UNIVERSITY OF PISA

Engineering PhD School "Leonardo da Vinci"

PhD Course in
"Applied Electromagnetism in Electrical and Biomedical Engineering, Electronics, Smart Sensors, Nano-Technologies"

PhD Thesis

Intrinsic variability of nanoscale CMOS technology for logic and memory

ING/INF-01

Advisor:

Prof. Giuseppe IANNACCONE

Author:

Valentina Bonfiglio

Contents

Acknowledgements

First of all, I would like to thank my tutor Prof. Giuseppe Iannaccone for the valuable guidance during the course of this work.

I also would like to thank the Ing. Alessandro Nannipieri for his support and for the stimulating discussions about Sentaurus TCAD.

I also wish to express my gratitude for my mother, my father, Deborah and Andrea, for the support they have always given me.

A special thanks for my boyfriend who is a person very important in my life.

Finally I thank my dearest friends who are always close to me.

Introduction

The continuous downscaling of CMOS technology, the main engine of development of the semiconductor Industry, is limited by factors that become important for nanoscale device size, which undermine proper device operation completely offset gains from scaling.

One of the main problems is device variability: nominally identical devices are different at the microscopic level due to fabrication tolerance and the intrinsic granularity of matter. For this reason, structures, devices and materials for the next technology nodes will be chosen for their robustness to process variability, in agreement with the ITRS (International Technology Roadmap for Semiconductors). Examining the dispersion of various physical and geometrical parameters and the effect these have on device performance becomes necessary.

In this thesis, I focus on the study of the dispersion of the threshold voltage due to intrinsic variability in nanoscale CMOS technology for logic and for memory. In order to describe this, it is convenient to have an analytical model that allows, with the assistance of a small number of simulations, to calculate the standard deviation of the threshold voltage due to the various contributions.

In the first chapter of the thesis will address the problem of variability of physical and geometrical parameters in CMOS technology and will analyze the

main factors of variability and the effect of variability on the performance of electronic circuits.

The second chapter will proposed an overview of the state of the art of research in variability, taking into account various mechanisms and their combined effects.

In the third chapter we will describe our model to investigate the dispersion of the threshold voltage based on sensitivity analysis. We have considered various structures, for which results from three-dimensional atomistic statistical simulations were available, in order to compare and verify the validity of our method: in particular a 32 nm ultra-thin body SOI MOSFET and a 22 nm double-gate MOSFET adopted within the *EC* PULLNANO Project, one bulk 45 nm NMOSFET within the ENIAC project MODERN and a 32/28 nm CMOS process developed by STM again for the MODERN project.

In the fourth chapter we consider the variability of Flash memory devices, focusing on a template of a flash memory cell obtained in collaboration with Micron for the 32 nm technology node. We shall see the specificity of studying device variability in the context of nonvolatile memories

1 Variability in CMOS technology

The semiconductor industry was born with the invention of the first bipolar transistor in 1948. In 1961 appeared the first planar circuit and in 1964 it was the turn of the first MOSFET. Today the progress of the semiconductor industry has led to microprocessors operating at GHz frequencies, to microprocessors with more than 1 billion transistors and GByte memory chips.

This rapid technological progress had been predicted in 1965 in a famous speech by Gordon Moore. Moore said that the number of transistors per square inch present on a chip would double every 18 months, achieved both by reducing the size of the transistor, and by increasing the size of the single circuit. This prediction became the so-called "Moore's Law" (Fig. 1.1).

The further reduction in the size of MOS transistors, however, requires the introduction of new materials and new architectures that take the place of conventional planar MOS.

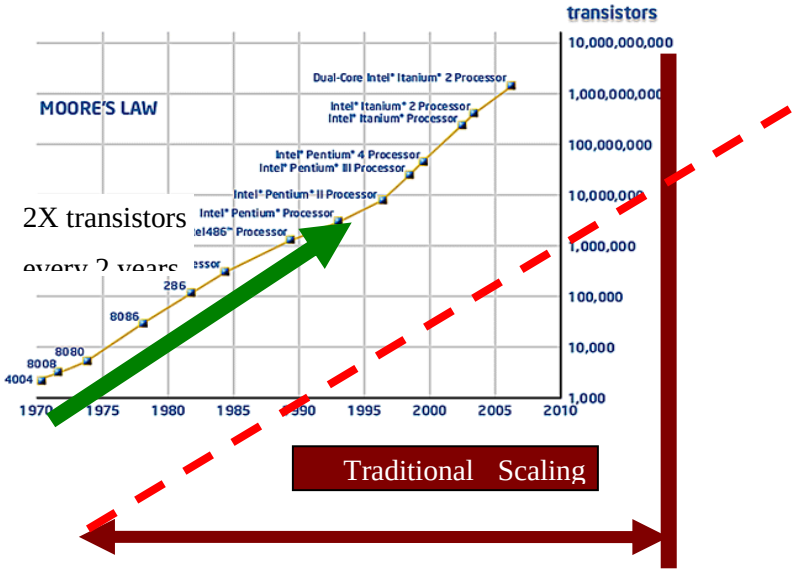


Fig. 1.1 The time progresses of Moore's law [1].

One of the main causes of the rapid improvements of integrated circuits have to be found in the excellent performance and scaling properties of MOS transistors, for the first time described by Dennard in 1974: he saw that reducing the horizontal size, vertical size and operating voltage, we can obtain simultaneous improvements in the density of transistors, the switching speed and switching energy. Since then, the recipes for scaling have been updated to modern processes. However, every 2-3 years we see a reduction in the minimum size of about a factor of 0.7 and each generation provides transistors that are smaller, faster and use less energy [2].

A method to estimate the delay of gates based on MOS transistors is the use of the CV/I metric. In this metric, the switching speed can be estimated when

the gate capacitance (C_{GATE}), the operating voltage (V_{DD}) and the "On Current" (I_{ON}) (i.e. the current when $V_{GS} = V_{DS} = V_{DD}$) are known.

This simple metric underestimates the gate delay because it omits some important performance factors, as the junction capacitance and the fact that the typical MOS circuits have a fan-out larger than 1 (so the load capacity is typically larger than the gate capacitance). However, a metric is especially useful when there is not a complete set of transistor parameters for a more rigorous comparison.

Another important characteristic of the MOS transistor is to reduce the amount of energy used during a switching event. The reduction in switching energy is due to the combination of several factors: lower parasitic capacitance, smaller feature sizes, lower supply voltage. A metric to estimate the switching energy of MOS transistor is $C_{GATE}V_{DD}^2$, using the gate capacitance of the transistor and the supply voltage. This is also a simplified metric, which omits some factors of the second order, but it is useful to estimate the trends of the various technology generations. The transistor switching energy has become increasingly important because of constraints on the overall power consumption.

Scaling also poses performance challenges. Scaled MOS transistors require reduced threshold voltages than in turn lead to higher subthreshold leakage (I_{OFF}). Another limitation is the gate oxide thickness reduction: the leakage current of the gate oxide increases exponentially with each new generation due to the reduction of oxide thickness and approaches the values of the subthreshold drain current (~ 1 nA / μm).

Metal gate with high-k dielectric have been implemented in the recent technology generations in order to allow scaling of the EOT (equivalent oxide

thickness), consistent with the overall transistor scaling while keeping gate leakage currents within tolerable limits.

The characteristics of the MOSFET for very small dimensions present variability issues because of the challenging lithography and because of intrinsic process variations, such as the random fluctuation of dopants in the channel regions that concern the control of the threshold voltage V_{th}.

To approach the issue of CMOS technology scaling, we refer to chapter *Process Integration, Devices, and Structures (PIDS)* of the International Technology Roadmap for Semiconductors (ITRS) [3], which addresses the subject of aggressive scaling of MOSFETs, treating the entire process flow for the realization of integrated circuits, considering the tradeoffs of reliability associated with new options.

This aggressive scaling drives the industry toward a series of important technological innovations, including material and process changes such as high-k gate dielectric, metal gate electrode, and at longer term, new structures such as ultra-thin, multi-gate MOSFETs (as well as FinFET).

The key objectives of the ITRS include both the identification of the main technical requirements and the key challenges in order to support the scaling of CMOS technology for the Moore's Law, and the encouragement to research and development necessary to meet the key challenges.

The ITRS provides potential solutions, which are intended as a stimulus and not as limitations to the research, exploring new and different approaches.

In the Tab. 1.1 the difficult challenges identified by the ITRS 2010 for process integration are shown.

Difficult Challenges for $L_g \geq 16$ nm	*Summary of Issues*
1. Scaling of logic MOSFETs	Scaling planar bulk CMOS Implementation of fully depleted SOI and multi-gate (MG) structures Controlling source/drain series resistance within tolerable limits Further scaling of EOT with higher κ materials (κ > 30) Threshold voltage tuning and control with metal gate and high-κ stack Inducing adequate strain
2. Scaling of DRAM and SRAM	DRAM— Adequate storage capacitance with reduced feature size; implementing high-κ dielectric Low leakage in access transistor and storage capacitor Low resistance for bit and word lines to ensure desired speed Improve bit density and to lower production cost in driving toward $4F^2$ cell size SRAM— Maintain adequate noise margin and control key instabilities and soft-error rate Difficult lithography and etch issues
3. Scaling high-density non-volatile memory	Endurance, noise margin, and reliability requirements Non-scalability of tunnel dielectric and interpoly dielectric in flash Difficult lithography and etch issues with pitch scaling Maintain high gate coupling ratio in floating-gate flash
4. Reliability due to material, process, and structural changes	Threshold voltage shifts due to traps, carrier injection, and program/erase cycling in memory cells Mobility degradation due to mechanical stress relaxation or interface states New or changed failure mechanisms resulting from high-κ/metal gate and new doping/activation processes New failure mechanism resulting from fundamental length scales or new device structures Process variability
Difficult Challenges for $L_g < 16$ nm	*Summary of Issues*
1. Implementation of advanced non-classical CMOS structures	Advanced non-planar multi-gate MOSFETs below 10 nm gate length Control of short-channel effects Drain engineering to control parasitic resistance Strain enhanced thermal velocity and quasi-ballistic transport
2. Implementation of non-classical CMOS channel materials	Identification and demonstration of alternate channel materials New issues from materials, devices, and processing Integration of alternate channel materials on Si platform
3. Identification and implementation of new memory structures	Density and voltage scaling of NVM 3-D integration of NVM Implementing non-charge-storage type of NVM Scaling storage capacitor for DRAM DRAM and SRAM replacement solutions
4. Reliability of novel devices, structures, materials, and applications	Reliability characterization of new devices Dealing with fluctuations and statistical process variations Impact of microscopic physical effects Need for Design for Reliability tools
5. Power scaling	V_{dd} scaling Controlling subthreshold current
6. Beyond CMOS	Identification and implementation of non-CMOS devices and architectures Integration onto Si-CMOS platform See *ERD* and *ERM chapters* for more discussions and details

Tab. 1.1 Process Integration Difficult Challenges [3].

In the ITRS Emerging Research Devices chapter, information on several new technologies proposed for beyond CMOS information processing, memory, and storage technologies was evaluated and discussed. In the Table 1.2 and Table 1.3 the difficult challenges of the emerging research devices and materials are reported.

Difficult Challenges ≥ 16 nm and < 16 nm	*Summary of Issues and opportunities*
Scale high-speed, dense, embeddable, volatile and non-volatile memory technologies to and beyond the 16 nm technology generation.	SRAM and FLASH scaling will reach definite limits within the next several years (see PIDS Difficult Challenges). These are driving the need for new memory technologies to replace SRAM and FLASH memories. Identify the most promising technical approach(es) to obtain electrically accessible, high-speed, high-density, low-power, (preferably) embeddable volatile and non-volatile RAM The desired material/device properties must be maintained through and after high temperature and corrosive chemical processing Reliability issues should be identified & addressed early in the technology development
Scale CMOS to and beyond the 16 nm technology generation.	Develop new materials to replace silicon as an alternate channel and source/drain to increase the saturation velocity and maximum drain current in MOSFETs while minimizing leakage currents and power dissipation for technology scaled to 16 nm and beyond. Develop means to control the variability of critical dimensions and statistical distributions (e.g., gate length, channel thickness, S/D doping concentrations, etc.) Accommodate the heterogeneous integration of dissimilar materials The desired material/device properties must be maintained through and after high temperature and corrosive chemical processing Reliability issues should be identified & addressed early in t
Extend ultimately scaled CMOS as a platform technology into new domains of application.	Discover and reduce to practice new device technologies and a primitive-level architecture to provide special purpose optimized functional cores heterogeneously integrable with silicon CMOS.
Continue functional scaling of information processing technology substantially beyond that attainable by ultimately scaled CMOS.	Invent and develop a new information processing technology eventually to replace CMOS Ensure that a new information processing technology is compatible with the new memory technology discussed above; i.e., the logic technology must also provide the access function in a new memory technology. Bridge a knowledge gap that exists between materials behaviors and device functions. Accommodate the heterogeneous integration of dissimilar materials The desired material/device properties must be maintained through and after high temperature and corrosive chemical processing Reliability issues should be identified & addressed early in the technology development

Tab. 1.2 Emerging Research Devices Difficult Challenges [3].

Difficult Challenges ≤16 nm	*Summary of Issues*
Integration of alternate channel materials with high performance	III-V has high electron mobility, but low hole mobility
	Germanium has high hole mobility, but electron mobility is not as high as III-V materials
	Demonstration of high mobility n and p channel alternate channel materials co-integrated with high κ dielectric
	Demonstration of high mobility n and p channel carbon (graphene or carbon nanotubes) FETs with high on-off ratio co-integrated with high κ dielectric and low resistance contacts
	Selective growth of alternate channel materials in desired locations with controlled properties and directions on silicon wafers (III-V, Graphene, Carbon nanotubes and semiconductor nanowires)
	Achieving low contact resistance to sub 16 nm scale structures (graphene and carbon nanotubes)
	Ge dopant thermal activation is much higher than III-V process temperatures
	Growth of high κ dielectrics with unpinned Fermi Level in the alternate channel material
Control of nanostructures and properties	Ability to pattern sub 16 nm structures in resist or other manufacturing related patterning materials (resist, imprint, self assembled materials, etc.)
	Control of CNT properties, bandgap distribution and metallic fraction
	Control of stoichiometry, disorder and vacancy composition in complex metal oxides
	Control and identification of nanoscale phase segregation in spin materials
	Control of surfaces and interfaces
	Control of growth and heterointerface strain
	Control of interface properties (e.g., electromigration)
	Ability to predict nanocomposite properties based on a "rule of mixtures"
	Data and models that enable quantitative structure-property correlations and a robust nanomaterials-by-design capability
Controlled assembly of nanostructures	Placement of nanostructures, such as CNTs, nanowires, or quantum dots, in precise locations for devices, interconnects, and other electronically useful components
	Control of line width of self-assembled patterning materials
	Control of registration and defects in self-assembled materials
Characterization of nanostructure-property correlations	Correlation of the interface structure, electronic and spin properties at interfaces with low-dimensional materials
	Characterization of low atomic weight structures and defects (e.g., carbon nanotubes, graphitic structures, etc.)
	Characterization of spin concentration in materials
	Characterization of vacancy concentration and its effect on the properties of complex oxides
	3D molecular and nanomaterial structure property correlation
Characterization of properties of embedded interfaces and matrices	Characterization of the roles of vacancies and hydrogen at the interface of complex oxides and the relation to properties
	Characterization of transport of spin polarized electrons across interfaces
	Characterization of the structure and electrical interface states in complex oxides
	Characterization of the electrical contacts of embedded molecule(s)
Fundamental thermodynamic stability and fluctuations of materials and structures	Geometry, conformation, and interface roughness in molecular and self-assembled structures
	Device structure-related properties, such as ferromagnetic spin and defects
	Dopant location and device variability

Tab. 1.3 Emerging Research Material Technologies Difficult Challenges [3].

The Front End Processes (FEP) Roadmap focuses on future process requirement and potential solutions related to scaled field effect transistors (MOSFETs), DRAM storage capacitors, and non-volatile memory (Tab. 1.4).

The purpose is to define comprehensive future requirements and potential solutions for the key front end wafer fabrication process technologies and the materials associated with these devices.

Difficult Challenges ≥ 16 nm (Metal 1 1/2-pitch)	*Summary of Issues*
	Strain Engineering - continued improvement for increasing device performance - application to FDSOI and Multi-gate technologies
	Achieving DRAM cell capacitance with dimensional scaling - finding robust dielectric with dielectric constant of ~60 - finding electrode material with high work function
	Achieving clean surfaces free of killer defects - with no pattern damage - with low material loss (<0.2 A)
	High-k/Metal Gate - introduction to full scale manufacturing for HP, LOP, and LSTP - scaling equivalent oxide thickness (EOT) below 0.8 nm
	450 mm wafers - meeting production level quality and quantity
Difficult Challenges < 16 nm (Metal 1 1/2-pitch)	*Summary of Issues*
	Continued scaling of HP multigate device in all aspects: EOT, junctions, mobility enhancement, new channel materials, parasitic series resistance, contact silicidation.
	Lowering required DRAM capacitance by 4F2 cell scheme or like, while continuing to address materials challenges
	Continued achievement of clean surfaces while eliminating material loss and surface damage and sub-critical dimension particle defects
	Continued EOT scaling below 0.7 nm with appropriate metal gates
	Continued charge retention with dimensional scaling and introduction of new non-charged based NVM technologies

Tab. 1.4 Front End Processes Difficult Challenges [3].

The limits of CMOS scaling have led researchers worldwide at the introduction of new device structures, such as ultra-thin body (UTB) silicon on insulator (SOI) devices, double-gate SOI, FinFETs. These new devices help eliminate some of the short-channel effects displayed by conventional MOSFETs. However, many old questions remain, and also new problems arise. For example, in UTB SOI devices, in double-gate MOSFET and FinFET, mechanical quantum effects significantly influence the overall behavior of the device. In addition, there are still considerable fluctuations in device parameters.

A potential solution to the unacceptable variation in threshold voltage in very small MOSFET, caused by the small number of dopants in the channel, is to use ultra-thin body, fully depleted SOI MOSFETs. For these devices the

channel doping is relatively low, and the threshold voltage can be adjusted acting on the gate work function, rather than on the doping of the channel as in conventional bulk MOSFETs.

As far as double-gate SOI MOSFETs are concerned, these devices have a potential profile quite different from that of conventional bulk MOSFETs, due to the symmetrical structure and the extremely low doping, so one can not apply the models developed for bulk planar MOSFETs.

In this regard, while for the conventional bulk MOSFET, there are several models in the literature related to analytical and numerical fluctuations of the parameters, which become evident with the reduction in size, relatively little has been done to these new components.

It is useful to analyze the dispersion of parameters in these new devices, to study the effects and find solutions to allow for the scaling while respecting the constraints imposed by the Roadmap.

1.1 Main factors of variability of the CMOS technology

The rapid growth of semiconductor industry over the past 40 years has been mainly the result of a constant reduction in the size of the CMOS switching elements, which form the basis of the logic circuits in almost all modern digital systems. When the size of the CMOS switches, and field effect transistors that implement them, are reduced, the integrated circuits make with them, to improve in terms of speed, density of the total circuit and cost for function.

But there are some physical limitations to the miniaturization process [4].

The International Roadmap for Semiconductors outguess that CMOS transistors with channel lengths of 7 nm become mass-produced since 2018; devices with channel lengths of 25 nm are already in production. The size of the

CMOS will continue to scale over the next two decades, but when they come near to the dimensions of the silicon lattice, the precise atomic configuration of the structure will become extremely important for their macroscopic properties.

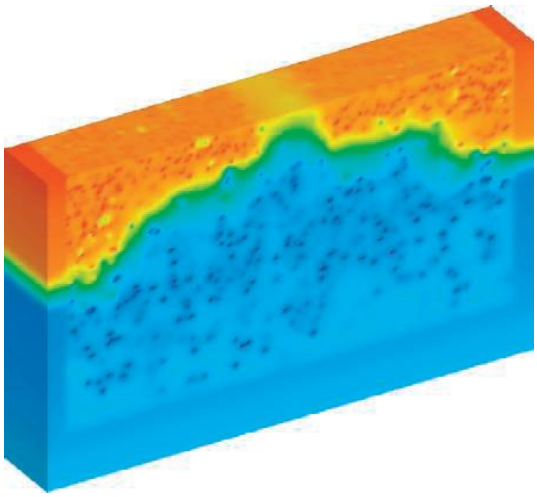


Fig. 1.2 Dopants in a small transistor: the electrostatic potential is mapped from red (1 V) to blue (0 V). The fluctuations of the potential in the channel associated with the random distribution of dopants lead to have different characteristics for each device [Fig of [4]].

Mead and Keyes recognized in 1970 [5], [6] that below a critical size, the devices cannot be described, designed, modeled, or referred as a discrete semiconductors with smooth boundaries and interfaces. At the nanoscale, the effect of the number and position of dopant atoms, introduced to alter the electrical properties of different regions of field effect transistors, (Fig. 1.2) will be larger because of the small number of dopants; at the nanoscale it will therefore be important to consider that every transistor is microscopically different. The variation in the position of dopants between devices leads to

measurable differences in the macroscopic parameters such as drive current, threshold voltage, and leakage current. In addition, the reduction in size reduces the number of dopants, and thus increase the variations and consequently worsens the differences in device performance.

In addition, by reducing the size of the device, the roughness of typical gate oxides (one or two atomic layers), becomes comparable with the thickness of the gate dielectrics itself. Thus, each device will have a different thickness of the gate, and a profile of the roughness of the interface unique.

The use of high permittivity gate insulators (high k), to replace the existing gate oxides, permits to obtain the gate thicknesses that can reduce this source of fluctuations for one or two technology generations. However, variations on the atomic scale of the position of impurity atoms, the local variations at the interface silicon / silicon dioxide over the canal, and local variations of the thickness of silicon dioxide, introduce variations in the electrostatics of the device, in the electronic transport, and in the leakage current. The granularity of the photoresist used to model the gate, will also introduce local variations in the shape of the gate itself.

With existing technology, it is impossible to associate the detailed atomic structure of individual nanoscale CMOS transistors corresponding to the characteristics of each device. Over the past decade, researchers have been focused to gain understanding of the fluctuations of the intrinsic parameters in nanoscale CMOS transistors, and they have therefore made use of numerical simulations, using more detailed mathematical models.

Key to any discussion of the fluctuations is to understand if the effect of a variation is fundamental (can be removed only with a change in the structure or in the operation of the device) or can be reduced with improvements in technology over the years [7].

A large number of effects of both types of variations have been documented in the literature. Examples include: highly random effects (random dopant fluctuations, RDF [8], line-edge roughness, LER [9], [10], line width roughness, LWR [10], [11], oxide thickness fluctuations, OTF [12], poly-silicon granularity PSG [13], interface trapped charges [14], non-uniformity of the charge at the interface [15]), proximity effects in the pattern (classic and OPC / RET [16]), proximity effects associated with stress (over layers, PMOS epitaxial, STI-induced [17]), proximity effects associated with the polish (STI and ILD [18]), proximity effects associated with the annealing (RTA-generated [19]), effects related to the device (pockets planted with grains of poly [20], oxide thickness [21]), and effects related to design (hot spots, falls [22]).

Of all the causes listed, it is assumed [23] that the main sources of fluctuations are:

Random Discrete Dopant (RDD)

Line Edge Roughness (LER) and Line Width Roughness (LWR)

Oxide-thickness variations (OTV)

1.1.1 Random discrete dopant (RDD)

The impact of the number and placement of dopant atoms on the characteristics of the device is crucial in determining the behavior of nanoscale semiconductor devices, because these factors cause random variations in the transistor threshold voltage (V_{th}) [24] - [26].

This can result in threshold voltage mismatch between transistors on die (intra-die variations) resulting in significant delay variation of logic gates and circuits [26]. The effect of random dopant fluctuations (RDF) on V_{th} increases

with technology scaling. This is due to the fact that the average number of dopant atoms in the channel of a transistor reduces with technology scaling. For example, assuming a doping density of $10^{18}/cm^3$, the average number of dopant atoms in the channel of a minimum size (width = 2 x length) 70 nm device (effective channel length of 40 nm) is approximately 100. The random variation in this small number of dopant atoms can result in significant variations in the V_{th} of the transistor. Since the V_{th} variation due to RDF can result in significant variation in the delay of an electronically circuit, a careful analysis of the effects of various sources of fluctuations is very important to make further progress in VLSI technology. In particular, threshold matching, which is important for some types of circuits such as SRAM and sense amplifiers, may be limited by these fluctuations [27].

For example in Fig. 1.3 the discrete dopants randomly distributed in a cube of 80 nm^3 with an average concentration of 1.48×10^{18} cm^{-3} are shown.

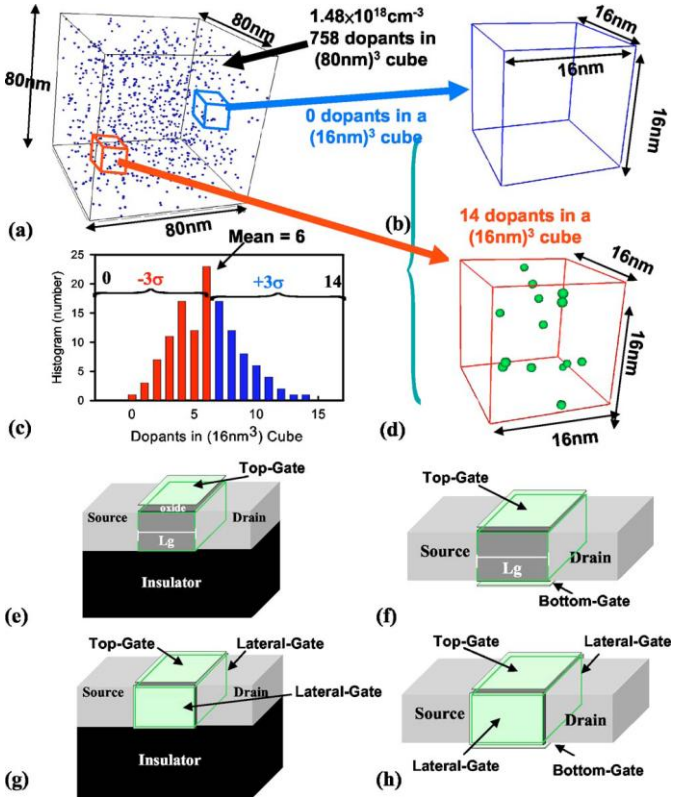


Fig. 1.3 (a) Discrete dopants randomly distributed in a cube of 80 nm^3 with an average concentration of 1.48×10^{18} cm^{-3}. There will be 758 dopants within the cube, but dopants vary from 0 to 14 (the average number is 6) within its 125 subcubes of 16 nm^3 [(b), (c) and (d)]. These 125 subcubes are then equivalently mapped in the corresponding regions of channel of the device single-gate (s), double-gate (f), triple-gate (g), and (square shaped) surrounded gate (h) for 3D simulations of sensitivity to the number-position of the dopants [Fig 1 of [28]].

It is seen that the random nature of the dopants causes appreciable dispersion of the threshold voltage, of the small signal parameters, and of the subthreshold

characteristics. Considerable efforts have been made to analyze the fluctuations of the threshold voltage induced by the random doping and to design structures resistant to the fluctuations of dopants [25], [29].

Different doping profiles (such as retrograde, halo and superhalo) have been proposed to reduce the threshold voltage fluctuations in ultra-small devices. The International Roadmap for Semiconductors has predicted a transition from conventional bulk devices to devices silicon-on-insulator (SOI) and then to the multi-gate SOI devices as a high-performance. As a result, nanoscale devices with vertical channel structures, such as double and triple gate and FinFET (fin-type field effect transistor) with gate surrounded, are of great interest. However, the channel doping must be used to alter the threshold voltage in the present semiconductor manufacturing processes.

Several approaches, such as small-signal analysis [30] - [32], drift-diffusion [33] - [35], and simulation with the Monte Carlo method [36] - [38] were adopted to study the issues concerning variations and fluctuations in semiconductor devices.

1.1.2 Line edge roughness (LER) and line width roughness (LWR)

Gate patterning is known to induce a no ideal (rough) edge herewith referred to as line-edge roughness (LER).

Current state of the art processes is able to consistently reproduce poly line widths below 100 nm. As the line width is scaled down, however, the roughness on the edge of the line does not scale. The total value of the LER is defined to be traditionally 3Δ, where Δ is the rms amplitude of LER that can be obtained statistically from inspection of the lines generated by a given lithography process [10]. The data collected by different processes, summarized in Fig. 1.4,

show that, at present, there is a minimum limit of the edge roughness of poly lines is typically on the order of 5 – 6 nm, but can have values much larger than that, depending on how the poly line was formed.

This value is larger than the requirements of the Roadmap for the devices below 100 nm, and it is alarming because the dimensional requirements are often more difficult to meet than the other specifications, for a given process.

The SIA national technology roadmap tells us that devices built as this scale are required to control gate length within approximately 8 nm [12].

A typical image of photoresist lines and spaces shows variation along the edge of the photoresist. Fig. 1.4 shows such edge variation. Measurements of the linewidth can be performed on such structures, and the resulting distribution of line widths can be determined.

Edge roughness in one of the primary concerns in controlling the gate length.

