid in presence of a larger number of electrons in the nitride. To further investigate this point, Fig. 3.3.1 also shows the $\Delta V_{T,N}$ distribution calculated from $N-1$ self-convolutions of the $\Delta V_{T,1}$ statistics (red dashed lines). This distribution is expected to describe the $\Delta V_{T,N}$ statistics in the case electrons add their individual contribution to the V_T shift independently one another. The predicted curve is in good agreement with the simulated $\Delta V_{T,N}$ distribution in the case of $N = 14$, revealing a statistically independent effect of the stored electrons on V_T for small N. However, this is no longer true for $N = 50$, where non-negligible differences appear between simulation results and calculations, highlighting a correlation among the stored electrons effect on V_T. This can be clearly seen from Fig. 3.10, where the average value and variance of $\Delta V_{T,N}$ ($\overline{\Delta V_{T,N}}$ and $\sigma^2_{\Delta V_{T,N}}$, respectively) are shown normalized to the average value and variance of $\Delta V_{T,1}$ ($\overline{\Delta V_{T,1}}$ and $\sigma^2_{\Delta V_{T,1}}$) as a function of N. While the relation $\overline{\Delta V_{T,N}} = N \times \Delta V_{T,1}$ correctly holds, $\sigma^2_{\Delta V_{T,N}}$ equals $13.03\times$ and $30.26 \times \sigma^2_{\Delta V_{T,1}}$ respectively for $N = 14$ and $N = 50$. This result confirms that stored electrons act independently for low N, thanks to the large distance among them. However, when their mutual distances decrease, their positions get closer and the spread is reduced, due to their correlated electrostatic control on substrate inversion.

Further insight can be obtained from the plots of the current density at threshold shown in Fig. 3.11. The left column refers to continuous doping for neutral (top) and programmed (bottom) cell; the right column is analogous but related to a case of atomistic doping. Results for continuous doping and neutral cell clearly show the non-uniform current density profile in the substrate determined by field enhancement at the cell corners, locally increasing the inversion charge. A different current density profile is instead present in the case of atomistic doping, due to the additional and random effect of discrete dopants on substrate inversion. However, in both cases, dicrete and localized electrons in the nitride slightly modify the current density profile in the substrate, due to the local nature of the electrostatic effect of each electron on substrate inversion. In this framework, the effect of stored electrons on substrate conduction is similar

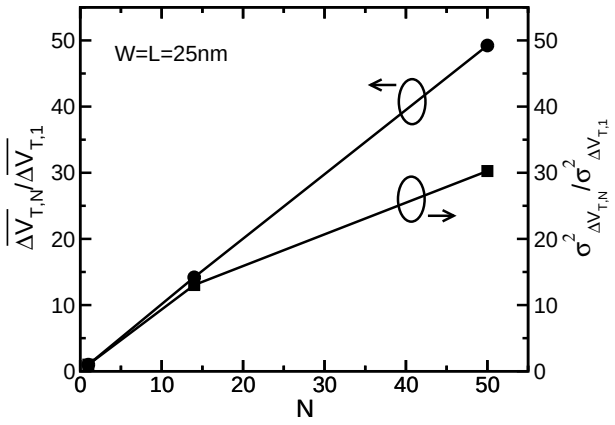


Figure 3.10: Normalized average and variance of $\Delta V_{T,N}$ as a function of the number of electrons stored in the nitride.

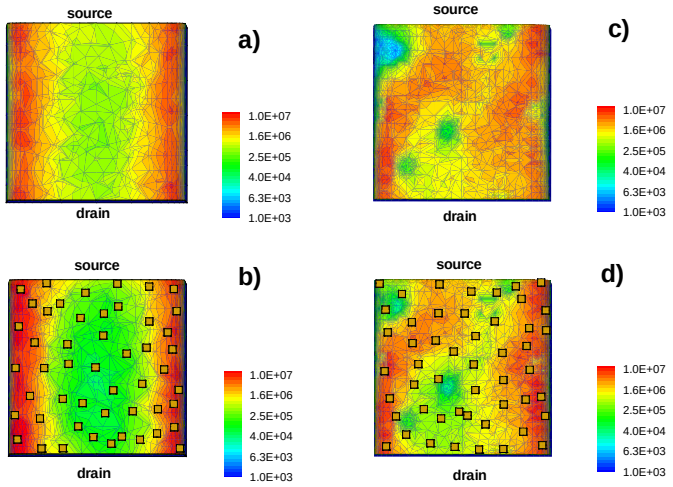


Figure 3.11: Plot of the current density at threshold for: a) continuous doping, neutral cell; b) continuous doping, 50 electrons in the nitride, c) atomistic doping, neutral cell; d) atomistic doping, 50 electrons in the nitride. Small fluctuations along L in case a) are due to numerical interpolations used to extract the current density profile at the substrate surface.

to that of atomistic doping, introducing randomness in the percolation paths connecting source-to-drain at threshold. Moreover, this means that a different current density profile is present in the substrate for neutral and programmed cells, changing, for instance, the impact on V_T of a single electron trapped in the bottom oxide. This point will be further discussed in the next section, where RTN in nitride cells will be addressed in detail.

Though the quantitative data obtained so far refer to the particular device structure investigated and should be cautiously applied to different (*e.g.* non-planar) cell geometries [103], the picture emerging from Figs. 3.4 and 3.3.1 is that of a significant V_T spread following electron injection. Such a spread impacts the V_T control, which must be retained for all cells in the array, *i.e.*, down to very low probability levels. For example, Fig. 3.4 shows that a single electron may give rise to a V_T shift of 50 mV with a probability of 10^{-2}, representing a high probability level for Flash arrays. This ΔV_T may, in turn, appear as a V_T-loss during data retention, determining a critical reliability issue for multi-level devices, where a severe control of V_T is required. Moreover, the statistical dispersion of ΔV_T may also influence the P/E operation of these devices. In fact, due to the relatively small separation among the V_T levels, multi-level arrays usually adopt a staircase algorithm for the program operation, resulting into very tight programmed V_T distributions. The width of these distributions is limited by many different physical phenomena, among which electron injection statistics [21, 22] and RTN [62, 63, 111–113] represent severe reliability constraints. In the case of nitride memory cells, the $\Delta V_{T,N}$ statistical spread may give an additional contribution to the dispersion of the programmed V_T distribution, as will be discussed in detail in the next section, where a scaling analysis of the $\Delta V_{T,N}$ statistical distribution will also be presented.

3.3 Scaling Analysis and Impact on Device Performance

This section presents a detailed analysis of the scaling of the $\Delta V_{T,N}$ distribution in nitride memories and of its impact on device performance. For the same electron density stored in the nitride layer, conventional cell scaling is shown to reduce the average $\Delta V_{T,N}$ ($\overline{\Delta V_{T,N}}$) and to increase its dispersion. The reduction of $\overline{\Delta V_{T,N}}$ is due to 3-D electrostatics, displaying fringing fields at the cell corners that reduce the impact of stored electrons on cell V_T [25]. The $\Delta V_{T,N}$ dispersion is shown to increase the width of the programmed V_T distribution when a staircase algorithm is used for a more accurate V_T placement [20]. A further burden for the staircase algorithm accuracy is given by random telegraph noise (RTN) instabilities, whose amplitude is enhanced by percolative substrate conduction in presence of atomistic doping [51, 52, 62, 63, 104, 109, 111–114]. As localized charge stored in the nitride was shown to affect the percolation paths in the substrate, differences between RTN instabilities of neutral and programmed cells may result. We will show at the end of the chapter, however, that despite cell programming may change the amplitude of RTN fluctuations of each single cell, the statistical distribution of RTN instability for the array is barely modified.

3.3.1 Scaling of the ΔV_T distribution

The statistical distribution of $\Delta V_{T,N}$ was shown to result from the different capability of localized electrons in the nitride to block source-to-drain conduc-

$\mathbf{W = L}$	**O/N/O**	$\mathbf{N_{sub}}$
32 nm	4/4.5/5 nm	1.5×10^{18} cm^{-3}
25 nm	4/4.5/5 nm	3×10^{18} cm^{-3}
18 nm	4/4.5/5 nm	4.5×10^{18} cm^{-3}

Table 3.1: Device parameters of the investigated nitride cells, showing the thicknesses of the gate stack (Oxide/Nitride/Oxide) and substrate doping (N_{sub}).

tion in presence of 3-D electrostatics and atomistic doping, both leading to non-uniform substrate inversion. These effects are expected to worsen with cell scaling, making the $\Delta V_{T,N}$ statistical dispersion a possible reliability issue for discrete-trap memories. In order to investigate this point, we performed numerical simulations of the planar cell structure presented in the previous section, scaling its width (W), length (L) and substrate doping according to Table 3.1. Thicknesses of the oxide/nitride/oxide layers used for the gate stack were not modified, as these cannot be reduced due to a number of constraining physical effects. First, 4 nm can be considered a lower limit for the bottom oxide thickness, as data retention and disturb immunity cannot be safely guaranteed below this value [115]. Second, reducing the nitride thickness has been shown detrimental for the achievable program/erase V_T window, with a critical reduction of the electron storage capability below 4 nm [116–119]. Finally, the top oxide should be thicker than the bottom oxide to endure the erasing capability reducing the tunneling current from the gate and to avoid thar this layer becomes the weak path for charge detrapping in data retention conditions, therefore we kept a constant 5 nm thickness when scaling the cell. Note, moreover, that assuming a relative dielectric constant nearly equal to 10 for Al_2O_3 [26], the 5 nm top oxide corresponds to nearly 13 nm of alumina, falling in the typical range of thicknesses considered for the top dielectric of TANOS cells [29, 96]. This, however, should be considered as a first-order equivalence between the two materials, as the pronounced 3-D electrostatic effects in scaled memory cells requires an accurate comparative analysis.

Single-electron storage results

Fig. 3.12 shows the statistical distribution of the V_T shift obtained when a single electron is randomly placed in a localized position inside the nitride volume ($\Delta V_{T,1}$), for the three technology nodes of Table 3.1. A clear increase of both the distribution average value ($\overline{\Delta V_{T,1}}$) and spread ($\sigma_{\Delta V_{T,1}}$) appears as cell dimensions are reduced, with, moreover, a larger slope (in mV/dec) of the exponential tail departing to high $\Delta V_{T,1}$. The increase of $\overline{\Delta V_{T,1}}$ is determined by the larger areal density of stored charge as the active area is reduced. However, Fig. 3.13a shows that the simulated $\overline{\Delta V_{T,1}}$ (symbols) are lower than what expected from the following simplified 1-D expression (dashed line in the figure):

$$\overline{\Delta V_{T,1}} = -\frac{qt_{eq}}{\epsilon_{ox} WL} \tag{3.1}$$

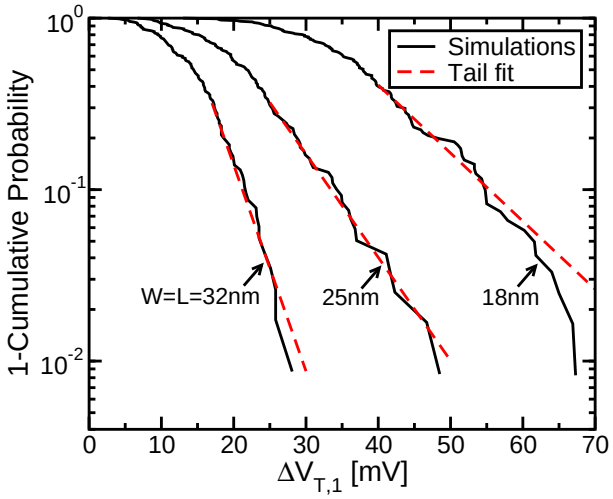


Figure 3.12: Simulated statistical distribution of $\Delta V_{T,1}$ for the three cell structures reported in Table 3.1.

where q is the electronic charge and $t_{eq} = t_{top} + t_N \epsilon_{ox}/2\epsilon_N$ represents the equivalent oxide thickness of the dielectric materials from the nitride central position to the gate (ϵ_{ox}, ϵ_N and t_{top} and t_N are the oxide and nitride dielectric constants and thicknesses, respectively). Note that simulation results for $\overline{\Delta V_{T,1}}$ as a function of $1/WL$ in Fig. 3.13a are actually lower than what predicted by (3.1). A ratio of 1.96, 2.09 and 2.31 can be extracted between the $\overline{\Delta V_{T,1}}$ expected from 1-D calculations and the value obtained from 3-D simulations for the 32 nm, 25 nm and 18 nm cells, respectively. This reveals that the control exerted by the stored electron on cell conduction is smaller than what predicted by 1-D calculations and differences increase as cell is scaled down. This effect is related to field fringing [25] and can be reproduced only in the case 3-D electrostatics and substrate conduction in presence of atomistic doping are accounted for in the simulations. This point will be discussed in more detail later, considering also the results from multi-electron storage.

Fig. 3.13b shows the simulated increase of $\sigma_{\Delta V_{T,1}}$ with scaling. This increase is attributed to more relevant percolative effects in the substrate when cell dimensions are reduced, due to enhanced 3-D electrostatics and a more relevant role played by the atomistic nature of substrate dopants. Note that the increase of the spread and of the exponential tail of the distributions in Fig. 3.12 allows, for a fixed probability level, for larger and larger $\Delta V_{T,1}$ as cell is scaled down. Assuming, for instance, a reference probability equal to 5×10^{-2} (a quite large probability value for state-of-the-art NAND memory arrays), Fig. 3.12 predicts that the maximum single-electron shift increases from nearly 25 mV to 40 mV to 60 mV when moving from the 32 nm to the 25 nm to the 18 nm technology.

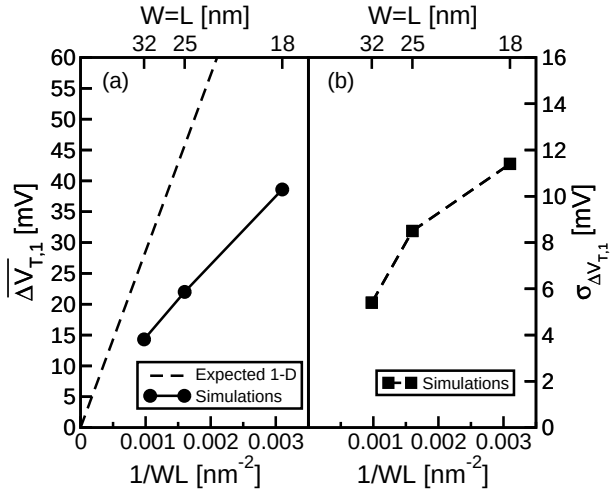


Figure 3.13: $\overline{\Delta V_{T,1}}$ (a) and $\sigma_{\Delta V_{T,1}}$ (b) from our numerical simulations as a function of the reciprocal of cell area. In (a) the expected linear dependence calculated by (3.1) is also shown.

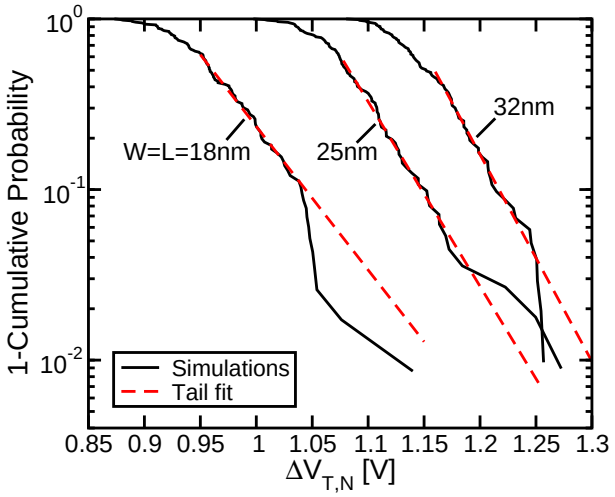


Figure 3.14: Simulated statistical distribution of $\Delta V_{T,N}$ in the case of $N = 26$, 50 and 82 electrons for the 18 nm, 25 nm and 32 nm technology, respectively.

Multi-electron storage results

In order to investigate the statistical distribution of the V_T shift determined by multi-electron storage in the nitride, we considered the case of $N = 26$, 50 and 82 electrons randomly placed in the nitride of the 18 nm, 25 nm and 32 nm technology cell, respectively. These N values result into the same trapped charge density in the nitride for the three cell dimensions reported in Table 3.1,

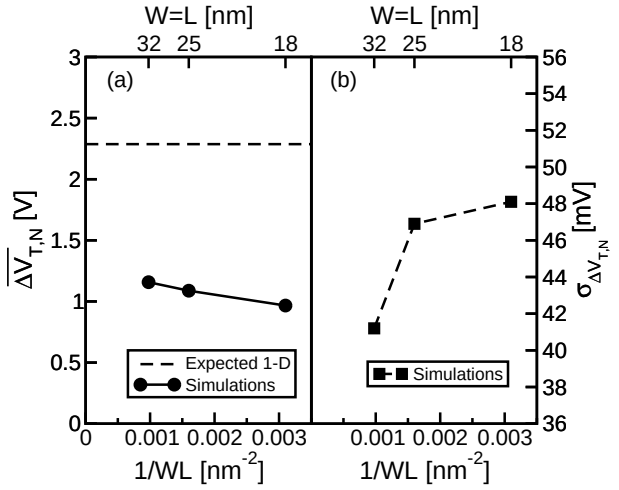


Figure 3.15: $\overline{\Delta V_{T,N}}$ (a) and $\sigma_{\Delta V_{T,N}}$ (b) from the numerical simulations of Fig. 3.14. In (a) results calculated by (3.2) are also shown.

nearly equal to 8×10^{12} electrons/cm^2. As a consequence, the same average $\Delta V_{T,N}$ ($\overline{\Delta V_{T,N}}$) should be obtained from simplified 1-D calculations:

$$\overline{\Delta V_{T,N}} = -\frac{N q t_{eq}}{\epsilon_{ox} W L} \tag{3.2}$$

However, Fig. 3.14 shows that the simulated statistical distribution of $\Delta V_{T,N}$ displays a reduction of $\overline{\Delta V_{T,N}}$ as cell dimensions decrease. The extracted $\overline{\Delta V_{T,N}}$ are reported in Fig. 3.15a and compared with the constant value nearly equal to 2.3 V predicted by (3.2): quite smaller average shifts are obtained from 3-D numerical simulations, confirming the reduced stored electron control on substrate conduction that already appeared from Fig. 3.13. In particular, the ratios between expected 1-D predictions and simulation results are the same that were obtained considering $\overline{\Delta V_{T,1}}$. Note, in fact, that the $\overline{\Delta V_{T,N}} = N \times \overline{\Delta V_{T,1}}$ holds both for 1-D calculations and for 3-D simulations [120].

Fig. 3.15b shows that the spread of the distributions in Fig. 3.14 increases when scaling cell size. This spread is, however, lower than what predicted assuming an uncorrelated superposition of N single-electron ΔV_T, as discussed in detail in the previous section. In particular, the ratio $\sigma_{\Delta V_{T,N}}/(\sqrt{N}\sigma_{\Delta V_{T,1}})$ is nearly equal to 0.8 for all the three cell dimensions.

Fringing–fields effect on ΔV_T

Results presented in previous paragraphs reveal that large differences exist between 3-D simulations and simplified 1-D calculations for the planar cell structure investigated in this work. Besides the inability to predict the $\Delta V_{T,N}$ statistical dispersion, 1-D calculations fail also in evaluating $\overline{\Delta V_{T,N}}$, as evident from Figs. 3.13-3.15. This is a result of 3-D electrostatics, contributing to the reduction of the average V_T shift in two different ways. First, the gate-to-channel

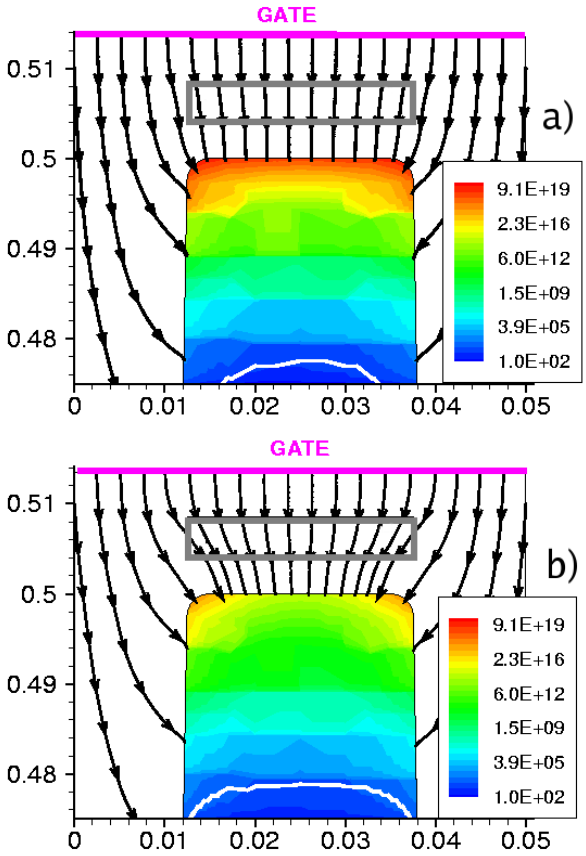


Figure 3.16: Electrostatics simulations for the 25 nm cell (along the W direction) in the case of 0 (a) and 50 electrons (b) stored in the nitride, with a 3 V bias applied to cell gate. Electric field lines (whose number is not related to the field strength) and electron concentration in the substrate (according to the color bar) are shown.

capacitance (C_{GC}) for the 3-D structure largely differs from what predicted by 1-D calculations. This is due to fringing fields at the active area edges, as shown in Fig. 3.16a for the 25 nm cell structure. A large spreading of the fields lines appears from the active area edges towards the gate in the W direction, determining an increase of the *active gate area* with respect to the substrate active area for C_{GC} evaluation. Referring for simplicity to the strong inversion regime, 1-D calculations predict $C_{GC} = 1.87$ aF, while 3-D simulations give $C_{GC} = 2.5$ aF. As a consequence of the larger C_{GC} value, in the real 3-D case the reduction of the substrate inversion charge that is determined by electron storage in the nitride can be balanced by an increase of the gate voltage that is lower than what predicted by 1-D electrostatics.

In addition to the C_{GC} increase, there is a second effect contributing to lowering $\overline{\Delta V_{T,N}}$, which is related to the amount of charge induced on the subtrate

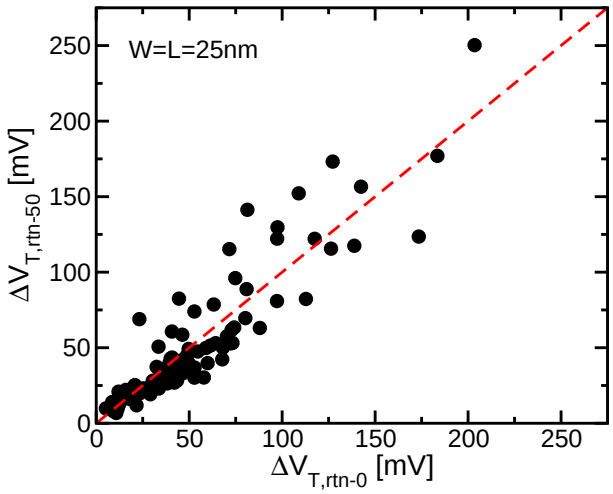


Figure 3.17: Simulated $\Delta V_{T,rtn-50}$ as a function of $\Delta V_{T,rtn-0}$ for our 25 nm cell.

(Q_{ind}) by the electrons stored in the nitride. When an electron is placed at the center of the active area the fraction of its charge that is induced on the channel is nearly equal to the ratio between t_{eq} and the equivalent oxide thickness of the whole gate stack. However, this is no longer true if the electron is placed at the nitride edges along L or W: in the former case some charge is induced on the junctions and not on the channel, in the latter more charge is induced on the gate by fringing fields. Both these phenomena contribute to the reduction of the charge induced on the channel and therefore to the reduction of $\overline{\Delta V_{T,N}}$. For example, Fig. 3.16b shows the field lines in the case 50 electrons are uniformly placed in the nitride volume of the 25 nm cell, showing their possibility to modify the fringing fields of Fig. 3.16a. Referring again to the strong inversion regime, the variation of the substrate inversion charge due to electron storage in the nitride from 3-D simulations is equal to $Q_{ind} = 2.84$ aC, and not to 4.3 aC as predicted by 1-D calculations.

As a result of the larger C_{GC} and the lower Q_{ind} with respect to 1-D calculations, 3-D simulations predict $\overline{\Delta V_{T,N}} \simeq Q_{ind}/C_{GC} = 2.84\text{aC}/2.5\text{aF} = 1.13$ V, which is in good agreement with the $\overline{\Delta V_{T,N}}$ value reported in Fig. 3.15a for the 25 nm cell. Moreover, these results explain also the lowering of $\overline{\Delta V_{T,N}}$ with cell scaling predicted by 3-D simulations in Fig. 3.15a. In fact, fringing-field effects are more relevant as active area is reduced, as the cell dielectric stack thicknesses are not modified. As a final remark, note that fringing–field effects reported in this Section are strictly related to the particular cell geometry investigated in this work, as different designs of the 3-D cell can increase or alleviate these effects [24, 25].

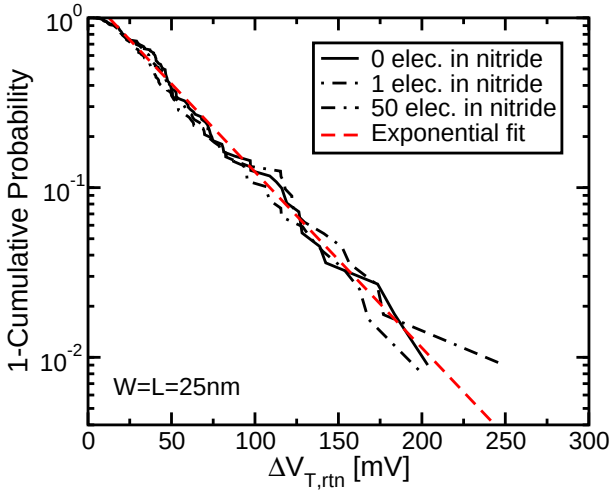


Figure 3.18: Simulated cumulative $\Delta V_{T,rtn}$ distribution in the case of 0, 1 and 50 electrons stored in the nitride of our 25 nm cell.

3.3.2 RTN instabilities

RTN represents an additional constraint for the accuracy of the programming operation. Differently from the $\Delta V_{T,N}$ dispersion previously investigated, RTN is related to the presence of traps in the cell bottom oxide, capturing/emitting electrons at random instants of time.

The statistical distribution of the amplitude of single-trap fluctuations was shown as one of the most important ingredients to characterize RTN instabilities in Flash memories [105, 108, 111]. An exponential behavior was shown for this distribution in the case of floating-gate Flash arrays, allowing quite large V_T fluctuations to exist, even if at low probability levels [66]. This was mainly attributed to atomistic doping and 3-D electrostatics, allowing a single RTN trap localized over a substrate percolation path to largely stop source-to-drain conduction, similarly to what reported for nitride traps in the previous section. Moreover, the exponential distribution decay constant λ (in mV/dec) was shown to increases with cell scaling [105], becoming a major issue for deca-nanometer Flash arrays.

RTN was shown to constrain the reliability of nitride-based memories as well [121]. Despite the basic phenomenology highlighted for floating-gate devices should also apply for SONOS/TANOS arrays, the localized nature of the stored charge in this devices may arise a noticeable difference with respect to conventional Flash memories [114]. In fact, electron storage in discrete nitride traps was shown in the previous section to modify the current density profile in the substrate, which critically affects the fluctuation amplitude of RTN traps. This means that, in principle, different RTN amplitudes can be observed for the neutral and the programmed cell state in the case of nitride-based memories, while no such behavior was observed for floating-gate cells [114]. In order to in-

vestigate this point, we performed numerical simulations of RTN amplitudes for our nitride cells. Using the same Monte Carlo procedure reported in the previous section, we evaluated cell V_T with and without a single negatively-charged RTN trap randomly placed at the substrate/bottom oxide interface. The resulting V_T difference ($\Delta V_{T,rtn}$) represents the trap RTN amplitude, which varies statistically due to its different position over the active area and the different dopant configuration in the substrate [105]. In order to investigate RTN amplitudes in the neutral and programmed cell state, for the same RTN trap position $\Delta V_{T,rtn}$ was calculated assuming 0 ($\Delta V_{T,rtn-0}$) and N ($\Delta V_{T,rtn-N}$) electrons stored in the nitride. Fig. 3.17 shows $\Delta V_{T,rtn-50}$ as a function of $\Delta V_{T,rtn-0}$ for our 25 nm cell (similar results were obtained for the other cell dimensions): the possibility for localized charge storage in the nitride to actually modify substrate conduction clearly appears from this figure as a change in the RTN amplitude for the neutral and the programmed cell state. Moreover, note that programming can either enhance or reduce the amplitude of an RTN trap and that this depends on how substrate conduction is modified by the localized nitride charge. However, Fig. 3.17 shows that there is a strong correlation between $\Delta V_{T,rtn-50}$ and $\Delta V_{T,rtn-0}$, meaning that the RTN amplitude variation after program is low when compared to the initial amplitude.

Despite the differences reported in Fig. 3.17 between the neutral and programmed cell state, Fig. 3.18 shows that the cumulative distribution of $\Delta V_{T,rtn}$ is barely modified by cell programming. In fact, Fig. 3.18 reveals that the exponential distribution [108] for $\Delta V_{T,rtn}$ has negligible changes when moving from 0 to 1 to 50 electrons in the nitride of our 25 nm cell. This is due to two main reasons, both related to the variations of $\Delta V_{T,rtn}$ obtained when moving from the neutral to the programmed cell state in Fig. 3.17 (vertical displacements of the simulation points from the dashed line): 1) the variations are relatively small if compared with the dispersion of $\Delta V_{T,rtn}$ for the neutral cell; 2) no preferential direction for the variations appears from the figure, *i.e.* they can be positive or negative. These results reveal that 3-D electrostatics and atomistic substrate doping play the main role in determining RTN statistics also for nitride-based memories, with a marginal role of electrons *locally* stored in the nitride.

This last observation is very important in justifying the work we will present in chapter 5. In that case we will report a doping engineering study realized to suppress the RTN instabilities in Flash memories. To carry out the study we have adopted a floating-gate device template instead of a charge-trap device template. The reasons for this choice are manifold: (i) no comprehensive doping optimization studies are reported in literature for the floating-gate technology (that is the current technology in production), (ii) the simulation of floating-gate devices is computationally less costly respect to the charge-trap case, (iii) as previously shown, despite in charge-trap memories the cell programming may change the amplitude of RTN fluctuations of each single cell, the statistical distribution of RTN instability for the array is barely modified.

3.4 Conclusions

This chapter presented a comprehensive investigation of statistical effects in deeply-scaled nitride memory cells, highlighting that 3-D electrostatics, atomistic substrate doping and charge localization in the nitride volume result into a statistical dispersion of ΔV_T. The local electrostatic effect of stored electrons and percolative substrate conduction were shown as the main reason for the ΔV_T spread.

A scaling analysis of the statistical distribution of ΔV_T in nitride memories was also provided. For fixed density of trapped charge, the average ΔV_T was shown to decrease as a consequence of fringing fields, not predictable by any 1-D simulation approach. Moreover, the distribution statistical dispersion was shown to increase with technology scaling due to a more sensitive percolative substrate conduction in presence of atomistic doping and 3-D electrostatics. The impact of these effects on RTN instabilities were then highlighted, showing that locally stored charges in nitride have a marginal role in determining RTN statistics.

Chapter 4

Programming variability in charge-trap memories

This chapter presents a detailed investigation of charge-trap memory programming by means of 3-D TCAD simulations accounting both for the discrete and localized nature of traps and for the statistical process ruling granular electron injection from the substrate into the storage layer. In addition, for a correct evaluation of the threshold-voltage dynamics, cell electrostatics and drain current are calculated in presence of atomistic doping, largely contributing to percolative substrate conduction. The first part of the chapter shows that the low average programming efficiency commonly encountered in nanoscale charge-trap memories mainly results from the low impact of locally stored electrons on cell threshold voltage in presence of fringing fields at the cell edges. The second part of the chapter addresses the programming variability arising from the discreteness of charge and matter. Results show that the injection variability source plays the dominant role in determining the statistical dispersion of cell threshold voltage during the program operation when compared to the random dopant and random trap variability sources.

4.1 Introduction

DISCRETENESS of charge represents a major variability source for the program operation of nanoscale floating-gate and charge-trap memory devices, compromising the tightness of the programmed threshold-voltage (V_T) distribution obtained by incremental step pulse programming (ISPP, see chapter 1) [21, 22, 122–126]. Cell scaling reduces, in fact, the number of electrons

to be transferred to the storage layer during programming and this makes the statistical process ruling their injection a non-negligible spread source for cell V_T transients. A detailed investigation of the impact of charge granularity on the programming performance of conventional floating-gate Flash arrays has been presented in [21, 22]. The possibility to extend the same analysis to charge-trap memory devices is, however, complicated by the additional variability contributions given by the discrete number of traps in the cell storage layer and by the localized electron storage therein. Percolative source-to-drain conduction, due to fringing fields at the cell corners and atomistic substrate doping [45, 51–53, 104, 105, 107], makes, in fact, not only the fluctuation of the trap number in the cell but also of the trap position over the channel a major variability source for nanoscale cells [54, 55].

Before dealing with the program variability effects, it is mandatory, however, to understand the physics underlying the programming efficiency of charge-trap memories, which has been shown to largely decrease when device dimensions are reduced to the deca-nanometer scale [24, 25, 103, 127]. Referring to the incremental step pulse programming (ISPP, see chapter 1) algorithm [20, 128], in fact, significant differences have been shown to appear in the ratio between the threshold-voltage increase per step ($\Delta V_{T,s}$) and the step amplitude (V_s) for large area capacitors and nanoscale cells. While the ratio is only slightly below 1 for the former, at least far from the saturation of the available traps in the storage layer [23], quite lower values in the 0.5-0.65 range are typically reported for the latter [24, 103, 127]. Despite the decrease of the ISPP slope for small area cells has been clearly correlated not only to cell dimensions but also to cell geometry [25], its origin has not be well assessed so far.

We will start the chapter trying to study and clarify the physics governing the programming efficiency, while we will address the programming variability topic in the second part of the chapter.

4.2 Programming dynamics and efficiency

In this section, we present a detailed simulation analysis of the ISPP dynamics on nanoscale charge-trap memory cells, highlighting the basic features of the electron storage process and its impact on cell threshold voltage (V_T). In order to account in detail for all the effects of discreteness, 3-D TCAD simulations have been performed accounting for the discrete nature both of traps in the storage layer and of the electron flow charging it [21,22,54,55,122]. Moreover, to carefully evaluate the impact of each single stored electron on cell V_T, substrate doping was treated as atomistic when solving for cell electrostatics and source-to-drain current conduction [45, 51–53, 104, 105, 107]. Statistical results were collected following a Monte Carlo approach, randomly changing the number and position of both the nitride traps and the substrate dopants and reproducing the stochastic process ruling discrete electron injection into the storage layer during programming. The average results from the Monte Carlo simulations show that the low programming efficiency of nanoscale charge-trap cells mainly results from the low impact of locally stored electrons on V_T due to fringing fields at the cell edges. This is also supported by results from a continuous 3-D

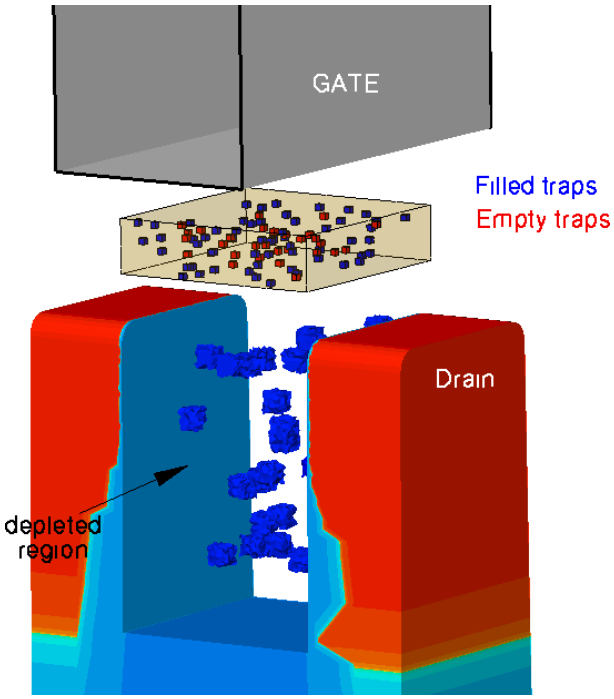


Figure 4.1: TCAD structure for the simulated 18 nm MONOS cell, highlighting discrete dopants in the substrate and discrete traps in the nitride. Red regions: n implants; blue regions: p substrate; oxide regions are not shown for clarity.

model for the program operation, including only the localized nature of charge storage but neglecting the discreteness of traps, dopants and electron flow.