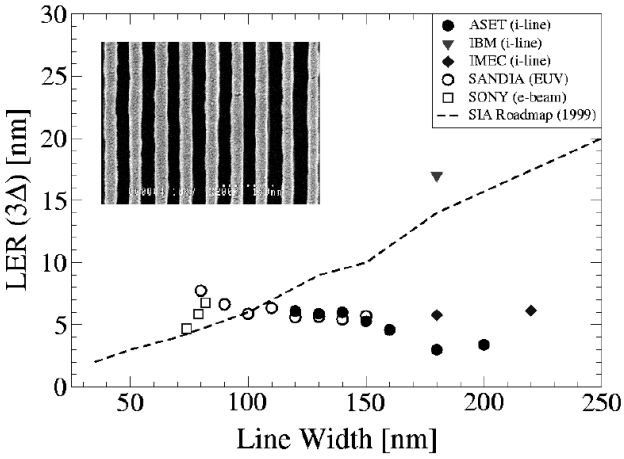


Fig. 1.4 LER found in advanced lithography processes and required by the SIA roadmap. The inset shows the LER found in the lines below 100 nm generated by electron beam [Fig 1 of [10]].

Line Width Roughness (LWR) is another source of variability. It is the phenomenon that the edges of device patterns irregularly wind and cause a variation in the pattern width [39] LWR is created during the etching and filling of the surrounding shallow trench insulation (STI) in the edges of the template flash cell [40]. The rough edges lead to a variation in device width along the length of the device altering the amount of current produced.

A larger gate LWR enhances the fluctuation in the subthreshold leakage current in short-channel n-MOSFETs even when the average gate length is maintained. Consequently, suppressing the gate LWR effectively reduces the variability in the threshold voltage of the scaled n-MOSFETs for a high drain voltage.

1.1.3 Oxide thickness variation (OTV) or Surface Roughness (SR)

The microelectronics industry owes much of its success to the existence of the thermal silicon oxide, i.e., the silicon dioxide (SiO_2). A thin layer of SiO_2, form the insulating layer between the control gate and the conducting channel of the transistors used in most modern integrated circuits. Since the circuits are more and more dense, all transistor sizes are scaled accordingly, so that today the thickness of silicon dioxide is 2 nm or less [41]. Over the years, the oxide thickness was decreased by a geometrical ratio with the next technology node, but clearly this trend has already saturated, because there are physical and practical limits on what can be done by a thin oxide film.

The essential physical limitations on the thickness of the insulator of gate, ignoring the "extrinsic" effects relating to the creation and production, are due to the exponential increase of gate current when the oxide thickness is reduced,

and the effect of this current feels, both on functionality, and reliability of devices and of circuits. For example in the Fig. 1.5 is illustrated a typical profile of the random interface Si/SiO_2 in a 30×30 nm^2 MOSFET.

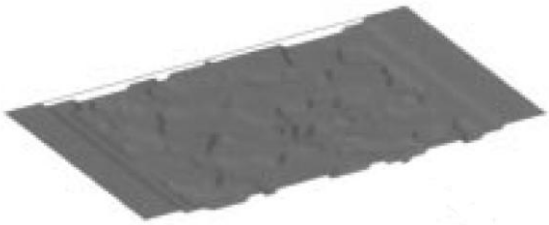


Fig. 1.5 A typical profile of the random interface Si/SiO_2 in a 30×30 nm^2 MOSFET [Fig 1 of [12]].

1.2 Mismatch

The design of analog circuits requires an in-depth understanding of the matching of components available in the various technologies. In MOS technology, the capacitors are widely used to design precision analog circuits such as data converters and filters, because of their excellent matching characteristics. The matching of MOS capacitors was discussed in detail [42] - [44]. However, all precision analog circuits cannot be designed using only capacitors. For applications such as high-speed data conversion, capacitive techniques are too slow. In addition, a digital VLSI process cannot offer linear capacitors. These factors push to study the matching of MOS transistors [45].

The mismatch is the process that causes time-independent random variations in physical quantities of identically designed devices [46]. The mismatch is a

limiting factor in the processing of general-purpose analog signals, but especially in "multiplexed" analog systems, analog-digital converters, reference generators, etc.. Matching may be also important in digital circuits, circuits for reading and writing digital memories, and in the noise margins of static RAM cells. The impact of (mis)match in the MOS transistor becomes more important due to the reduction of device dimensions.

The matching of the devices has been treated for years as an empirical method due to the absence of a modeling and systematic analysis. The mismatch is the time-independent variation in device parameters observed between two or more identically designed devices. This is because each step, necessary for the manufacture of integrated circuits, has several uncontrolled variations, related to the discrete nature of matter, fluctuation of temperature, mechanical stress, etc..

One can usually distinguish a global aspect and a local aspect. It is typically the result of gradients in the process, i.e., quantities which change progressively over the wafer. These are caused by instrumental variations and spatial derivatives, i.e., distortions in the photomask, lens aberrations, variation of the thickness of the photoresist, mechanical stress and variation in oxide thickness. The global variations produce systematic mismatch for a group of identically designed devices. Therefore, this can be minimized by using some "tricks", as the technique of the common centroid, by locating "matched" devices as close as possible, while maintaining the same orientation of the current, etc.. The local aspect of the mismatch on the variations that occur in the range of short-range, reflect changes in the total value of the component with reference to an adjacent component on the same chip, and it is related to the discrete nature of matter. Some causes are the diffusion and clustering of the dopants, interface states and fixed charges, edge roughness and effects of grain polysilicon. The

local variations produce a random mismatch that depends on the parameters of the process, the size of the device and the polarization. It must be clear to the designers a way to avoid the limitations imposed by the project.

The models of mismatch or use simple models of the drain current limited to a specific region of work [44] - [46] - [47] or complex expressions [48]. However, in general, is widely accepted that the matching can be modeled by random variations in the geometry, process and / or device parameters, and the effect of these parameters on the drain current can be quantified using the model of the dc transistor. As noted above in [47] and [49], there is an important flow of the current used in the dc model to analyze the mismatch that leads to inconsistent formulas. The implicit models of mismatch assume that the actual values of lumped model's parameters can be obtained from the integration of distributed parameters position-dependent area of the channel region of the device. As discussed in [47], the application of this concept of series or parallel combination of transistors leads to a result inconsistent due to the nonlinear nature of MOSFETs. Consequently, the simple use of fluctuations in parameters of the dc lumped model (V_{th}, β, etc.) is not appropriate to develop models of matching and new formulas must be derived from basic principles.

Several models of matching are proposed in the literature [45], [48], [50], [51].

In analog circuit blocks, such as A / D converters, differences of the threshold voltage of millivolts or less can set the performance and / or the yield of a product. Fig. 1.6 shows an example of how the physical effect of transistor matching influences the performance of an analog / digital converter.

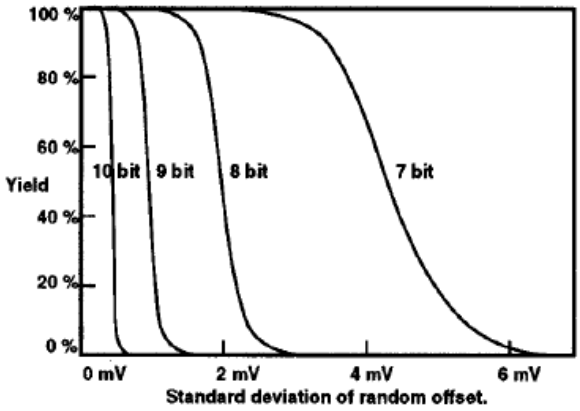


Fig. 1.6 Performance of analog-digital converters from 7... 10 bits depending on the standard deviation of the mismatch of the pair of input transistors [Fig 1 of [52]].

One effect of the mismatch of components is the offset voltage of operational amplifiers. This is the input differential voltage required to set the output to 0.

Since the operational amplifiers are typically used as part of more complex circuits, their uncertainty of the offset appears as a limitation of the specific circuit. Fig. 1.7 shows the standard deviation of the random offset in the pair of transistors that form the input of the comparator. Although the random offset is of the order of millivolts, it has a significant effect on the performance of the circuit.

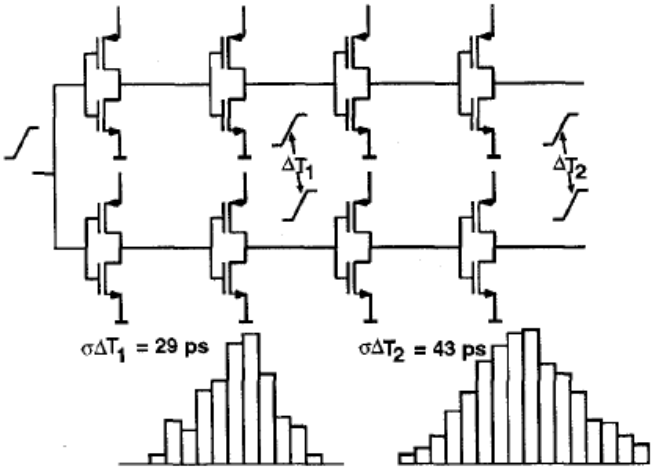


Fig. 1.7 Because of the matching of MOS, the clock signal propagates differently on both chains of inverters. The histograms show the simulated distribution in the 200 tests [Fig 8 of [52]].

ICs based on high performance CMOS are required for parallel processing. As a consequence, the quality of these paths in parallel (i.e., multiplexers, comparators, input stages, etc.) is important. Fig. 1.7 gives an example of how transistor matching influences the differences in the clock delay of the clock trees: different paths lead to losses in performance or yield in analog circuits or reduce the strength in digital circuits.

Fig. 1.7 shows the random portion of the clock skew between two branches of a tree clock, which is built in a CMOS process to 0.25 μm with 1/0.25 2/0.25 μm transistors. In digital circuits the amplitude of the variations of the skew of GHz clock will be comparable with the clock cycle.

1.3 Delays of multi-level logic

The variability of parameters due to the scaling of devices has, among other effects, the threshold voltage mismatch between transistors on the die (intra-die variations) resulting in significant variations of the delay in logic gates and other circuits. In addition, the effect of variations of V_{th} on the distribution of delay of a circuit depends strongly on the geometry of the device (channel length, width, oxide thickness, etc...) and the doping profile.

So, one needs a statistical model and analysis of the delay of logic gates (considering the variation of V_{th} due to RDF) both for the circuit and for the phase of the design of the device, in order to increase the efficiency of logic circuit in nanoscale systems.

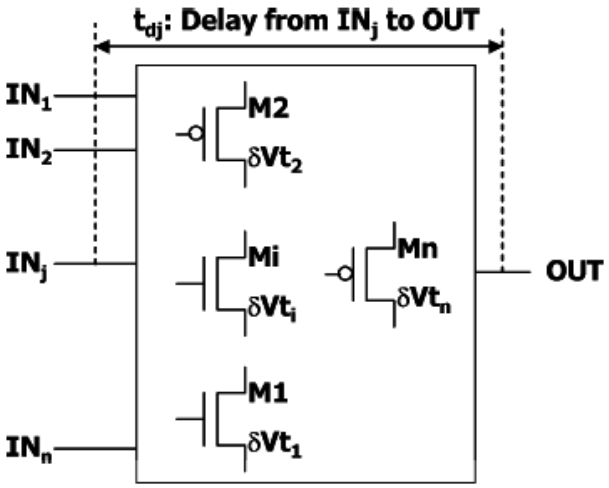


Fig. 1.8 General circuit with n transistors [Fig 1. of [8]].

We consider a typical logic gate with *n* transistors (Fig. 1.8). In general, the propagation delay from the input IN_J t_{dj} to output depends on the V_{th} of the n transistors (i.e., V_{Ti}). Therefore, considering the fluctuations of each transistor (δV_{Ti}) from their nominal value (V_{Ti0}), t_{dj} can be written as:

$$t_{dj} = f(V_{T1}, \dots V_{Tn}) = f(V_{t10} + V_{T1}, \dots, V_{tn0} + V_{Tn}) \tag{1.1}$$

Since fluctuations in V_{th} of the different transistors are independent from those in the other transistors, δV_{T1}, ..., δV_{Tn} are considered Gaussian random variables with zero mean.

There are two possible output transitions: from low to high (LH) and high-low (HL). Although you can design the gates with the same delay in the nominal case for both transitions, due to random variations of the process, these delays can be different. Therefore the total delay from INj at the output is given by $t_{dj} = Max(t_{djLH}, t_{djHL})$.

The distributions of the delay of logic gates in a library of standard cells can be obtained using the semi-analytical models proposed in [8]. Here we consider results for two basic logic gates, i.e., that are designed using the Berkeley Predictive Technology Models (BPTM) for 70 nm technology [52].

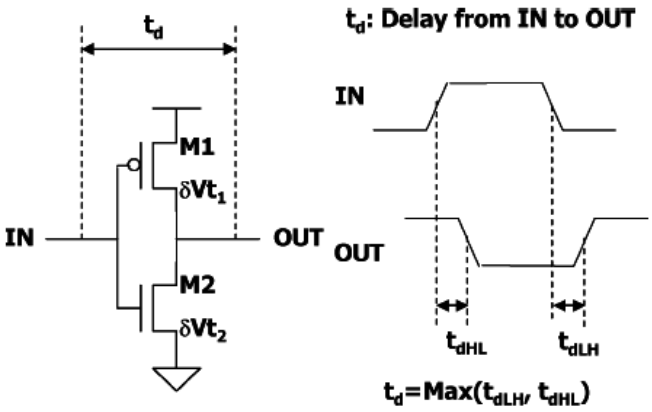


Fig. 1.9 Definitions of inverter and delay parameters [Fig. 2 of [8]].

Fig. 1.9 shows the definitions of inverter and delay parameters. The inverter is designed to have the same delay both for LH (t_{dLH}) and for HL (t_{dHL}) in the nominal case ($\delta V_{T1} = \delta V_{T2} = 0$). It may be noted that δV_T of the PMOS (δV_{T1}) has a strong impact on t_{dLH} (Fig. 1.10). On the other hand, it is sensitive mainly to t_{dHL} δV_T of the nMOS (δV_{T2}) (Fig. 1.10). The distributions of t_{dLH}, t_{dHL}, and t_d $(= Max(t_{dLH}, t_{dHL}))$ estimated using the proposed method are closer to the distributions obtained with the Monte Carlo method in SPICE (Fig. 1.11). It is noted that the application of the standard deviation of 30% of V_{th} ($\sigma_{V_{th}} = 30\%$ of the nominal V_{th}) leads to a dispersion of 5% (STD / Mean) in the overall delay of an inverter. Increasing the variation of V_{th} results in a larger dispersion of the delay [8].

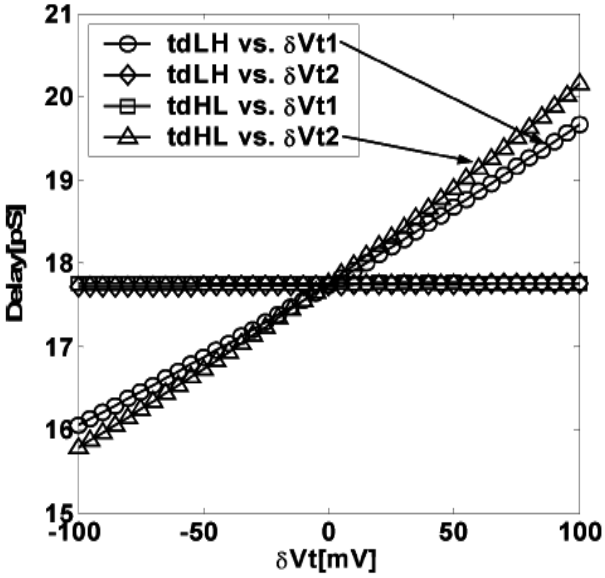


Fig. 1.10 Delay vs. δV_{th} for an inverter [Fig 3. of [8]].

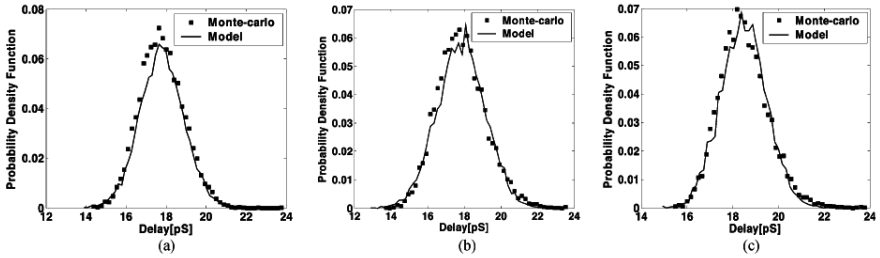


Fig. 1.11 Verification of the model: PDF (a) t_{dLH}, (b) t_{dHL}, and (c) t_d = Max (t_{dLH}, t_{dHL}) for an inverter (σ_{VT0} = 60 mV is chosen to obtain a considerable dispersion in the delay distributions; the Monte Carlo SPICE simulations are made to 10000 points) [Fig 4.of [8]].

The performance of CMOS logic circuits is significantly influenced by the amplitude of the deviations of the delay of the critical path due to the fluctuations of the intrinsic and extrinsic parameters. To estimate the impact of

these fluctuations of the parameters, a delay distribution of a static CMOS critical path is calculated from a rigorously derived device and circuit models [54]. Two possible options to achieve the desired performance are: 1) reduce the performance by operating at a lower clock frequency, 2) increase the supply voltage and, consequently, power dissipation, to satisfy the nominal critical path delay. For the 50 nm technology generation, delay increases by 12% - 29% and power dissipation of 22% - 46% have been estimated. They are due only to fluctuations of the extrinsic parameters such as effective channel length, gate oxide thickness and doping concentration of the channel. In comparison, when both intrinsic and extrinsic fluctuations have included in the analysis, delay and power dissipation increase to 18% - 32% and 31% - 53%, respectively, which demonstrates the importance of including intrinsic fluctuations in projects of the future CMOS logic circuits.

1.4 Noise margins of SRAM

The CMOS SRAM has a key role in modern Systems on Chip (SoC) [55]. However, the fluctuation of the intrinsic parameters of the devices increases with scaling. The new generations of SRAMs are therefore more sensitive to fluctuations at the atomic level, which are unavoidable through an external control of the manufacturing process.

Another important aspect in the design of an SRAM cell is its stability, which determines the probability of error and the sensitivity of memory to the operating conditions [56].

There is a tradeoff between the stability of an SRAM cell and area, for which a design for the improvement of stability leads to higher circuit area. In recent years many efforts have been made to create a model of stability of the cell flip-

flop. The stability of an SRAM cell is typically associated with the static noise margin (SNM), defined as the minimum static noise of the voltage needed to change the state of the cell. The static noise may be due to offset and mismatch due to manufacturing process and variations in operating conditions. An SRAM cell should be designed in such a way that is not modified by dynamic disturbances caused by alpha particles, crosstalk, power supply voltage variations and thermal noise.

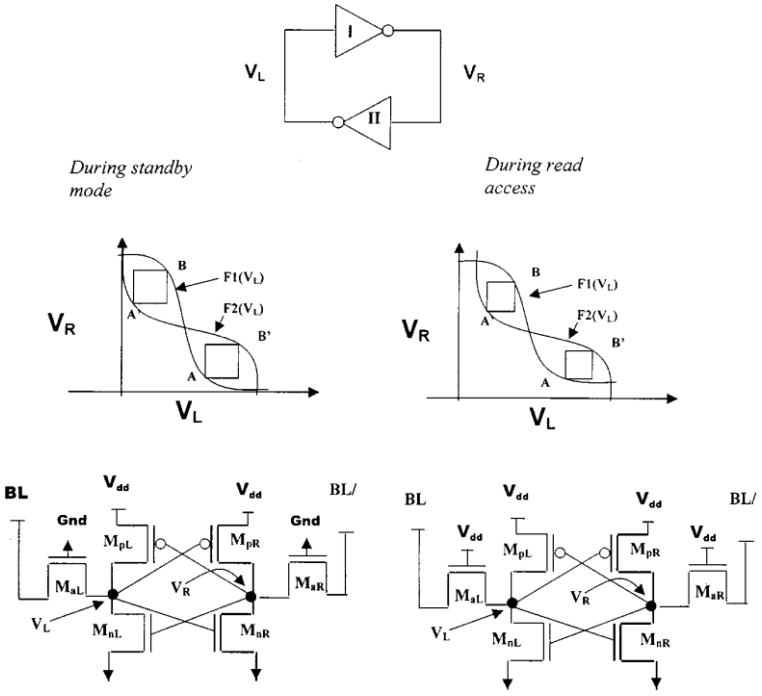


Fig. 1.12 Schematic of a CMOS SRAM cell during the standby mode and during the read access. The subscript "R" indicates the right side of the cell, the subscript "L" the left. The cell is more vulnerable at the noise during a read access [Fig 1 of [24]].

As shown in Fig. 1.12, there is a graphic method to determine the value of the static noise margin from the characteristics of inverter that constitute the cell: Static noise margin is the side of the largest square that can be included in each "wing" of the butterfly diagram. In the tracking of the features shown in Fig. 1.12 is used as a MOS basic model with constant threshold voltage and exponential subthreshold current.

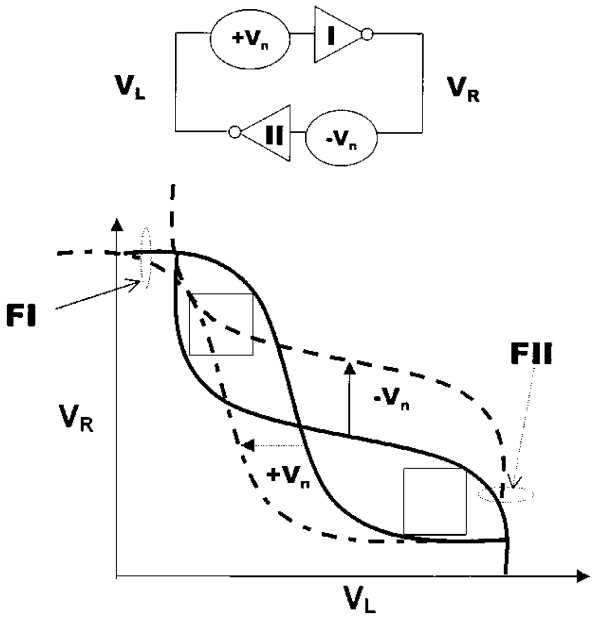


Fig. 1.13 The static noise margin is defined as the minimum noise voltage present at each of the cell storage nodes necessary to flip the state of the cell. Graphically, this may be seen as moving the static characteristics vertically or horizontally along the si side of the maximum nested square until the curves intersect at only one point [Fig. 2 of [24]].

From Fig. 1.12 and Fig. 1.13 we can see how one can build two squares within the butterfly cell, whose you can produce the noise margins and SNM_L and SNM_R. Only for a perfectly symmetrical unperturbed cell, SNM_L is equal to

SNM_R. In order to evaluate the SNM in the worst case, the static noise margin should be evaluated as follows:

$$SNM = min[SNM_L, SNM_R] \tag{1.2}$$

The optimization of the static noise margin is one of the most important aspects in the design of SRAM cells[24]. One of the possible solutions to improve the scalability of SRAM cells, providing immunity to the fluctuations of the intrinsic parameters, is the technique of bias control. In [55], in particular, it focuses on the approaches of polarization of the bit line and the polarization of the gate of the access transistor.

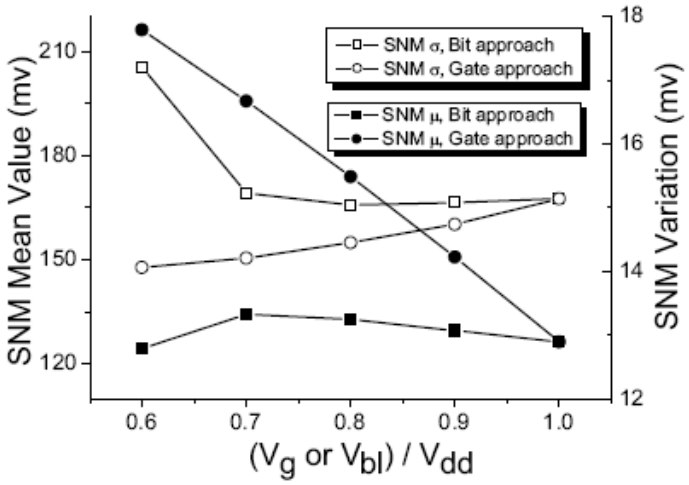


Fig. 1.14 Mean value μ and standard deviation σ of the SNM as a function of the polarization of the bit line and the polarization of the gate of access transistor [Fig 4 of [55]].

Fig. 1.14 shows the improvements of the performance of the static noise margin following these approaches. These studies indicate that the approach of the bit line bias leads to moderate improvements in the performance of the SNM, with an increase of the mean value of static noise margin μ of 6%. In contrast, the approach of biasing the gate of the access transistor improves noticeably the performance in terms of static noise margin. In fact, a decrease of 10% of the gate voltage of access transistor introduces an increase in the average value of the SNM (μ) by about 18%. Note that the standard deviation (σ) of the SNM decreases with decreasing gate voltage, with consequent benefits in terms of yield [55].

Fig. 1.15 shows the improvement of the static noise margin using a combination of the polarization of the bit line and the gate of the access transistor. In this way we can achieve improvements of over 40% of the SNM.

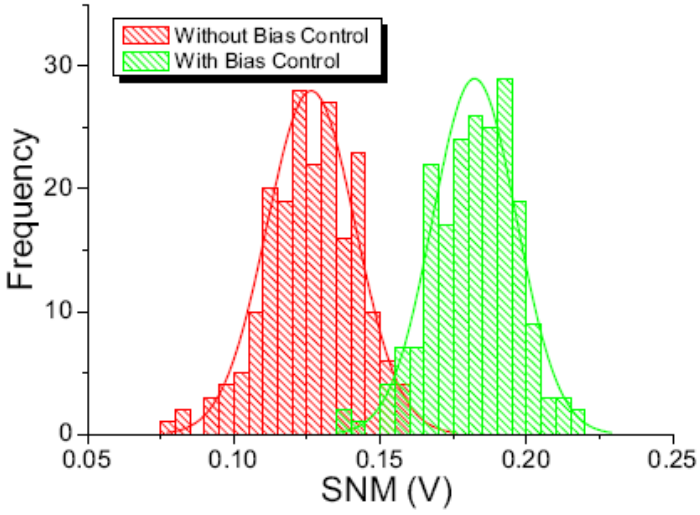


Fig. 1.15 Distribution of the noise margin with and without the control of the bias voltage [Fig. 5 del [55]].

At the device level, a proposed solution to alleviate the effects of intrinsic fluctuations [57] is to use the retrograde doping profiles in transistors and also use the channel lengths that are marginally larger than the minimum feature size. This alleviates the problem but does not eliminate it entirely. Circuit techniques that push away the tensions of the memory cell nodes from each other, improve the SNM of the cell, making it more immune to noise, are proposed in [58] as a solution that addresses the stability of the cell, the subthreshold leakage, and performance degradation in scaled CMOS SRAM.

2 State of the art

2.1 Random discrete dopant (RDD)

Wong and Taur were the first to propose a full 3D simulation of field effect transistors under the influence of random discrete dopants (RDD) [59]. They used a drift-diffusion simulator, which models the electronic transport as an incompressible fluid flow, considering the area under the gate like a checkerboard of smaller devices connected together, each with a different density of dopant atoms. The results show the two main effects of randomly distributed discrete dopants: a dispersion of the threshold voltage of the device, and a reduction of the average threshold voltage compared to the threshold voltage of the system with constant doping [4].

These 3D simulations prefigured the current techniques and used a 3D device simulator that is computationally efficient, but they were not taken immediately in commercial simulators. A 3D simulation requires significant computational resources; the simulations are even more important because of the need to have accurate models of atomic-scale lengths.

The dependence of the threshold voltage caused by the random fluctuation in the number of dopants in the channel of the MOSFET and of the random distribution at the microscopic level of atoms random discrete dopants in the channel of the MOSFET [59] has been studied, always using the drift-diffusion

approach, with the devices simulator FIELDAY [60]. Fig. 2.1 shows an example of a set of I-V curves of 24 MOSFETs with different random distributions of atoms for W = 50 nm, L = 100 nm, t_{ox} = 30 Å, and a uniform doping average of the substrate of $8.6 \times 10^{17}\,cm^{-3}$. When compared it with the I-V characteristic of the same MOSFET simulated using the conventional continuous doping model, the simulation of discrete doping shows: 1) a dispersion of I-V curves along the axis of the gate voltage of about 20-30 mV, 2) a average shift of the I-V in the direction of the negative gate voltages of about 30 mV in the subthreshold region and about 15 mV in the linear region, and 3) a light degradation (<3 mV/dec) and fluctuation of the subthreshold slope. The shift of V_{th} in the subthreshold region is lower than in the linear region due to the logarithmic dependence. The asymmetry of the threshold is about 20-40 mV, and this can be attributed to the discrete and random nature of the dopant atoms, resulting in an inhomogeneous channel potential [27].