4.2.1 Physics-based numerical model

Fig. 5.1 shows the TCAD structure of the charge-trap memory cell investigated in this work, featuring an aluminum metal gate, a bottom-oxide/nitride/top-oxide (ONO) dielectric stack with thicknesses equal to 4/4.5/5 nm, and STI trenches at the active area edges. Note that the nitride layer is patterned over the channel area, having width (W) and length (L) equal to 18 nm. The gate extends, instead, beyond the active area in the W direction, remaining completely planar. A uniform substrate doping $N_a = 4.2 \times 10^{18}$ cm^{-3} was assumed, discretizing the acceptor atoms in the channel region down to 25 nm from the bottom-oxide interface. A trap density equal to $N_t = 6 \times 10^{19}$ cm^{-3} was assumed for the nitride. Criteria for doping and nitride traps discretization are the same adopted in the previous chapter.

Fig. 4.2 shows the simulation procedure adopted for the statistical analysis of the ISPP dynamics. After the definition of the deterministic features of the device, a first Monte Carlo loop was used to gather information on cells having different stochastic configurations of atomistic dopants and nitride traps,

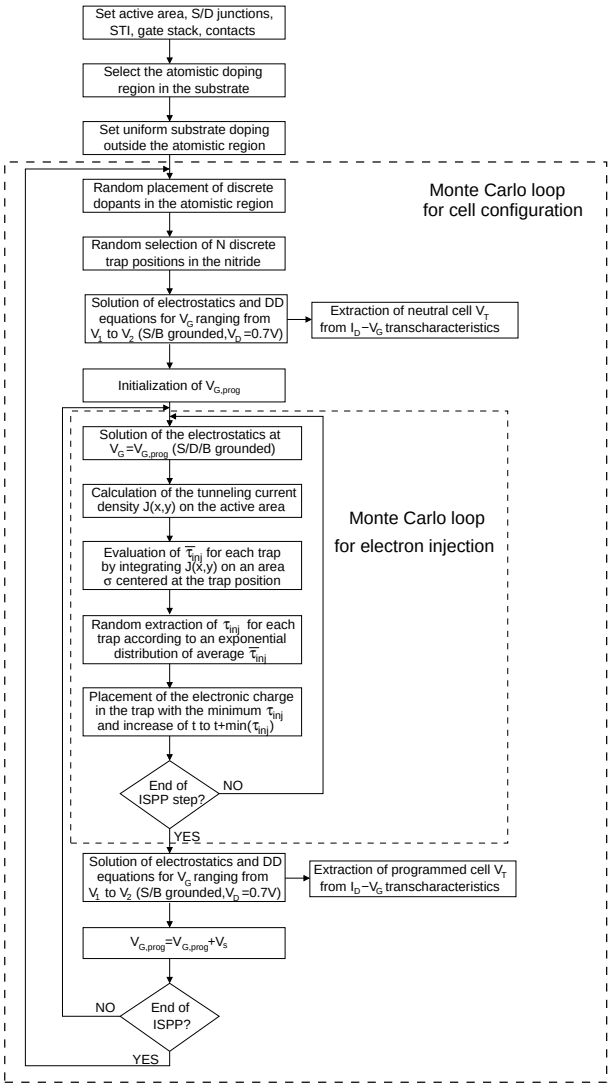


Figure 4.2: Block diagram for the simulation procedure used to statistically investigate ISPP on nanoscale charge-trap memories.

obtained drawing their number from a Poisson statistics and their spatial position from a uniform distribution in their discretization region. Assuming empty traps in the nitride, neutral cell V_T was extracted from the drain current vs. gate voltage (I_D–V_G) transcharacteristics, calculated solving the drift-diffusion equations as discussed in chapter 3, keeping source and bulk grounded and drain at $V_D = 0.7$ V. An inner Monte Carlo loop was then used to simulate the electron injection process during each ISPP step: once cell electrostatics is solved for $V_G = V_{G,prog}$ (with source, drain and bulk grounded), the tunneling current

density $J(x, y)$ is calculated over the channel area in the WKB approximation and taking into account the local electron supply in each point of the channel, then computing the average electron injection time from the substrate to each trap as:

$$\overline{\tau}_{inj} = \frac{q}{\int_{\sigma} J(x,y)\ dxdy}, \tag{4.1}$$

where q is the electron charge and the integral is evaluated on an area equal to the trap capture cross-section σ (assumed equal to 10^{-14} cm^2 throughout this work [26]) centered at the trap position. Then, for each trap, the stochastic electron injection time τ_{inj} was drawn from an exponential distribution with average value equal to its corresponding $\overline{\tau}_{inj}$, and a single electron is placed in the nitride trap having the smallest τ_{inj}. This value is added to the total time t elapsed since the beginning of the program operation, then going back to the solution of cell electrostatics with the new electron in the nitride and repeating the calculations for the stochastic injection of the next electron until the end of the ISPP step. When this happens, the I_D–V_G transcharacteristics is calculated again to extract cell V_T after the programming step, then reentering the Monte Carlo loop for electron injection after adding V_s to $V_{G,prog}$. The simulation flow reaches its end when $V_{G,prog}$ equals the maximum value selected for the ISPP algorithm.

As a final remark, note that the model does not include the possibility for electron emission from filled nitride traps. This process is surely very important in the late stages of the program operation, when the number of electrons stored in the nitride is large and the electric field within the top oxide is high [26]. The analysis is therefore expected to hold for V_T values far from the saturation of the available traps in the nitride. Moreover, no trapping was included in the bottom- and in the top-oxide layers, which were considered as ideal trap-free dielectrics. Despite this choice allows the investigation of the ultimate programming performance of the charge-trap technology, trapping in the top dielectric should be carefully taken into account when a high-k material is adopted. Similarly to electron emission, results on TANOS memories revealed, in fact, that trapping in the top alumina layer has a non-negligible impact on the late stages of the programming transient [129–131].

4.2.2 Electron injection

In order to highlight the basic features of the statistical process leading to discrete electron injection into the nitride traps, ISPP was first investigated on a single stochastic cell, having a random configuration of substrate dopants and nitride traps, whose number was set nearly equal to the corresponding average value. Fig. 4.3 shows $J(x, y)$ over the channel area in the case of $V_{G,prog} =$ 13 V and empty nitride, representing the initial condition for the simulated ISPP. A largely non-uniform tunneling current density clearly appears, with a higher $J(x, y)$ at the STI corners where field peaks occur (see chapter 3). As a result, $\overline{\tau}_{inj}$ strongly depends on the trap position over the channel, as shown in Fig. 4.4: traps placed near the active area edges along W display a smaller average capture time than traps located near the channel center, with only a

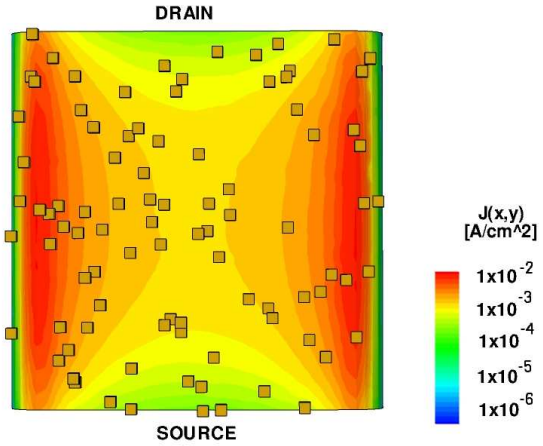


Figure 4.3: Tunneling current density $J(x, y)$ at the channel surface of the stochastic cell investigated in Sections 4.2.2-4.2.3, for $V_G = 13$ V and empty nitride (beginning of ISPP). Squares are trap positions in the nitride.

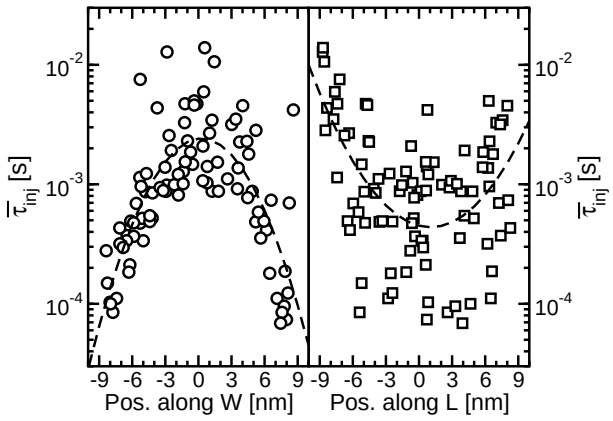


Figure 4.4: Average injection time into the nitride traps ($\overline{\tau}_{inj}$) as a function of trap position along W (left) and L (right), for $V_G = 13$ V and empty nitride. Channel center is at $(0, 0)$. Dashed-lines are average trend guidelines.

relatively small statistical dispersion of the scatter plot. This dispersion is due, first of all, to a weak $\overline{\tau}_{inj}$ dependence on the trap position along L, resulting into a slightly smaller capture time when the trap is placed half-way between source and drain than near the junctions. Moreover, $\overline{\tau}_{inj}$ also depends on the trap position along the nitride thickness when the trap is vertically aligned to other traps. In this case, in fact, the injected electron is assumed to be captured by the lowest trap in the stack, and $\overline{\tau}_{inj}$ for the higher traps was calculated by limiting the integral in (4.1) only to the fraction of the σ projection not shadowed by the lowest traps.

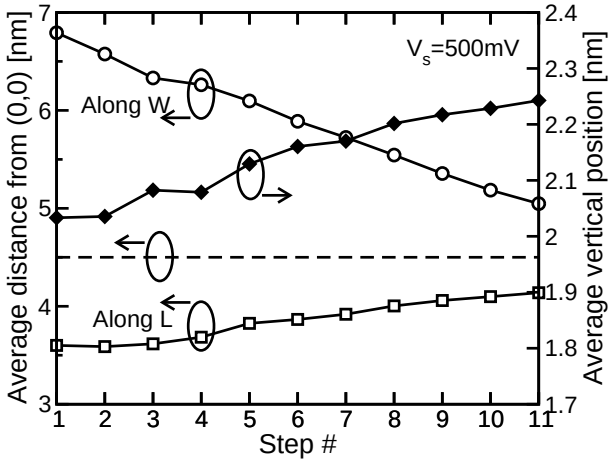


Figure 4.5: Average distance of trapped electrons from the channel center (*i.e.*, $(0,0)$) in the W and L direction as ISPP proceeds (dashed line represents the expected average distance in the case electrons were uniformly distributed over the channel). The average vertical position of the trapped electrons in the nitride is also shown (0 is the bottom-oxide/nitride interface).

From the results of Figs. 4.3-4.4, electron storage in the nitride layer is not uniform during programming. This is shown in Fig. 4.5, where the average distance of stored electrons from the channel center is reported along the L and W directions. If electrons were uniformly stored in the nitride, their average distance from the channel center in our 18 nm cell should be 4.5 nm (dashed line in the figure). Fig. 4.5 shows, instead, that the real distance in the W direction is larger than this value during the first steps of ISPP, due to a larger electron injection/capture near the cell STI corners. However, such a trapped charge reduces the tunneling probability in the same regions, making electron trapping in the central part of the channel more and more important as programming proceeds. As a result, the average electron distance from the channel center decreases and approaches 4.5 nm. The same Fig. 4.5 also shows a weak increase of the stored charge position along L: this trend can be explained considering the results in Figs. 4.3-4.4, showing that during the initial ISPP steps electrons are nearly stored half-way between source and drain. Note, however, that disuniformities in the electron distribution along L are recovered as ISPP proceeds, with the mean stored charge position approaching 4.5 nm. Finally, Fig. 4.5 also shows that the average distance of stored electrons from the nitride/bottom-oxide interface slightly increases as programming proceeds.

4.2.3 ΔV_T transients and ISPP efficiency

The impact of electrons stored in the nitride layer on cell V_T is strongly affected by fringing fields and atomistic substrate doping [54, 55]. In fact, both these effects result into a non-uniform source-to-drain current density on the active area during read, as shown in Fig. 4.6, where percolative substrate conduction

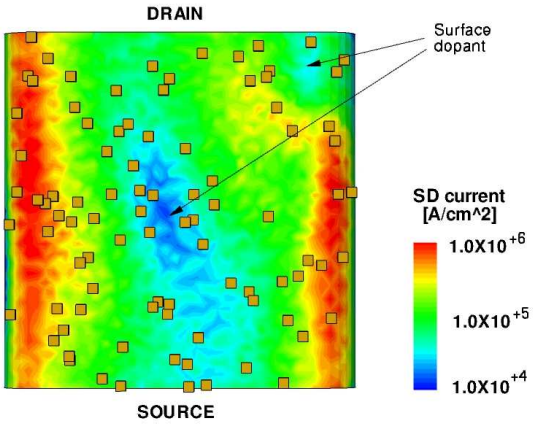


Figure 4.6: Source-to-drain current density at the channel surface of the stochastic cell investigated in Sections 4.2.2-4.2.3, during read at $V_G = V_T$ with empty nitride. Squares are trap positions in the nitride.

clearly appears. Note that the current density profile is not symmetrical in the W direction, as expected in presence only of field intensifications at the cell STI corners, due to the variability contribution given by atomistic doping on substrate inversion. In these conditions, the V_T shift (ΔV_T) obtained by a single electron stored in the nitride depends on the electron position over the active area, with electrons placed above a current percolation path having a larger impact on V_T than the others, thanks to their larger possibility to stop source-to-drain conduction. This is shown in Fig. 4.7, where the ΔV_T produced by a single electron placed in the different nitride traps is reported as a function of the trap position along W and L. Besides the dispersion of the scatter plot, matching the current density profile in Fig. 4.6, the resulting ΔV_T in this graph are all below the 1-D prediction $\Delta V_T^{1-D} = q/C_{NG}^{1-D} \simeq 88$ mV, where $C_{NG}^{1-D} = 1.81$ aF is the 1-D capacitance from the nitride center to the gate, given by:

$$C_{NG}^{1-D} = \epsilon_{ox} \frac{WL}{t_{eq}} \tag{4.2}$$

with $t_{eq} = t_{top} + t_N \epsilon_{ox}/2\epsilon_N$ representing the equivalent oxide thickness of the dielectric materials from the nitride central position to the gate (ϵ_{ox}, ϵ_N and t_{top} and t_N are the oxide and nitride dielectric constants and thicknesses, respectively). The smaller ΔV_T in Fig. 4.7 with respect to the 1-D prediction is the result of fringing fields in the 3-D electrostatics, increasing the gate coupling both with the nitride stored charge and with the channel, as discussed in [54].

In order to correctly quantify the stored charge effect on cell V_T, Fig. 4.8 shows the average number of electrons in the nitride ($\overline{n_t}$) as a function of the average V_T shift ($\overline{\Delta V_T}$) at the end of the ISPP steps ($V_s = 500$ mV, step duration $\tau_s = 10$ μs), as resulting from more than 100 Monte Carlo simulations on the investigated stochastic cell. From the slope of this graph, an effective 3-D

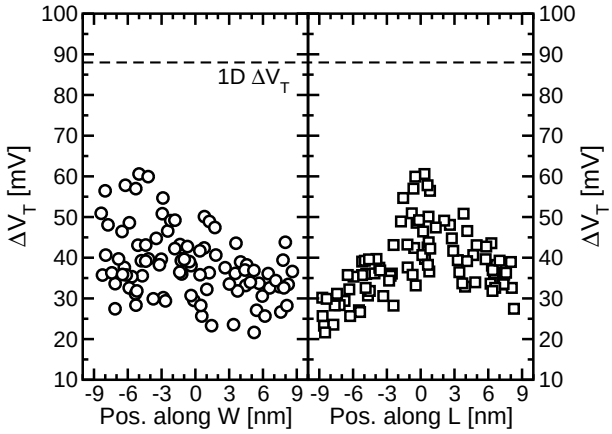


Figure 4.7: ΔV_T given by a single electron stored in the nitride layer as a function of the trap position along W (left) and L (right). The channel center is at $(0, 0)$.

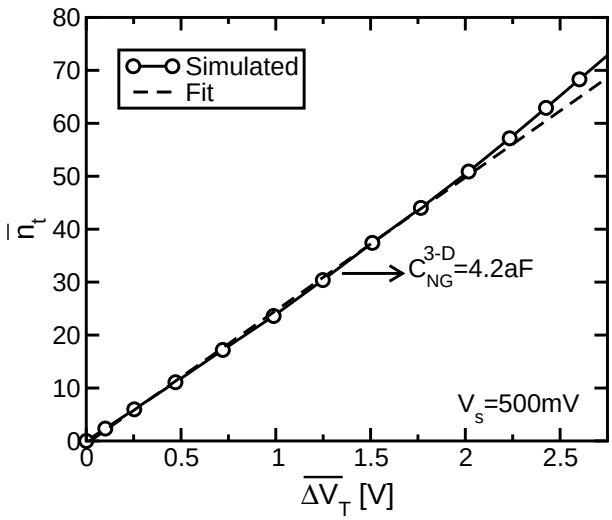


Figure 4.8: $\overline{n}_t$ vs. $\overline{\Delta V_T}$ relation during ISPP with $V_s = 500$ mV and $\tau_s = 10$ μs. The linear fit of the curve allows the evaluation of C_{NG}^{3-D}.

electrostatic capacitance $C_{NG}^{3-D} = 4.2$ aF can be extracted, allowing a correct evaluation of $\overline{\Delta V_T}$ as $q\overline{n}_t/C_{NG}^{3-D}$. Note, first of all, that the effective C_{NG}^{3-D} is larger than C_{NG}^{1-D}, as required to give rise to ΔV_T values lower than the 1-D predictions in Fig. 4.7. However, C_{NG}^{3-D} cannot be used to explore the variability of the scatter plot, as C_{NG}^{3-D} is defined from Fig. 4.8 using the integral of n_t and ΔV_T as programming proceeds, therefore averaging the different effect of stored electrons on cell V_T. Anyway, the displacement of the $\overline{n}_t$ vs. $\overline{\Delta V_T}$ relation from the linear behavior for large numbers of stored electrons in Fig. 4.8 clearly

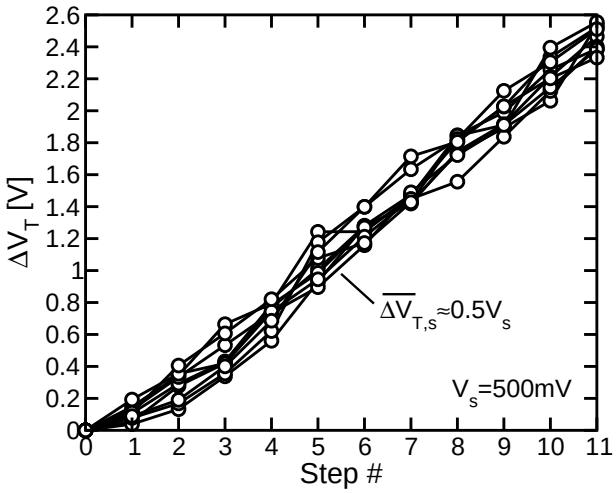


Figure 4.9: Monte Carlo simulation results for the ΔV_T transients of the investigated nitride cell.

highlights a change in the average impact of the stored charge on V_T. This reflects the displacement of the storage position from the STI cell corners to the central channel regions as ISPP proceeds, in agreement with the results of Fig. 4.5 and determining a consequent reduction of the stored electron control on V_T.

Fig. 4.9 shows some Monte Carlo simulations for the ΔV_T transients during ISPP on the same stochastic cell. Besides the statistical dispersion of the curves, which will be addressed in detail in the second part of this chapter, an average increase per step $\overline{\Delta V_{T,s}} \simeq 0.5 V_s$ appears, confirming the low ISPP efficiency commonly observed on charge-trap memories [24, 103, 127]. The low $\overline{\Delta V_{T,s}}/V_s$ is mainly the result of the low impact exerted by stored electrons on V_T and, in turn, of the large effective C_{NG}^{3-D}. Note that assuming for the electrons an electrostatic control as in the 1-D case, the resulting ISPP slope would be increased by the ratio $C_{NG}^{3-D}/C_{NG}^{1-D} \simeq 2.3$, *i.e.* it would be even a little bit larger than 1. This means that the non-uniform $J(x, y)$ and $\overline{\tau}_{inj}$ profiles over the active area not only do not degrade the electron injection process, but enhance indeed the process with respect to the 1-D case.

In order to exclude that the previous results are a specific feature of the single stochastic cell investigated, we simulated the ISPP transient on a large number of different cells, following the complete Monte Carlo procedure of Fig. 4.2. Fig. 4.10 shows the comparison between the average results from many ISPP transients on the same cell and from a single ISPP transient on many stochastically different cells, in terms of $\overline{V_T}$ and $\overline{\Delta V_T}$. Despite the number of atomistic dopants in the substrate of the previously considered single cell was selected nearly equal to the average value expected from N_a, a higher neutral V_T (*i.e.*, V_T at step 0) appears for this cell with respect to the average value of the cells statistics. This is due to a significant impact of dopants position on cell neutral

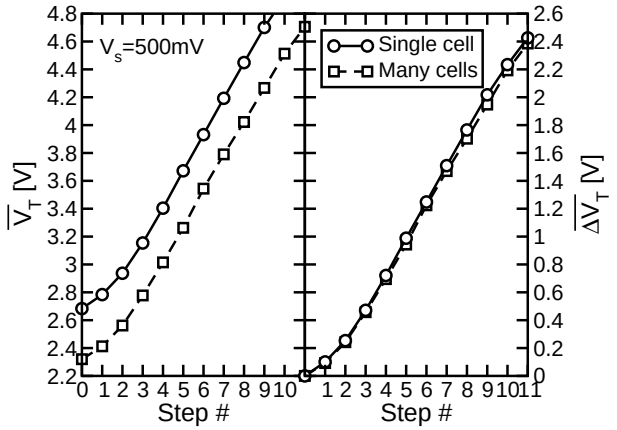


Figure 4.10: Average V_T and ΔV_T transients during ISPP simulated from many Monte Carlo runs on the same single cell or from a single Monte Carlo run on many stochastic cells.

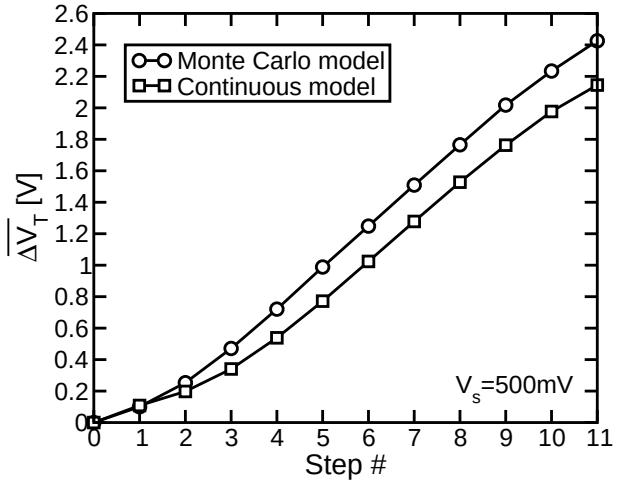


Figure 4.11: Comparison between the average ΔV_T transient during ISPP from our Monte Carlo model and from a 3-D numerical tool treating substrate doping, nitride traps and the programming electron flow as continuous.

V_T, as shown in chapter 3. However, no significant difference appears between the single- and the many-cell $\overline{\Delta V_T}$ transients in Fig. 4.10. This confirms in more general terms the low programming efficiency of the investigate charge-trap cell, leading an average $\overline{\Delta V_{T,s}}/V_s \simeq 0.5$.

4.2.4 Scaling analysis

Results of Fig. 4.10 revealed that the low ISPP efficiency is a general feature of nanoscale charge-trap memories, clearly appearing in the average program-

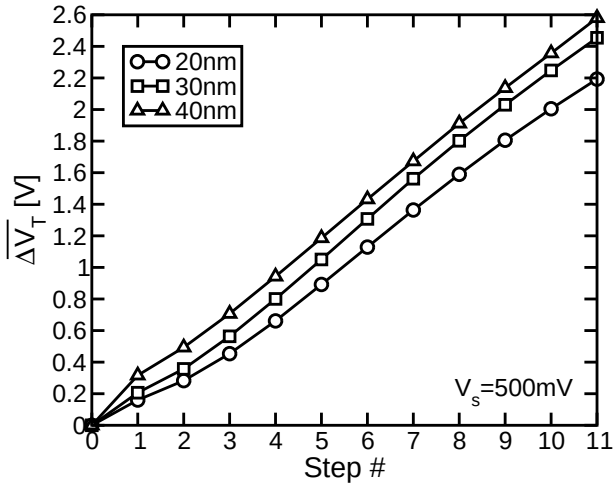


Figure 4.12: Simulated average ΔV_T transients during ISPP for different cell sizes.

ming behavior extracted from single- and many-cells statistics. Fig. 4.11 shows that this average behavior can be reproduced, as reasonably expected, even neglecting the discrete nature of charge and matter, *i.e.* using 3-D simulations with continuous substrate doping, continuous trap density in the nitride and continuous electron flow from the substrate to the nitride during programming. These simulations were obtained by a numerical tool extending to 3-D geometries the model for charge-trap memory programming that we presented in [26], implementing electron trapping as [101, 132]:

$$\frac{dn_t'(x,y)}{dt} = \frac{J(x,y)}{q}\,\sigma\left[N_t - n_t'(x,y)\right] \tag{4.3}$$

where $n_t'(x,y)$ is the trapped electron density in the nitride. Note that, for a correct comparison of the Monte Carlo and the continuous model, no emissivity from filled electron traps was included in (4.3). Moreover, in the continuous tool cell V_T was obtained from the simulation of cell I_D-V_G transcharacteristics as discussed in Fig. 4.2. Despite a small displacement of the ΔV_T curves, mainly attributed to numerical differences between the two simulation codes, Fig. 4.11 confirms, first of all, the correctness of the Monte Carlo approach for the program operation we presented in Fig. 4.2. Moreover, the agreement between the results of Fig. 4.11 makes possible the use of the continuous 3-D model to investigate the programming performance when only the average results are of interest. However, note that the Monte Carlo model allows a more complete analysis of the program operation, including variability effects representing the topic of the second part of this chapter.

Fig. 4.12 shows a scaling analysis of the ISPP $\overline{\Delta V_T}$ transient, obtained using the continuous 3-D model for cell programming and assuming a reduction of the cell area keeping the same gate stack investigated in the previous sections. A reduction of the ISPP efficiency with cell scaling clearly appears, in terms

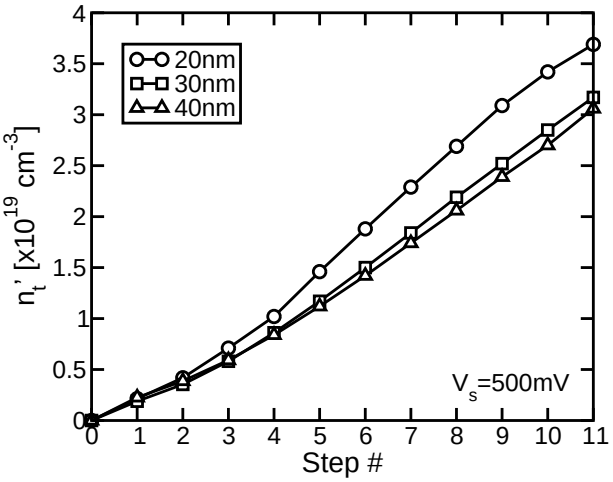


Figure 4.13: Simulated average trapped electron density n_t' in the nitride during ISPP for different cell sizes.

both of slope and of horizontal delay of the curves. This is due to a decreasing impact of electrons stored in the nitride on V_T as cell dimensions are reduced, as discussed in Section 4.2.3. Fig. 4.13 shows, in fact, that the average n_t' curves display a faster electron injection and storage in the nitride as scaling proceeds, due to a larger field enhancement at the corners of the cell area.

4.3 Programming variability

This section presents a comprehensive analysis of the main statistical variability sources affecting the program operation of nanoscale charge-trap memories. Results were obtained by the 3-D TCAD model presented in the first part of this chapter, running Monte Carlo simulations to deal with discrete traps in the storage layer, atomistic doping in the substrate and granular electron injection from the substrate to the storage layer. In so doing, the effect of three main variability sources impacting charge-trap memory programming was investigated: the statistical process ruling electron injection and trapping, the fluctuation in the number and position of the trapping sites and the statistical distribution of the V_T shift (ΔV_T) induced by stored electrons in presence of percolative substrate conduction. We show that discrete electron injection represents the dominant effect for the statistical dispersion of cell ΔV_T during programming of nanoscale charge-trap memories.

4.3.1 Single-cell variability

Programming variability of charge-trap memories was investigated by means of the Monte Carlo simulation tool and the template 18 nm MONOS cell presented in the first part of the chapter. Fig. 4.14a shows some simulated V_T transients

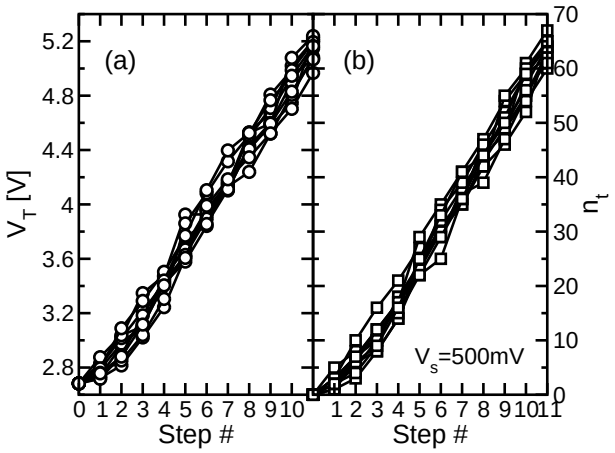


Figure 4.14: Monte Carlo simulation results for the V_T (a) and n_t (b) transients during ISPP with $V_s = 500$ mV and $\tau_s = 10$ μs, on the same stochastic cell. The initial gate bias is $V_G = 13$ V.

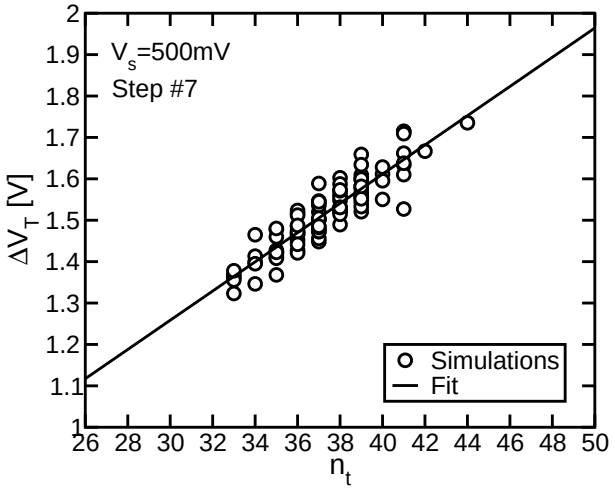


Figure 4.15: Scatter plot for the simulated ΔV_T and n_t achieved after 7 ISPP steps on the same stochastic cell considered in Fig. 4.14.

during ISPP (step amplitude $V_s = 500$ mV, step duration $\tau_s = 10$ μs) on a single cell, having a random configuration of substrate dopants and nitride traps. Despite cell configuration is fixed, a statistical dispersion of V_T clearly appears during programming, due to the statistics of electron injection and capture in the nitride traps. Similarly to the case of floating-gate cells [21,22,122,123,126], this process introduces fluctuations in the number of stored electrons (n_t) at the end of the ISPP steps, as shown in Fig. 4.14b. Note, however, that in the case of charge-trap memories the fluctuation arises from a more complex physics,

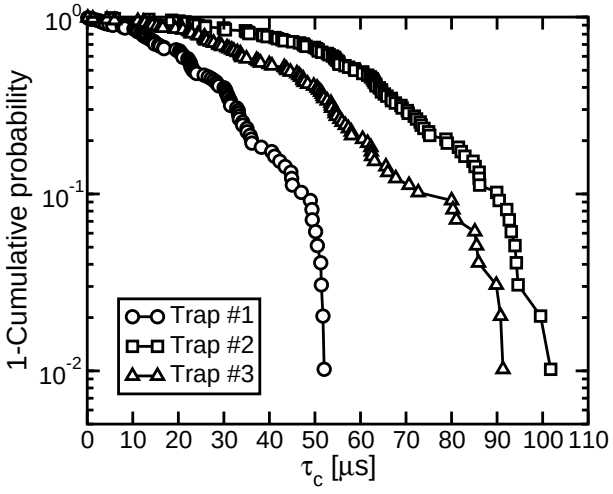


Figure 4.16: Simulated statistical distribution of τ_c for three traps of the MONOS cell investiagted in Fig. 4.14.

involving the statistical charging of each single trap in the nitride. The discrete and localized nature of traps makes not only the time but also the spatial distribution of the electron injection events relevant, leading to an important difference with respect to floating-gate devices. Despite the first part of this chapter has shown that traps are on average filled starting from the active area edges along W and half-way between source and drain along L, there is a non-negligible statistical dispersion of the time τ_c at which each single trap captures an electron. In other words, the sequence in time of trap filling adds further fluctuations. This is confirmed by Fig. 4.15, showing ΔV_T as a function of the n_t after 7 ISPP steps: at constant n_t obtained from the Monte Carlo simulations of the program operation, a different ΔV_T may result. This is due to the different locations of captured electrons, resulting in a different ΔV_T as discussed in chapter 3. Note, however, that the ΔV_T dispersion at fixed n_t is significantly lower than that caused by the n_t spread, thus confirming that the statistical process ruling electron injection and capture is the main variability source for program in the single-cell case.

The statistical distribution of τ_c for each nitride trap is not purely exponential, due to two main reasons. The first is the rise of the gate bias V_G as ISPP proceeds, increasing the tunneling current density $J(x, y)$ at the substrate surface and, in turn, reducing the average value of the electron injection time $\overline{\tau}_{inj}$ towards each trap. This contributes to an hyper-exponential bahavior of τ_c, with a smaller probability of long τ_c with respect to the pure exponential case, as clearly appearing in Fig. 4.16. This figure shows the statistical distribution of the time τ_c at which three of the traps of the stochastic cell investigated in Fig. 4.14 capture their electron during the ISPP transient, as obtained from many Monte Carlo simulations of the program operation, highlighting that the increase of the gate bias during programming forces trapping to take place. Be-

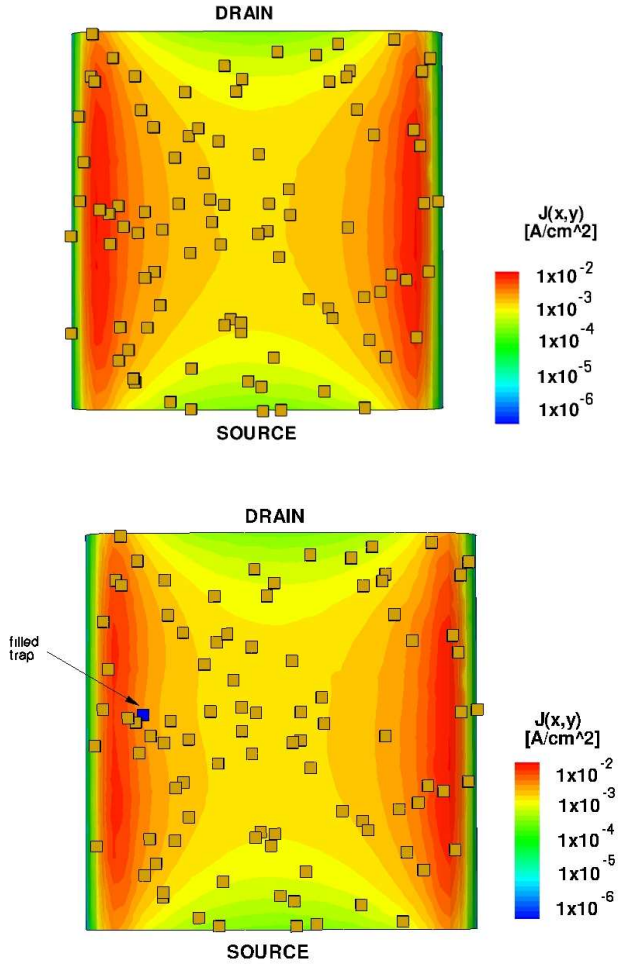


Figure 4.17: Tunneling current density $J(x, y)$ at the channel surface of the stochastic cell investigated in Section 4.3.1, in the case of empty nitride (top) and when a single electron is stored in the nitride (bottom). The gate voltage is $V_G = 13$ V. Squares are trap positions in the nitride.

sides this effect, deviations of the τ_c distribution from the exponential behavior are also due to electron storage in neighboring traps. When an electron is stored in the nitride its negative charge gives rise to a reduction of the bottom-oxide electric field and of $J(x, y)$ in its close proximity. This is shown in Fig. 4.17, where $J(x, y)$ at $V_G = 13$ V is reported in the case nitride traps are all empty (top) and when an electron is stored in a single trap (bottom). A deformation of the contour plot appears near the negatively charged trap, highlighting an increase of $\overline{\tau}_{inj}$ in these regions. The variation of $\overline{\tau}_{inj}$ for the traps in Fig. 4.17 that results from the single electron storage is quantitatively shown in Fig. 4.18 as a function of traps position along W (left) and L (right). A clear local ef-

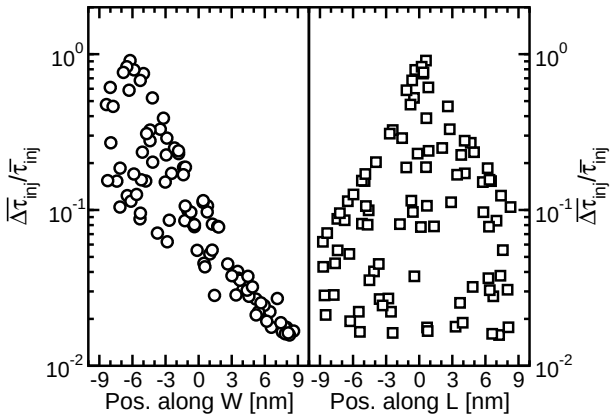


Figure 4.18: $\overline{\tau}_{inj}$ increase for the traps of the cell of Fig. 4.17 that is determined by the storage of a single electron in the nitride (results are normalized to the $\overline{\tau}_{inj}$ value for the case of empty nitride), as a function of trap position along W (left) and L (right). $(0, 0)$ is the channel center.

fect appears, with a significant $\overline{\tau}_{inj}$ increase only for those traps located close to the electron storage point. This effect represents a non-negligible difference between the statistical charging process of charge-trap and floating-gate memories. When a single electron is stored in a floating-gate cell, in fact, the electric field feedback reducing the tunneling current from the substrate acts over the whole active area, resulting into a global increase of the average time required for the next electron injection [21, 22]. Figs. 4.17-4.18 show, instead, that the local nature of the field feedback in the case of charge-trap memories makes traps close to the first electron storage position to largely increase their $\overline{\tau}_{inj}$, while traps far from it to have their $\overline{\tau}_{inj}$ barely modified. This makes the field feedback effect more complex to investigate than in the case of floating-gate cells, preventing a simple extension of the analysis presented in [22] and making Monte Carlo 3-D simulations the most suited approach for the programming spread analysis of charge-trap devices.