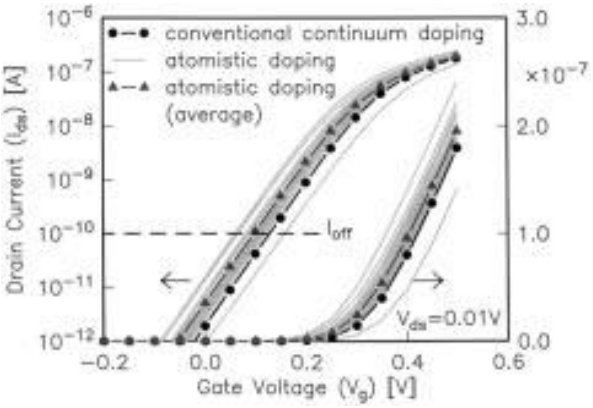


Fig. 2.1 Drain current as a function of gate voltage for a conventionally doped MOSFET (circles) and 24 devices with different distributions of discrete dopants in the channel (gray lines). The current average of all 24 devices is indicated with triangles. The shift of the threshold voltage in the subthreshold region was defined as the shift of gate voltage to a level of constant current (I_{off}) [Fig 18 of [27]].

Successively, Asenov *et al.* have carried out studies on the 3D atomistic simulations [61], [62]. For the first time was made of the systemic analysis of the effects of random doping in 3D on a scale sufficient to provide quantitative statistical predictions. It was adopted an approach to hierarchical atomistic simulations, always based on the drift-diffusion method. To reduce processing time and memory required for high drain voltages has been developed a self-consistent option based on a solution of the continuity equation of the current restricted to a thin slice of the channel. At low drain voltages, the single solution of the Poisson equation is sufficient to extract the current with satisfactory accuracy [61].

Dependencies and $<V_{th}>$ (threshold voltage of the device with constant doping profile) on the concentration of the dopants are compared in Fig. 2.2 for transistors with $L_{eff} = W_{eff} = 50$ nm and $t_{ox} = 3$ nm. The inset in the same figure

shows that the reduction of the threshold voltage induced by the random doping, increases almost linearly with increasing of the concentration of dopants [62]. In Fig. 2.3 is shown instead the dependence of $\sigma_{V_{th}}$ of the same concentration.

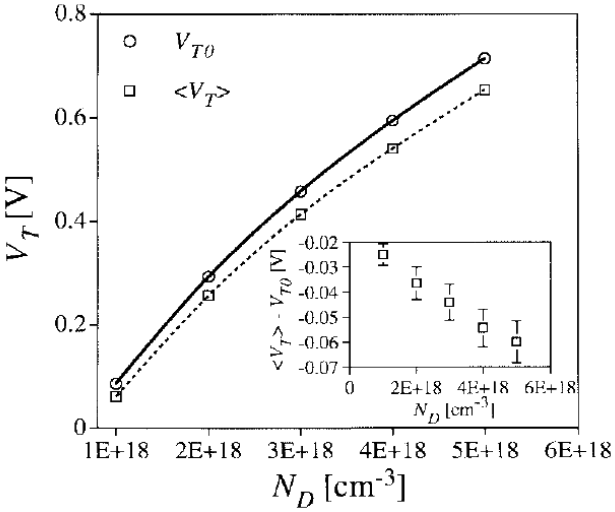


Fig. 2.2 Comparison of the dependence of the concentration of the dopants of $<V_{th}>$ for transistor with $L_{eff} = W_{eff} = 50$ nm and $t_{ox} = 3$ nm. Samples of 200 transistors [Fig 6 of [62]].

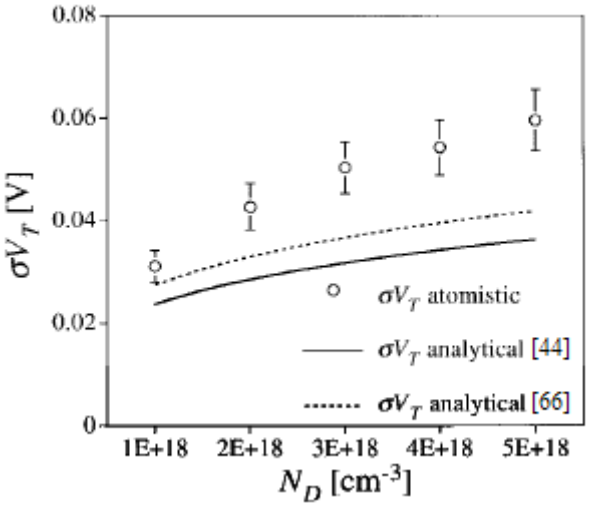


Fig. 2.3 Comparison of the dependence of the concentration of dopants by $\sigma_{V_{th}}$ calculated atomistically and the analytical models [45] and [63], with L_{eff} = W_{eff} = 50 nm and t_{ox} = 3 nm. Samples of 200 transistors [Fig 7 of [62]].

In 2003, Andrei and Mayergoyz included quantum-mechanical effects in the analysis, suggesting a very fast technique for the calculation of the threshold voltage fluctuations induced by variations of the dopants [30]. This technique is based on linearization of the equations of transport in respect of the fluctuating quantities, and it is, from point of view computational, very efficient, since it avoids several simulations for various doping (as in the case of Monte Carlo techniques).

The results for the standard deviations of the threshold voltage obtained for a MOSFET with a channel length of 50 nm are shown in Fig. 2.4 and compared with those obtained by Asenov *et al.* for various oxide thicknesses. In [64], $\sigma_{V_{th}}$ is calculated by simulating N = 200 MOSFETs, which implies errors of $\sigma_{V_{th}}$ of

about $1/\sqrt{2N} = 5\%$. The vertical bars in Fig. 2.4 correspond to the absolute value of these errors and show the range in which it has a probability of 68%. There is good agreement between the results extracted and those obtained using the statistical method in the case of classical computing. In the case of quantum calculations, the values are somewhat smaller than those reported in [64] due to the different electronic mass used in the simulations. The effective electron mass used in [64] $m_n^* = 0.18$ is smaller than that used in these simulations, therefore, the values reported on, are approximately 15% larger.

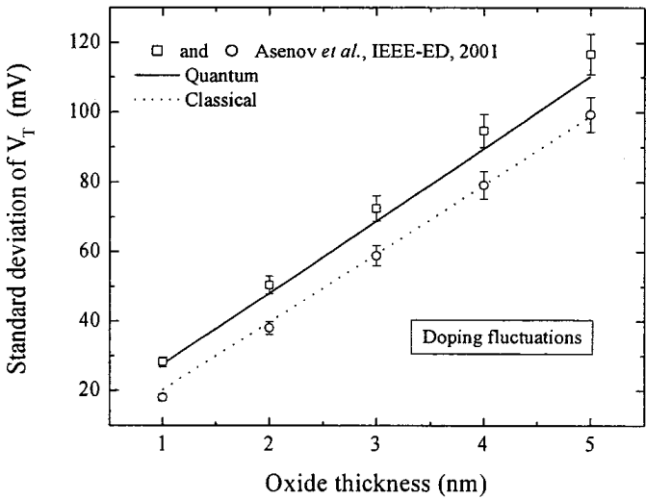


Fig. 2.4 Comparison of the $\sigma_{V_{th}}$ with Asenov *et al.* [64]. The effective electron mass used in the simulations presented in [64] is $m_n^* = 0.18m_0$ and it is different from the value $m_n^* = 0.21m_0$ used in [30], which justifies the difference in the calculations [Fig 5 of [30]].

By the comparison of data on the devices to 65 nm, can be seen that the RDF is about 65% of the total $\sigma_{V_{th}}$. Similar results are obtained when comparing the

results for the devices to 45 nm, where the RDF is about 60% of the total $\sigma_{V_{th}}$ [7] (Fig. 2.5).

To eliminate the effects caused by the RDF is necessary to introduce a destructive invention: one possibility is to use fully depleted devices (UTB or Trigate devices), where one can maintain control in the channel with a significantly lower channel doping.

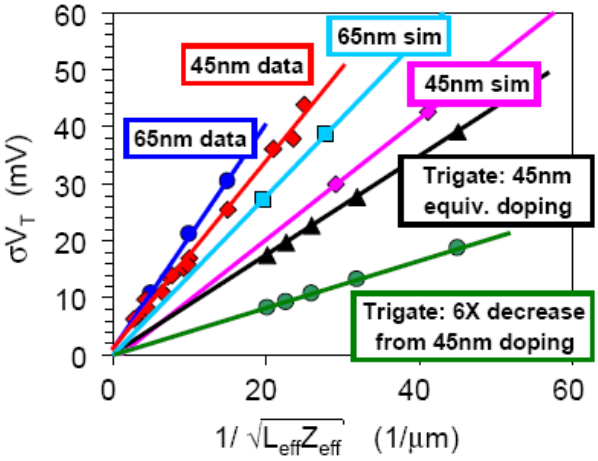


Fig. 2.5 Variations of the transistor 65 nm and 45 nm, the data are compared with the RDF simulations under conditions of equivalent doping [Fig 4 of [7]].

Recently, Reid *et al.*, using the Glasgow "atomistic" simulator, have performed 3-D statistical simulations of random-dopant-induced threshold voltage variation in state-of-the-art- 35- and 13-nm bulk MOSFETs consisting of statistical samples of 10^5 or more microscopically different transistors [65]. Simulations on such an unprecedented scale has been enabled by grid technology, which allows the distribution and the monitoring of very large

ensembles on heterogeneous computational grids, as well as the automated handling of large amounts of output data.

The simulator self-consistently solves the nonlinear Poisson and current continuity equations in a drift diffusion approximation. The resolution of the individual discrete dopants is enabled by employing density gradient (DG) quantum corrections for both electrons and holes [66], [67].

The simulated 35-nm n-channel MOSFET (based on a device originally proposed by Toshiba) has a complex doping profile featuring retrograde indium channel doping and source/drain pockets [68]. The microscopically different devices in the $>10^5$ simulated statistical samples are generated using a continuous doping profile, which has been extracted from carefully calibrated process and device simulation using commercial TCAD tools [69]. The simulated 13-nm MOSFET is a scaled version of the 35-nm device, based on generalized scaling rules with structural parameters and doping profiles guided by the requirements of the ITRS, and represents a limiting case for conventional bulk MOSFET scaling [70][23].

In order to better understand the physical mechanisms whereby RDD affects V_{th}, the study the surface potential distributions of devices from close to the mean, and from the upper and the lower tails of the distributions, at the identical gate voltages. There surface potential plots for both the 35- and 13-nm devices are shown in Fig. 2.6a and Fig. 2.6b, respectively. At both channel lengths, the behavior of the devices with higher V_{th} is determined by the clustering of the dopants across the channel width at the location of the maximum of the potential barrier between the source and the drain. At this position, the dopants have the maximum impact on V_{th} by almost completely blocking the current path. Conversely, the behavior of the transistors is determined by the lack of

dopants in the part of the channel that is near the potential barrier maximum, creating an open current path that is responsible for the low V_{th}.

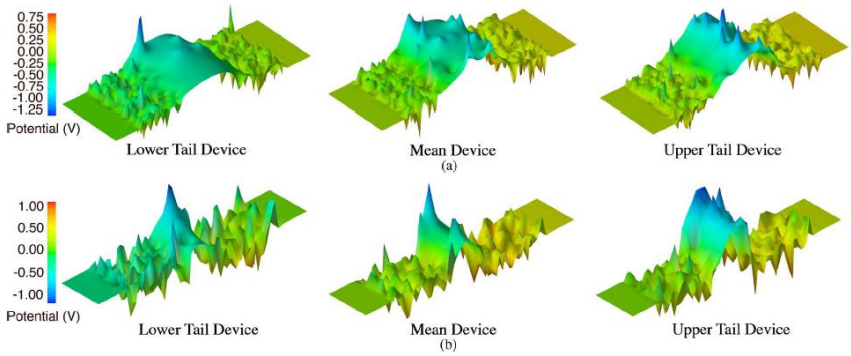


Fig. 2.6 Raw electrostatic surface potential profiles for the devices in the lower part, the middle, and the upper part of the distributions. (a) 35-nm devices, (b) 13 nm devices [Fig 6 of [65]].

In order to study the asymmetry in the V_{th} distribution induced by the random dopant distribution, they have done a more detailed analysis to determine the statistically significant region (SRR) of the transistor that dominates the statistical behavior of the device ensemble. Moreover it is necessary to fix the number of dopants within the SSR (N_{SSR}): this is possible by estimating the distribution of the threshold voltage caused by the random partitioning of the dopants and calculating their mean and standard deviation.

From their analysis it becomes clear that the asymmetry in the random dopant induced threshold voltage distribution is due to factors: first, the Poisson distribution for a fixed value of N_{SSR} is asymmetric with a positive skew, and this asymmetry increases as N_{SSR} is reduced, and, second, the standard deviation σ_{NSSR} increase with N_{SSR}.

This statistical analysis has identified the SRR of the devices, in which the number of dopants and their positions are closely correlated to the threshold voltage variation. The asymmetry of the distribution stems from the linear dependence of the mean and the standard deviation of the threshold voltage on the number of dopants in the statistically important region, as shown in Fig. 2.7.

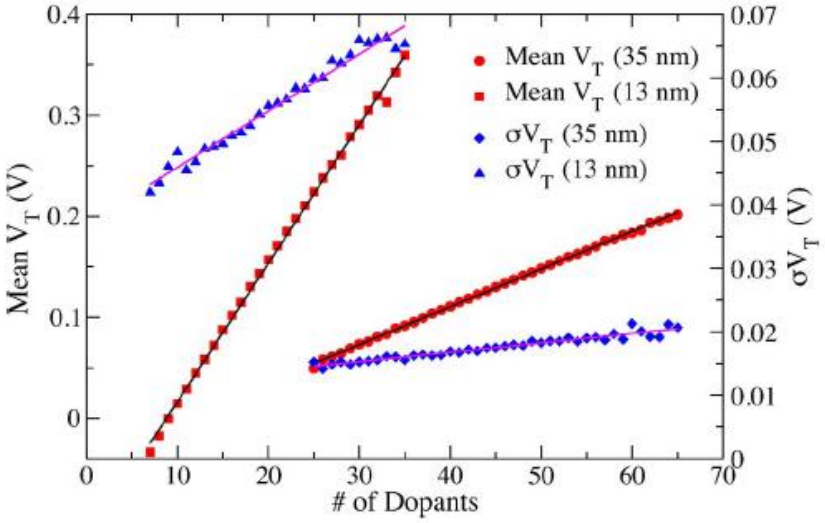


Fig. 2.7 Dependence of the V_{th} mean and standard deviation as a function of N_{SSR} for both devices. The linear dependence allows the positional effects on V_{th} to be extrapolated out to larger values of σ [Fig 10 of [65]].

2.2 Line edge roughness (LER)

The impact of line edge roughness (LER) on the performance becomes increasingly important with size reduction. In fact, due to the rapid technology development, the process of defining the gate are not yet mature and the LER is typically high [71].

Recently, attention aimed at the study of the effects of line edge roughness of the gate on the performance of short-channel MOSFET is increasing. Device simulations have suggested that the LER of a gate in a short-channel MOSFET with large gate width results in a significant increase in leakage current, while causing a minimum increase of current in strong inversion [72]. The effect can be observed with the experiments comparing the currents of MOSFETs with different gate LER [73].

The experimental study of the doping profile and the extraction of relevant information from real devices is difficult and expensive and therefore almost exclusively numerical simulations are used [72].

The effects of LER depend on the technological process. In order to identify processes that provide intrinsic performance improvements, it is important to be able to separate the effects of LER by other factors.

In the ideal case, without taking into account the effects of LER, two transistors with the same gate length L, would have the same I_{on} / I_{off} graphs. However, when including the effects of LER, the transistors, while having the same length L, are characterized by different edge profiles. The resulting dispersion of the characteristics could also cause a shift of the mean value compared to the I_{on} / I_{off} ideal curve. In Fig. 2.8 the Gate LER of the gate in a traditional MOSFET is illustrated.

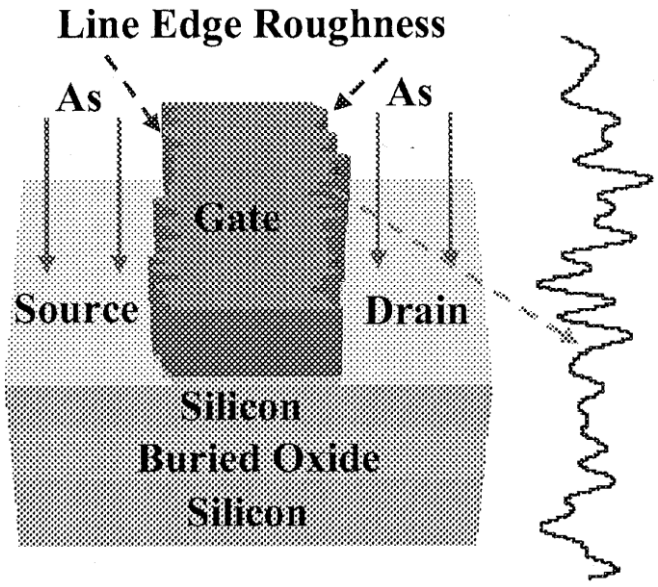


Fig. 2.8 Illustration of a MOSFET with LER at the gate [Fig 1 of [72]].

Simulations for the study the LER are made using the approximation of a two-dimensional device [71]. The width of the transistor is divided in different segments of the two-dimensional devices with a width equal to the characteristic spatial period of the LER (called correlation length Λ), as shown in Fig. 2.9. Initially are determined I_{ds} and I_{off} for each portion of the device and are then added together to obtain the ratio I_{on} / I_{off} for the entire device.

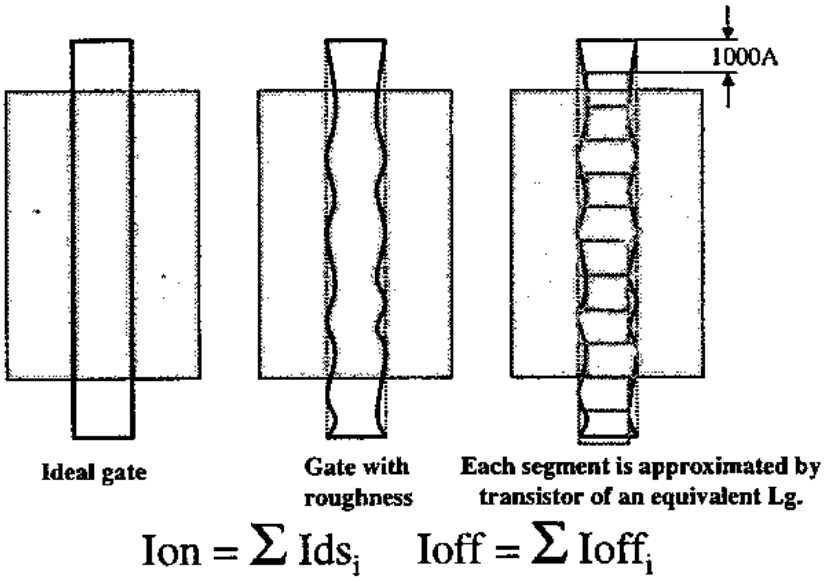


Fig. 2.9 A wide MOSFET with gate LER significant can be schematized with the parallel of many gate with equal width to the correlation length of the LER and gate length constant [Fig 1 of [71]].

Xiong *et al.* use a quick and convenient experimental method to extract and characterize the LER of the polysilicon gate [72]. This method consists in finding a controllable approach to produce a variable LER of the poly gate and integrate it into the process flow of MOSFET.

Software was then developed to extract the trends of the profile of the gate from the processing of data recorded on the poly lines using a scanning electron microscope (SEM). These measurements are performed on each line of poly (Fig 2.10).

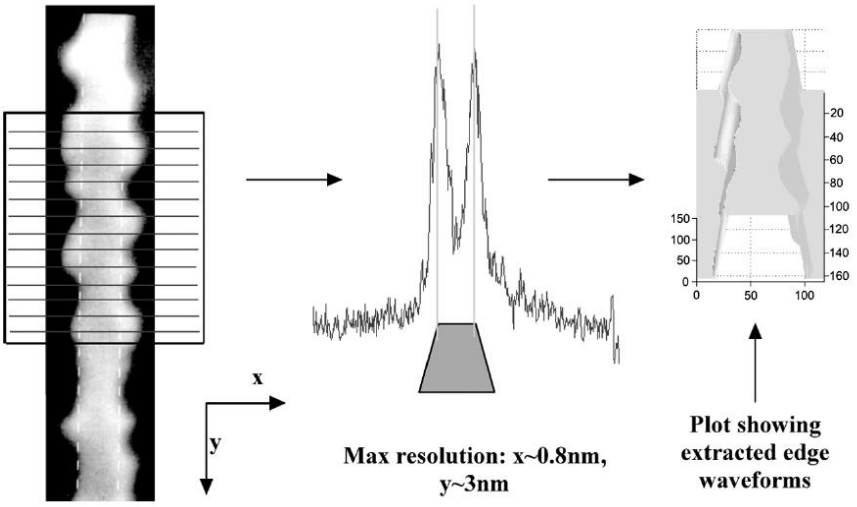


Fig. 2.10 Extraction of the waveform of the line edge of the current data recorded by SEM [Fig 2 of [72]].

This method requires a large amount of data recorded by SEM, but does not require special tools for adjustment except for some considerations related to the resolution. The extent of the LER at this point can be obtained from analysis of the data that describe the shape of the contour of the poly. A Gaussian distribution cut at ± 3 dB can be used to approximate the statistical deviation of the edge. RMS values (indicated with Δ) of the profiles within the day are fairly consistent, if compared with the relatively large variations within the wafer.

Xiong *et al*. found an effect of LER observable on the curves I_{on} / I_{off} .[72]. However, it was shown that this effect is significantly smaller than those caused by other changes in the process of the wafers where LER is generally quite small. With the help of numerical simulations it was noted that the main effect of gate LER on the device occurs in the doping profile. The scattering of dopant

implantation and diffusion smoothes partially "roughness" of the edges of the gate introduced by LER and reduce its effect on the curves I_{on} / I_{off}.

Finally, it was concluded that, by minimizing the gate LER (mean square value of the border lines less than 2 nm), its effect is secondary to other process variations for devices with gate lengths of 40 nm or larger.

In 2001 it was proposed an analytical model of the LER [73], which has thus provided an efficient and accurate estimation of the effects related to it.

It is possible to distinguish two types of LER: short radius and long radius. The short radius LER has a high spatial frequency (i.e., a characteristic length of ~ 1 nm) and it is mainly attributable to the conditions of the lithography process and of the resist. On the other hand, we speak of long range LER for characteristic lengths greater than 10 nm; this type of LER is mainly due to surface roughness of polysilicon. The Tab. 2.1 summarizes the impact of various lithographic processes on the LER.

Lithography approach	Short range LER [nm]	Long range LER [nm]
193 BIM	8.3	9.3
248 APSM	5.9	6.5
248 Alt PSM	3.5	4.1

Tab. 2.1 LER for 100-nm resist lines [Table II of [73]].

LER-induced variability has been the subject of numerous modelling and simulation studies of different degrees of complexity and sophistication. The use of 2-D simulations of devices with different channel lengths in combination with the statistics of different channel length occurrences in the presence of LER has been popular due to the low computational burden [74], [75].

Comprehensive 3-D simulations vary in the complexity of the LER description from square wave approximations [76], [77] to realistic statistical descriptions of the gate edge based on different autocorrelation functions fitted to experimental LER data [78]. More sophisticated 3-D simulation studies include the confluence of LER and atomic-scale process simulation [79] and the impact of LER-induced strain variations [80]. However, a common denominator in all of the published 3-D simulations studies is the relatively small statistical sample, which rarely exceeds 200 microscopically different devices.

In [81], Reid *et al.* present a comprehensive 3-D simulation study of LER-induced MOSFET threshold voltage variability using statistical samples of more than 10^4 transistors. Contemporary, bulk, ultrathin-body (UTB) SOI, and double-gate (DG) MOSFETS have been simulated and analyzed. The large size of the simulated statistical samples allows accurate estimation of the higher order moments and the shape of the distributions of V_{th}. Intensive statistical data mining is also used in order to explain the specific shape of the simulated distributions.

The simulations presented in [81] were carried out using the well-established Glasgow 3-D "atomistic" statistical device simulator [23]. Random gate LER patterns are introduced into the simulations using 1-D Fourier synthesis, as described in [78]. A Gaussian autocorrelation function characterized by an RMS amplitude (Δ) and correlation length (Λ) has been adopted to describe LER, as in previous LER simulation studies [78]. In all simulations reported in [81], values of $\Delta = 1.6667$ nm and $\Lambda = 30$ nm have been used to generate random source/drain and gate edges introduced by roughness of the resist and the following gate patterning process. This corresponds to LER patterns with magnitude 5 nm, which is representative of the state-of-the-art 193 nm

lithography expected to be used in manufacturing of the 32- and 22-nm technology generations.

The 3-D doping profile in the presence of LER is generated using 2-D process simulation results and by assuming that the *p-n* junctions follow the same pattern as the gate edge. This is a reasonable approximation in contemporary CMOS technology, where laser scan annealing is progressively applied for doping activation.

In order to confirm the trends observed in the simulations of the bulk 35-nm MOSFET, and to examine the potential impact of LER on different device architectures, smaller ensembles of three other devices have been simulated. The selection includes an LP 42-nm physical gate length bulk MOSFET with an oxide thickness of 1.7 nm, developed by ST Microelectronics and described in detail in [83]; a 32-nm physical gate-length UTB SOI MOSFET with a body thickness of 7 nm and equivalent oxide thickness (EOT) of 1.2 nm; and a 22-nm DG MOSFET with a body thickness of 10 nm and EOT of 1.1 nm. The last two devices were developed by the PULLNANO consortium [84] and are described in detail in [85]. The distributions of threshold voltage for 35- and 45-nm bulk; 32-nm SOI and 22-nm DG devices at low drain voltage (V_D = 100 mV) are presented in Fig. 2.11. For all devices, the shape of the distribution of V_{th} is similar and all are negatively skewed. It should be noted that the SOI MOSFET in particular exhibits good immunity to LER-induced variability, having a standard deviation of V_{th} that is much lower than the other three devices. This is due to better electrostatic integrity and hence reduced short-channel effects.

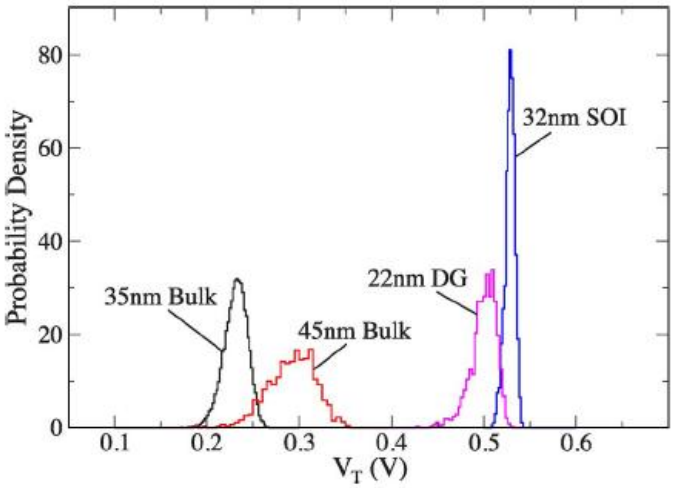


Fig. 2.11 Comparison of the distribution of V_{th} due to LER in the four simulated devices at V_{DS} = 100 mV [Fig 8 of [81]].

The skew and kurtosis values for the four simulated devices are given in Tab. 2.2, along with the other moments of the statistical distributions. The values of the moments also confirm the visual observation that the SOI device has significantly better immunity to LER-induced fluctuations.

	35 nm Bulk	45 nm Bulk	32 nm SOI	22 nm DG
# Simulated	25,000	1,000	1,000	1,280
Min (mV)	159.4	187.3	508.8	427.7
Max (mV)	271.9	351.9	541.6	529.3
Mean (mV)	231.1	292.2	528.3	499.7
St. Dev. (mV)	12.75	24.91	5.25	13.84
Skew	-0.4066	-0.3854	-0.5199	-0.9615
Kurtosis	0.2547	0.1121	0.1989	1.444

Tab. 2.2 Summary of the statistical moments of the distribution of V_{th} at low drain in all four transistors [Table III of [81]].

Analysis of the simulation results of large statistical samples of a 35-nm MOSFET subject to LER has revealed that the distribution of the threshold voltage is asymmetrical with a negative skew, which increases with drain bias. There is a very strong nonlinear correlation between the threshold voltage and the average channel length of the LER transistors that very closely follows the channel length dependence of the threshold voltage in transistors with uniform gate edges. Increasing the channel width reduces the threshold voltage standard deviation more slowly than $1/\sqrt{W}$ and improves the symmetry of the distribution of threshold voltage. The asymmetry of the distribution and the strong nonlinear correlation between the threshold voltage and the average channel length was also confirmed in the simulation of a 42-nm physical channel-length bulk LP MOSFET, a 32-nm channel-length thin-body SOI MOSFET, and a 22-nm channel-length DG MOSFET [81].

2.3 Oxide thickness variation (OTV) or Surface Roughness (SR)

In past years, the threshold voltage fluctuations induced by random variations of oxide thickness have not received the same attention to fluctuations induced by random doping. However, Andrei and Mayergoyz in their analysis [30], considered in addition to the threshold voltage fluctuations induced by variations of the dopant (RDD), even those induced by the oxide thickness variation (OTV).

The surface oxide was initially characterized by a Gaussian autocorrelation function. However, measurements made more recently have shown that fluctuations in the oxide thickness are better described by a distribution function of exponential type [30].