4.3.2 Many-cells variability

Fig. 4.19 shows some simulated programming transients on stochastically different 18 nm MONOS cells, as obtained from the Monte Carlo model presented in the first part of this chapter. Cells have different configurations of atomistic dopants in the substrate and of traps in the nitride, obtained drawing their number and position from a Poisson and a uniform distribution, respectively. Atomistic doping gives rise, first of all, to a statistical dispersion of neutral (empty nitride) cell V_T, corresponding to the V_T value at step number 0 in the figure. This is due to the randomness introduced by atomistic doping in the non-uniform substrate inversion, leading to the possibility for cells to reach the threshold condition with stochastically different current percolation paths between source and drain [45, 51–55, 104, 105, 107].

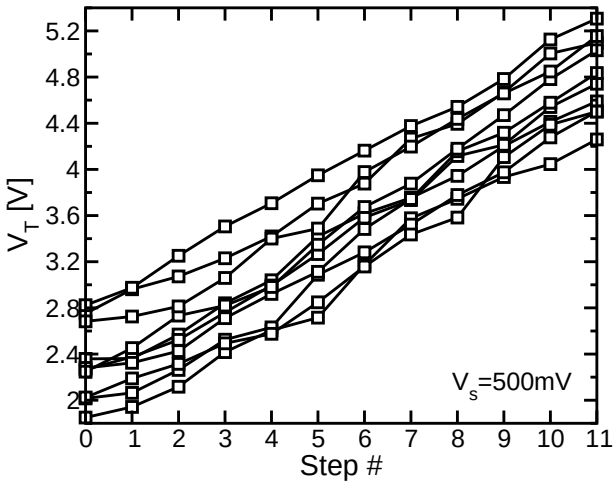


Figure 4.19: Monte Carlo simulation results for the V_T transients during ISPP (V_s = 500 mV, $\tau_s = 10$ μs, initial $V_G = 13$ V) on stochastically different 18 nm MONOS cells.

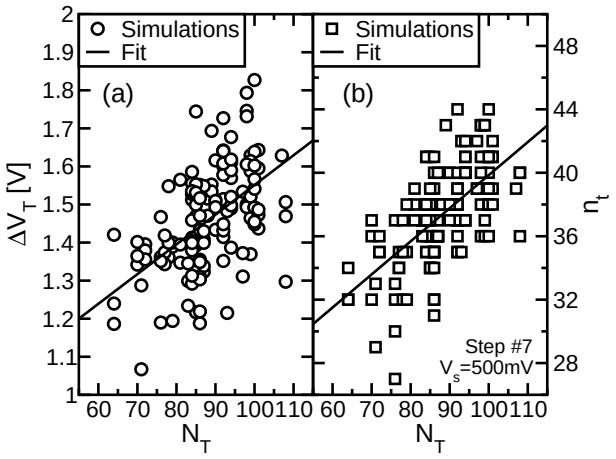


Figure 4.20: Simulated ΔV_T (a) and n_t (b) achieved after 7 ISPP steps on stochastically different cells, as a function of the number of traps in the cell (N_T).

In addition to the dispersion of neutral cell V_T, Fig. 4.19 highlights statistical fluctuations in the V_T increase as programming proceeds, similarly to Fig. 4.14. For a comprehensive analysis of these fluctuations in the case many stochastically different cells are considered, it should be noted that besides the electron injection statistics and the randomness of the ΔV_T resulting from electron storage in traps placed at different spatial positions over the channel, the statistical dispersion of the number of traps N_T in the nitride represents an additional spread source for programming. In order to investigate this point, Fig. 4.20

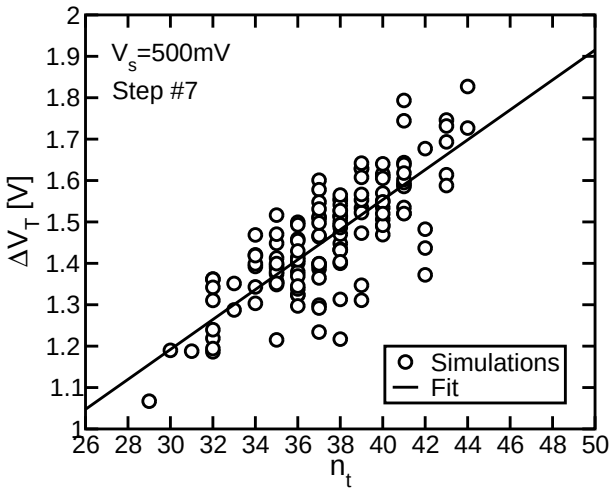


Figure 4.21: Same as in Fig. 4.15, but for many stochastically different cells.

shows, for each cell, the simulated ΔV_T (a) and n_t (b) achieved after 7 ISPP steps as a function of N_T. Despite the correlation leading to a larger ΔV_T and n_t for cells having higher N_T (see the fit lines), the wide spread of the points reveals that the fluctuation of the initial number of traps is not a main source of variability during ISPP. Moreover, Fig. 4.21 directly shows the scatter plot for ΔV_T as a function of n_t after 7 ISPP steps. Despite the vertical dispersion of the points is larger than that of Fig. 4.15, due to the additional randomness given by the change in the atomistic dopants configurations in the substrate when different stochastic cells are considered, this contribution is however still much lower than the total ΔV_T dispersion in the graph. The result highlights that the statistical spread coming from the different locations of trapped electrons over the channel plays a minor role in the total ΔV_T variability.

In summary, the results of Figs. 4.20-4.21 show that the statistics of electron injection and capture in the nitride traps is the dominant spread source for charge-trap based nanoscale cells, with minor contributions coming from the statistical fluctuations of trap number and positions in the nitride and from atomistic doping and percolative substrate conduction during read. Note, however, that this conclusion relies on considering the ISPP transients far from the saturation of all the available nitride traps. If the program operation is extended closer to the limit when all the available traps are filled, the resulting ΔV_T spread would be mainly determined by the N_T spread.

4.3.3 Sub-poissonian nature of the charge injection statistical process

In order to further confirm that the statistics of electron injection and capture represents the dominant variability source for the program operation of nanoscale charge-trap memories, the cumulative distributions of ΔV_T from single-

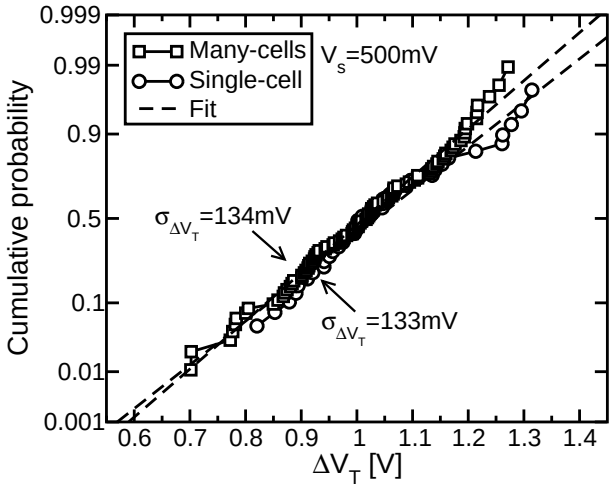


Figure 4.22: Simulated statistical distribution of ΔV_T evaluated between step 7 and step 3 of the ISPP algorithm, in the case of single-cell (*i.e.*, fixed trap configuration) and many-cells (*i.e.*, variable trap configuration) statistics. Nearly 100 Monte Carlo simulations were used to gather the statistical results.

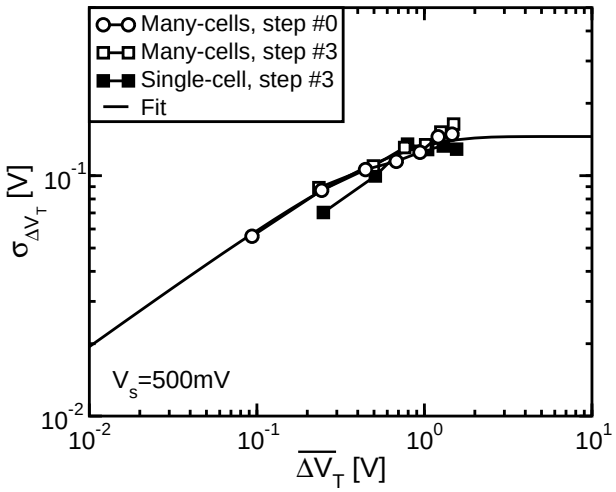


Figure 4.23: Simulated $\sigma_{\Delta V_T}$ vs. $\overline{\Delta V_T}$ results from single- and many-cells statistics. Step #0 or #3 has been used to evaluate ΔV_T.

cell and many-cells statistics are compared in Fig. 4.22. ΔV_T was calculated as the V_T shift between step 7 and step 3 of the ISPP transients shown in Fig. 4.14 and Fig. 4.19 for the single- and the many-cells case, respectively. Note that the choice not to evaluate ΔV_T with respect to the neutral cell state (*e.g.*, V_T at step 7 minus V_T at step 0) for this comparison is required by the deterministic initial condition of the single-cell transients. In fact, even if both single-cell and

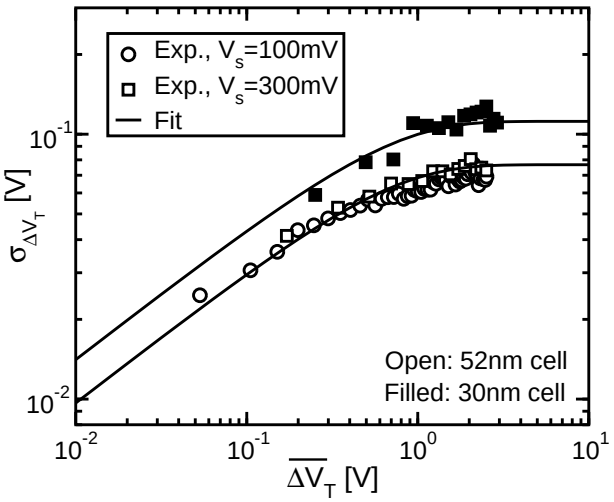


Figure 4.24: Experimental $\sigma_{\Delta V_T}$ vs. $\overline{\Delta V_T}$ from many ISPP ramps on the same charge-trap memory cell. Results are shown for a 52×52 nm^2 and a 32×32 nm^2 cell, using ISPP ramps with $V_s = 300$ mV and 100 mV.

many-cells Monte Carlo ISPP simulations start from a neutral nitride condition, in the former case trap configuration in the nitride is fixed, while it is random in the latter case. As a consequence, the statistical dispersion of ΔV_T from many-cells statistics would surely be larger than that from single-cell simulations if ΔV_T were evaluated with respect to step 0. In order to obtain a meaningful comparison of ΔV_T, therefore, this has to be evaluated between steps in the central part of the ISPP transients, *i.e.* considering a reference step for the ΔV_T evaluation after some trapping has already taken place, to have a random configuration of empty traps at the beginning of the ISPP steps used for ΔV_T evaluation in both cases.

Fig. 4.22 shows that similar ΔV_T spread ($\sigma_{\Delta V_T}$) values are obtained from single- and many-cells statistics, confirming that the additional randomness introduced by the fluctuation of N_T, trap positions and atomistic doping, when different cells are considered, plays a minor role with respect to the electron injection and trapping statistics. This is also shown in Fig. 4.23, where $\sigma_{\Delta V_T}$ is reported as a function of the average ΔV_T ($\overline{\Delta V_T}$). Curves are obtained by cumulating some ISPP steps starting from a reference one, namely step #0 and step #3 in the figure. Results are quite similar independently of being collected from single-cell or from many-cells statistics. They look, moreover, like those previously reported for floating-gate devices [21, 22, 122]. In fact, for low-$\overline{\Delta V_T}$, results reveal that the electron injection process behaves as poissonian, with $\sigma_{\Delta V_T}$ depending on the square root of $\overline{\Delta V_T}$. This is due to the negligible effect of field-feedback in the bottom oxide when a small number of electrons are stored in the nitride layer, making electron injections events almost uncorrelated. However, for increasing $\overline{\Delta V_T}$, the field-feedback becomes more and more important and gives rise to a saturation of $\sigma_{\Delta V_T}$, with the electron injection

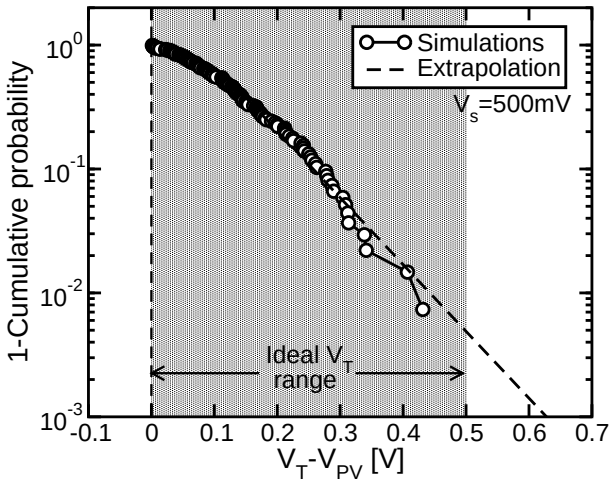


Figure 4.25: Simulation results for the cumulative distribution of the programmed V_T obtained after ISPP with $V_s = 500$ mV and assuming a verify level at $V_{PV} = 3.7$ V. Results are obtained from many-cells statistics.

process becoming sub-poissonian [21, 22, 122]. The whole $\sigma_{\Delta V_T}$ vs. $\overline{\Delta V_T}$ curve may be described by the following relation, originally derived for floating-gate cells [22]:

$$\sigma_{\Delta V_T} = \sqrt{\frac{q}{\gamma C_{NG}} \left(1 - e^{-\gamma \overline{\Delta V_T}}\right)}. \tag{4.4}$$

where q is the electron charge and C_{NG} is the 3-D electrostatic capacitance between the gate and the nitride relating the average number of trapped electrons $\overline{n}_t$ with $\overline{\Delta V_T}$, as discussed in the first part of this chapter. The parameter γ in (4.4) quantifies the field-feedback exerted by electron trapping in the cell. Despite this parameter has a well-defined expression in the case of floating-gate devices [22], for nitride cells it has been used as a fitting parameter, due to the strong differences in the field-feedback process in the case of charge-trap memories, as discussed in Section 4.3.1. By fitting the simulation results of Fig. 4.23, it turns out $\gamma = 1.8$ V^{-1}.

As a validation of the simulation results for the programming spread, Fig. 4.24 shows that the same $\sigma_{\Delta V_T}$ vs. $\overline{\Delta V_T}$ behavior is experimentally observed on 52 and 30 nm charge-trap cells. Results were obtained from many ISPP operations on the same single-cell, in the case of $V_s = 300$ mV and $V_s = 100$ mV. Moreover, the comparison of the results from samples with different dimensions reveals the increase of $\sigma_{\Delta V_T}$ as scaling proceeds. This is due to the more severe variability introduced by the electron injection process in the program operation as the number of electrons transferred from the substrate to the nitride decreases. This is accounted for in (4.4) by the reduction of C_{NG} as cell size decreases [133].

4.3.4 Accuracy limitations to the programmed V_T

ΔV_T variability during ISPP represents a severe limitation to the accuracy of the program operation, even if this is accomplished by a program-and-verify algorithm [20,21,123]. To obtain high programming accuracies, these algorithms make use of a verify operation after each ISPP step, comparing cell V_T with a selected program-verify level V_{PV} and determining the end of the program operation if the condition $V_T > V_{PV}$ is met; otherwise, another ISPP pulse is applied. In the case of negligible programming variability, this algorithm assures all cells fall between V_{PV} and $V_{PV}+V_s$, thus making possible high programming accuracies if small V_s are adopted (see also chapter 1). However, programming variability compromises this possibility, as first reported on floating-gate Flash arrays [21, 122, 123, 133]. In fact, if the ΔV_T resulting by each ISPP step is statistically distributed, cells may show a V_T shift larger than V_s at the last ISPP step, then overcoming V_{PV} by more than V_s. This effect reduces the tightness of the programmed V_T distribution with respect to the ideal case, introducing design constraints for multi-level devices.

Fig. 4.25 shows the simulated V_T distribution from the many-cells ISPP transients at $V_s = 500$ mV when $V_{PV} = 3.7$ V. Extrapolations show that the distribution exceeds $V_{PV} + V_s$ with a quite high probability of 5×10^{-2}, due to the large electron injection spread of Fig. 4.23. This effect leads to a more severe limitation of the programming accuracy considering that the average V_T increase per step during ISPP is only $\simeq 0.5V_s$ in the case of our simulated charge-trap cells, as largely discussed in the first part of this chapter. This means that a correct evaluation of the programming accuracy, taking into account also the time required to accomplish the program operation, should compare the resulting V_T distribution with $V_{PV} + 0.5V_s$ in Fig. 4.25. This assesses programming variability as a serious constraint for nanoscale charge-trap memories, strongly limiting the tightness of the V_T distribution which, instead, is rather immune to parasitic cell-to-cell interference [125].

4.4 Conclusions

This chapter presented a detailed simulation analysis of charge-trap memory programming, carefully reproducing the discrete and localized nature of storage traps and the statistical process ruling granular electron injection into the storage layer. The average results for ISPP on single- and many-cells statistics revealed that the low programming efficiency of nanoscale charge-trap cells mainly results from the low impact of locally stored electrons on cell V_T in presence of fringing fields and 3-D electrostatics.

A comprehensive 3-D TCAD Monte Carlo investigation of the main variability sources affecting the ISPP operation of nanoscale charge-trap memories was also presented in the second part of the chapter, comparing the effect of the statistical electron injection and trapping process, the fluctuation in the number and position of the trapping sites and the statistical distribution of the V_T shift induced by stored electrons in presence of percolative substrate conduction. Results showed that discrete electron injection is dominant over the other

variability sources, due to a lower and lower number of electrons controlling cell state as scaling proceeds.

Chapter 5

Doping Engineering for RTN suppression in Flash memories

This chapter presents a thorough numerical investigation of the effect of non-uniform doping on random telegraph noise in nanoscale Flash memories. For fixed average threshold voltage, the statistical distribution of the random telegraph noise fluctuation amplitude is studied with non-constant doping concentrations in the length, width or depth direction in the channel, showing that doping increase at the active area corners and retrograde and δ-shape dopings appear as the most promising profiles for random telegraph noise suppression. In particular, the improvements offered by retrograde and δ-shape dopings increase the more the high doping regions are pushed far from channel surface, thanks to a more uniform source-to-drain conduction during read. Finally, the suppression of random telegraph noise by engineered doping profiles is correlated with the reduction of cell threshold-voltage variability.

5.1 Introduction

THE statistical dispersion of the amplitude of random telegraph noise (RTN) fluctuations in deca-nanometer MOS devices has been clearly recognized in the last years and attributed to the localized nature of electrons trapped into oxide defects close to the channel surface in presence of three-dimensional (3-D) electrostatics and atomistic doping [62, 76, 105, 108, 109, 134–137]. In particular, the threshold-voltage shift (ΔV_T) given by single RTN traps was shown

to follow an exponential distribution in [76, 105, 108], with standard deviation proportional to the square root of the channel doping concentration N_a when a uniform doping profile is adopted. These results reveal that channel doping is one of the most important parameters for RTN in MOS devices, opening the possibility for technology optimizations by engineered doping profiles.

In section 3.3.2 we have seen that 3-D electrostatics and atomistic substrate doping play the main role in determining RTN statistics also for nitride-based memories, with a marginal role of electrons *locally* stored in the nitride.

With this in mind, and considering that no comprehensive doping optimization studies are reported in literature for the floating-gate technology (which is the current technology in production), we will carry out our investigation on the effect of non-uniform channel doping on RTN, using a general template floating-gate Flash memory cell. By means of 3-D numerical simulations accounting for the atomistic nature of doping and using a Monte Carlo procedure to gather statistical results, the magnitude of RTN fluctuations is evaluated with the doping concentration changing in the length (L), width (W) or depth direction in the channel region. Results reveal that for fixed average threshold voltage (V_T), non-constant doping concentrations in the L or W direction deviate the RTN distribution from a pure exponential behavior. In particular, the increase of the doping concentration at the source side of the channel enlarges the statistical distribution of RTN amplitudes at low probabilities, while a narrowing appears when doping is increased near the shallow trench isolations (STI) corners of the active area. A reduction of the RTN magnitude can also be achieved when the doping concentration near the channel surface is reduced, as shown referring to retrograde and δ-shape profiles. In these two latter cases, the exponential distribution of RTN amplitudes is shown to be preserved and, more important, narrower and narrower distributions are obtained as the high doping regions are pushed deeper and deeper in the substrate. This comes from a more uniform potential profile at the channel surface, leading to a more uniform electron concentration and source-to-drain conduction during read, when the Coulomb peaks due to ionized dopants are far from the channel surface. Finally, the effect of engineered dopings on RTN and V_T variability is compared, highlighting their clear correlation.

5.2 Numerical model

Fig. 5.1 shows the template Flash cell investigated in this work, featuring an 8 nm tunnel oxide, a 70 nm polysilicon floating gate and a 4-3-5 nm oxide-nitride-oxide (ONO) stack for the interpoly dielectric. Cell W and L were set to 32 nm and doping was implemented atomistically in the channel region, as discussed in chapter 3. Different types of non-uniform doping profiles were considered, changing the doping concentration in the L, W or depth direction in the channel. A constant continuous doping equal to $xx \times 10^{20}$ cm^{-3} was used for the source/drain n^+ regions.

In order to investigate the effect of the channel doping profile on the statistical distribution of RTN amplitudes, a Monte Carlo procedure was used. First, a large number of atomistically different cells were generated for each se-

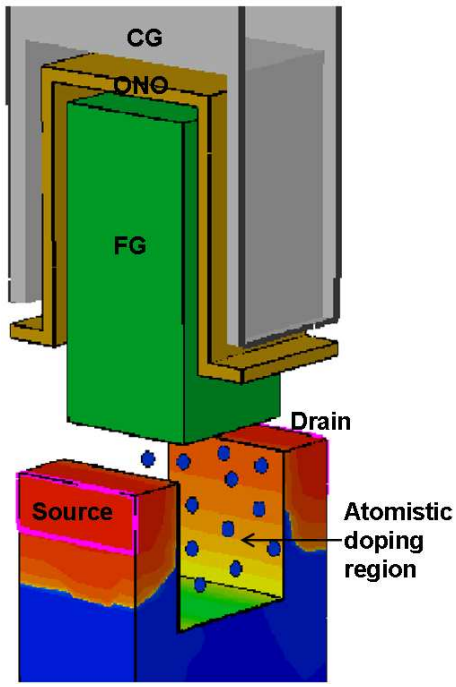


Figure 5.1: Schematics for the template floating-gate Flash cell investigated in this work. The atomistic doping region is highlighted.

lected profile by replacing the continuous doping with discrete atoms, drawing their number from the Poisson statistics and choosing their spatial distribution according to the continuous profile. Then, the drain current vs. control-gate voltage ($I_D - V_{CG}$) transcharacteristics was calculated for each cell by solving the Poisson and drift-diffusion equations for drain bias $V_D = 0.7$ V and increasing V_{CG}, defining cell V_T as the V_{CG} value corresponding to $I_D = 100$ nA. Finally, a single electron was randomly placed over the channel of each atomistic device to reproduce the filled-state of an RTN trap, calculating again cell $I_D - V_{CG}$ and the V_T shift corresponding to the RTN fluctuation amplitude ΔV_T.

All the 3-D numerical simulations were performed by means of the commercial software SDevice [138],

5.3 Simulation results

Fig. 5.2 shows the non-constant doping profiles investigated in this work and compared with a reference uniform case of value $N_a = 2 \times 10^{18}$ cm^{-3}. For a meaningful comparison, the doping concentrations of Fig. 5.2 were chosen to obtain the same average neutral V_T in all cases.

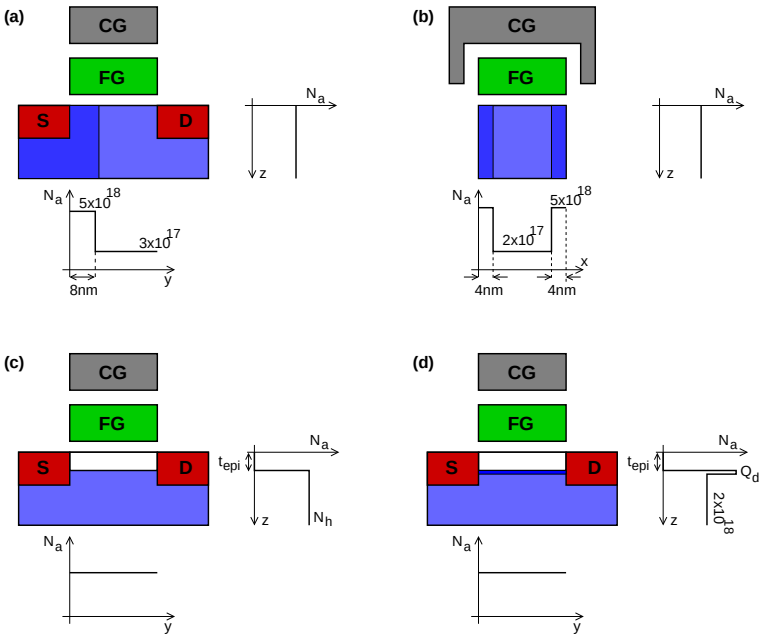


Figure 5.2: Schematics for the doping profiles investigated in this work: non-uniform doping along L (a), non-uniform doping along W (b), retrograde doping (c) and δ-shape doping (d). Case (b) shows cell cross-section in the W direction (x coordinate), all other cases in the L direction (y coordinate). All the doping concentrations are in cm^{-3}. Values of N_h, Q_d and t_{epi} for case (c) and (d) will be given in Tab. 5.1.

5.3.1 Doping variations along L and W

Non-uniform doping along L was explored considering the profile of Fig. 5.2a: the doping concentration has a step increase from 3×10^{17} to 5×10^{18} cm^{-3} in the last 8 nm of the channel near the source junction. This profile can be considered representative of the effect of a halo implant near the source and aims at largely reducing the doping concentration in a wide portion of the channel area, while increasing it only in a small region near the junction. Fig. 5.3 shows the simulation results for the ΔV_T statistical distribution: while a clear exponential behavior is obtained for the uniform reference device [76,105,108], a double slope in the semi-log plot appears for the distribution of the non-uniform doping profile. In particular, this latter distribution displays a slope higher than that of the uniformly-doped sample in the low-ΔV_T regime, but lower than it at high-ΔV_T. As a consequence, the RTN distribution for the profile of Fig. 5.2a results narrower at high probabilities than the uniform doping case, but enlarges more than it at low probability levels, which are of more interest for technology reliability.

Results of Fig. 5.3 can be explained considering the cell of Fig. 5.2a as the series of two devices having different local V_T. The higher doping concentration at the source side of the channel increases, in fact, the control-gate bias required

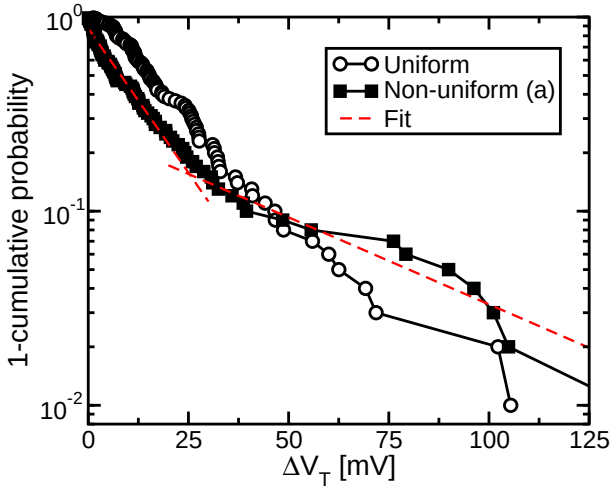


Figure 5.3: Simulation results for the ΔV_T statistical distribution of the non-uniform doping profile of Fig. 5.2a, compared to that of the reference uniform sample.

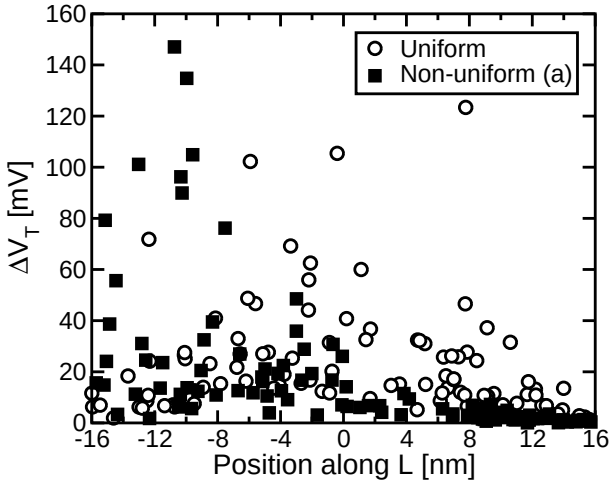


Figure 5.4: Simulated ΔV_T as a function of trap position along L for the non-uniform doping sample of Fig. 5.2a and for the uniform reference device. Channel center, source and drain are at 0, -16 and 16 nm, respectively.

to invert this region with respect to the rest of the channel which, in turn, has a low control on the source-to-drain conduction. As a consequence, RTN traps placed over the low-doped region give rise to lower ΔV_T than traps placed at the same positions in the reference device, while higher ΔV_T result for traps over the high-doped channel area near the source. This is clearly shown in Fig. 5.4, where the ΔV_T obtained from the Monte Carlo procedure are reported as a function of trap position along L, in the case of the uniform and non-uniform

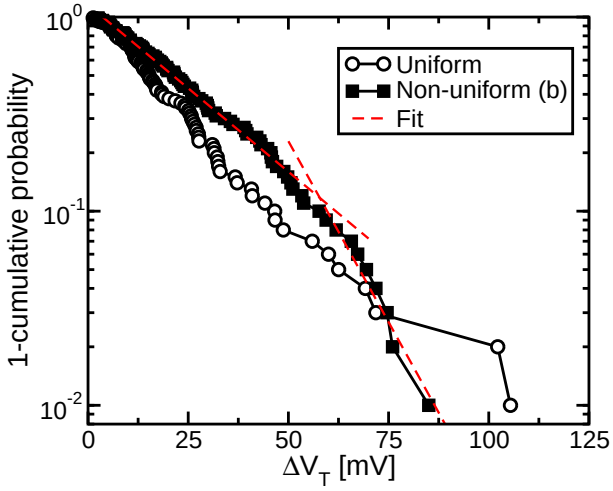


Figure 5.5: Same as in Fig. 5.3, but for the profile of Fig. 5.2b.

device. As traps are uniformly placed over the channel, leading to a higher probability to have them over the low-doped region than over the high-doped region of the sample of Fig. 5.2a, the reduction of the RTN amplitude of traps over the former region results in the narrowing of the ΔV_T distribution at high probability levels with respect to the reference sample. At low probabilities, instead, the ΔV_T distribution is mainly determined by traps over the high-doped area of the channel, resulting flatter than the uniform case due to the larger impact of traps over this region on V_T. As a result, the ΔV_T distribution crosses that of the uniform sample, making the investigated non-uniform profile disadvantageous from the RTN standpoint.

Fig. 5.5 shows the ΔV_T statistical distribution for the non-uniform doping profile of Fig. 5.2b. An increase of the doping concentration from 2×10^{17} to 5×10^{18} cm^{-3} was assumed in this case in the last 4 nm of the active area near the STI edges, with no changes in the L and depth direction. This doping profile aims at increasing the local V_T at the cell corners, reducing, in turn, current crowding that occurs in these regions due to field intensifications in the 3-D electrostatics [76, 105, 108]. Fig. 5.5 shows that the resulting ΔV_T distribution is slightly larger than the reference uniform sample at high probabilities, but narrows rapidly at low probability levels. This can be explained considering that in the uniform cell high ΔV_T mainly result from traps at the corners of the active area, where strong source-to-drain percolation paths exist. This is shown in Fig. 5.6, where the ΔV_T from the Monte Carlo procedure are shown as a function of trap position along W. The same figure shows that the non-uniform doping profile of Fig. 5.2b is highly effective in reducing the ΔV_T of traps near the STI corners, thanks to the reduction of current crowding in these regions. This narrows the ΔV_T distribution in Fig. 5.5 at low probability levels with respect to the uniform doping case, leading to clear benefits in terms of RTN impact on technology reliability.

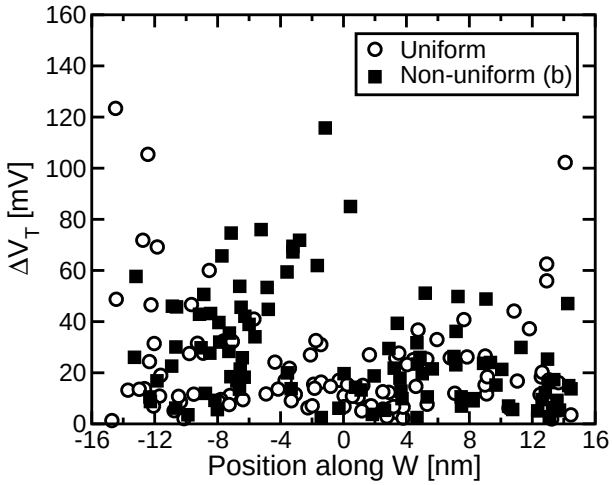


Figure 5.6: Simulated ΔV_T as a function of trap position along W for the non-uniform doping sample of Fig. 5.2b and for the uniform reference device. 0 is channel center.

Doping Type	$\mathbf{t_{epi}}$ [nm]	$\mathbf{N_h}$ [cm^{-3}]	$\mathbf{Q_d}$ [cm^{-2}]
Retrograde	2.5	2.5×10^{18}	–
Retrograde	5	3.4×10^{18}	–
Retrograde	10	5.7×10^{18}	–
Retrograde	16	1×10^{19}	–
δ-shape	10	–	3.5×10^{12}
δ-shape	16	–	7.5×10^{12}

Table 5.1: Parameters for the retrograde and δ-shape doping profiles (Figs. 5.2b-5.2c) investigated in this work.

5.3.2 Vertically non-uniform dopings

In order to explore the possibility for vertically non-uniform doping profiles to statistically suppress the amplitude of RTN fluctuations, we considered the simplified retrograde and δ-shape dopings of Figs. 5.2c and 5.2d. Different thicknesses t_{epi} of the undoped epitaxial region were investigated, as reported in Tab. 5.1, changing N_h and Q_d to set the same average V_T of the uniform reference sample. Figs. 5.7 and 5.8 show the simulated ΔV_T statistical distributions: differently from the non-uniform profiles considered in the previous section, a clear exponential behavior is preserved for both the dopings, with only a reduction of the distribution slope λ (units: [mV/dec]) with respect to the reference device. From Figs. 5.7-5.8, the reduction of λ is larger when increasing t_{epi}, as quantitatively shown in Fig. 5.9. From this figure, a stronger suppression of RTN appears also for the retrograde than for the δ-shape doping when t_{epi} is low, while nearly the same benefits are obtained when high t_{epi} are adopted.