Asenov *et al*. have studied the intrinsic threshold voltage fluctuations introduced by local variations in oxide thickness (OTV) in decanano MOSFETs, using three-dimensional numerical simulations on statistical scale [12]. Si/SiO_2 random interface is generated by the power spectrum corresponding to the autocorrelation function of the interface roughness. The simulations showed that the intrinsic fluctuations of the threshold voltage induced by OTV become significant when the device size becomes comparable to the correlation length (Λ) of the interface.

The dependence of $\sigma_{V_{th}}$ by the average oxide thickness is shown in Fig. 2.12; in the simulations has considered only the Si/SiO_2 interface roughness. This trend is related to the linear dependence between V_{th} and t_{ox}, resulting in a constant variance with respect to the oxide thickness.

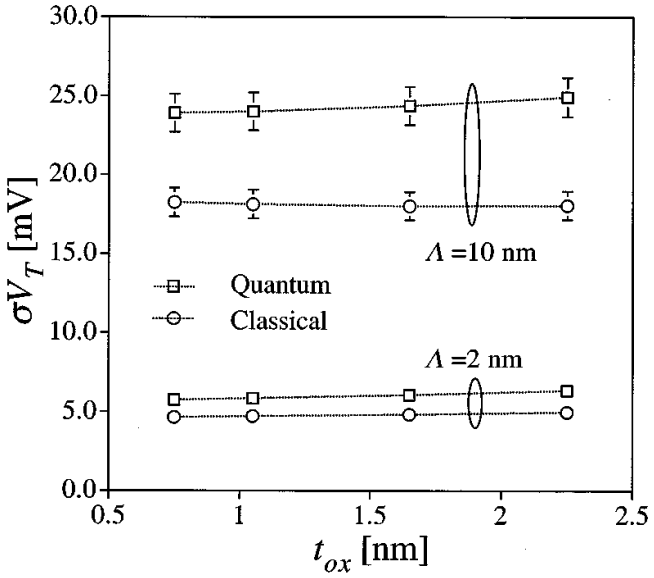


Fig. 2.12 Dependence of threshold voltage standard deviation of the average oxide thickness <t_{ox}> for a 30 × 30 nm^2 MOSFET with a random Si/SiO_2 interface [Fig 6 of [12]].

The Fig. 2.13 shows the dependence of the mean threshold voltage <V_{th}> vs. <t_{ox}> calculated from classical and quantum method, for Λ = 10 nm. For comparison the dependence of the threshold voltage V_{th} of the device with uniform oxide thickness (t_{ox}) is plotted as a function of t_{ox}. It is clear that the threshold voltage <V_{th}> of the device with random thickness is very close to the corresponding threshold voltage of the MOSFET with uniform oxide. The slight reduction of <V_{th}> in the classic case is associated with the increase of the current density at the boundaries between the regions of thinner oxide and thicker region due to the effects of the intensification of the field [12].

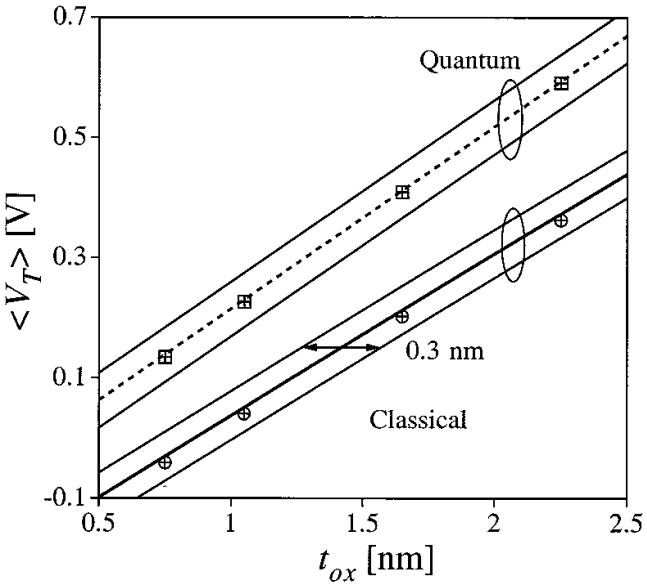


Fig. 2.13 Dependence of the average threshold voltage $< V_{th} >$ from the medium thickness oxide $<t_{ox}>$ for a 30 × 30 nm^2 MOSFET with a random Si/SiO_2 interface (symbols) and of the threshold voltage V_{th} on the oxide thickness t_{ox} for a similar device with uniform oxid (lines) [Fig 7 of [12]].

2.4 Combined effects

Asenov *et al.* carried out some 3D simulations to study the dispersion of the intrinsic parameters, not considering only the effects individually, but also the combined effects of RDD, LER and OTV on the fluctuations of the threshold voltage [23].

The study of the fluctuations of the threshold voltage was done using 3D drift-diffusion simulations (DD). The DD approximation does not capture the effects of the non-equilibrium transport of carriers and therefore underestimates the drain current in the ON state. However, this is adequate for the calculation

of the threshold voltage and its variations based on the criterion of current in the subthreshold region when the Poisson equation is decoupled from the equation of continuity of current; the electrostatics dominates the device behavior and the density current depends exponentially by the surface potential and its fluctuations. Even at very short channel lengths, the quantum corrections "Density Gradient" implemented in the simulator capture well the effects associated with the direct source-drain tunneling, which become apparent in the channel lengths of 10 nm, and accurately reproduces the results of simulations of the non-equilibrium Green's function (NEGF).

In the case of RDD simulations, the generation of distribution of random doping is based on continuous distribution of dopant obtained from the whole simulation process of the device of reference and of the scaled device: all sites in the lattice of silicon that covers the simulated device are controlled one by one. The dopants are introduced randomly in sites with a probability given by the corresponding dopant concentration ratio of silicon, using a rejection technique. Each dopant is assigned to the eight surrounding nodes of the grid using the cloud-in-cell technique (CIS) commonly used in Monte Carlo simulations.

With less than 10 dopants in the region emptied of the channel of most of the simulated devices, the resolution of each individual dopant becomes very important. However, the resolution of individual charges in the simulations "atomistic" using a dense grid creates problems. Due to the use of Boltzmann statistics or Fermi-Dirac in the approach to drift-diffusion classic, the electron concentration follows the electrostatic potential obtained from the solution of the Poisson equation. Consequently, a significant amount of mobile charge can be trapped (localized) in Coulomb potential wells created by discrete dopants assigned to the dense grid. The trapped charge artificially increases the

resistance of the source-drain regions and change the depletion layer that results by the reduction of the threshold voltage. Another damaging effect of this trapping of charges in classical simulations is the strong sensitivity of the amount of charge trapped on the size of the grid.

If using a denser mesh, you get a solution to the single Coulomb potential well and increases the amount of trapped charge.

Attempts were made to correct these problems in the "atomistic" simulations both with the charge distribution on more grid points and with the decomposition of the Coulomb potential in the short-and long-range components, based on considerations of screening. The charge-smearing approach is however purely empirical and may result in a loss of resolution on the effects of "atomistic" scale. The splitting of the Coulomb potential in the short and long range components suffers from some disadvantages including the arbitrary choice of cutoff parameters and for double counting of the potential of screening of the mobile charge.

Typically the approach adopted in the "atomistic" DD simulations is the DG approximation. Typically, for example, in the DG simulations of n-channel MOSFETs, this is enough to solve the Poisson equation self-consistent, the current continuity equation for electrons and the equation of state for electrons written in approximation DG.

$$2b_n \frac{\nabla^2\sqrt{\mathrm{n}}}{\sqrt{\mathrm{n}}} = \phi_n - \Psi + \frac{k_B T}{q} ln\left(\frac{n}{n_i}\right) \tag{2.1}$$

where $b_n = \hbar^2/12qm_n^*$. This takes care of the quantization of the electrons in the Coulomb potential well associated with the donors and it remedies at the

problems associated with the trapping of electrons in the source-drain regions doped atomistically. In order to avoid the trapping of holes in the substrate non depletion atomistically doped, is added to the DG equation of state of the holes, but without solving the equation of continuity of the current for the holes.

$$2b_p \frac{\nabla^2\sqrt{\mathrm{p}}}{\sqrt{\mathrm{p}}} = \Psi - \phi_p + \frac{k_B T}{q} ln\left(\frac{p}{p_i}\right) \tag{2.2}$$

This modified system of equations is solved self-consistently by using Gummel iterations, solving first the Poisson equation together with the modified equations of state for electrons and holes, and substituting this result into the continuity equation of the current in the form of a quantum potential correction.

The allocation of dopants in the "atomistic" simulations and the strong variation of the electric field around the individual discrete dopants make the direct use of traditional models of the concentration and of the mobility field-dependent unfeasible in RDD simulations. To obtain reasonable values of mobility in the simulations, they adopted the following approach. The mobility of the simulation of a MOSFET with a constant doping profile is mapped and stored for relevant combinations of gate and drain voltages. In atomistic simulations for a particular set of applied voltages, the corresponding mobility map from simulations of continuous doping is loaded and used in the atomistic devices. This approach is good enough when one estimates the changes in threshold voltage, which are virtually independent of mobility, but may be less appropriate for simulations of the variation of the current.

The line edge roughness is introduced in the simulations using a method based on one-dimensional Fourier synthesis, which generates random gate edges of a power spectrum corresponding to the Gaussian autocorrelation

function. The parameters are the correlation length Λ and the rms amplitude Δ. Similarly, the generation of the random interface that determines the OTV is also based on the Fourier synthesis, first used to generate a surface two-dimensional (2D) randomly from a power spectrum corresponding to the exponential autocorrelation function. This random surface is then generated and then quantized at the Si/SiO_2 interface in steps with a height equal to the atomic spacing.

The reference 35-nm MOSFET has a complex doping profile characterized by a retrograde doped with indium in the channel and source-drain pockets. The continuous doping profile was been obtained from a simulation process carefully calibrated using Taurus Process.

From the dopant concentration profile shown in Fig. 2.14 we can see that the very low concentration of dopant is beneficial not only to avoid the reduction of mobility, but also to reduce the RDD which induces fluctuations of the threshold voltage.

The Glasgow atomistic simulator was carefully calibrated in accordance with the measured characteristics of the device, using a constant doping profile in the simulations.

The comparison of the experimental and simulated I_D-V_G characteristics can be detected a good deal.

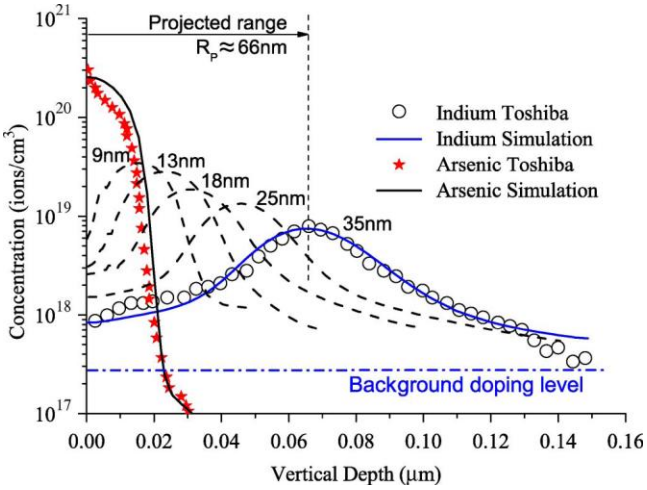


Fig. 2.14 Doping profiles of the one-dimensional channel of the MOSFET calibrated at 35 nm and of the scaled devices with gate length of 25, 18 , 13, 9 nm [Fig 3 of [23]].

The reduction in size is based on the generalized scaling rules and it follow closely the "guide" of the ITRS in terms of equivalent oxide thickness (EOT), junction depths, doping and supply voltage. The intention is also to preserve the main characteristics of the 35-nm reference MOSFET and, in particular, to maintain as low as possible the concentration of dopant in the channel at the interface.

Samples of 200 microscopically different square gate devices are been simulated for each channel length and for each source of fluctuations in the intrinsic parameters in order to extract the mean and standard deviation of the threshold voltage. Fig. 2.15 summarizes the channel length as a function of the standard deviation of threshold voltage $\sigma_{V_{th}}$ and the lowering of the threshold voltage ΔV_{th} associated with RDD.

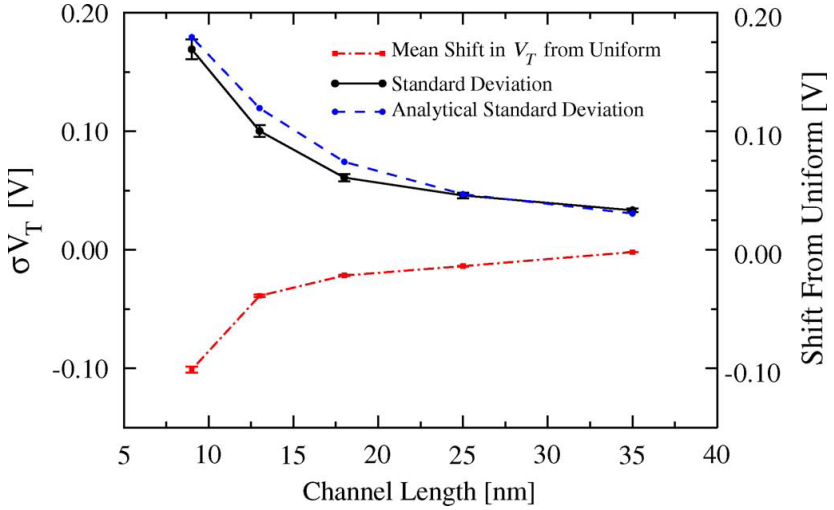


Fig. 2.15 Standard deviation of the threshold voltage and shift of the threshold voltage for transistors with gate lengths of 35 -, 25 -, 18 -, 13 -, and 9-nm due to the RDD [Fig 5 of [23]].

In the same figure we show the standard deviation of the threshold voltage obtained using the empirical expression for $\sigma_{V_{th}}$ based on 3D statistical simulations of idealized MOSFET with uniform doping profile of the channel and compensated to take into account the quantization in the inversion layer. The standard deviation of the threshold voltage is approximated by:

$$\sigma V_{th} = 3.9x10^{-8} \frac{(t_{ox}+z_0\varepsilon_{ox}/\varepsilon_{Si})N_A{}^{0.4}}{\sqrt{L_{eff}W_{eff}}} \quad (2.3)$$

where z_0 is the center of the inversion layer charge. The expression is coincident with the results obtained from numerical simulations.

The LER simulations can follow two scenarios. In the first scenario, shown below, LER = 3Δ = 4 nm was used for all channel lengths that represent the

current state of lithography and following the assumption that scaling of the LER is a very difficult task due to the molecular structure of photoresist. In this case the relative standard deviation of the LER, which is initially lower than that of the RDD, exceeds it at the channel length of 18 nm and reaches the absurdity of almost 400 mV to 9 nm. This is complemented by the reduction of the threshold voltage approximately at 100 mV. In the second scenario, the LER follows the prescribed values of the ITRS 2003 of 1.2, 1.0, 0.75 and 0.5 nm for the technology node to 65 -, 45 -, 32 - and 22-nm, respectively. In this case, the threshold voltage corresponding fluctuations are better controlled, reaching approximately 35 mV at the channel length of 9-nm, and the reduction of the threshold voltage is negligible.

When the oxide thickness is reduced to a few atomic layers of silicon, the interface roughness on the atomic scale requires significant changes in the oxide thickness (OTV) in the gate region of a single MOSFET. As done for the simulations that take into account the effects of random fluctuations of dopants [61], the effects of local fluctuations in the oxide thickness, require three-dimensional statistical simulations of samples of MOSFETs with identical design parameters but with different profiles of the oxide thickness at the microscopic level.

The results were obtained using a three-dimensional hierarchical drift-diffusion simulator, originally designed to study the effects of random dopant fluctuations.

In most simulations they used a constant doping profile, rather than a discrete random doping, in order to isolate the effects of fluctuations of the threshold voltage associated with the OTV. The requirements for statistical analysis and interpretation of results transforms a 1D problem in a 4D problem, where the fourth dimension is the measure of the statistical sample of devices

with identical design parameters but different profiles of the interface of the thickness oxide at the microscopic level. Statistical samples of MOSFETs have been simulated and analyzed, merely considering low drain voltages.

It is possible to plot the standard deviation of threshold voltage and the shift due to the OTV for transistors with different gate lengths (Fig. 2.16).

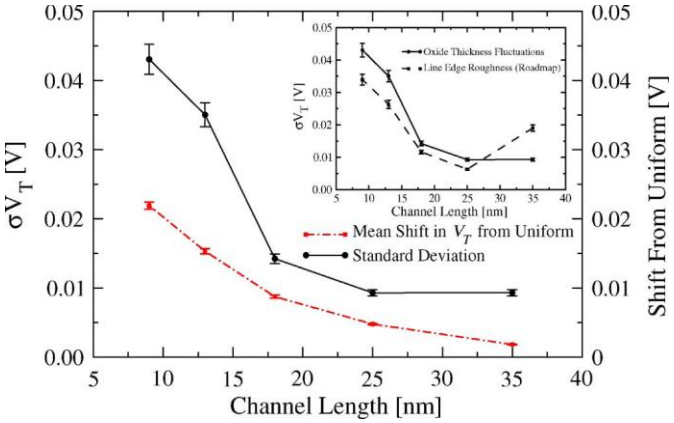


Fig. 2.16 Threshold-voltage standard deviation and threshold-voltage shift for 35-, 25-, 18-, 13-, and 9-nm gate-length transistors due to LER, assuming the values of LER prescribed by the ITRS and using Λ = 30 nm. [Fig 7 of [23]].

To test the statistical interactions of the various sources of intrinsic parameter fluctuations in MOSFETs for the 35 nm reference, the combinations LER, OTV and LER with OTV have been simulated (Tab. 2.3).

In this case the different sources of intrinsic parameter fluctuations are virtually statistically independent, and standard deviation of the threshold voltage resulting from the combined action of different sources of intrinsic parameter fluctuations can be approximated as:

$$\sigma V_T = \sqrt{\sigma V_{th1}{}^2 + \sigma V_{th2}{}^2 + \cdots + \sigma V_{thn}{}^2} \tag{2.4}$$

where $\sigma V_{th1}, \sigma V_{th2}, ..., \sigma V_{thn}$, are the standard deviations of the threshold voltage obtained from independent simulations of each source of intrinsic parameter fluctuations. The result of the combined effects of RDD, LER and OTV in the 35 nm MOSFET is $\sigma_{V_{th}} \approx 40$ mV.

Fluctuation	$\overline{V}_T$	σV_T	**Calc.** σV_T
RDD	133 mV	33.2 mV	-
LER	126 mV	19.0 mV	-
OTV	122 mV	1.8 mV	-
RDD & LER	126 mV	38.7 mV	38.2 mV
RDD & OTV	123 mV	33.9 mV	33.3 mV
LER & OTV	113 mV	22.8 mV	19.1 mV

Tab. 2.3 Standard deviation of the threshold voltage for the 35 × 35 nm MOSFET caused by single and combined sources of intrinsic parameter fluctuations, compared with the theoretical values [Table II of [23]].

It is difficult to compare the simulation results presented in [23] with experimental data for a variety of reasons. The first reliable data on the fluctuations of the parameters can be obtained only for mature technologies that have been fine-tuned in terms of mass production. Even in this case, it is difficult to separate the fluctuations of the intrinsic parameters from the

fluctuations of the extrinsic parameters associated with the process variations and the corresponding size, thickness and variations of the dopant.

When the special test structures are used to access only to the fluctuations of the intrinsic parameters, it is impossible to separate the individual contributions of different sources of intrinsic parameter fluctuations.

In Tab. 2.4 are compared the resulting $\sigma_{V_{th}}$ with the combined action of RDD, LER and OTV with published data. In any case, in order to scale the data for transistors square 35 × 35 nm, it was assumed that $\sigma_{V_{th}} \propto 1/\sqrt{LW}$.

Source	σV_T mV	Comment
Combining RDD, LER, and OTV	40	Simulation
H. Fukutome, *et al.* [39]	49	40 nm MOSFETs
F. Arnaud, *et al.* [40]	63	65 nm technology node

Tab. 2.4 Comparison of the standard deviation of the simulated threshold voltage of 35 × 35 nm MOSFET induced by the combined presence of RDD, LER and OTV [Table III of [23]].

In both cases the experimental results are very close but slightly higher than the results of the simulations. The reasons for this may be related to differences in technology, fluctuations in the contribution of extrinsic parameters of the measurements, as well as the omission of the intrinsic fluctuations of the parameters associated with grain boundaries of polysilicon in the simulations.

Recently, Reid *et al,* using full-scale 3-D simulations of 100000 devices, have studied in detail statistical threshold voltage variability in a state-of-the-art n-channel MOSFET introduced by the combined effect of random-discrete

dopants (RDD) and line edge roughness (LER) and they have demonstrated that the resulting distribution is non-normal [86].

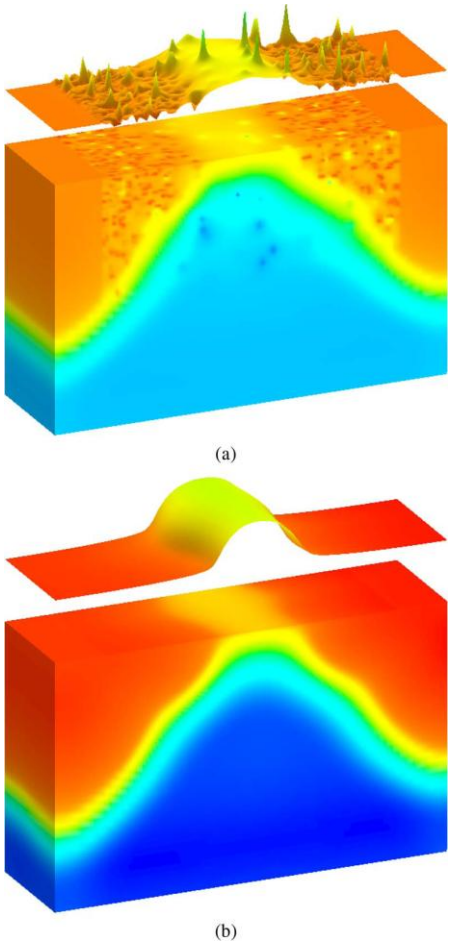


Fig. 2.17 Electron concentration in two example devices showing the impact of (a) random dopants and (b) LER [Fig. 1 of [86]].

The simulated 35-nm gate length n-channel MOSFET (based on a device originally published by Toshiba) has a complex doping profile. The simulated

MOSFET is identical to the one used in their previous individual studies of random dopant and LER effects to allow a fair comparison between the souces of variability and with our previously published results.

To examine the combined impact of RDD and LER on transistor variability, they have performed simulations of 100000 microscopically different devices, in which both RDD and LER are present, which allows the distribution of V_{th} due to their combined effects to be accurately characterized up to the higher moments required to determine the degree of normality of the statistical distribution.

Numerical values for the moments of the V_{th} variations introduced by RDD and LER both in isolation and in combination are given in Tab. 2.5. Errors are obtained by bootstrapping, which is a technique for estimating the variation in a statistic by resampling with replacement [88].

Statistic	RDD	LER	RDD+LER
# of Simulations	100,000	25,000	100,000
Minimum (mV)	112.7±1.5	159.4±5.2	75.35±6.6
Maximum (mV)	370.5±2.3	271.9±2.0	384.0±5.7
Mean (mV)	225.9±0.1	231.1±0.1	225.6±0.1
St. Dev. (mV)	30.28±0.07	12.75±0.06	33.08±0.07
Skewness	0.159±0.008	-0.407±0.02	0.0623±0.008
Kurtosis	0.0486±0.02	0.255±0.06	0.0527±0.02

Tab. 2.5 Summary of the statistical moments and standard errors of the data for the combined RDD and LER simulations at V_{DS} = 100 mV. The standard errors are obtained by bootstrapping the simulation data [Table I of [86]].

They have demonstrated that the distribution due to combined RDD and LER can accurately be reproduced using the statistically enhanced approaches, it is possible to characterize the combined effects of random dopants and LER on the threshold voltage and benefit from the reduced computational time needed to individually characterize the two sources by applying the developed computationally efficient statistical enhancement strategies.

The effects of random dopants can accurately be characterized from the simulation of the order of 1000 devices. Similarly, LER can be characterized by simulating a small number of devices, e.g., 20, with uniform gate edges and different channel lengths. This way, a complete characterization of the effects of RDD and LER, both in isolation and in combination, can be achieved at the expense of ≈ 5000 CPU hours of simulation, compared to approximately 300000 CPU hours to perform the complete low-drain brute-force characterization of the 35-nm transistor.

3 Analysis of the threshold voltage dispersion in different MOSFET structures

We propose an approach to evaluate the effect on the threshold voltage dispersion of nanoscale MOSFETs of line edge roughness, surface roughness, and random dopant distribution [89]. The methodology is fully based on parameter sensitivity analysis, performed by means of a limited number of TCAD simulations or analytical modeling. We apply it to different nanoscale transistor structures: a 32 nm ultra-thin body SOI MOSFET and a 22 nm double-gate MOSFET adopted within the EC PULLNANO project as template devices, and one bulk 45 nm NMOSFET within the project *MODERN* reported [83]. The 32 nm and 22 nm templates are shown in

Fig. 3.1(details can be found in [90]), whereas the 45 nm template is illustrated in [83]. The choice of the template devices is due to the availability of data from statistical atomistic simulations on the very same templates [4], [83], which enables us to compare results obtained with our proposed approach.

Moreover we have used this method to analyze the main causes of intrinsic threshold voltage dispersion in the 32/28 nm CMOS process developed by STM for MODERN WP2 project reported in [91]]. The investigation has been

focused on minimum size RVT transistors (NMOS and PMOS) i.e. a 30 nm n-channel RVT MOSFET and a 30 nm p-channel RVT MOSFET, and in particular on the effect of random dopants and line edge roughness.

In all cases our approach is capable to reproduce with very good accuracy the results obtained through 3D atomistic statistical simulations at a small computational cost. We believe the proposed approach can be a powerful tool to understand the role of the main variability sources and to explore the device design parameter space**.**

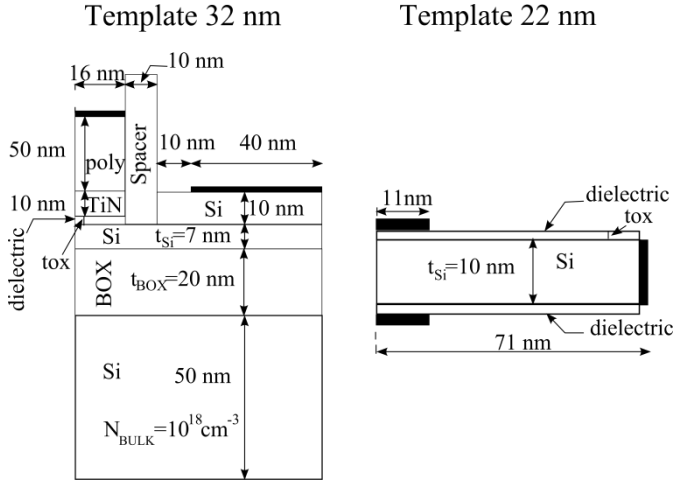


Fig. 3.1 Template structures for the 32 nm UTB SOI MOSFET (left) and the 22 nm double-gate MOSFET (right). The device is symmetrical. Doping profiles for source and drain are described in [90]. The effective oxide thickness t_{ox} is 1.2 nm for the 32 nm template and 1.1 nm for the 22 nm template.

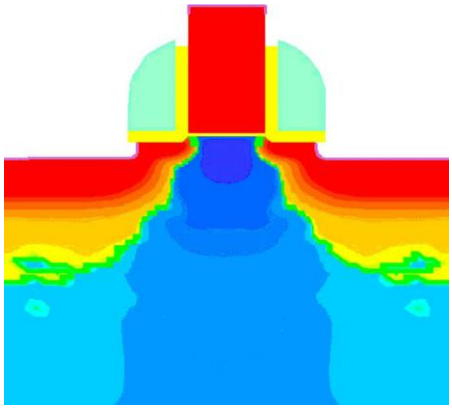


Fig. 3.2 Template 45 nm from [83].

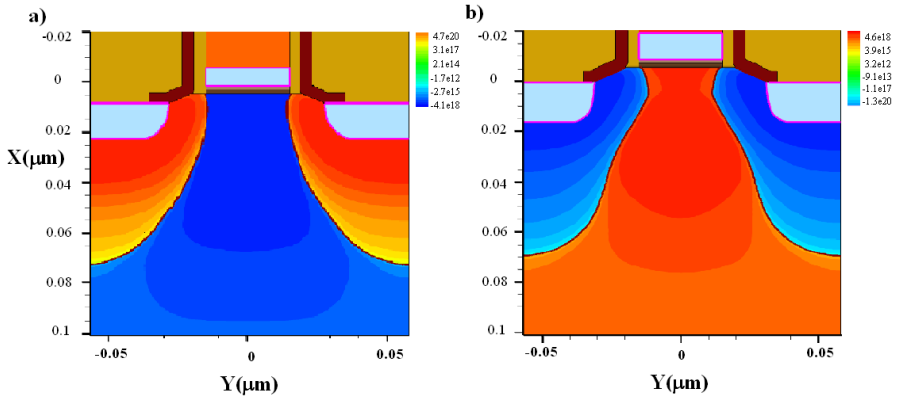


Fig. 3.3 Doping profiles of 30 nm RVT NMOS (a) and PMOS (b) as obtained by STM after calibration with electrical characteristics.

3.1 Methodology

The approach we propose requires the identification of the relevant quantities that translate process variability into the dispersion of electrical parameters. It involves the following three steps:

First, we need to express all process and geometry variability sources in terms of a set of synthetic parameters.

Then, we need to identify the *independent* parameters.

Finally, we use sensitivity analysis to evaluate the contribution to the dispersion of electrical parameters (e.g. the threshold voltage V_{th}) of each independent source. This step is based on the assumption that the effect of each source is sufficiently small that first-order linearization is applicable. In the literature non linear and cross terms have been explicitly evaluated, for example through statistical simulations on 35 nm MOSFETs [23]: the variance of the threshold voltage due to combined effect has been shown to be equal to the sum of the variances due to individual effects, giving us confidence in the linear approximation.

3.2 Variability due to line edge roughness and surface roughness

As an example, let us consider the 32 nm device shown in Fig. 3.4, where the *y* axis runs along the channel length direction, the *x* axis is is perpendicular to the device plane and the *z* axis runs along the channel width.

We can translate line edge roughness in terms of the dispersion of the average position of both gate edges along the y axis ($y = 0 + y_1$ and $y = L + y_2$), as illustrated in Fig. 3.4a. This in turn translates into gate length dispersion. In practice the two rough edges are not completely independent, but here for simplicity we can assume they are. Surface roughness is translated into the dispersion of the average position of the interface between adjacent layers: the offsets are x_1, x_2, x_3 in Fig. 3.4b.

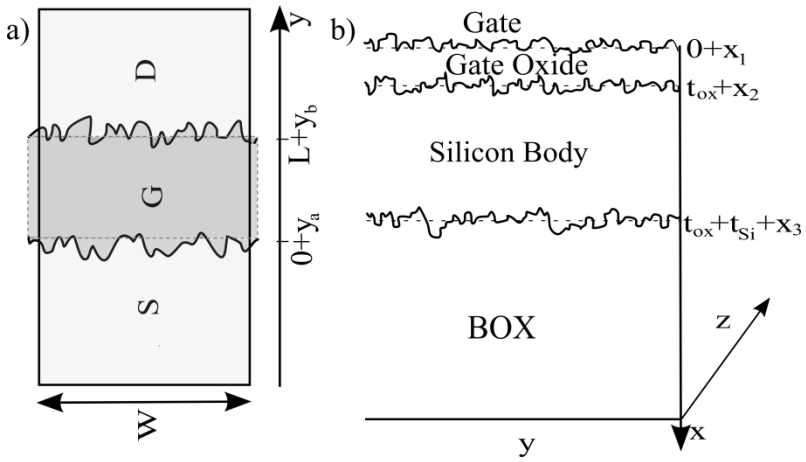


Fig. 3.4 a) top view of the active area highlighting the gate LER; b) layered structure highlighting the interface roughness between adjacent layers in the 32 nm template. *y*-axis runs along the channel length direction, *x*-axis is perpendicular to the device plane and the *z* axis runs along the channel width.

We assume that parameters y_1, y_2, x_1, x_2, x_3 are only affected by LER and SR and are physically independent. We start by considering the effect of the

offset of the position between two adjacent Si-SiO_2 layers $(x_i, i = 1,2,3)$. The first step is to evaluate the variance $\sigma_{x_i}^2$ of x_i. In the case of the ultra-thin body SOI MOSFET we should consider the fluctuations present in the bottom interface of the buried oxide, but these are practically irrelevant for our calculation.

Interface roughness leads to a deviation from the nominal position of the interface between two layers that we can describe statistically as a random function $f(y,z)$ with zero mean value and exponential autocorrelation $r(y_1,z_1,y_2,z_2) \equiv < f(y_1,z_1) f(y_2,z_2) >$, characterized by mean square amplitude Δ_S and correlation length Λ_S.

$$r(y_1,z_1,y_2,z_2) = \Delta_S^2 \exp\left(- \frac{\sqrt{(y_1 - y_2)^2 + (z_1 - z_2)^2}}{\Lambda_S} \right). \qquad (3.1)$$

The average position of the interface $\overline{f}$ for a given occurrence of a rough interface is:

$$\overline{f} = \frac{1}{LW} \int_0^L dy \int_0^W dz\, f(y,z); \qquad (3.2)$$

$\overline{f}$ has zero mean value and variance given by:

$$\sigma_f{}^2 = < \overline{f}^2 > = \frac{1}{L^2 W^2} \int_0^L dy_1 \int_0^W dz_1 \int_0^L dy_2 \int_0^W dz_2 < f(y_1,z_1) f(y_2,z_2) >, \qquad (3.3)$$

which, using (3.1) can be written as:

$$\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_{x_3}^2 = \sigma_{SR}^2 = \frac{2\pi\Delta_S^2}{LW}\left[\Lambda_S^2 - e^{-\frac{\sqrt{L^2+W^2}}{\Lambda_S}}\Lambda_S\left(\Lambda_S + \sqrt{L^2+W^2}\right)\right]. \quad (3.4)$$

where x_1, x_2, x_3 in (3.4) are the average position of interfaces between adjacent layers, as indicated in Fig. 3.4.

In the common case $L, W >> \Lambda_S$, Eq. (3.4) reduces to

$$\sigma_{SR}^2 = \frac{2\pi\Lambda_S^2\Delta_S^2}{LW}. \quad (3.5)$$

If instead we consider a Gaussian autocorrelation, as in [4], [83], expressed as

$$r(x_1, y_1, x_2, y_2) = \Delta_S^2 \exp\left(-\frac{(x_2 - x_1)^2 + (y_2 - y_1)^2}{2\Lambda_S{}^2}\right). \quad (3.6)$$

Replacing (3.6) in (3.3), we find

$$\sigma_{SR}^2 = \frac{2\pi\Lambda_S^2\Delta_S^2}{L^2W^2}\left[L\cdot erf\left(\frac{L}{\sqrt{2}\Lambda_S}\right) + \sqrt{\frac{2}{\pi}}\Lambda_S\left(e^{-\frac{L^2}{2\Lambda_S^2}} - 1\right)\right]\cdot\left[Werf\left(\frac{W}{\sqrt{2}\Lambda_S}\right) + \sqrt{\frac{2}{\pi}}\Lambda_S\left(e^{-\frac{W^2}{2\Lambda_S^2}} - 1\right)\right], \quad (3.7)$$

which again reduces to (3.5) if $L,W >> \Lambda_S$.

We can express the variation of the threshold voltage in terms of thickness variations of different layers, and then consider as independent physical quantities only x_1, x_2, x_3:

$$\begin{aligned} dV_{th} &= \frac{\partial V_{th}}{\partial t_{ox}} dt_{ox} + \frac{\partial V_{th}}{\partial t_{Si}} dt_{Si} + \frac{\partial V_{th}}{\partial t_{BOX}} dt_{BOX} \\ dV_{th} &= \frac{\partial V_{th}}{\partial t_{ox}} (dx_2 - dx_1) + \frac{\partial V_{th}}{\partial t_{Si}} (dx_3 - dx_2) - \frac{\partial V_{th}}{\partial t_{BOX}} (-dx_3). \end{aligned} \tag{3.8}$$

Then, using linearization and the hypothesis of independence of the different parameters, we can write:

$$\sigma^2_{V_{th}SR} = \left(\frac{\partial V_{th}}{\partial x_1}\right)^2 \sigma^2_{x1} + \left(\frac{\partial V_{th}}{\partial x_2}\right)^2 \sigma^2_{x2} + \left(\frac{\partial V_{th}}{\partial x_3}\right)^2 \sigma^2_{x3}. \tag{3.9}$$

The partial derivatives can be expressed as:

$$\begin{aligned} \frac{\partial V_{th}}{\partial x_1} &= -\frac{\partial V_{th}}{\partial t_{ox}}, \\ \frac{\partial V_{th}}{\partial x_2} &= \frac{\partial V_{th}}{\partial t_{ox}} - \frac{\partial V_{th}}{\partial t_{Si}}, \\ \frac{\partial V_{th}}{\partial x_3} &= \frac{\partial V_{th}}{\partial t_{Si}} - \frac{\partial V_{th}}{\partial t_{BOX}}. \end{aligned} \tag{3.10}$$

As far as LER is concerned, we assume that the average edge position is a random function g(z) with zero mean value and an exponential autocorrelation function $r(d) \equiv < g(z)g(z+d) >$ characterized by correlation length Λ_L and mean square amplitude Δ_L.

$$r(d) = \Delta_L^2 e^{-|d|/\Lambda_L}, \tag{3.11}$$

from which we can write

$$\sigma_f^2 = < \bar{g}^2 > = < \frac{1}{W^2} \int_0^W g(z_1)dz_1 \cdot \int_0^W g(z_2)dz_2 > . \tag{3.12}$$

Therefore we find:

$$\sigma_{ya}^2 = \sigma_{yb}^2 = \sigma_{LER}^2 = \frac{2\Lambda_L \Delta_L^2}{W} \left\{ 1 - \frac{\Lambda_L}{W} \left[1 - \exp\left(-W/\Lambda_L\right) \right] \right\}. \tag{3.13}$$

where y_a, y_b in Eq. 13 are the average gate edges indicated in Fig. 3.4.

If instead we consider function Gaussian autocorrelation,

$$r(d) = \Delta_L^2 e^{-\frac{d^2}{2\Lambda_L^2}} \tag{3.14}$$

Solving the integral (3.12), we find:

$$\sigma_{LER}^2 = \frac{2\Delta_L^2 \Lambda_L}{W^2}\left[\Lambda_L\left(e^{-\frac{W^2}{2\Lambda_L^2}} - 1\right) + \sqrt{\frac{\pi}{2}} W erf\left(\frac{W}{\sqrt{2}\Lambda_L}\right)\right]. \qquad (3.15)$$

The variance of V_{th} due to line edge roughness is:

$$\sigma_{V_{th}LER}^2 = \left(\frac{\partial V_{th}}{\partial y_1}\right)^2 \sigma_{y_1}^2 + \left(\frac{\partial V_{th}}{\partial y_2}\right)^2 \sigma_{y_2}^2 = 2\left(\frac{\partial V_{th}}{\partial L}\right)^2 \sigma_{LER}^2. \qquad (3.16)$$

All required derivatives can be computed with TCAD simulations or – if the device structure is simple enough - with an appropriate analytical model.

As can be seen, if proper independent parameters are identified, the evaluation of the dispersion of the threshold voltage only requires the computation of a limited number of derivatives, each obtainable from a single device simulation. Even using derivatives obtained from TCAD, the computational cost of the procedure is extremely reduced with respect to a statistical simulation. The price to pay is the initial analysis of variability sources and the consequent assumptions.

The threshold voltage of the 32 nm template MOSFET as a function of gate length is plotted in Fig. 3.5, where it is compared with results from the analytical model presented in the Appendix. V_{th} is defined as the V_{GS} corresponding the current of 10^{-5} A/μm as in [4]. The agreement is very good.

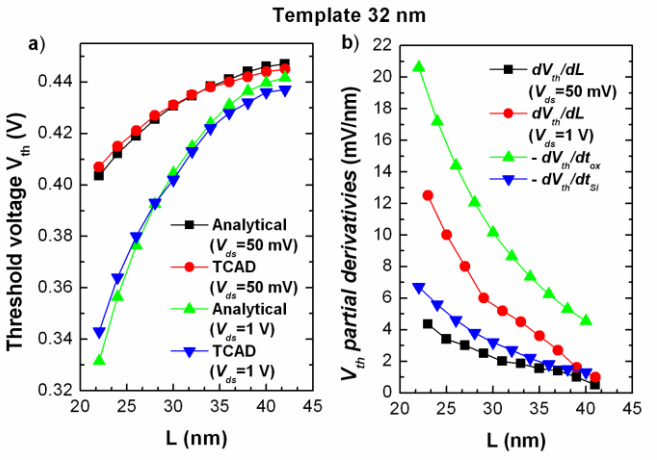


Fig. 3.5 Threshold voltage as a function of L from analytical model and TCAD (a) and analytical partial derivatives of V_{th} (b) of the 32 nm template MOSFET.

In Fig. 3.6 we show the same comparison for the 22 nm template device.

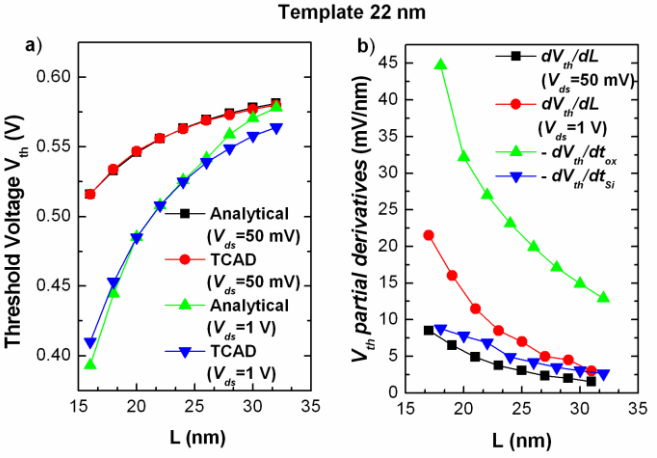


Fig. 3.6 Threshold voltage as a function of L from analytical model and TCAD (a) and analytical partial derivatives of V_{th} (b) for the 22 nm template MOSFET.

The dependence on gate length of the threshold voltage of the 45 nm MOSFET is shown in Fig. 3.7 for V_{DS} = 50 mV and V_{DS} = 1.1 V. In the latter case there is no analytical model, because the doping profiles for different lengths are directly obtained from process simulations and cannot be described by a simple expression.

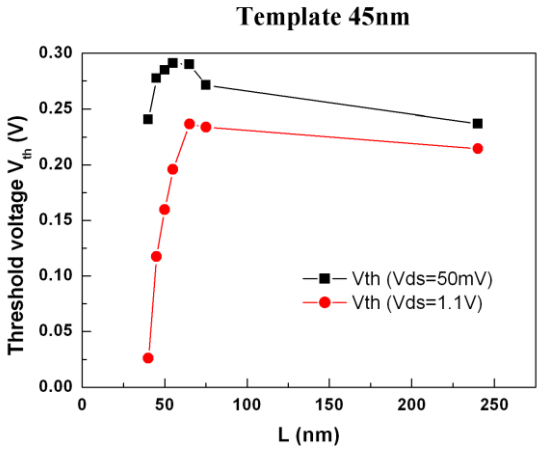


Fig. 3.7 Threshold voltage as a function of L for the 45 nm template MOSFET, for two different values of V_{DS}.

Now we can use Eq. (3.9) and (3.16) to compute the variance of V_{th} due to LER and to SR, and to compare our results with those obtained with atomistic statistical simulations in [4], [83].

For the sake of comparison, we assume as in [4] for all rough interfaces a Gaussian autocorrelation function with mean square amplitude $\Delta_S = 0.15$ nm and correlation length $\Lambda_S = 1.8$ nm, which are close to values observed from TEM measurements [92]. For LER, we assume a Gaussian autocorrelation

function with $\Delta_L = 1.3$ nm, $\Lambda_L = 25$ nm for 32 nm and 22 nm templates and, as in [83]; we assume a Gaussian autocorrelation function with $\Delta_L = 1.3$ nm and $\Lambda_L = 30$ nm for 45 nm template.

Results are shown in Tab. 3.1. The columns *An* and *TCAD* indicate results from our approach where the partial derivatives of V_{th} are computed analytically (as described in the Appendix) or with TCAD [93], respectively. The column *Stat. Sim.* indicates results reported in [4], [83]. As can be seen the agreement, for the LER data, is always extremely good. Very good agreement is obtained between columns *An* and *TCAD*, for the effect of SR, for which data from [4] are not available.

	V_{DS} = 50 mV	Approach		
		An.	***TCAD***	***Stat.Sim [4], [83]***
32nm	σ_{Vth} LER (mV)	*3.36*	*3.45*	*3.3*
	σ_{Vth} SR (mV)	*0.32*	*0.38*	*N/A*
22 nm	σ_{Vth} LER (mV)	*6.7*	*6.2*	*5.8*
	σ_{Vth} SR (mV)	*2.07*	*2.03*	*N/A*
45 nm	σ_{Vth} LER (mV)	-	*7.4*	*7*
	V_{DS} = 1 V (32nm; 22nm), 1.1 V (45nm)	*Approach*		
		An.	***TCAD***	***Stat.Sim. [4], [83]***
32 nm	σ_{Vth}LER (mV)	*9.25*	*9.47*	*8.6*
	σ_{Vth}SR (mV)	*0.94*	*0.85*	*N/A*
22 nm	σ_{Vth}LER (mV)	*15.8*	*15*	*13*
	σ_{Vth} SR (mV)	*6.1*	*6.26*	*N/A*
45 nm	σ_{Vth} LER (mV)	-	*22*	*25*

Tab. 3.1 Standard deviation of the threshold voltage due to LER and SR for the 32 nm and 22 nm template MOSFETs, and to LER for the 45 nm MOSFET, obtained with different methods.

For the other 30 nm template, assuming reasonable values for the two parameters, such as Λ_L = 30 nm and $3\Delta_L$ = 4 nm, results for W = L = 30 nm RVT MOSFETs are shown in. Results have been obtained using three device structures with slightly different gate lengths (28, 30, and 32 nm), very useful for computing the derivatives of threshold voltage with respect to gate length.

With the parameters considered, also for the shortest gate length considered, the contribution of LER to threshold voltage variability is much smaller than that due to random dopants. For example for $V_{DS} = 1$ V in the case of the NMOS, the presence of LER increases the standard deviation of V_{th} by 13% with respect to the value due to only random dopants.

Given that in planar CMOS process gate length scaling is always accompanied by an increase of channel doping, it is very likely that LER will always have a minor effect on threshold voltage variability with respect to random dopants. It can become dominant for multiple gate or ultra thin body solutions, where channel doping might be significantly reduced.

	RVT NMOS 30 nm		
		Our approach [89]	***Stat. Sim. [91]***
50 mV	σ_{Vth} LER (mV)	*13*	*12.7*
1 V	σ_{Vth} LER (mV)	*25*	*24.9*
	RVT PMOS 30 nm		
		Our approach	***Stat. Sim. [91]***
-50 mV	σ_{Vth}LER (mV)	*10*	*12.8*
-1 V	σ_{Vth} LER (mV)	*29*	*33.3*

Tab. 3.2 Standard deviation of the threshold voltage due to LER for the RVT 30 nm obtained with different methods.

3.3 Effect of random dopant distribution

In the case of random dopant distribution, the source of threshold voltage dispersion is the fluctuation of the dopant distribution in the active area. What matters is not only the total number of dopants in the active area, but also their position. In any case, it is pretty intuitive that we do not need to know with atomistic precision the effect of dopant distribution on the threshold voltage.

First, we can acknowledge that the mechanism is mainly governed by electrostatics, therefore impurity position along the width direction is of minor relevance. This allows us to simplify our analysis considering only 2D device structures. Indeed, statistical simulations with random dopants typically yield a family of parallel transfer characteristics, corresponding to threshold voltage dispersion independent of the inversion level in the channel. This means that percolation is hardly effective, since it should strongly depend on the Debye length and therefore on the mobile charge density in the channel. An ex-post verification of this assumption will be provided by comparing our results with 3D statistical simulations

Moreover, we can assume that the effects of fluctuations of the number of dopants in different regions are small enough to add up linearly. For a given variation of dopant distribution $\Delta N_A(x,y,z)$ with respect to the nominal value we can write the following expression for the variation of V_{th}:

$$\Delta V_{th} = \int K(x,y)\Delta N_A(x,y,z)dxdydz, \tag{3.17}$$

where $K(x,y)$ has the role of a propagator, or Green's function [94]. The expression requires the linearity assumption to hold. Let us notice that we are neglecting the dependence of K on *z* according to the hypothesis above.

To conveniently compute the propagator K, we can assume that K is a smooth function of x and y, and move from the continuum to a discrete space, partitioning the active area in small rectangular boxes, as shown in Fig. 3.8a. Now we can write:

$$\Delta V_{th} = \sum_i \Delta V_{th_i} = \sum_i K_i \Delta N_i \tag{3.18}$$

The sum runs over all boxes, ΔN_i is the variation of the number of dopants in box i, and ΔV_{th_i} is the threshold voltage variation if only dopants in box i are varied.

In practice, we multiply doping in box i by a factor $(1+\alpha)$ and compute ΔV_{thi} with TCAD simulations. Therefore we have

$$\begin{aligned} \Delta N_i &= \alpha N_i \\ \Delta V_{th_i} &= \alpha K_i N_i \end{aligned} \tag{3.19}$$

so that (3.18) becomes,

$$\Delta V_{th} = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha} \right) \alpha = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha} \right) \frac{\Delta N_i}{N_i} \tag{3.20}$$

We know need another reasonable assumption: doping variations in different boxes are independent Poissonian processes. Therefore from (3.20) we can write

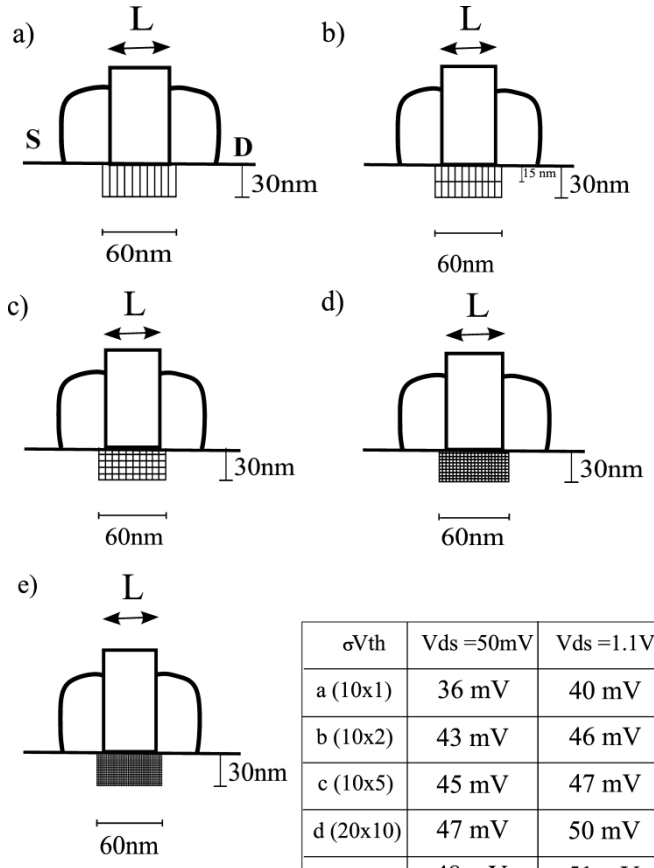
$$\sigma^2_{V_{th}RDD} = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha} \right)^2 \frac{\sigma^2_{N_i}}{N_i^2} = \sum_i \sigma^{2\,[i]}_{V_{th}RDD} \tag{3.21}$$

Since N_i is a Poisson process is $N_i = \sigma_{N_i}$ we finally have

$$\sigma^2_{V_{th}RDD} = \sum_i \left(\frac{\Delta V_{th_i}}{\alpha} \right)^2 \frac{1}{N_i} = \sum_i \sigma^{2\,[i]}_{V_{th}RDD} \tag{3.22}$$

The threshold voltage dispersion due to RDD only requires a single TCAD simulation for each box, and an integral of the doping profile in each box. Box partitioning is shown in Fig. 3.8a and covers a region smaller than the whole active area, because one can easily check that far from the channel the impact of doping fluctuations on V_{th} rapidly goes to zero.

To evaluate the granularity of partition required to obtain reasonably accurate results we have used different partitions for the 45 nm MOSFET, shown in Fig. 3.8 : 10x1 (a), 10x2 (b), 10x5 (c), 20x10 (d), 40x20 (e). The table in the inset of Fig. 3.8 shows the standard deviation of the threshold voltage obtained for V_{DS} = 50 mV and V_{DS} = 1.1 V. If the device is symmetric with respect to a source-drain swap, for low V_{DS} we can reduce to half the number of simulations required, since also the propagator is symmetric.



σVth	Vds =50mV	Vds =1.1V
a (10x1)	36 mV	40 mV
b (10x2)	43 mV	46 mV
c (10x5)	45 mV	47 mV
d (20x10)	47 mV	50 mV
e (40x20)	48 mV	51 mV

Fig. 3.8 Different partitions used to evaluate the variance of the threshold voltage for the 45 nm MOSFET: 10x1 boxes (a), 10x2 (b), 10x5 (c), 20x10 (d), 40x20 (e). In the inset: table with results of σ_{Vth} due to RDD for the 45 nm template and comparison with atomistic simulations.

Results show that only few simulations (in case Fig. 3.8b 20 for high V_{DS}, or 10 for low V_{DS}) are sufficient to obtain reasonably accurate results. Very accurate results can be obtained in case Fig. 3.8d with a factor 10 more simulations. In Tab. 3.3 results for the three template devices are compared with

results from statistical simulations [4], [83] for two different values of V_{DS}: 50 mV, which corresponds to quasi equilibrium, and 1.1 V, which corresponds to far from equilibrium transport. In all cases - except the 32 nm device which posed convergence problems upon the application of the method - the agreement is rather good.

	45 nm		
		Our approach [89]	***Stat. Sim. [83]***
50 mV	σ_{Vth} RDD (mV)	*47*	*44*
1.1 V	σ_{Vth} RDD (mV)	*50*	*50*
	22 nm		
		Our approach [89]	***Stat. Sim. [4]***
50 mV	σ_{Vth}RDD (mV)	*6.6*	*6.4*
1 V	σ_{Vth} RDD (mV)	*7.9*	*8.1*
	32 nm		
		Our approach [89]	***Stat. Sim. [4]***
50 mV	σ_{Vth} RDD (mV)	*7.4*	*5.3*
1 V	σ_{Vth} RDD (mV)	*11*	*6.1*

Tab. 3.3 Standard deviation of the threshold voltage due to RDD for the 45 nm, 32 nm and 22 nm template MOSFETs.

Such investigation is particularly interesting because if we plot the individual region contributions $\sigma^{2\,[i]}_{V_{th}RDD}$ on a 2D plot, we can understand which part of the device region contributes the most to the threshold voltage dispersion.

In Fig. 3.9, we show the doping profiles of the 30 nm RVT NMOS and PMOS. Doping profiles have been calibrated by partner STMicroelectronics on the basis of device characteristics.

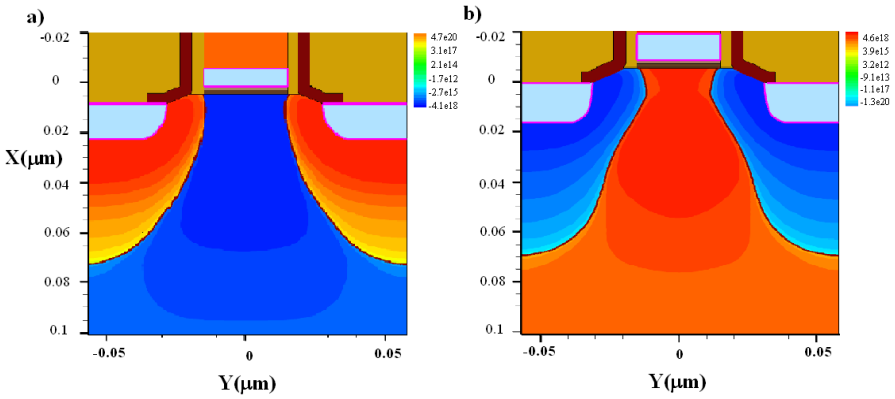


Fig. 3.9 Doping profiles of 32 nm RVT NMOS (a) and PMOS (b) as obtained by STM after calibration with electrical characteristics.

In Fig. 3.10, on the same coordinates, we show a color map of $\sigma^2_{V_{th}RDD}{}^{[i]}$ for the two devices in the linear region ($|V_{DS}|$ = 50 mV). We limit our consideration to boron doping in the NMOS and phosphorus doping for the PMOS. The effect of other impurity profiles on the threshold voltage variation is negligible.

As can be seen in Fig. 3.10, basically all contribution comes from a region smaller than 10 nanometers in the central part of the channel and within few nanometers from the silicon-dielectric interface. This is a significant observation, not known a priori, since doping is pretty high in the whole device region.

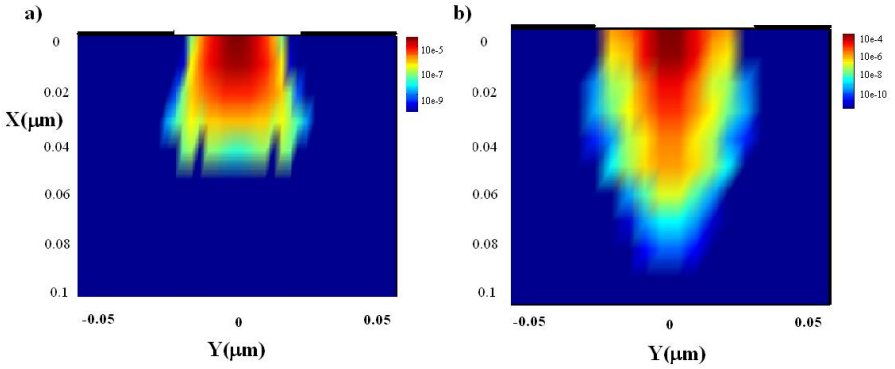


Fig. 3.10 Colour map of the local contribution to the variance of the threshold voltage indicated with $\sigma^2_{VthRDD}{}^{[i]}$ as a function of position: a) effect of boron doping for the 30 nm RVT NMOS for V_{DS} = 50 mV, b) effect of phosphorus doping for the 30 nm RVT PMOS for V_{DS} = - 50 mv. Units are V^2.