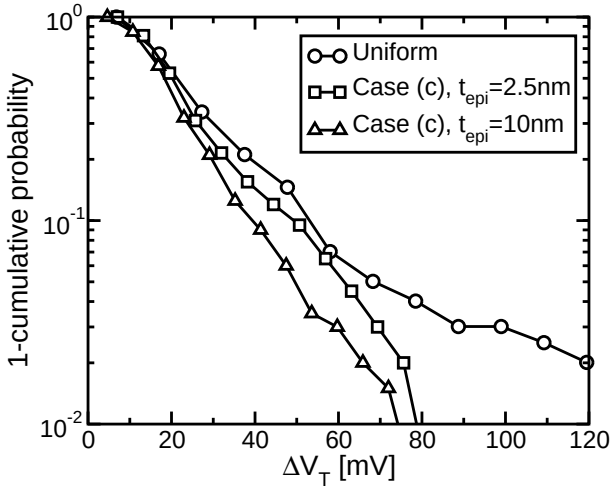


Figure 5.7: Same as in Fig. 5.3, but for the profile of Fig. 5.2c. Results for different t_{epi} are shown.

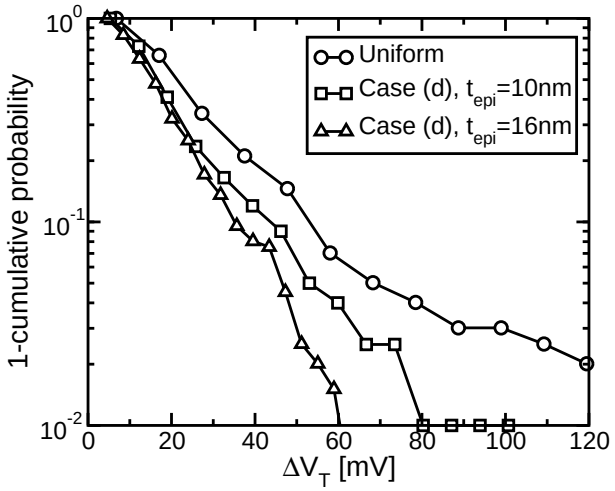


Figure 5.8: Same as in Fig. 5.3, but for the profile of Fig. 5.2d. Results for different t_{epi} are shown.

The effectiveness of retrograde and δ-shape dopings in reducing the amplitude of RTN fluctuations can be explained considering that dopants placed near the channel surface represent a major source of percolative conduction during read, giving traps over points where current crowding occurs the possibility to result into quite large ΔV_T [62, 76, 105, 108]. The absence of dopants in the epitaxial layer of the devices of Figs. 5.2b and 5.2c allows, therefore, the benefit of a more uniform source-to-drain conduction, which is less affected by locally

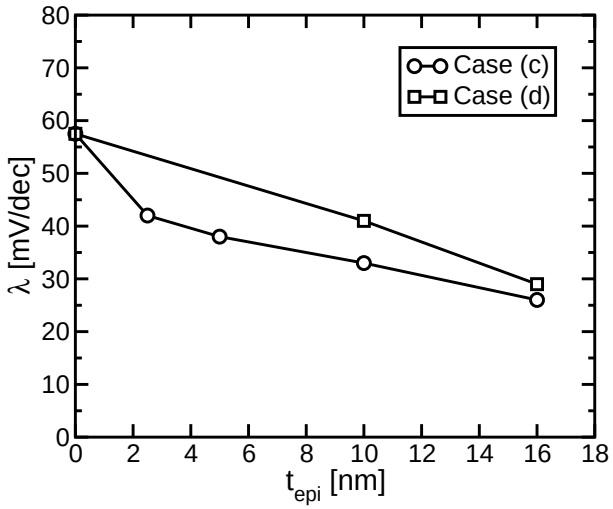


Figure 5.9: Slope of the ΔV_T distribution for the retrograde and δ-shape dopings of Figs. 5.2b-5.2c, as a function of t_{epi}.

stored electrons in RTN traps, resulting into a decrease of λ. This is far more true when t_{epi} is increased, as confirmed by the results of Fig. 5.9, because the larger distance of the ionized dopants from the silicon/oxide interface reduces the impact of their Coulomb field on surface potential. This explains also why, for small t_{epi}, a retrograde profile allows a lower λ than a δ-shape doping in Fig. 5.9: the latter profile, in fact, maximizes the number of ionized dopants at the interface of the doped region with the epitaxial layer, *i.e.*, at the minimum distance from the channel surface, resulting into a more percolative conduction than retrograde doping. For large t_{epi}, instead, the two doping profiles result in nearly the same λ, as the increase of Q_d and N_h required by Tab. 5.1 makes the depletion layer width nearly equal to t_{epi}, with all the ionized dopants placed at the same distance from the channel surface.

5.4 Correlation between RTN and V_T variability

Besides being a major source of statistical dispersion for ΔV_T, atomistic doping gives a fundamental contribution to V_T variability in nanoscale MOS devices [45,51–53,107,139]. The statistical fluctuation of the number and position of dopants in the channel introduces, in fact, a spread σ_{V_T} in the V_T of nominally identical devices, growing with technology scaling as $N_a^{0.4}/\sqrt{WL}$ [53]. The possibility to reduce this spread by engineered doping profiles has been already shown [46, 48, 77], highlighting that large reductions of σ_{V_T} can be obtained when retrograde or δ-shape dopings are adopted. This is confirmed by our Monte Carlo results in Fig. 5.10, showing that the V_T distribution of the retrograde (left) and the δ-shape (right) dopings with $t_{epi} = 10$ nm are tighter than that of the reference uniform sample. Note, moreover, that similarly to

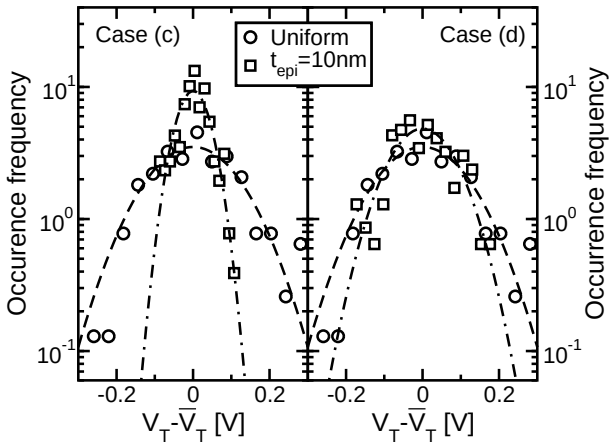


Figure 5.10: Simulation results for the V_T statistical distribution of retrograde (left) and δ-shape (right) dopings, in the case of $t_{epi} = 10$ nm. Results for the reference uniform device are also ahown for comparison.

what shown in Fig. 5.9 for the slope λ of the RTN exponential distribution, the reduction of σ_{V_T} is larger for the retrograde than for the δ-shape doping.

Fig. 5.11 shows σ_{V_T} as a function of λ for all the cases of retrograde and δ-shape doping investigated in this work: a very good correlation of the two parameters clearly appears, confirming the strong connection between V_T variability and RTN. This not only reveals that workable solutions to reduce the impact of both the reliability issues on Flash technology can be devised, but shows engineered channel doping and non-uniform vertical profiles as one of the most practical ways for technology optimization.

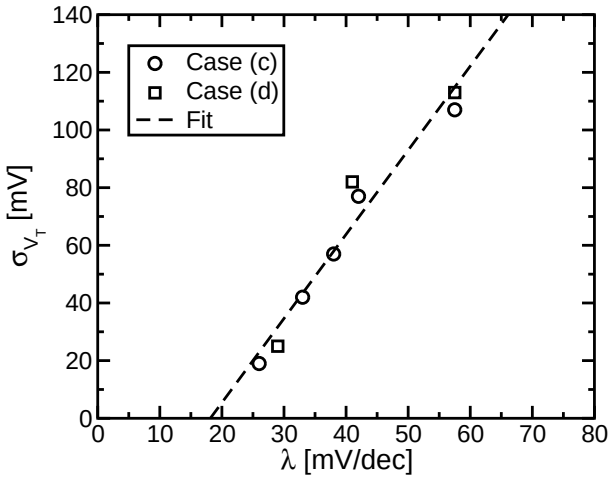


Figure 5.11: σ_{V_T} as a function of λ for the retrograde and δ-shape doping profiles investigated in this work.

5.5 Conclusions

This Chapter presented a detailed numerical investigation of the effect of non-uniform channel doping on RTN in nanoscale Flash memories. The RTN amplitude distribution was shown to enlarge at low probabilities when increasing the doping concentration at the source side of the channel, while its narrowing appears when higher doping concentrations are used at the STI corners. Clear benefits from the RTN standpoint were shown when retrograde and δ-shape dopings are adopted, with improvements that increase the more the high doping regions are pushed far from the channel surface. Finally, RTN suppression by doping engineering was shown to be strongly correlated to the reduction of cell V_T variability.

Conclusions

THIS thesis deals with the modeling of statistical variability affecting the program and reading operation of nanoscale charge-trap Flash memories.

We have presented (Chapter 1) the charge-trap technology as one of the most promising solution to the floating-gate scaling issues. We have furthermore seen that the electrical behavior of the charge-trap memory device is affected by several statistical variability sources, among which the most important are the the random dopant fluctuation, the random trap fluctuation, the discrete charge injection process, the random telegraph noise fluctuation.

In order to study these statistical variability effects, charges (e.g. ionized dopants in channel or trapped electrons in the storage layer) have to be treated as discrete entities (Chapter 2). This represents a non-trivial issue in the framework of TCAD simulation, where a drift-diffusion formalism is usually employed to study the carrier conduction. Indeed, the Coulomb potential associated with each discrete charge becomes physically inconsistent with the concepts of electrostatic potential presumed in drift-diffusion device simulations. It has been pointed out that to cut off the short-range part of the Coulomb potential associated with each discrete dopant is essential in correctly simulating the device properties under the atomistic regime. To this aim, three dopant models, namely (i) the smearing method, (ii) the Sano model and (iii) the quantum correction, have been introduced showing their range of validity and their limitations. In this context we have proposed a new mobility model for atomistic simulation in order to improve the accuracy and reliability of the quantum correction method.

A comprehensive investigation of statistical variability in deeply-scaled nitride memory cells has been then presented in the remainder of the thesis.

The study of the threshold voltage shift variability (Chapter 3) has highlighted that 3-D electrostatics, atomistic substrate doping and charge localization in the nitride volume result into a statistical dispersion of ΔV_T. In particular we have shown that the local electrostatic effect of stored electrons and percolative substrate conduction are the main reasons for the ΔV_T spread. A scaling analysis of the statistical distribution of ΔV_T has been also provided, showing that, for fixed density of trapped charge, the average ΔV_T decreases as a consequence of fringing fields, not predictable by any 1-D simulation approach. Moreover, the distribution statistical dispersion has been shown to increase with technology scaling due to a more sensitive percolative substrate conduction in presence of atomistic doping and 3-D electrostatics. The impact of these effects on RTN instabilities has been then highlighted, showing that locally stored

charges in nitride have a marginal role in determining RTN statistics.

A detailed simulation analysis of charge-trap memory programming dynamics (Chapter 4), carefully reproducing the discrete and localized nature of storage traps and the statistical process ruling granular electron injection into the storage layer, has revealed that the low programming efficiency of nanoscale charge-trap cells mainly results from the low impact of locally stored electrons on cell V_T in presence of fringing fields and 3-D electrostatics. By means of a 3-D TCAD Monte Carlo we have also investigated the main variability sources affecting the programming operation of nanoscale charge-trap memories, comparing the effect of the statistical electron injection and trapping process, the fluctuation in the number and position of the trapping sites and the statistical distribution of the V_T shift induced by stored electrons in presence of percolative substrate conduction. Results showed that discrete electron injection is dominant over the other variability sources, due to a lower and lower number of electrons controlling cell state as scaling proceeds.

Finally (Chapter 5) we presented a detailed numerical investigation of the effect of non-uniform channel doping on RTN in nanoscale Flash memories, considering both discrete RTN traps and discrete channel dopants. In fact we have shown (Chapter 3) that a further limit for the programming accuracy is given by RTN instabilities, whose amplitude is enhanced by percolative substrate conduction in presence of atomistic doping. On the other hand the threshold-voltage shift given by single RTN traps is known to follow an exponential distribution [76, 105, 108], with standard deviation proportional to the square root of the channel doping concentration when a uniform doping profile is adopted. These results reveal that channel doping is one of the most important parameters for RTN in MOS devices, opening the possibility for technology optimizations by engineered doping profiles. Our simulation analysis showed that the RTN amplitude distribution increase at low probabilities when increasing the doping concentration at the source side of the channel, while its narrowing appears when higher doping concentrations are used at the STI edges. Clear benefits from the RTN standpoint were shown when retrograde and δ-shape dopings are adopted, with improvements that increase the more the high doping regions are pushed far from the channel surface. In particular, RTN suppression by doping engineering was shown to be strongly correlated to the reduction of cell V_T variability.

Future works

Looking at the future directions of this research field, several suggestions may be given. For example, in this thesis we have focused our attention on the impact of statistical variability on the reading and programming operations' performances. A natural continuation of this work is therefore represented by the study of statistical variability affecting the erase/retention operations.

We have also seen in chapter 1 that the 3D memory structure represents the most viable solution to allow the Flash memory technology to reach the terabit

level of integration. It is then of great interest to extend the analysis presented in this thesis to this advanced memory structures based on non-planar cell geometries. In particular, 3D structures have a different electrostatic affecting in a different way the threshold voltage statistics. Moreover 3D structures could not suffer from the random dopant variability (because their channel is usually undoped), but they could suffer from the variability associate to the discrete nature of poly-silicon grains (because their channel is usually deposited and not mono-crystalline): the presence of grains boundary in the channel gives a complicated pattern of the Fermi-level pinning [78], resulting in a threshold voltage variability related to the fluctuation in the size and in the number of grains [140]. This variability in the Fermi-level pinning can have also a impact on the variability of the RTN time constants, as they strongly depend on the value of the Fermi level in the channel in correspondence of the RTN trap.

Bibliography

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

[2] P. Olivo and E. Zanoni, "Flash memories: an overview," in *Flash memories*, P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds. Kluwer, 1999.

[3] S. T. Wang, "On the I-V characteristics of floating-gate MOS transistors," *IEEE Trans. Electron Devices*, vol. 26, pp. 1292–1294, Sep. 1979.

[4] M. Wada, S. Mimura, H. Nihira, and H. Iizuka, "Limiting factors for programming eprom of reduced dimensions," in *IEDM Tech. Dig.*, 1980, p. 38.

[5] A. Kolodny, S. T. K. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of floating-gate EEPROM cells," *IEEE Trans. Electron Devices*, vol. 33, pp. 835–844, June 1986.

[6] K. Prall, W. I. Kinney, and J. Macro, "Characterization and suppression of drain coupling in submicrometer EPROM cells," *IEEE Trans. Electron Devices*, vol. 34, pp. 2463–2468, Dec. 1987.

[7] M. Wong, D. K.-Y. Liu, and S. S.-W. Huang, "Analysis of the subthreshold slope and the linear transconductance techniques for the extraction of the capacitance coefficients of floating-gate devices," *IEEE Electron Device Lett.*, vol. 13, pp. 566–568, Nov. 1992.

[8] W. L. Choi and D. M. Kim, "A new technique for measuring coupling coefficients and 3-D capacitance characterization of floating-gate devices," *IEEE Trans. Electron Devices*, vol. 41, pp. 2337–2342, Dec. 1994.

[9] R. Bez, E. Camerlenghi, D. Cantarelli, L. Ravazzi, and G.Crisenza, "A novel method for the experimental determination of the coupling ratios in submicron eprom and flash eeprom cells," in *IEDM Tech. Dig.*, 1990, pp. 99–102.

[10] K. Tamer San, C. Kaya, D. K. Y. Liu, T.-P. Ma, and P. Shah, "A new technique for determining the capacitive coupling coefficients in Flash EPROM's," *IEEE Electron Device Lett.*, vol. 13, pp. 328–331, June 1992.

[11] B. Moison, C. Papadas, G. Ghibaudo, P. Mortini, and G. Pananakakis, "New method for the extraction of the coupling ratios in flotox eprom cells," *IEEE Trans. Electron Devices*, vol. 40, pp. 1870–1872, 1993.

[12] M. Woods, "An e-proms integrity starts with its cell structure," in *Nonvolatile Semiconductor Memories: Technologies, Design, and Application.* IEEE Press, 1991.

[13] G. A. Baraff, "Distribution functions and ionization rates for hot-electrons in semiconductors," *Phys. Rev.*, vol. 128, pp. 2507–2517, 1962.

[14] C. Hu, "Lucky-electron model for channel hot-electron emission," in *IEDM Tech. Dig.*, 1979, p. 22.

[15] Z. A. Weinberg, "On tunneling in metal-oxide-silicon structures," *J. Appl. Phys.*, vol. 53, pp. 5052–5056, 1982.

[16] J. Suñé, P. Olivo, and B. Riccò, "Self-consistent solution of the Poisson and Schrödinger equations in accumulated semiconductor-insulator interfaces," *J. Appl. Phys.*, vol. 70, pp. 337–345, 1991.

[17] ——, "Quantum-mechanical modeling of accumulation layers in MOS structure," *IEEE Trans. Electron Devices*, vol. 39, pp. 1732–1739, July 1992.

[18] S. M. Amoroso, C. Monzio Compagnoni, A. Mauri, A. Maconi, A. S. Spinelli, and A. L. Lacaita, "Semi-analytical model for the transient operation of gate-all-around charge-trap memories," *IEEE Trans. Electron Devices*, vol. 58, pp. 3116–3123, Sep. 2011.

[19] W. Brown and J. Brewer, *Nonvolatile semiconductor memory technology.* IEEE Press, 1997, pp. 163–166.

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

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

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

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

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

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

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

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

[28] S. Kim and J. Choi, "Future outlook of nand flash technology for 40nm node and beyond," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2006, pp. 9–11.

[29] C.-H. Lee, J. Choi, C. Kang, Y. Shin, J.-S. Lee, J. Sel, J. Sim, S. Jeon, B.-I. Choe, D. Bae, K. Park, and K. Kim, "Multi-level NAND Flash memory with 63 nm-node TANOS (Si-Oxide-SiN-Al_2O_3-TaN) cell structure," in *Symp. VLSI Tech. Dig.*, 2006, pp. 21–22.

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

[31] S. Jeon, J. H. Han, J. Lee, S. Choi, H. Hwang, and C. Kim, "Impact of metal work function on memory proprieties of charge-trap memory devices using fowler-nordheim p/e mode," *IEEE Electron Device Lett.*, vol. 27, pp. 486–488, 2006.

[32] H.-T. Lue, S.-Y. Wang, E.-K. Lai, Y.-H. Shih, S.-C. Lai, L.-W. Yang, K.-C. Chen, J. Ku, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "BE-SONOS: a bandgap engineered SONOS with excellent performance and reliability," in *IEDM Tech. Dig.*, 2005, pp. 547–550.