Similar observations can be drawn fron Fig. 3.11, which shows the colorplot of the same partial contributions to the variance of the threshold voltage $\sigma^2_{V_{th}RDD}{}^{[i]}$ for the same two devices but in saturation ($|V_{DS}|$ = 1 V). As can be seen also in Tab. 3.4, the standard deviation of the threshold voltage slightly increases in saturation, but contribution comes from a smaller region, closer to the interface and to the center of the channel. Indeed, the curvature of the potential is larger for larger drain-to-source voltage; therefore the region close to the conduction barrier edge peak in the channel has a higher relative importance. This observation explains both the increase of the threshold voltage variability and the relative larger weight of the central region. In [91], results for the standard deviation of the threshold voltage obtained from sensitivity analysis have been demonstrated to be consistent with those obtained from statistical atomistic simulations.

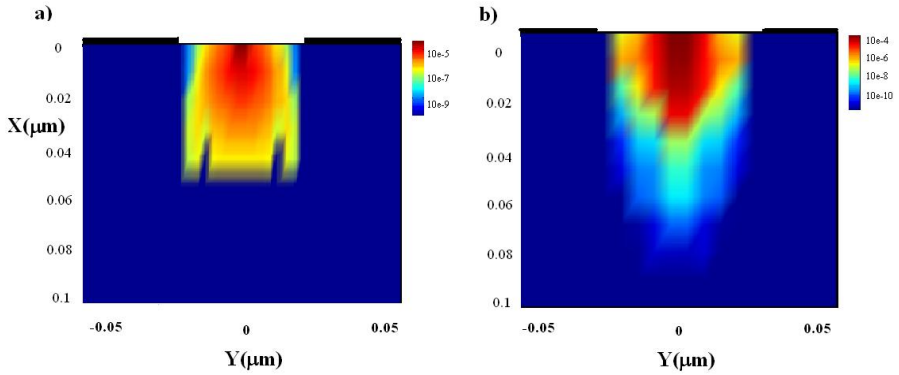


Fig. 3.11 Colour map of the local contributions to the variance of the threshold voltage indicated with $\sigma^2_{VthRDD}{}^{[i]}$ as a function of position: a) effect of boron doping for the 30 nm RVT NMOS for V_{DS} = 1 V, b) effect of phosphorus doping for the 30 nm RVT PMOS for V_{DS} = - 1 V. Units are V^2.

	RVT NMOS 30 nm		
		Our approach [89]	***Stat. Sim.[91]***
50 mV	σ_{Vth} RDD (mV)	*43*	*44.4*
1 V	σ_{Vth} RDD (mV)	*48*	*49.8*
	RVT PMOS 30 nm		
		Our approach [89]	***Stat. Sim. [91]***
-50 mV	σ_{Vth}RDD (mV)	*46*	*42.3*
-1 V	σ_{Vth} RDD (mV)	*58*	*54.4*

Tab. 3.4 Standard variation of the threshold voltage due to random dopants for minimum feature size (W=30) 32/28 nm CMOS STM process obtained with different methods.

The partial contributions to the variance of the threshold voltage indicated with $\sigma^{2}_{V_{th}RDD}{}^{[i]}$ are plotted as a function of positions in the color maps in Fig. 3.12, for the three template devices. In Fig. 3.12a the effect of the acceptor doping of the 45 nm MOSFET is shown, whereas in Fig. 3.12b and Fig. 3.12c the effect of the donor doping of the contacts of the 32 nm and 22 nm templates are shown, respectively. In all cases, it is pretty clear that a limited part of the active area has a practical impact on threshold voltage dispersion.

The total variance of the threshold voltage is computed by summing the variances due to all independent physical effects.

$$\sigma^{2}_{V_{th}TOT} = \sigma^{2}_{V_{th}RDD} + \sigma^{2}_{V_{th}LER} + \sigma^{2}_{V_{th}SR} \tag{3.23}$$

As mentioned previously the cross terms are negligible even when they are considered. An ex-post evaluation of results obtained with 3D atomistic statistical simulations of separate and combined variability sources (for example [4]) confirms such assumption.

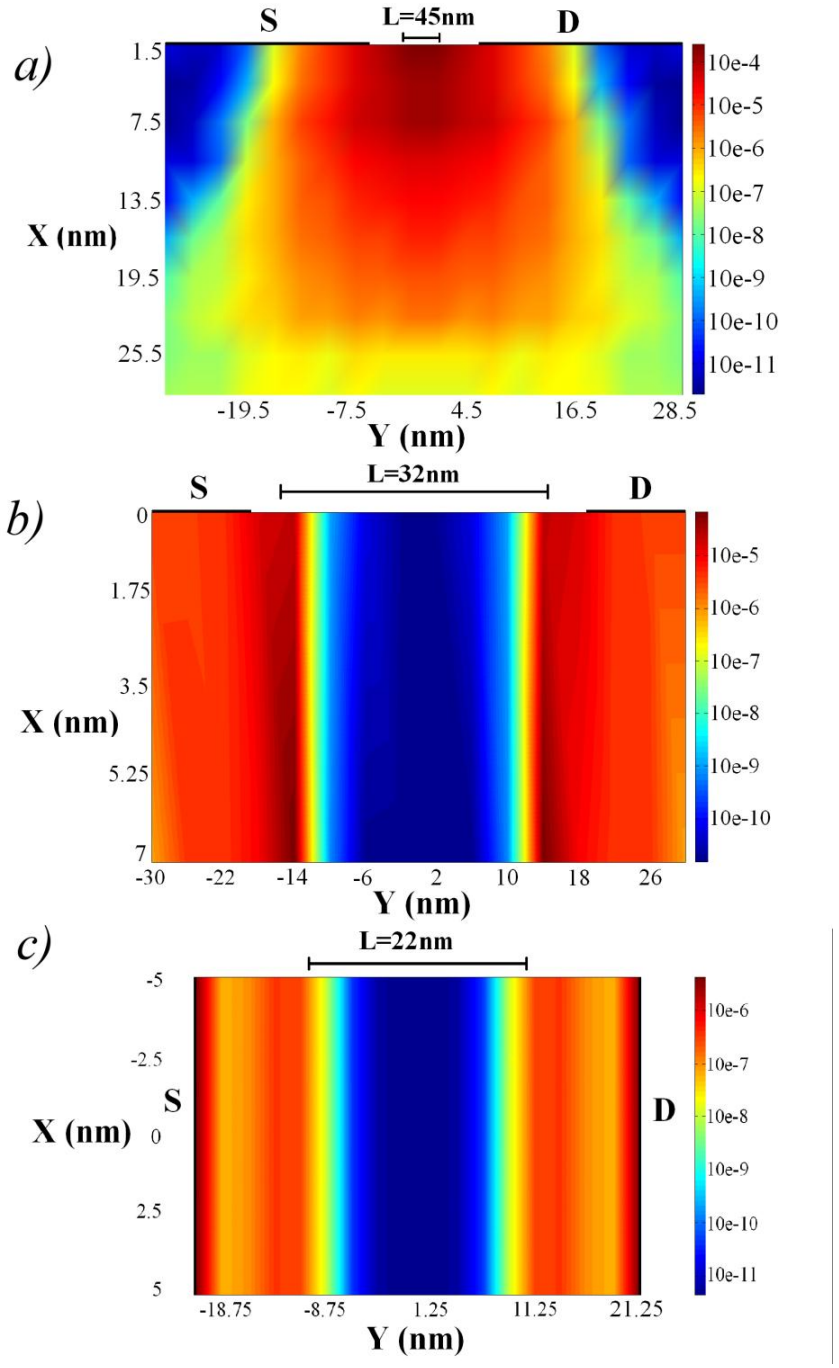

a)
S
L=45nm
D
1.5
7.5
13.5
19.5
25.5
X (nm)
-19.5
-7.5
4.5
16.5
28.5
Y (nm)
10e-4
10e-5
10e-6
10e-7
10e-8
10e-9
10e-10
10e-11
b)
S
L=32nm
D
0
1.75
3.5
5.25
7
X (nm)
-30
-22
-14
-6
2
10
18
26
Y (nm)
10e-5
10e-6
10e-7
10e-8
10e-9
10e-10
c)
L=22nm
-5
-2.5
S
X (nm)
0
2.5
5
D
-18.75
-8.75
1.25
11.25
21.25
Y (nm)
10e-6
10e-7
10e-8
10e-9
10e-10
10e-11

Fig. 3.12 Colour maps of the partial contributions to the variance of the threshold voltage indicated with $\sigma^2_{VthRDD}{}^{[i]}$ as a function of position: (a) effect of acceptor doping of the 45 nm MOSFET; (b) effect of donor doping of the 32 nm MOSFET, (c) effect of the donor doping of the 22 nm MOSFET.

3.4 Conclusions

We have proposed a methodology for the quantitative evaluation of the effect of line edge roughness, surface roughness, and random dopant distribution, that is based on the careful analysis of the main independent physical parameters affecting threshold voltage variability. The approach requires the calculation of partial derivatives of V_{th} with respect to device structure parameters that can be obtained with a limited number of two-dimensional TCAD simulations or – for simple doping profiles – with analytical models. We have shown that in all cases we are able to obtain results in very good agreement with 3D atomistic statistical simulations [4], [83].

Let us stress the fact that one of the main tenets of our approach is that 3D properties have no specific effect on the threshold voltage of MOSFETs for logic. Device width - as we have seen - has only an effect in determining the variance of the average doping, gate edge, or interface position. Such approximation is based on the assumption that MOSFET behavior is governed mainly by electrostatics. The best validation of our approximation is the very good agreement between results from statistical simulations (based on 3D modeling) and our sensitivity approach (based on 2D modeling) for all device structures considered. We qualify this statement to the threshold voltage of MOSFETs for logic, because when quantities are associated to deep subthreshold bias, as for example in the case of the threshold voltage of non-

volatile memories or the off current of MOSFETs, peculiar 3D effects such as percolation might play a significant role. In addition, our approach in the present form cannot provide information on the far tails of the distribution, which might be particularly important for evaluating device/circuit yield, and would require extension to higher order terms.

We believe that our approach has multiple advantages over statistical modeling, obviously in terms of computational requirements (by several orders of magnitude), but also in terms of providing a good framework for understanding the physical relevant effects affecting device variability and in the possibility of providing a quick way to evaluate V_{th} variability of candidate devices.

The main advantages of statistical simulation, on the other hand, are that it does not require preliminary device analysis and assumptions, and that it would work even when the linear approximation does not hold, for example in the presence of very large and critical variability. We firmly believe that the method presented here is a powerful tool to quickly evaluate variability of device parameters in the context of technology developments, using simulations tools already available and routinely used by CMOS technology developers. It also provides a better understanding of the effect of single physical parameters on the overall device behaviour, and can therefore be a useful guide for device design.

3.5 Appendix: Analytical model for the dependence of threshold voltage on gate length.

The analytical model for the threshold voltage of ultrathin body SOI and double gate MOSFETs is obtained from a simple derivation of surface potential profile $\phi_S(y)$ at the interface between the silicon body and the gate dielectric. We devise a simple extension of Liu's approach [95] along the lines proposed in [96]. Let us consider, for example, Fig. 3.13, in which an ultrathin body SOI MOSFET is considered. We assume that the channel can be divided in three regions: the central undoped region under the gate, and two external highly doped source and drain regions where the influence of the gate voltage is negligible.

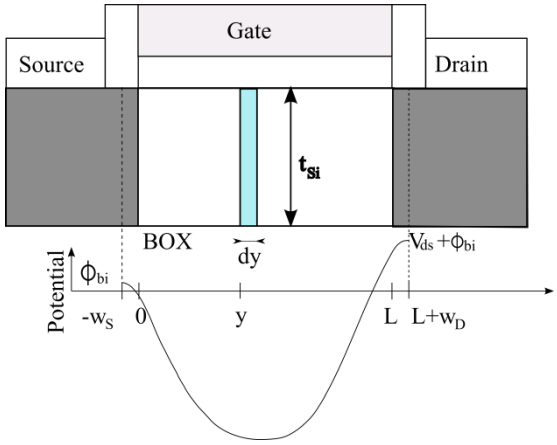


Fig. 3.13 Illustration of the method derived from [94] to obtain an analytical expression of the surface potential profile.

In the external regions ($y<0$ and $y>L$) we assume complete depletion, and therefore a parabolic potential profile given by

$$\frac{d^2\phi_S}{dy^2} = -\frac{qN_D}{\varepsilon_{Si}}, \tag{3.24}$$

where q is the electron charge, N_D the average doping in the source and drain regions, and ε_{Si} is silicon dielectric permittivity. In the central region $(0 < y < L)$ we use Gauss' theorem to write that the electric field flux through the surface of a slice of thickness dy (shown in Fig. 3.13) is zero [90], [93]. This allows us to write

$$\frac{t_{Si}}{\eta}\frac{dE_S(y)}{dy} + \frac{\varepsilon_{ox}}{\varepsilon_{Si}}\frac{V_{GS} - V_{FB} - \phi_S(y)}{t_{ox}} = 0, \tag{3.25}$$

where $E_S(y)$ is the lateral surface electric field, $\phi_S(y)$ is the channel potential at the SiO_2 interface, ε_{ox} is the oxide electric permittivity, V_{FB} is the flatband voltage, V_{GS} is the gate-to-source voltage. η is a fitting parameter that takes into account the fact that the electric field is not constant along x and that is not zero in the buried oxide in the case of single gate MOSFETs [96]. Its value usually varies between 1.0 and 1.3 [97].

The first term of (3.25) is the flux entering the Gaussian box along the y direction; the second term is the electric flux entering the top surface of the Gaussian box. Since we adopt the guess that the electric field is constant along the x direction, we can write $E_S(y) = -\frac{d\phi_S(y)}{dy}$.

Therefore the (3.25) becomes:

$$\frac{d^2\phi_S(y)}{dy^2} - \frac{\eta}{\varepsilon_{Si}t_{Si}}\frac{\varepsilon_{ox}}{t_{ox}}\phi_S(y) = \frac{\eta}{\varepsilon_{Si}t_{Si}}\frac{\varepsilon_{ox}}{t_{ox}}(V_{FB} - V_{GS}). \tag{3.26}$$

Solving (3.26) we have

$$\phi_S(y) = \phi_P + C\frac{\sinh(y/\lambda)}{\sinh(L/\lambda)} + D\frac{\sinh[(L-y)/\lambda]}{\sinh(L/\lambda)}, \tag{3.27}$$

where ϕ_P is the particular solution of the equation and result equal at $\phi_P = V_{GS} - V_{FB}$, and λ is the characteristic length defined as $\lambda = \sqrt{\varepsilon_{Si}t_{Si}t_{ox}/\eta\varepsilon_{ox}}$. Unknown terms w_S, w_D, C, D are obtained by enforcing continuity of ϕ_S and its derivative and by the boundary conditions:

$$\begin{cases} \phi_S(-w_S) = \phi_{bi} \\ \dfrac{d\phi_S}{dy}(-w_S) = 0 \end{cases}; \begin{cases} \phi_S(L + w_D) = V_{DS} + \phi_{bi} \\ \dfrac{d\phi_S}{dy}(L + w_D) = 0 \end{cases}. \tag{3.28}$$

Once $\phi_S(y)$ is known we can extract its minimum value in the channel $\phi_{SMIN}(V_{GS}, V_{DS})$ and obtain the threshold voltage as the gate voltage required to have $\phi_{SMIN} = \phi^*$, corresponding to the drain current used in the definition of V_{th}.

The minimum potential along the longitudinal direction can be obtained from

$$\left.\frac{d\phi_S(y)}{dy}\right|_{y_{\min}} = 0. \tag{3.29}$$

At low drain-source voltage value, we can be assuming that $y_{\min} = L/2$.

First, we want to validate our analytical model for the surface potential and the threshold voltage by comparison with TCAD results [93] on the template devices with the full doping profiles. In Fig. 3.14 the surface potential profiles for the 32 nm template MOSFET are compared for V_{DS} = 50 mV and 1 V, and for different values of V_{GS}. We use a single fitting parameter ($\eta = 1.2$) with the same value for the 32 nm and the 22 nm templates at the price of a suboptimal fitting. Nevertheless, agreement is very good. Very good agreement is obtained also for the 22 nm template MOSFET (Fig. 3.15).

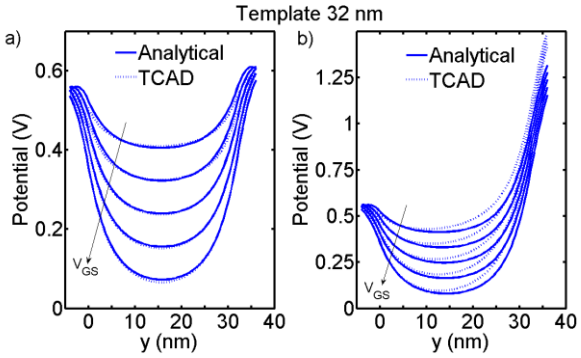


Fig. 3.14 Surface potential profile of the 32 nm template MOSFET as a function of y for different V_{GS} (0-0.5 V in steps of 0.1 V) for V_{DS} = 50 mV (a) or V_{DS} = 1 V (b). Comparison between analytical model and TCAD simulations. The arrow indicates the increase of V_{GS}.

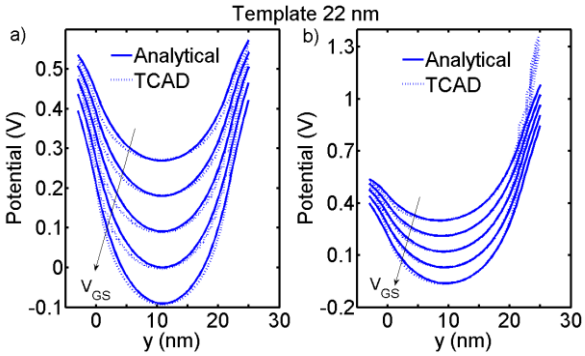


Fig. 3.15 Surface potential profile of the 22 nm template MOSFET as a function of *y* for different V_{GS} (0-0.5 V in steps of 0.1 V) for V_{DS} = 50 mV (a) or V_{DS} = 1 V (b). Comparison between analytical model and TCAD simulations. The arrow indicates the increase of V_{GS}.

4 Variability in Flash memory cells

4.1 Flash memory

With the progress of microelectronics technology and computer technology, the demand for information storage has been increasing rapidly. In the past 30 years, the semiconductor memory market continued to grow at high rate and give birth to flourishing market segments. One of such segments is represented by nonvolatile semiconductor memories (NVSMs, in particular Flash EEPROMs (Electrically Erasable Programmable Read-Only Memory) [98]

In a Flash EEPROM the content of the whole memory array, or a memory sector, is erased in one step. The first Flash EEPROM was proposed by Masuoka *et al*. at 1984. Flash has the advantage of good no volatility, and high density, and has found many application areas.

Different Flash implementations exist: NOR and NAND are the most common, and will be described later in the chapter [99]. NAND architectures, in particular, are used in memory cards and USB Flash drives for storage and data transfer between computers and digital systems. Other applications include personal digital assistants, notebooks, MP3 players, digital cameras, and cellular phones. In the last years it also gained some popularity in the game console

market. In the future, Flash-based systems such as solid-state disks (SSDs) are now replacing conventional hard disk drives (HDDs) [100].

Flash memories have strong limitations to perform random access in writing and reading. A Flash card like an SD-card is a small system in package built around the Flash memory and combined with a microcontroller capable of overcoming these limitations, being able to perform random access in both read and write, featuring a technology-independent interface.

The popularity of the Flash memory for applications such as storage on portable devices is mainly based on its distinctive characteristic of being nonvolatile (i.e., stored information is retained even when not powered, differently from other kinds of memories, like DRAM). Flash memory also offers fast access times in read (although not as fast as DRAM) and better mechanical shock resistance compared to hard-disks. The high storage density of a Flash-based system is typically achieved by means of advanced packaging techniques.

Furthermore, the smart memory management carried out by the microcontroller allows a high endurance and an appealing reliability of the resulting system.

As manufacturers increase the density of data storage in Flash devices, the size of individual memory cells becomes smaller and the number of electrons stored in the cell decreases. Moreover, coupling between adjacent cells and quantum effects can change the write characteristics of cells, making it more difficult to design devices able to guarantee reasonable data integrity. Therefore countermeasures such as improved error correction codes are applied.

A flash memory contains an array of floating-gate transistors: each of them is acting as memory cell. In single level cell (SLC) devices, each memory cell

stores one bit of information; for multi level cell (MLC) devices, more than one bit per cell can be stored [101]. Fig. 4.1 shows a schematic cross-section of a floating-gate memory cell. The floating gate is used to store electrons: changing the number of electrons results in a different threshold voltage of the transistor and, therefore, in a different drain current under fixed biasing conditions.

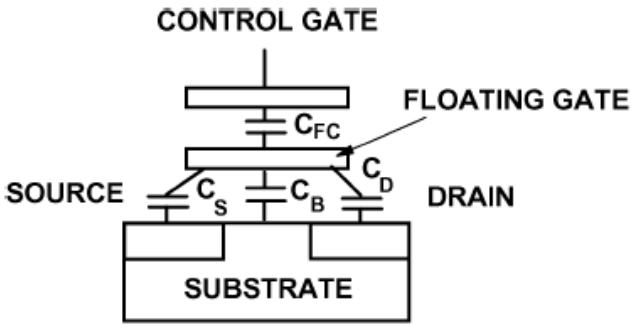


Fig. 4.1 Schematic representation of a floating-gate memory cell [Fig. 1 of [100]].

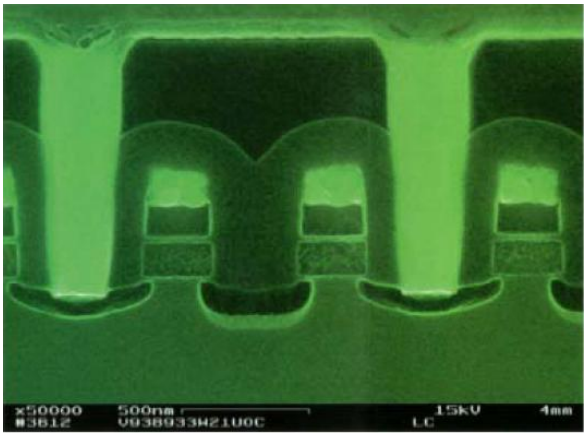


Fig. 4.2 Scanning electron microscope image showing the section of memory cells in Flash technology from 0.18 µm.

There are two dominant kinds of Flash memories: NAND and NOR. The name is related to the topology used for the cell array. NAND Flash memory is the core of the removable USB interface storage devices known as USB drives, as well as most memory card formats available today on the market.

NAND Flash uses Fowler–Nordheim (FN) tunnel injection and Channel Hot- Electron (CHE) for writing and Fowler–Nordheim tunnel release for erasing.

Due to their construction principles, NAND Flash memories are accessed like hard disks, and therefore are very suitable for use in mass-storage devices such as memory cards.

While programming is performed on a page basis, erase can only be performed on a block basis (i.e., a group of pages). Pages are typically 2048 or 4096 bytes in size.

Associated with each page there are a few bytes that can be used for storage of error detection and checksum as well as for administration as requested by the external microcontroller. Fig. 4.3 shows a schematic representation of the memory organization of a NAND Flash memory.

Another limitation is the finite number of write-erase cycles (manufacturers usually guarantee 10^5 write-erase cycles for SLC NAND). Furthermore, in order to increase manufacturing yield and reduce NAND memory cost, NAND devices are shipped from the factory with some bad blocks (i.e., blocks where some locations do not guarantee the standard level of reliability when used), which are identified and marked according to a specified bad-block marking strategy.

On the other neither hand, NOR Flash memories are capable of a very fast read access time (less than 100 ns), they offer a programming time comparable to that of the NAND (but the amount of programmed bit per operation is considerably smaller), but they feature an erase time which is some order of magnitudes higher than the NAND.

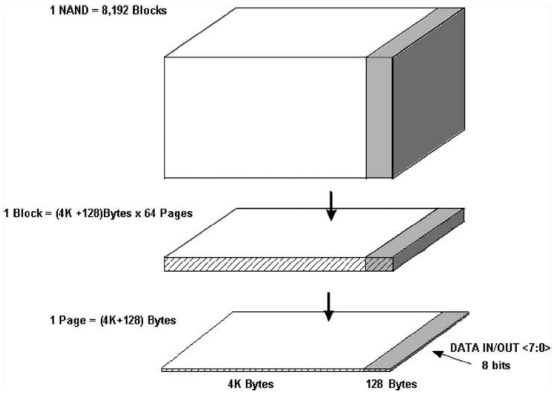


Fig. 4.3 Schematic representation of the memory organization of a NAND Flash memory [Fig. 2 of [100]].

Since different Flash cell structures require different programming methods, there are several physical principles for write and erase operation.

Currently, main mechanisms are the Fowler-Nordheim (F-N) Tunneling and Channel Hot- Electron (CHE). Both are shown in Fig. 4.4, Fig. 4.5 and Fig. 4.6. These are the ETOX cell programming method.

Fowler-Nordheim Tunneling is a field-assisted electron tunneling effect. During erase, a large voltage is applied between source and gate, which increases the transparency of the gate oxide.

The electrons can tunnel through the barrier from the poly-silicon conduction band and through the oxide conduction band into source. This lets the floating gate positively charge. The barrier of SiO_2 is about 3.2 eV for electrons in the conduction band of silicon and 4.8 eV for holes in the valence band. When a high field of > 10 MV/cm is applied across the ≈ 100A oxide, the electrons could directly tunnel the barrier and cause a significant tunneling current. The current density is given by [102]:

$$J = \alpha \cdot E_{inj}^2 \cdot e^{\left(-\frac{E_c}{E_{inj}}\right)} \tag{4.1}$$

wherein *J*: tunneling current density, E_{inj},: field of silicon -SiO_2 interface, E_c: the barrier energy, α: a field-independent coefficient.

Generally, Fowler-Nordheim current is for Flash cell erase operation. CHE is a generally used for Flash cell write operation. When the high voltage is applied to both drain and gate simultaneously, the high voltage across the drain to source gives a high channel field to generate high energy "hot" electrons and the control gate with high voltage attracts the part of these hot electrons (1uck electrons) to the floating gate. This process lets floating gate negatively charge and means write operation. The CHE current expression is following (4.2):

$$ln\left(\frac{I_g}{I_d}\right) = C_1 + \left(\frac{\varphi_b}{\varphi_l}\right) ln\left(\frac{I_{sub}}{I_d}\right) \tag{4.2}$$

wherein,

$$\varphi_b(E_{ox}) = 3.2 - \beta(E_{ox})^{1/2} - \vartheta(E_{ox})^{2/3} \tag{4.3}$$

The CHE current depends on both the horizontal and vertical field, the determined operation condition **is** a tradeoff between them.

In Fig. 4.4 and Fig. 4.5 the mechanisms to program the cell are shown: injection of hot electrons channel and Fowler - Nordheim tunneling. In Fig. 4.6 is illustrated the mechanism of erase the cell: Fowler - Nordheim tunneling.

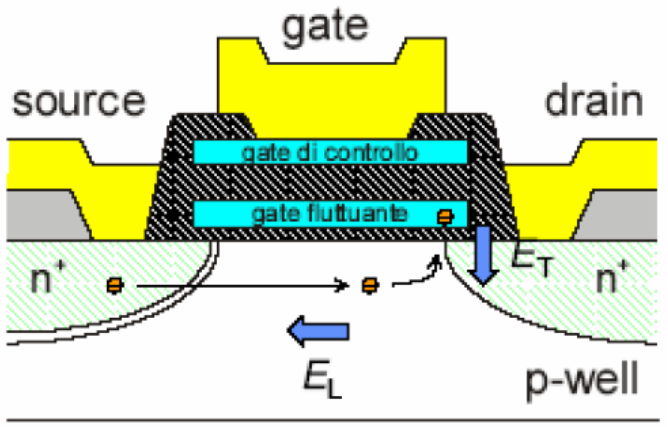


Fig. 4.4 Programming of the cell: injection of hot electrons channel.

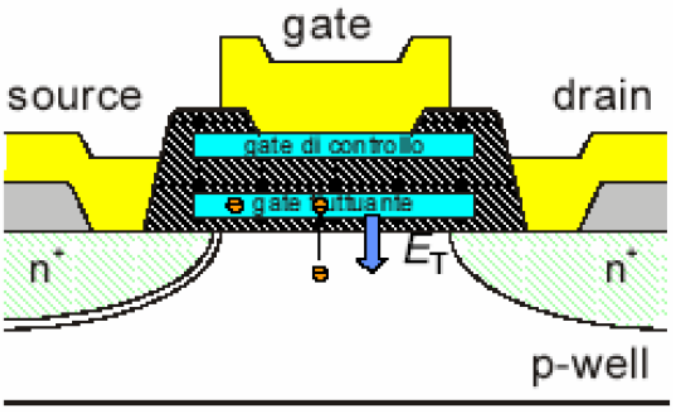


Fig. 4.5 Programming of the cell: Fowler - Nordheim tunneling.