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

[34] J. Jang, H.-S. Kim, W. Cho, H. Cho, J. Kim, S. Shim, Y. Jang, J.-H. Jeong, B.-K. Son, D. W. Kim, K. Kim, J.-J. Shim, J. S. Lim, K.-H. Kim,

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

[35] J. Kim, A. J. Hong, S. M. Kim, E. B. Song, J. H. Park, J. Han, S. Choi, D. Jang, J.-T. Moon, and K. L. .Wang, "Novel vertical-stacked-array-transistor (VSAT) for ultra-high-density and cost-effective NAND Flash memory devices and SSD (solid state drive)," in *Symp. VLSI Tech. Dig.*, 2009, pp. 186–187.

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

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

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

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

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

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

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

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

[44] K. Nishinohara, N. Shigyo, and T. Wada, "Effects of microscopic fluctuations in dopant distributions on mosfet threshold voltage," *IEEE Trans. Electron Devices*, vol. 39, pp. 634–639, 1992.

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

[46] K. Takeuchi, T. Tatsumi, and A. Furukawa, "Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuation," in *IEDM Tech. Dig.*, 1997, pp. 89–92.

[47] D. Vasileska, W. J. Gross, and D. K. Ferry, "Modeling of deepsubmicrometer mosfets: random impurity effects, threshold voltage shifts and gate capacitance attenuation," in *Proc Int Workshop Computational Electronics*, 1998, pp. 259–262.

[48] P. Stolk, F. Widdershoven, and D. Klassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Trans. Electron Devices*, vol. 45, pp. 1960–1971, 1998.

[49] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: a 3-D "atomistic" simulation study," *IEEE Trans. Electron Devices*, vol. 45, pp. 2505–2513, Dec. 1998.

[50] D. J. Frank, Y. Taur, M. Ieong, and H. S. P. Wong, "Monte carlo modeling of threshold variation due to dopant fluctuations." in *Proc VLSI Tech Symp*, 1999, pp. 169–170.

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

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

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

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

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

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

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

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

[59] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete resistance switching in submicrometer silicon inversion layers: individual interface traps and low-frequency (1/f?) noise," *Phys. Rev. Lett.*, vol. 52, pp. 228–231, 1984.

[60] M.-H. Tsai, T. P. Ma, and T. B. Hook, "Channel length dependence of random telegraph signal in sub-micron MOSFET's," *IEEE Electron Device Lett.*, vol. 15, pp. 504–506, Dec. 1994.

[61] S. T. Martin, G. P. Li, E. Worley, and J. White, "The gate bias and geometry dependence of random telegraph signal amplitudes," *IEEE Electron Device Lett.*, vol. 18, pp. 444–446, Sep. 1997.

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

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

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

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

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

[67] A. Cathignol, B. Cheng, D. Chanemougame, A. R. Brown, K. Rochereau, G. Ghibaudo, and A. Asenov, "Quantitative evaluation of statistical variability sources in a 45-nm technological node LP N-MOSFET," *IEEE Electron Device Lett.*, vol. 29, pp. 609–611, June 2008.

[68] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometer MOSFETs introduced by gate line edge roughness," *IEEE Trans. Electron Devices*, vol. 50, pp. 1254–1260, May 2003.

[69] A. R. Brown, G. Roy, and A. Asenov, "Poly-Si-gate-related variability in decananometer MOSFETs with conventional architecture," *IEEE Trans. Electron Devices*, vol. 54, pp. 3056–3063, Nov. 2007.

[70] D. Reid, C. Millar, S. Roy, and A. Asenov, "Understanding LER-induced MOSFET V_T variability: Part I: three-dimensional simulation of large statistical samples," *IEEE Trans. Electron Devices*, vol. 57, pp. 2801–2807, Nov. 2010.

[71] ——, "Understanding LER-induced MOSFET V_T variability: Part II: reconstructing the distribution," *IEEE Trans. Electron Devices*, vol. 57, pp. 2808–2813, Nov. 2010.

[72] A. R. Brown, N. M. Idris, J. R. Watling, and A. Asenov, "Impact of metal gate granularity on threshold voltage variability: a full-scale three-dimensional statistical simulation study," *IEEE Electron Device Lett.*, vol. 31, pp. 1199–1201, Nov. 2010.

[73] B. Hoeneisen and C. A. Meed, "Fundamental limitations in microelectronics i. mos technology," *Solid-State Electron*, vol. 15, pp. 819–829, Feb. 1972.

[74] T. Tanaka, T. Usuki, T. Futatsugi, Y. Momiyama, and T. Sugii, "V_{th} fluctuation induced by statistical variation of pocket dopant profile," in *IEDM Tech. Dig.*, 2000, pp. 271–274.

[75] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov, "Analysis of threshold voltage distribution due to random dopants: a 100000-sample 3-D simulation study," *IEEE Trans. Electron Devices*, vol. 56, pp. 2255–2263, Oct. 2009.

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

[77] A. Asenov and S. Saini, "Suppression of random dopant-induced threshold voltage fluctuations in sub-0.1 μm MOSFET's with epitaxial and δ-doped channels," *IEEE Trans. Electron Devices*, vol. 46, pp. 1718–1724, Aug. 1999.

[78] A. Asenov, A. R. Brown, G. Roy, B. Cheng, C. Alexander, C. Riddet, U. Kovac, A. Martinez, N. Seoane, and S. Roy, "Simulation of statistical variability in nano-cmos transistors using drift-diffusion, monte carlo and non-equilibrium greens function techniques," *J Comput Electron*, pp. 349–373, 2009.

[79] N. Sano, M. Tomizawa, and K. Natori, "Statistical threshold fluctuations in si-mosfets: jellium vs atomistic dopant variations," in *Ext Abst Int Conf Solid State Devices Materials*, 2000, pp. 216–217.

[80] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, "Role of long-range and short-range coulomb potentials in threshold characteristics under discrete dopants in sub-0.1 μm si-mosfets," in *IEDM Tech. Dig.*, 2000, pp. 275–278.

[81] A. Asenov, G. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, "Increase in the random dopant induced threshold fluctuations and lowering in sub-100 nm MOSFETs due to quantum effects: a 3-D density-gradient simulation study," *IEEE Trans. Electron Devices*, vol. 48, pp. 722–729, April 2001.

[82] G. Roy, "Simulation of intrinsic parameter fluctuations in nano-cmos devices," Ph.D. dissertation, Glasgow University, 2005.

[83] D. Chattopadhyay and H. J. Queisser, "Electron scattering by ionized impurities in semiconductors," *Rev Mod Phys*, vol. 53, pp. 745–768, 1981.

[84] R. J. V. Overstraeten and R. P. Mertens, "Heavy doping effects in silicon," *Solid-State Electron.*, vol. 30, pp. 1077–1087, 1987.

[85] D. Bohm, "A suggested interpretation of the quantum theory in terms of hidden variables," *Phys. Rev.*, vol. 85, p. 166, 1952.

[86] ——, "Precise modeling framework for short-channel double-gate and gate-all-around MOSFETs," *IEEE Trans. Electron Devices*, vol. 85, p. 180, 1952.

[87] A. R. Brown, J. R. Watling, G. Roy, C. Riddet, C. L. Alexander, U. Kovac, A. Martinez, and A. Asenov, "Use of density gradient quantum corrections in the simulation of statistical variability in mosfets," *J Comput Electron*, pp. 187–196, 2010.

[88] H. Tsuchiya and T. Miyoshi, "Quantum mechanical monte carlo approach to electron transport at heterointerface," *Superlattices Microstruct.*, vol. 27, pp. 529–532, 2000.

[89] A. R. Brown, "Tcad simulation of statistical variability," *SISPAD Workshop, Bologna*, 2010.

[90] R. Hockney and J. Eastwood, *Computer Simulation Using Particles*. IoP Publishing, Bristol, 1999.

[91] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, and boron-doped silicon," *IEEE Trans. Electron Devices*, vol. 30, pp. 764–769, 1983.

[92] D. Caughey and R. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proc. of the IEEE*, vol. 55, pp. 2192–2193, 1967.

[93] C. Lombardi, S. Vanzini, A. Saporito, and M. Vanzi, "A physically-based mobility model for numerical simulation of non planar devices," *IEEE Trans. on CAD*, vol. 7, Nov. 1988.

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

[95] T. Y. Chan, K. K. Young, and C. Hu, "A true single-transistor oxide-nitride-oxide EEPROM device," *IEEE Electron Device Lett.*, vol. 8, pp. 93–95, 1987.

[96] Y. Shin, J. Choi, C. Kang, C. Lee, K.-T. Park, J.-S. Lee, J. Sel, V. Kim, B. Choi, J.Sim, D. Kim, H.-J. Cho, and K. Kim, "A novel NAND-type MONOS memory using 63 nm process technology for multi-gigabit Flash EEPROM," in *IEDM Tech. Dig.*, 2005, pp. 337–340.

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

[98] J. S. Sim, J. Park, C. Kang, W. Jung, Y. Shin, J. Kim, J. Sel, C. Lee, S. Jeon, Y. Jeong, Y. Park, J. Choi, and W.-S. Lee, "Self aligned trap-shallow trench isolation scheme for the reliability of TANOS (TaN/AlO/SiN/Oxide/Si) NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 110–111.

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

[100] A. Furnemont, M. Rosmeulen, A. Cacciato, L. Breuil, K. De Meyer, H. Maes, and J. Van Houdt, "A consistent model for SANOS programming operation," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 96–97.

[101] E.-S. Choi, H.-S. Yoo, K.-H. Park, S.-J. Kim, J.-R. Ahn, M.-S. Lee, Y.-O. Hong, S.-G. Kim, J.-C. Om, M.-S. Joo, S.-H. Pyi, S.-S. Lee, S.-K. Lee, and G.-H. Bae, "Modeling and characterization of program/erasure speed and retention of TiN-gate MANOS (Si-Oxide-SiN_x-Al_2O_3-Metal gate) cells for NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 83–84.

[102] E. Vianello, F. Driussi, A. Arreghini, P. Palestri, D. Esseni, L. Selmi, N. Akil, M. J. van Duuren, and D. S. Golubovic, "Experimental and simulation analysis of program/retention transients in silicon nitride-based NVM cells," *IEEE Trans. Electron Devices*, vol. 56, pp. 1980–1990, Sep. 2009.

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

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

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

[106] *Sentaurus device user guide*, 2007th ed., Synopsys, Mountain View (CA), 2007.

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

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

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

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

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

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

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

[114] J. P. Chiu, Y. L. Chou, H. C. Ma, T. Wang, S. H. Ku, N. K. Zou, V. Chen, W. P. Lu, K. C. Chen, and C.-Y. Lu, "Program charge effect on random telegraph noise amplitude and its device structural dependence in SONOS Flash memory," in *IEDM Tech. Dig.*, 2009, pp. 843–846.

[115] C.-H. Lee, C. Kang, J. Sim, J.-S. Lee, J. Kim, Y. Shin, K.-T. Park, S. Jeon, J. Sel, Y. Jeong, B. Choi, V. Kim, W. Jung, C.-I. Hyun, J. Choi, and K. Kim, "Charge trapping memory cell of TANOS (Si-Oxide-SiN-Al_2O_3-TaN) structure compatible to conventional NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2006, pp. 54–55.

[116] T. Ishida, T. Mine, D. Hisamoto, Y. Shimamoto, and R. Yamada, "Anomalous electron storage decrease in MONOS nitride layers thinner than 4 nm," *IEEE Electron Device Lett.*, vol. 29, pp. 920–922, Aug. 2008.

[117] H.-T. Lue, P.-Y. Du, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh, R. Liu, and C.-Y. Lu, "A novel gate-sensing and channel-sensing transient analysis method for real-time monitoring of charge vertical location in SONOS-type devices and its applications in reliability studies," in *Proc. IRPS*, 2007, pp. 177–183.

[118] C.-H. An, M. Soo Lee, and H. Kim, "Effects of Si_3N_4 thickness on the electrical properties of oxide-nitride-oxide tunneling dielectrics," *J. Electrochem. Soc.*, vol. 155, no. 11, pp. G247–G252, 2008.

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

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

[121] S. H. Gu, C. W. Li, T. Wang, W. P. Lu, K. C. Chen, J. Ku, and C.-Y. Lu, "Read current instability arising from random telegraph noise in localized storage, multi-level SONOS Flash memory," in *IEDM Tech. Dig.*, 2006, pp. 487–490.

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

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

[124] H.-T. Lue, T.-H. Hsu, S.-C. Lai, Y. J. Chen, K. F. Chen, C. Lo, I. J. Huang, T. T. Han, M. S. Chen, W. P. Lu, K. C. Chen, C. S. Chang, M. H. Liaw, K.-Y. Hsieh, and C.-Y. Lu, "Study of electron and hole injection statistics of BE-SONOS NAND Flash," in *Proc. IMW*, 2010, pp. 92–95.

[125] R. Liu, H.-T. Lue, K. C. Chen, and C.-Y. Lu, "Reliability of barrier engineered charge trapping devices for sub-30 nm NAND Flash," in *IEDM Tech. Dig.*, 2009, pp. 745–748.

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

[127] M. F. Beug, T. Melde, M. Czernohorsky, R. Hoffmann, J. Paul, R. Knoefler, and A. T. Tilke, "Analysis of TANOS memory cells with sealing oxide containing blocking dielectric," *IEEE Trans. Electron Devices*, vol. 57, pp. 1590–1596, July 2010.

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

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

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

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

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

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

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

[135] M. Tanizawa, S. Ohbayashi, T. Okagaki, K. Sonoda, K. Eikyu, Y. Hirano, K. Ishikawa, O. Tsuchiya, and Y. Inoue, "Application of a statistical compact model for random telegraph noise to scaled-SRAM Vmin analysis," in *Symp. VLSI Tech. Dig.*, 2010, pp. 95–96.

[136] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *IEDM Tech. Dig.*, 2010, pp. 628–631.

[137] K. Abe, A. Teramoto, S. Sugawa, and T. Ohmi, "Understanding of traps causing random telegraph noise based on experimentally extracted time constants and amplitude," in *Proc. IRPS*, 2011, pp. 381–386.

[138] *Sentaurus device user guide*, 2010th ed., Synopsys, Zurich, Switzerland, 2010.

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

[140] Y. Kitahara, S. Takagi, and N. Sano, "Statistical study of subthresold characteristics in polycristallyne silicon thin-film transistors," *J. Appl. Phys.*, vol. 94, pp. 7789–7795, 2003.

List of publications

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

2. C. Scozzari, G. Albini, M. Alessandri, **S.M. Amoroso**, P. Bacciaglia, A. Del Vitto, G.Ghidini, "Al_2O_3 optimization for Charge Trap memory application", Ultimate Integration of Silicon - ULIS, 191-194, 2008.

3. C. Monzio Compagnoni, A. Mauri, **S. M. Amoroso**, A. Maconi, A. S. Spinelli, "Physical modeling for programming of TANOS memories in the Fowler-Nordheim regime", IEEE Transaction on Electron Devices, 2008-2015, 2009.

4. A. Mauri, C. Monzio Compagnoni, **S. M. Amoroso**, A. Maconi, A. Ghetti, A. S. Spinelli, A.L. Lacaita, "Comprehensive investigation of statistical effects in nitride memories - Part I: Physics-based modeling", IEEE Transaction on Electron Devices, 2116- 2123, 2010.

5. C. Monzio Compagnoni, A. Mauri, **S. M. Amoroso**, A. Maconi, E. Greco, A. S. Spinelli, A.L. Lacaita, "Comprehensive investigation of statistical effects in nitride memories - Part II: Scaling analysis and impact on device performance", IEEE Transaction on Electron Devices, 2124- 2131, 2010.

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

7. **S. M. Amoroso**, A. Mauri, N. Galbiati, C. Scozzari, E. Mascellino, E. Camozzi, A. Rangoni, T. Ghilardi, A.Grossi, P. Tessariol, C. Monzio Compagnoni, A. Maconi, A.L. Lacaita, A. S. Spinelli, G. Ghidini, "Reliability constraints for TANOS memories due to alumina trapping and leakage", International Reliability Physics Symposium, 966- 969, 2010.

8. C. Miccoli, C. Monzio Compagnoni, **S. M. Amoroso**, A. Spessot, P. Fantini, A. Visconti, A. S. Spinelli, "Impact of neutral threshold-voltage

spread and electron-emission statistics on data retention of nanoscale NAND Flash", IEEE Electron Device Letters, 1202- 1204, 2010.

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

10. A. Mauri, **S. M. Amoroso**, C. Monzio Compagnoni, A. Maconi and A. S. Spinelli, "Comprehensive Numerical Simulation of Threshold Voltage Transients in Nitride Memories", Solid state Electronics, 23-30, 2010.

11. G. Ghidini, N. Galbiati, E. Mascellino, C. Scozzari, A. Sebastiani, **S. M. Amoroso**, C. Monzio Compagnoni, A. S. Spinelli, A. Maconi, R. Piagge, A. Del Vitto, M. Alessandri, I. Baldi, E. Moltrasio, G. Albini, A. Grossi, P. Tessariol, E. Camerlenghi and A. Mauri, "Charge retention phenomena in CT silicon nitride: impact of technology and operating conditions", Journal of Vacuum Science and Technology B, 2011

12. A. Ghetti, **S. M. Amoroso**, A. Mauri, C. Monzio Compagnoni, "Doping Engineering for Random Telegraph Noise Suppression in Deca-nanometer Flash Memories", IEEE International Memory Workshop, 978-981, 2011

13. A. Maconi, **S. M. Amoroso**, C. Monzio Compagnoni, A. Mauri, A. S. Spinelli and A. L. Lacaita, "Three-Dimensional Simulation of Charge-Trap Memory ProgrammingPart II: Variability", IEEE Transaction on Electron Devices, 1872-1878, 2011.

14. **S. M. Amoroso**, A. Maconi, A. Mauri, C. Monzio Compagnoni, A. S. Spinelli and A. L. Lacaita, "Three-Dimensional Simulation of Charge-Trap Memory ProgrammingPart I: Average Behavior", IEEE Transaction on Electron Devices, 1864-1871, 2011.

15. **S. M. Amoroso**, C. Monzio Compagnoni, A. Mauri, A. Maconi, A. S. Spinelli and A. L. Lacaita, "Semianalytical Model for the Transient Operation of Gate-All-Around Charge-Trap Memories", IEEE Transaction on Electron Devices, 2011.

16. **S. M. Amoroso**, C. L. Alexander, S. Markov, G. Roy, and A. Asenov, "A Mobility Model Correction for Atomistic Drift-Diffusion Simulation", IEEE SISPAD, 2011.

Index

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