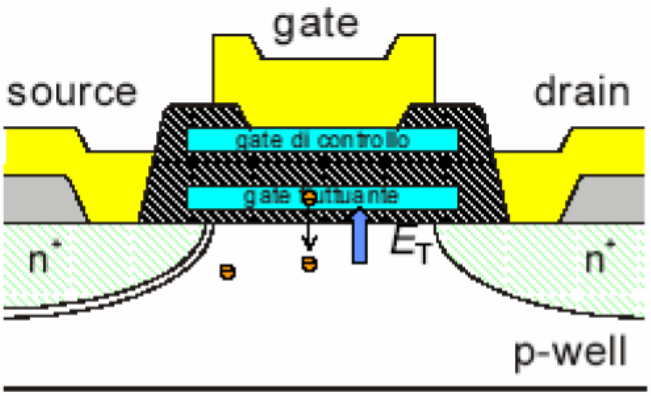


Fig. 4.6 Erased of the cell: Fowler - Nordheim tunneling.

As mentioned earlier Flash memory can be divided into two categories:

standard cells, in which each cell contains 1-bit, and multi-level, in which each cell contains n bits of information ($n = log_2 l$, l = number of levels). In these, the manufacturing process, procedures, and the circuitry for reading the data are much more complex. Most MLC NAND flash memory has four possible states per cell, so it can store two bits per cell. This reduces the amount of charge in the floating gate separating the states and results in the possibility of more errors. Multi-level cells which are designed for low error rates are sometimes called enterprise MLC (eMLC).

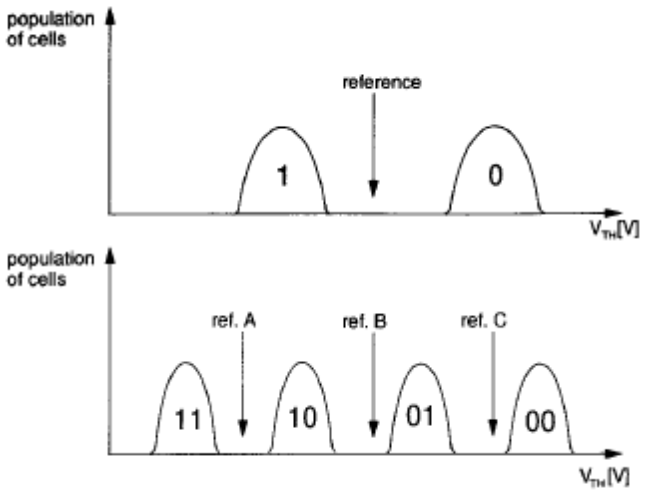


Fig. 4.7 Flash cells: threshold distribution for SLC and MLC.

NAND array is shown in Fig. 4.8. Its cell size is scaled down largely by connecting the bit of one byte in series to reduce the contact hole to 1/2n (n: bit number). Its both program and erase operations use Fowler-Nordheim tunneling. The select gate 1 and 2, selected control gate, unselected control gate, source and substrate are applied by different voltage to program or erase the selected cell. Its drawbacks are long access time and high programming voltage.

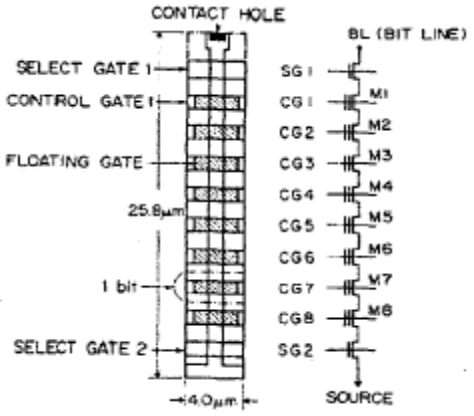


Fig. 4.8 NAND [Fig. 8 of [98]].

The array is presented in Fig. 4.9. Two cells in the array shares one common bit line. The write operation uses CHE, and erase operation uses Fowler-Nordheim tunneling at source. NOR has high programming speed and short access time. But the power consumption of programming is higher.

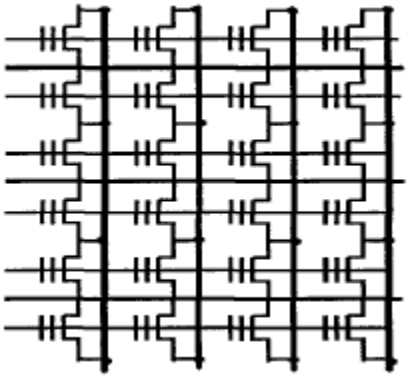


Fig. 4.9 NOR [Fig. 9 of [98]].

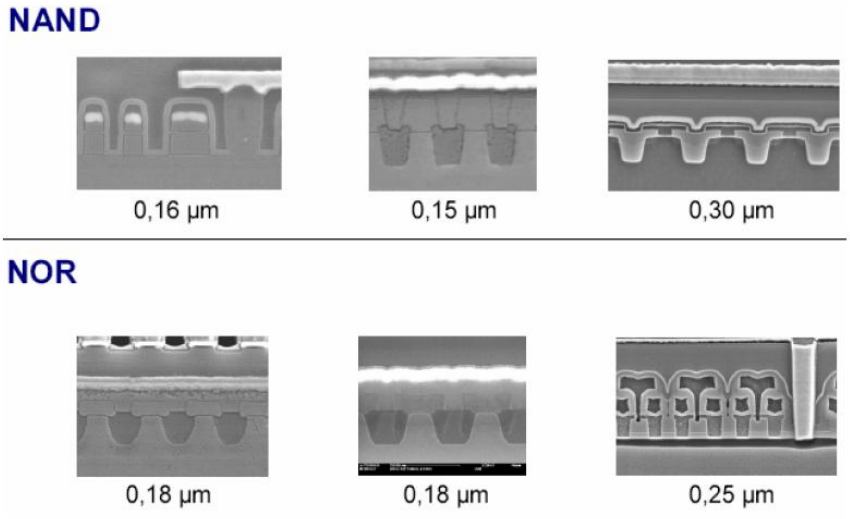


Fig. 4.10 NAND and NOR: various dimensions.

4.2 Variability

The aggressive scaling trend of the NAND Flash technology recently led to the development of storage cells with dimensions of few tens of nanometers. One of the consequences of the reduced device dimensions is the increasing role played by variability effects at the single-device level [103], [104], that strongly influence the threshold voltage (V_{th}) distribution of nanoscaled NAND arrays, affecting their performance and reliability. The standard way of assessing the impact of these phenomena is by 3D numerical simulations [83], but the computational load required makes this approach unpractical when extensive statistical analyses under different operating conditions are needed [105].

Variability effects are becoming more and more important for NAND Flash memories, affecting the uniformity and reproducibility of device characteristics and operations as cell dimensions scale down [105] - [108]. In this respect, two

main variability sources can be identified: The first is related to cell-to-cell parameter variations, due to process variations[105], and the second accounts for the statistics of few electron tunneling and is related to the discrete electron transfer to/from the floating gate (FG) during cell operation [106], [108], [109].

With the impressive shrinking of Non-Volatile Memory (NVM) device down to nanoscale dimensions atomistic level fluctuations play a significant role in the threshold voltage distribution of a memory array. However, multi-level bits storage in a single cell is required to reach a high degree of integration and it consequently requires an improved control of V_{th}. Thus, predicting scaling limits of Flash memories on the basis of signal/noise ratio requires both physical models and the analysis of the statistical distributions of the cell threshold voltage [110].

The paper [111] deals with the problem of threshold voltage fluctuations due to the Discrete Dopant Effects (DDE) but also introducing the Non-Uniform Conduction concept (NUC), including the Edge Effects (EE), the Trap and oxide fixed charge Distribution (TD) with the related Random Telegraph Signal (RTS) for a nanoscale Flash memory array. The capability to capture the relevant physical phenomena and their own weight with a simple analytical approach appears a strength point of the model otherwise supported by a physical understanding [112].

The resulting bit distributions are compared against experimental data and a good agreement in between is found.

4.3 State of art

The threshold voltage fluctuation in flash memory cells is one a critical issue for cell characterization. While several geometrical variations influence V_{th} for all cells in an array, a trap in tunnel oxide and dopant fluctuation mainly impact to the V_{th} variation of one memory cell. Thus, it is known that random telegraph noise (RTN) and dopant scattering are the major reason of V_{th} fluctuations, as reported in several research papers [113] - [116]. The V_{th} variation is more critical in multi-level-cell (MLC) operation due to the even narrower memory window.

In [117], the V_{th} fluctuation of one memory cell is investigated for NOR flashes cell scaling. From the experimental results, the influence of RTN and dopants on V_{th} fluctuation is estimated. Furthermore, it is proposed that this can be suppressed by the reduction of maximum RTN amplitude as a result of channel engineering and optimization.

In [105], a compact model is presented for studying the impact of variability effects on the V_{th} distribution of nanoscale NAND Flash memories not only when cells are in the neutral state but also after program/erase and retention. Both intrinsic and technological sources of fluctuation have been investigated and implemented in the model, assessing their importance under real operating conditions. The starting point for their analysis is the compact model for the NAND Flash memory array presented in [118].

The SPICE compact model was used in a Monte Carlo framework, running simulations to obtain the V_{th} distribution from the calculated string current under read conditions. In each simulation, the device parameters were randomly changed, to account for the different variability effects. In particular, the authors considered both process-induced fluctuations in the cell geometry and more

fundamental ones, due for example to the discrete nature of the charge. The former include W, L, tunnel and interpoly dielectric thickness fluctuations as well as fluctuations in the control-to-floating gate coupling coefficient; the latter account for random dopants (RDF) and oxide trap fluctuations (OTF). Process-induced fluctuations are directly inserted in the compact model by changing the device parameters (W, L, etc.) in each Monte Carlo run, according to Gaussian distributions whose spreads are extracted from process data. The implementation of the so-called intrinsic contributions is instead carried out as follows: RDF effect on V_{th} was accounted for by the analytical formula reported in [103], while the V_{th} variability due to OTF was implemented as $\sigma_{OTF} = K_{ox} Q_{ox}^{\alpha} / \sqrt{WL}$ with $\alpha \approx 0.5$ and K_{ox} and Q_{ox} fitted on cycled distribution data.

4.4 Our method

We have studied the effects of variability inherent in the case of Flash memory.

As part of the ENIAC MODERN project, we studied the effects of intrinsic parameter variability on the electrical characteristics of flash technology. As template device is used a Flash Memory Cell provided by Micron for the analysis of variability sources in non-volatile technologies in 32 nm. Glasgow University has carried out large-scale statistical simulations of threshold voltage variability of the template device.

The test device is basically a conventional floating gate NAND cell with simplified boundary and doping profiles. Dimensions are chosen to be representative of a 32 nm half pitch (F) technology.

4.5 32 nm Flash Cell

4.5.1 Device geometry

The cell layout is shown in Fig. 4.11. It is the typical layout of a cross-point device defined by the overlap of two orthogonal lines with minimum feature size (F) that represents the simplest and smallest lithographic pattern.

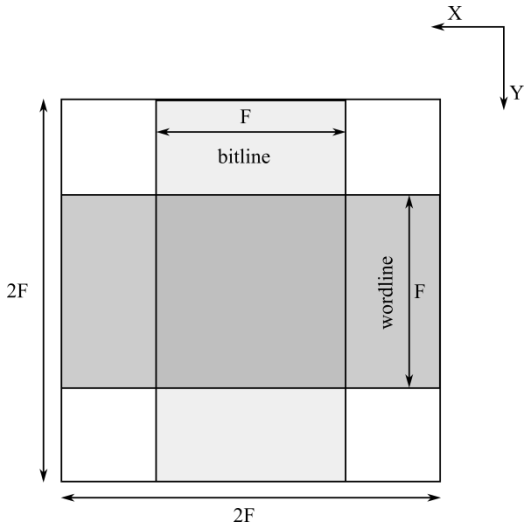


Fig. 4.11 $4F^2$ cell layout.

The cross sections along the wordline and bitline are shown in Fig. 4.12.

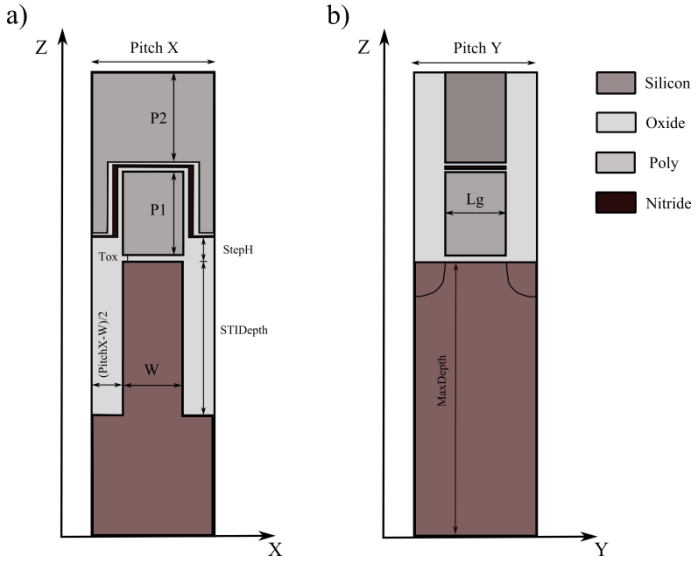


Fig. 4.12 a) wordline cross-section; b) bitline cross-section.

The template NVM cell is a generic floating-gate non-volatile memory (FG-NVM) cell with simplified boundary and doping profiles and a printed gate length of 32 nm. The simplified device is not indicative of any specific process technology or product but is a generalized structure that has been created to study the effects of variability on NVM. The gate stack of the FG-NVM cell consists of a polysilicon gate material with a 4-3-5-nm ONO layer and an 8-nm-thick tunneling oxide. The source–drain doping is symmetric and is created using an Arsenic implant at 1 x 10^{20} cm^{-3} with a Gaussian distribution, which results in a junction depth of 25 nm. The substrate has a uniform Boron doping concentration of 2 x 10^{18} cm^{-3}.

In Tab. 4.1 the symbols and the values of the geometrical parameters are reported.

Geometrical parameters	**Symbols and dimensions of layers**	
	Symbol	***Value***
Cell X dimension	PitchX	64 nm
Cell Y dimension	PitchY	64 nm
Active Area Width	W	32 nm
Gate length	Lg	32 nm
Silicon substrate thickness	MaxDepth	0.5 μm
Isolation Depth	STIDepth	0.2 μm
Tunnel oxide thickness	Tox	8 nm
	StepH	20 nm
Poly1 thickness	P1	70 nm
Poly2 thickness	P2	100 nm
ONO bottom oxide	ONO_bot	4 nm
ONO nitride	ONO_nit	3 nm
ONO top oxide	ONO_top	5 nm
Junction depth	xj	25 nm

Tab. 4.1 Numerical values of the different geometrical parameters [119].

The relative dielectric constants considered in the simulation are reported in the Tab. 4.2.

Material	Relative dielectric constant
Silicon	11.7
Oxide	3.9
Nitride	7.5

Tab. 4.2 The relative dielectric constants used in [119].

The electrodes are placed at the outer boundary of the structure:

• Control Gate (CGate) at the top face of Poly2;

• Substrate (Bulk) at the bottom face of the Silicon substrate;

• Source at the active area boundary at y = 0 for a depth of $x_j/2$;

• Drain at the active area boundary at y=PitchY for a depth of $x_j/2$;

In order to save mesh points, the FGate and CGate electrodes are located at the interface of Poly1 and Poly2 regions with other materials, respectively.

4.5.2 Doping profile

The Silicon substrate is doped with a constant profile of Boron with concentration $N_a = 2 \times 10^{18}\ cm^{-3}$. Poly1 (floating gate region) is doped with a constant concentration of Phosphorus of $3 \times 10^{19}\ cm^{-3}$, whereas Poly2 (wordline) is doped with a Phosphorus concentration of $10^{20}\ cm^{-3}$. Source and drain junction are symmetric and are made by an Arsenic Gaussian profile with peak concentration of $1e^{20} cm^{-3}$ and junction depth x_j. In the lateral direction the arsenic profile is still Gaussian with a "ratio" of the standard deviation with respect to the one of the vertical direction of 0.5. The Gaussian profiles extend

up to 0.5 x_j from the gate stack so that the drain (source) junction is aligned with the gate stack (Fig. 4.13).

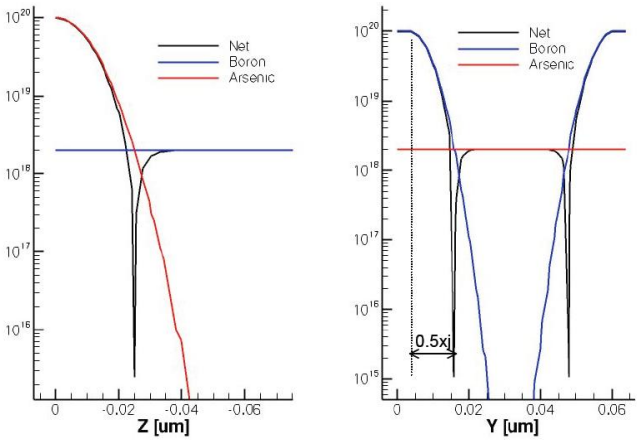


Fig. 4.13 Doping profiles in the vertical (left) and channel (right) directions.

4.6 Threshold voltage variability

To accurately characterise the intrinsic variability potentially of the template 32 nm flash cell the threshold voltage Roy *et al*. [118] perform a simulation of an ensemble of 1000 microscopically different devices. The sources of variability included in their study are Random Discrete Dopants (RDD), Line Edge Roughness (LER), Line Width Roughness (LWR), Oxide Thickness Fluctuations (OTF), Poly-silicon Granularity (PSG) and random interface-trapped charge (ITC). Each source is studied individually and then combined with RDD to observe the combined effect on the mean ($<V_{th}>$) and on the standard deviation (σV_{th}). This results in 13000 simulations being run with a total of 39000 CPU hours of computation.

We have extended our approach presented in the previous chapter, to the case of memories and in particular to this cell flash memory

The threshold voltage is calculated using a current criterion; in this case it is calculated using the equation:

$$I_{V_{th}} = 100 * \frac{W}{L} \text{ nA}$$

This gives a threshold voltage for the uniform device of 1.04 V for a drain current of 100 mV. This is selected to be below the "knee" of the transfer characteristics and at the top of the subthreshold region.

The methodology used by us to evaluate threshold voltage variability is described in detail in [89] and in Chapter 3.

The first step is critical, and requires close inspection of device fabrication and physics, the second is based on the assumption that first-order linearization is applicable. We have verified this to be the case in all cases, through detailed comparison with statistical atomistic simulations also for the non-volatile-memory technology described in [119].

The analysis for the memory is more complicated than that related to planar MOSFETs because it is necessary to carry out 3D simulations, since the shape of the control gate does not allows us to simply consider a two-dimensional structure.

4.6.1 Random Discrete Dopants

The first source of variation considered in [118] is Random Discrete Dopants (RDD) [120]. Variations due to RDD are caused by the random nature of the processing steps involved with implantation, diffusion, and annealing.

In order to study RDD, we adopt the same approach based on a propagator that is illustrated in Chapter 3. As a difference with respect to the previous situation, we here have to perform 3D simulations, since the Flash memory cells cannot be reduced to 2D structures.

The electron concentration of a template 32 nm flash cell with the inclusion of RDD is shown in Fig. 4.14; the oxide has been removed in the visualization so that the profile can be more easily seen.

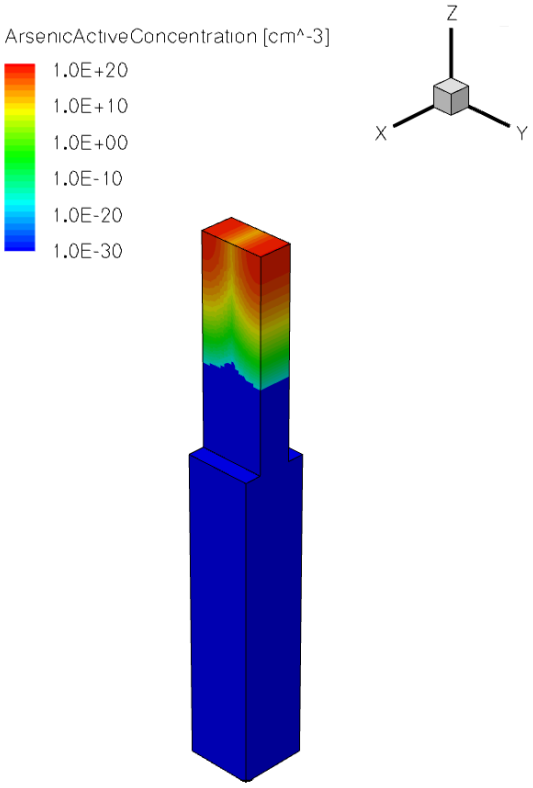


Fig. 4.14 Arsenic concentration in the cell obtained with Sentaurus TCAD.

In this case, we subdivide the device region into parallelepipeds. For each parallelepiped we multiply doping of each species by a factor (1 + α) (in our case α = 0.1) and compute the corresponding variation of the threshold voltage ΔV_{th} with respect to the nominal value.

Such investigation is particularly interesting because observing the results obtained for each parallelepiped, we can understand which part of the device region contributes the most to the threshold voltage dispersion.

We have evaluated that a partition of the three dimensional silicon body in boxes of size 8 x 8 x 25 nm^3 represents a good trade-off between computing time and accuracy. Considering the fact that we can exploit the symmetry of the structure also along the transport direction at very low drain-to-source voltage, only sensitivities corresponding to 16 boxes must be computed with TCAD simulations (Fig. 4.15).

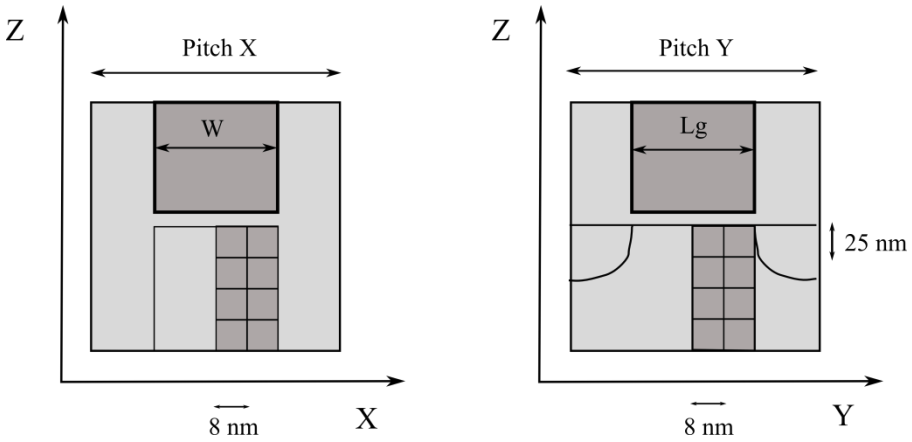


Fig. 4.15 Zoom of the regions which give contribution to the standard deviation of the threshold voltage.

The effect of RDD on the V_{th} have been compared with direct simulation of a statistical ensemble done at the University of Glasgow through GARAND [118] obtained simulating samples of 1000 microscopically different devices (Tab. 4.3).

RDD	Our method [89]	Stat. Sim. [118]
σ_{Vth} (mV)	137	141

Tab. 4.3 Standard variation of the threshold voltage due to random dopants.

4.6.2 Line Edge Roughness

The second source of variability considered in [118] is Line Edge Roughness (LER).

Also for this source we use the same approach described in Chapter 3. The results are compared with the values obtained from the statistical simulation done to study the V_{th} distribution of 1000 microscopically different devices.

In the simulations presented here Gaussian autocorrelation function has been considered.

The curve in the Fig. 4.16 shows the trend of the threshold voltage as a function of channel length.

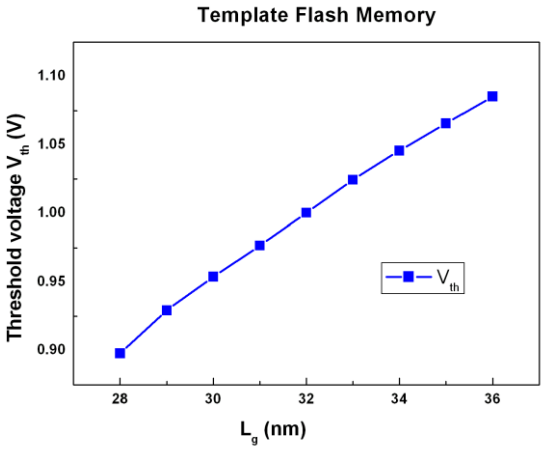


Fig. 4.16 Threshold voltage as a function of Lg for the template Flash Memory.

Assuming reasonable values for the two parameters, such as correlation length Λ_L = 20 nm and RMS amplitude Δ_L = 1.5 nm, results for W = L= 32 nm are shown in

Tab. 4.4. Results have been obtained using three device structures with slightly different gate lengths (31, 32, and 33 nm), very useful for computing the derivatives of threshold voltage with respect to gate length. Our results are compared with the values obtained by the University of Glasgow through GARAND [118].

LER	Our method [89]	Stat. Sim. [118]
σ_{Vth} (mV)	46	48

Tab. 4.4 Standard deviation of the threshold voltage due to LER obtained from sensitivity analysis.

4.6.3 Line Width Roughness

Line Width Roughness (LWR) is another source of variability included in [118]. LWR is created during the etching and filling of the surrounding shallow trench insulation (STI) on the edges of the template flash cell [40]. The rough edges lead to a variation in device width along the length of the device altering the amount of current produced. LWR is introduced in GARAND in a similar method as LER with randomly generated lines that can be applied to either front or back edges of the device. They are generated from the power spectrum

corresponding to a Gaussian autocorrelation function and are defined by RMS amplitude and correlation length.

For the LWR we adopt the same method and the same expression used to study the LER. We translate line width roughness in terms of the dispersion of the average position of both gate edges along the x axis ($x = 0 + x_a$ and $x = W + x_b$).

Assuming the same values for the two parameters we considered for LER (correlation length Λ_L = 20 nm and RMS amplitude Δ_L = 1.5 nm) we obtain results in Tab. 4.5. Threshold voltage sensitivity has been obtained using three device structures with slightly different gate width (30, 32, and 34 nm), very useful for computing the derivatives of threshold voltage with respect to gate width. These are compared with the results obtained using statistical simulation [118].

LWR	Our method [89]	Stat. Sim. [118]
σ_{Vth} (mV)	28	26

Tab. 4.5 Standard deviation of the threshold voltage due to LWR obtained from sensitivity analysis.

4.6.4 Oxide Thickness Fluctuations

The fourth source of variation considered in [118] is Oxide Thickness Fluctuations (OTF). OTF is caused by atomic scale roughness of the Si/SiO_2 and gate/SiO_2 interfaces on the scale of one inter-atomic layer and can introduce substantial variation in sub-1 nm EOT gate oxides [122].

Even for the OTF we applied the same method explained in Chapter 3 for MOSFETs. We then compared our results with those obtained from GARAND through the generation of a random 2D surface from the power spectrum corresponding to a Gaussian or Exponential autocorrelation function: as with the other sources of variation 1000 microscopically different device are simulated with OTF..

In Fig. 4.17 the dependence is shown of the threshold voltage as a function of oxide thickness.

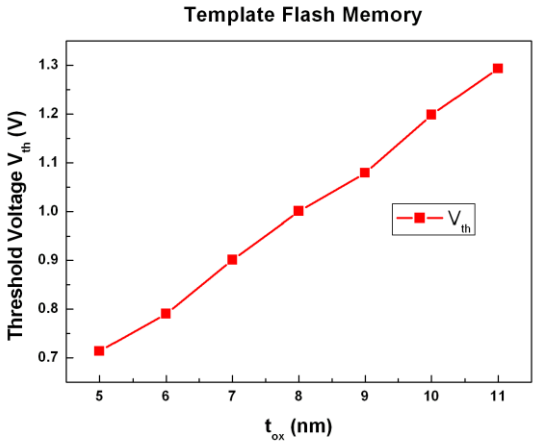


Fig. 4.17 Threshold voltage as a function of t_{ox} for the template Flash Memory.

Assuming reasonable values for the two parameters, such as correlation length Λ_S = 18 nm and RMS amplitude Δ_L = 0.2 nm, results for W = L = 32 nm are shown in Tab. 4.6, compared with the results in [118].

OTF	Our method [89]	Stat. Sim. [118]
σ_{Vth} (mV)	14	14

Tab. 4.6 Standard deviation of the threshold voltage due to OTV.

4.6.5 Interface-Trapped Charge (ITC)

Interface- Trapped Charge (ITC) in the insulator material is another significant source of variation considered in [118]. Interface trapped charge can be result of poor oxide quality or can be generated as a result of degradation associated with injection and trapping of carriers in the gate stack. Due to the discrete nature of the fixed charges and trapped carriers, reliability problems in contemporary CMOS have a statistical nature [123], [124]. For a particular average trapped charge density, the actual number of trapped charges will vary from transistor to transistor, and their actual position in each transistor will be unique.

For the ITC we used the same method used to solve the RDD, considering both top and lateral gate oxides.

First we did a simulation considering large parallelepipeds order to identify regions that effectively given contribution to the variation of the threshold voltage. Once the region to consider has been determined we reduced the size of

the parallelepipeds until the $\sigma_{V_{th}}$ does not decrease with decreasing size of the parallelepipeds.

Initially I considered a single region of size 16 x 64 x 200 nm^3 (Fig. 4.18). Simulating half device I have obtained: $\sigma_{V_{th}} = 27$ mV.

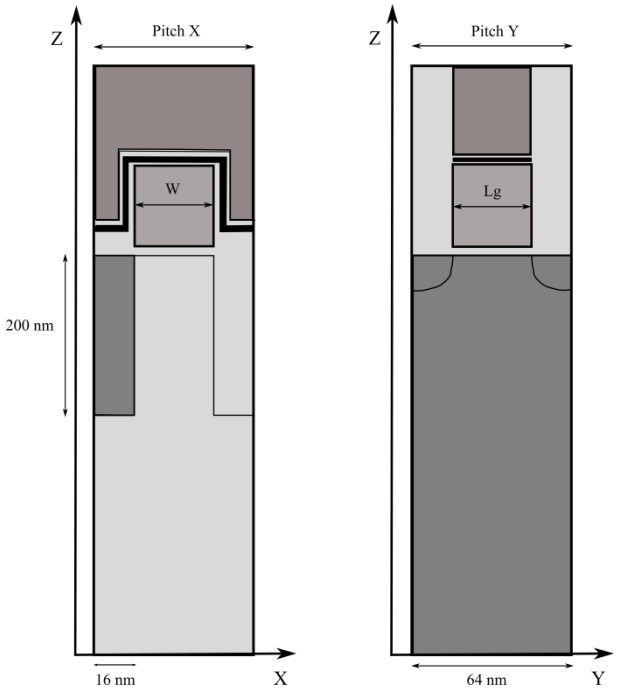


Fig. 4.18 One regions 16 x 64 x 200 nm^3.

After that I have further divided the region between 0 and 200 nm in the *z* direction in 2 parts, and also along *x* axis have divided the region into 2 parts. Basically I got parallelepipeds of dimensions 8 x 64 x 100 nm^3 (Fig. 4.19). I have obtained: $\sigma_{V_{th}} = 59$ mV.

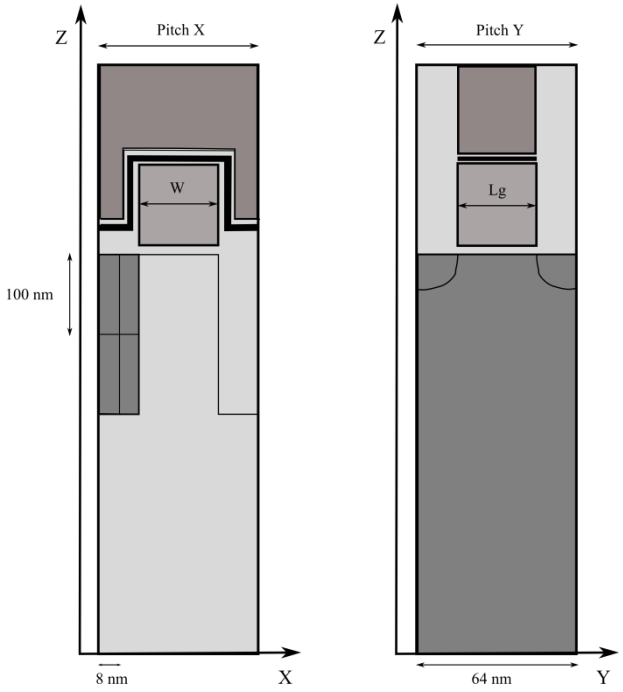


Fig. 4.19 Regions 8 x 64 x 100 nm^3.

I also tried to divide the same region into 4 rectangles vertically, that is, each of size equal to 16 x 64 x 50 nm^3 (Fig. 4.20). I have obtained: $\sigma_{V_{th}} = 56\ mV$.

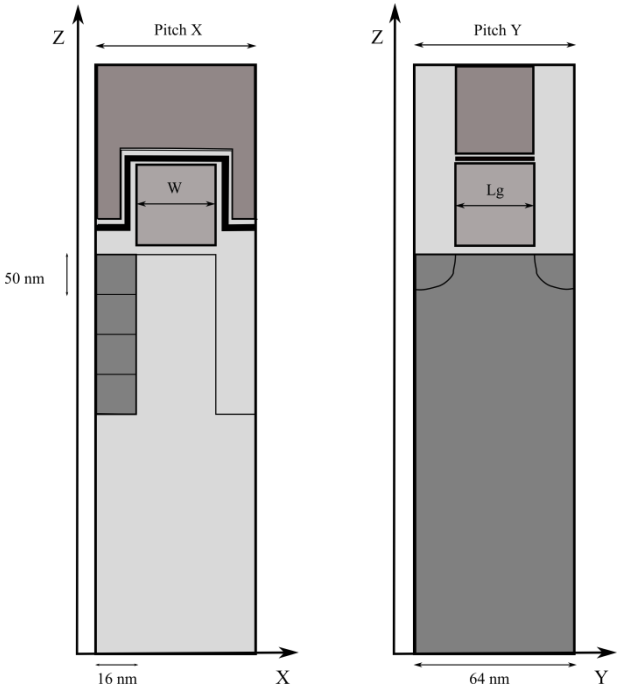


Fig. 4.20 Regions 16 x 64 x 50 nm^3.

Finally I tried to halve the size along the three directions, thus obtaining parallelepipeds of size 8 x 32 x 50 nm^3 (

Tab. 4.7) and I got a $\sigma_{V_{th}} = 59$ mV. It is thus evident that the value tends to stabilize.

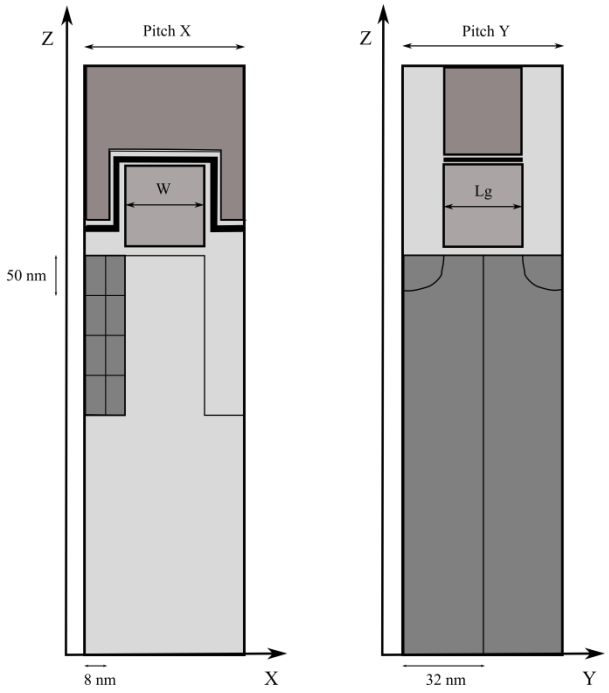


Fig. 4.21 Regions 8 x 32 x 50 nm^3.

We assume an average trap density of 5 x 10^{11} cm^{-2} and partition the tunnel oxide in tales of 8 x 64 x 100 nm^3, for a total of only four simulations, if the symmetry of the nominal structure is exploited.

The results obtained from us, compared with those obtained with the atomistic simulations by the University of Glasgow through GARAND [118], are shown in the Tab. 4.7 Standard deviation of the threshold voltage due to ITC.

ITC	Our method [89]	Stat. Sim. [118]
σ_{Vth} (mV)	59	67

Tab. 4.7 Standard deviation of the threshold voltage due to ITC.

4.7 Conclusions

We have proposed a methodology for the quantitative evaluation of the effects of the main mechanisms affecting threshold voltage variability, based on the careful identification of the main independent and relevant physical quantities. Our approach requires the calculation of partial derivatives of with respect to device structure parameters that can be obtained with a very limited number of TCAD simulations. We have shown that in all cases we are able to obtain results in good agreement with 3D atomistic statistical simulations [118] at a much smaller computational cost.

We qualify this statement to the second order moment of the threshold voltage distribution, because the proposed approach does not provide information on the far tails of the distribution, which are important for large Flash memory arrays, and would require extension of the method to higher order terms.

Conclusions

In this thesis we have addressed the problem of quantitative evaluation of device variability in CMOS technology, analyzing the main causes of variability and how its importance increases with technology downscaling. Among the different factors affected by parameter variations, we have chosen to focus exclusively on the dispersion of the threshold voltage.

In the first part we have proposed a methodology for the quantitative evaluation of the effect of line edge roughness, surface roughness, and random dopant distribution, that is based on the careful analysis of the main independent physical parameters affecting threshold voltage variability.

The analysis has been done considering various template devices: a 32 nm ultra-thin body SOI MOSFET and a 22 nm double-gate MOSFET adopted within the *EC PULLNANO* project, one bulk 45 nm NMOSFET within the project *MODERN* and 32/28 nm CMOS process developed by STM for *MODERN WP2* project.

In the second part we have used the model to investigate variability of memory devices. In particular, we have considered a Flash Memory Cell provided by Micron for the analysis of variability sources in non-volatile 32 nm technology.

We have shown that in all cases we are able to obtain results in good agreement with 3D atomistic statistical simulations at a much smaller computational cost.

In conclusion, in this thesis has been proposed a method for the quantitative assessment of the effect of the main sources of variability, which has a much lower computational cost than the statistical simulations typically used in the literature.

Bibliography

[1] T. N. Theis and P. M. Solomon, *Science* 327, pg. 1600, 2010.

[2] M. T. Bohr, "Nanotechnology Goals and Challenges for Electronic Applications", *IEEE Transactions on Nanotechnology*, vol. 1, NO. 1, March 2002.

[3] "Process Integration, Devices, and Structures", ITRS edition 2010.

[4] S. Roy and A. Asenov, "Where do the dopants go?", *SCIENCE* vol. 309, 15 July 2005.

[5] B. Hoeneisen, C. A. Mead, *Solid State Electron. 15*, pg. 819, 1972.

[6] R. W. Keyes, *Appl. Phys. 8*, pg. 251, 1975.

[7] K. J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS", *Electron Devices Meeting (IEDM)*, pp. 471-474, 2007.

[8] H. Mahmoodi, S. Mukhopadhyay and K. Roy, "Estimation of delay variations due to random-dopant fluctuations in nanoscale CMOS circuits", *IEEE J. of Solid State Circuits,* 40:9, pp. 1787-1795.

[9] H. Fukutome et al., "Direct Evaluation of Gate Line Edge Roughness Impact on Extension Profiles in Sub-50-nm n-MOSFETs", *IEEE Trans. Elec. Dev.*, 53:11, pp. 2755-2763, Nov. 2006.

[10] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometre MOSFETs introduced by gate line edge roughness", *IEEE Trans. Electron Devices*, vol. 50, no. 5, pp. 1254-1260, May 2003.

[11] Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*, 1999 Edition p.89.

[12] A. Asenov, S. Kaya, and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations", *IEEE Trans. Electron Devices*, vol. 49, no. 1, pp. 112-119, Jan. 2002.

[13] R. Difrenza, J. C. Vildeuil, P. Llinares and G. Ghibaudo, "Impact of grain number fluctuations in MOS transistor gate on matching performance", *Proc. International Conference on Microelectronic Test Structures, Monterey*, CA, USA, pp.244-249, 2003.

[14] C. L. Alexander, A. R. Brown, J. R. Watling, and A. Asenov, "Impact of single charge trapping in nano-MOSFETs—Electrostatics versus transport effects," *IEEE Trans. Nanotechnol.*, vol. 4, no. 3, pp. 339–344,May 2005.

[15] J. R. Brews, "Surface potential fluctuations generated by interface charge inhomogeneities in MOS devices", *J. Appl. Phys.*, vol. 43, pp. 2306-2313, 1972.

[16] L. Capodieci, "From optical proximity correction to lithography-driven physical design (1996-2006): 10 years of resolution enhancement technology and the roadmap enablers for the next decade", Proc. SPIE vol. 6154.

[17] C. T. Liu et al., "Severe thickness variation of sub-3 nm gate oxide due to Si surface faceting, poly-Si intrusion and corner stress", 2006 Sym. VLSI Tech., pg. 75.

[18] T. K. Yu et al., "A two-dimensional low pass filter model for die-level topography variation resulting from chemical mechanical polishing of ILD films", IEDM 1999, pp. 909-912.

[19] I. Ahsan et al., "RTA-Driven Intra-Die Variations in Stage Delay, and Parametric Sensitivities for 65 nm Technology", 2006 Sym. VLSI Tech., pp. 170-171.

[20] T. Tanaka et al., "Vth fluctuation induced by statistical variation of pocket dopant profile", IEDM 2006, pp. 271-274.

[21] A. Asenov, "Simulation of Statistical Variability in Nano MOSFETs", 2007 Sym. VLSI Tech., pp. 86-87.

[22] E. Fetzer, "Using Adaptive Circuits to Mitigate Process Variations in a Microprocessor Design", IEEE Design and Test of Computers, 23:6, pp. 476-483, June 2006.

[23] 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 Transactions on Electron Devices*, vol. 53, NO. 12, December 2006.

[24] A. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuatios on CMOS SRAM cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658-665, Apr. 2001.

[25] Y. Taur and T. H. Ning,, "Fundametals of Modern VLSI Devices," New York: Cambridge Univ Press, 1998.

[26] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variation and impact on circuits and microarchitecture ," in *Proc. Design Automation Conf.*, 2003, pp. 338-342.

[27] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime", *Proceedings of the IEEE,* vol. 85, NO. 4, April 1997.

[28] Y. Li and C.-H. Hwang, "Discrete-dopant-induced characteristic fluctuations in 16 nm multiple-gate silicon-on-insulator devices", *Journal of Applied Physics* 102, Dec. 2007.

[29] K. Takeuchi, T. Tatsumi, and A. Furukawa, Tech. Dig. – Int. Electron Devices Meet. 1997, pg. 841.

[30] P. Andrei and I. Mayergoyz, "Quantum mechanical effects on random oxide thickness and random doping induced fluctuations in ultrasmall semiconductor devices", *J.Appl. Phys.*, vol.94, no. 11, pp. 7163-7172, Dec. 2003.

[31] Y. Li and S.-M. Yu, J. Comput. Electron. 5, pg. 125, 2006.

[32] Y. Li and S.-M. Yu, Jpn. J. Appl. Phys., Part 1 45, pg. 6860, 2006.

[33] A. Asenov, IEEE Trans. Electron Devices 45, pg. 2505, 1998.

[34] D. J. Frank, Y. Taur, M. Ieong, and H.-S. Wong, Tech. Dig. VLSI Symp., pg. 169, 1999.

[35] R. Brown, A. Asenov, and J.R. Watling, IEEE Trans. Nanotechnol. 1, pg. 195, 2002.

[36] C. L. Alexander, G. Roy, and A. Asenov, Tech. Dig. – Int. Electron Devices Meet. 1, 2006.

[37] W. J. Gross, D. Vasileska, and D. K. Ferry, IEEE Electron Devices Lett.20, pg. 463, 1999.

[38] C. J. Wordelman and U. Ravaioli, IEEE Trans. Electron Devices 47, pg. 410, 2000.

[39] *Test Method for Evaluation of Line-Edge Roughness and Line-Width Roughness*, SEMI Std. Rep. P47-0307, 2007.

[40] Ji-Young Lee, Jangho Shin, Hyun-Woo Kim, Sang-Gyun Woo, Han-Ku Cho, Woo-Sung Han, and Joo-Tae Moon, "Effect of line-edge roughness (LER) and line-width roughness (LWR) on sub-100-nm device performance", Proc. SPIE 5376, 426 (2004).

[41] J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology", *IBM Journal of Research and Development*, Mar/May 2002.

[42] J. L. McCreary, "Matching properties and voltage and temperature dependence of MOS capacitors", *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 608-616, Dec. 1981.

[43] J. B. Shyu, G. C. Temes, and K. Yao, "Random errors in MOS capacitors", *IEEE J. Solid-State Circuits*, vol. SC-17, pp.1070-1076, Dec. 1982.

[44] J. B. Shyu, G. C. Temes, and F. Krummenacher, "Random error effects in matched MOS capacitors and current sources", *IEEE J. Solid-State Circuits*, vol. SC-19, pp. 948-955, Dec.1984.

[45] K. R. Lakshmikumar, R. A. Hadaway, and M. A. Copeland, "Characterization and Modeling of Mismatch in MOS Transistors for Precision Analog Design", *IEEE J. Solid-State Circuits*, vol. SC-21, pp. 1057-1066, 1986.

[46] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of MOS transistors", *IEEE J.Solid-State Circuits*, vol. SC-24, pp. 1433-1440, 1989.

[47] M-F. Lan and R. Geiger, "Impact of model errors on predicting performance of matching-critical circuits", *Proc. 43 rd IEEE Midwest Symp. on Circuits and System*, pp. 1324-1328, 2000.

[48] P. G. Drennan, and C. C. McAndrew, "Understanding MOSFET mismatch for analog design", *IEEE J.Solid-State Circuits*, vol. 38, no. 3, pp. 450-456, Mar. 2003.

[49] H. Yang et al., "Current mismatch due to local dopant fluctuations in MOSFET channel", *IEEE Trans. Electron Devices*, vol. 50, no. 11, pp. 2248-2254, Nov. 2003.

[50] H. Klimach, A. Arnaud, C. Galup-Montoro, and M. C. Schneider, "MOSFET Mismatch Modeling: A New Approach", *IEEE Design & Test of Computers*, Jan.-Feb. 2006.

[51] H. Klimach, A. Arnaud, M. C. Schneider and C. Galup-Montoro, "Characterization of MOS Transistor Current Mismatch", *SBCCI* , Pernambuco, Sept. 2004.

[52] M. J. M. Pelgrom, H. P. Tuinhout, and M. Vertregt, "Transistor matching in analog CMOS applications", *International Electroil Device Meeting Teclnrical Digest (IEDM)*, pp. 915-918, 1998.

[53] Berkeley Predictive Technology Model [Online]. Available: http:// www-device.eecs.berkeley.edu/~ptm/.

[54] K. A. Bowman, X. Tang, J. C. Eble, and J. D. Meindl, "Impact of Extrinsic and Intrinsic Parameter Fluctuations on CMOS Circuit

Performance", *IEEE J.Solid-State Circuits*, vol. 35, NO. 8, August 2000.

[55] B. Cheng, S. Roy, and A. Asenov, "Impact of Intrinsic Parameter Fluctuations on SRAM Cell Design", *Solid-State and Integrated Circuit Technology*, IEEE 2006.

[56] E. Seevinck, F. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells", *IEEE J.Solid-State Circuits*, vol. SC-22, pp. 748-754, Oct. 1987.

[57] X. Tang, V. K. De, and J. D. Meindl, "MOSFET fluctuation limits on gigascale integration (GSI) ", in *1998 Eur. Solid-State Device Research Conf. (ESSDERC'98)*, Bordeaux, France, pp. 508-511, Sept. 1998.

[58] A. J. Bhavnagarwala and J. D. Meindl, "Dynamic threshold CMOS SRAM cells for fast portable applications", in *Proc. 14th IEEE Int. ASIC/SOC Conf.*, Arlington, VA, pp. 359-363, Sept. 2000.

[59] H.-S. Wong and Y. Taur, "Three-Dimensional 'Atomistic' Simulation of Discrete Random Dopant Distribution Effects in Sub-0.1μm MOSFET's", *Tech. Digest IEDM*, pp. 705-708, 1993.

[60] E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, "A new 3-D device simulation formulation", in *NASCODE VI: Sixth Int. Conf. on*

Numerical Analysis of Semiconductor Devices and Integrated Circuits, Dublin, Boole Press, pg. 291, 1989.

[61] A. Asenov, A. R. Brown, J. H. Davies, and S. Saini, "Hierarchical Approach to 'Atomistic' 3-D MOSFET Simulation", *IEEE Transactions on computer-aided design of integrated circuits and systems*, vol. 18, NO. 11, November 1999.

[62] A. Asenov, "Random Dopant Induced Threshold Voltage Lowering and Fluctuations in Sub-0.1 μm MOSFET's: A 3-D 'Atomistic' Simulation Study", *IEEE Transactions on Electron Devices*, vol. 45, NO.12, December 1998.

[63] K. Takeuchi, T. Tatsumi, and A. Fukurawa, "Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuations", in *IEDM Tech. Dig.*, 1996.

[64] A. Asenov, S. Slavcheva, A. R. Brown, J. H. Davies, and S. Saini, *IEEE Trans. Electron Devices* 48, pg. 722, 2001.

[65] D. Reid, C. Millar, G. Roy, S. Roy, and A. Asenov, "Analysis of threshold voltage distribution due to random dopants: A 100 000 sample 3D simulation study," *IEEE Trans. Electron Devices*, vol. 56, no. 10, pp. 2255–2263,Oct. 2009.

[66] A. Asenov, A. Brown, B. Cheng, J. R.Watling, G. Roy, and C. Alexander, "Simulation of nano-CMOS devices: From atoms to architecture," in *Nanotechnology for Electronic Materials and Devices*. Berlin, Germany: Springer-Verlag, 2006.

[67] S. Rafferty, B. Biegel, M. G. Ancona, Z. Yu, J. Bude, and R.W. Dutton, "Multi-dimensional quantum effects simulation using a density-gradient model and script-level programming technique," in *Proc. SISPAD*, 1998, pp. 137–140.

[68] S. Inaba, K. Okano, S. Matsuda, M. Fujiwara, A. Hokazono, K. Adachi, K. Ohuchi, H. Suto, H. Fukui, T. Shimizu, S. Mori, H. Oguma, A. Murakoshi, T. Itani, T. Iinuma, T. Kudo, H. Shibata, S. Taniguchi, M. Takayanagi, A. Azuma, H. Oyamatsu, K. Suguro, Y. Katsumata, Y. Toyoshima, and H. Ishiuchi, "High performance 35 nm gate length CMOS with NO oxynitride gate dielectric and NI salicide," *IEEE Trans. Electron Devices*, vol. 49, no. 12, pp. 2263–2270, Dec. 2002.

[69] *Taurus Process and Device*, Synopsys, Mountain View, CA, Sep. 2004. 2004.09.

[70] 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. 53, no. 12, pp. 3063–3070, Dec. 2006.

[71] J. Wu, J. Chen, and K. Liu, "Transistor Width Dependence of LER Degradation to CMOS Device Characteristics", *Simulation of Semiconductor Processes and Devices*, pp. 95-98, 2002.

[72] S. Xiong and J. Bokor, "A Simulation Study of Gate Line Edge Roughness Effects on Doping Profiles of Short-Channel MOSFET Devices", *IEEE Transactions on Electron Devices*, vol. 51, NO. 2, Feb. 2004.

[73] C. H. Diaz, H.- J. Tao, Y.- C. Ku, A. Yen, and K. Young, "An Experimentally Validated Analytical Model For Gate Line-Edge Roughness (LER) Effects on Technology Scaling", *IEEE Electron Device Letters*, vol. 22, NO. 6, June 2001.

[74] P. Oldiges, Q. Lin, K. Pertillo, M. Sanchez, M. Ieong, and M. Hargrove, "Modelling line edge roughness effects in sub 100 nm gate length devices," in *Proc. SISPAD*, 2000, pp. 131–134.

[75] T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto, "Random threshold voltage variability induced by gate-edge fluctuations in nanoscale metal–oxide–semiconductor field-effect transistors," *Appl. Phys. Express*, vol. 2, no. 2, p. 024 501, Feb. 2009.

[76] T. Linton, M. Giles, and P. Packan, "The impact of line edge roughness on 100 nm device performance," in *Proc. Ext. Abs. Silicon Nanoelectron. Workshop*, 1998, pp. 82–83.

[77] T. D. Linton, S. Yu, and R. Shaheed, "3D modeling of fluctuation effects in heavily scaled VLSI devices," *VLSI Des.*, vol. 13, pp. 103–109, 2001.

[78] A. Asenov, S. Kaya, and A. R. Brown, "Intrinsic parameter fluctuations in decananometreMOSFETs introduced by gate line edge roughness," *IEEE Trans. Electron Devices*, vol. 50, no. 5, pp. 1254–1260, May 2003.

[79] M. Hane, T. Ikezawa, and T. Ezaki, "Atomistic 3D process/device simulation considering gate line-edge roughness and poly-Si random crystal orientation effects," in *IEDM Tech. Dig.*, 2003, pp. 951–954.

[80] X. Wang, S. Roy, and A. Asenov, "Impact of strain on the performance of high-k/metal replacement gate MOSFET," in *Proc. ULIS*, 2009, pp. 289–292.

[81] D. Reid, C. Millar, S. Roy, and A. Asenov, "Understanding LER-induced MOSFET *VT* variability—Part I: Three-dimensional simulation of large statistical samples," *IEEE Trans. Electron Devices*, vol. 57, no. 11, pp. 2801–2807, Nov. 2010.

[82] 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. 53, no. 12, pp. 3063–3070, Dec. 2006.

[83] 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, no. 6, pp. 609–611, Jun. 2008.

[84] FP6 Integrated Project PULLNANO, Deliverable D6.4.5.1.

[85] B. Cheng, S. Roy, A. R. Brown, C. Millar, and A. Asenov, "Evaluation of statistical variability in 32 and 22 nm technology generation LSTP MOSFETs," *Solid State Electron.*, vol. 53, no. 7, pp. 767–772, Jul. 2009.

[86] D. Reid, C. Millar, S. Roy, and A. Asenov, "Statistical Enhancement of the Evaluation of Combined RDD- and LER- Induced V_T Variability: Lessons From 10^5 Sample Simulations," *IEEE Trans. Electron Devices*, vol.58, no.8, pp.2257-2265, Aug. 2011.

[87] 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. 53, no. 12, pp. 3063–3070, Dec. 2006.

[88] Z. Qin and S. T. Dunham, "Atomistic simulations of effect of coulombic interactions on carrier fluctuations in doped silicon," in *Proc. Mater. Res. Soc. Symp.*, 2003, p. 765.

[89] V. Bonfiglio and G. Iannaccone, "An Approach Based on Sensitivity Analysis for the Evaluation of Process Variability in Nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2266-2273, 2011.

[90] Pullnano Deliverable D4.5.1. [Online]. Available: www.pullnano.eu

[91] X. Wang, G. Roy, O. Saxod, A. Bajolet and A. Juge, "Simulation Study of Dominant Statistical Variability Sources in 32-nm High-k/Metal Gate CMOS", *IEEE Electron Device Letter*, Jan. 2012.

[92] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, "Surface roughness at the Si (100)– SiO_2 interface," *Phys. Rev.*, vol. 32, no. 12, pp. 8171–8186, Dec. 1985.

[93] Manual of TCAD Sentaurus, ver. 12 Synopsys, Inc. Mountain View, CA, 2007.

[94] F. Bonani, S. D. Guerrieri, F. Filicori, G. Ghione, and M. Pirola, "Physicsbased large-signal sensitivity analysis of microwave circuits using technological parametric sensitivity from multidimensional semiconductor device models," *IEEE Trans. Microw. Theory Tech.*, vol. MTT-45, no. 5, pp. 846–855, May 1997.

[95] H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko, and Y. C. Cheng, "Threshold voltage model for deep-submicrometer MOSFETs," *IEEE Trans. Electron Devices*, vol. 40, no. 1, pp. 86–95, Jan. 1993.

[96] L. Perniola, S. Bernardini, G. Iannaccone, P. Masson, B. De Salvo, G. Ghibaudo, and C. Gerardi, "Analytical model of the effects of a nonuniform distribution of stored charge on the electrical characteristics of discrete-trap nonvolatile memories," *IEEE Trans. Nanotechnol.*, vol. 4, no. 3, pp. 360–368, May 2005.

[97] P. K. Ko, "Hot-electron effects in MOSFETs," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, 1982.

[98] Z. Jun, "Flash memory technology development," Solid-State and Integrated-Circuit Technology, 2001.

[99] P. Cappelletti et al., "Flash Memories." Norwood, MA: Kluwer Academic, 1999.

[100] R. Micheloni, M. Picca, S. Amato, H. Schwalm, M. Scheppler and S. Commodaro, "Non-Volatile Memories for Removable Media," in *Proc. of the IEEE*, vol. 87, no. 1, Jan. 2009.

[101] G. Campardo, R. Micheloni, and D. Novosel, " VLSI-Design of Non-Volatile Memories." Berlin, Germany: Springer-Verlag, 2005.

[102] C. Park et al., "A high performance controller for NAND Flash-based Solid State Disk (NSSD)," in *Proc. IEEE Non-Volatile Semiconduct. Memory Workshop (NVSMW)*, Feb. 2006, pp. 17–20.

[103] 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, Sept. 2003.

[104] J.-D. Lee, S.-H. Hur, and J.-D. Choi, "Effects of floating-gate interference on NAND Flash memory cell operation," *IEEE Electron Device Lett.*,vol. 23, pp. 264–266, May 2002.

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

[106] 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, no. 10, pp. 2695–2702, Oct. 2008.

[107] 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, no. 11, pp. 3192–3199, Nov. 2008.

[108] G. Molas, D. Deleruyelle, B. De Salvo, G. Ghibaudo, M. Gely, L. Perniola, D. Lafond, and S. Deleonibus, "Degradation of floating-gate memory reliability by few electron phenomena," *IEEE Trans. Electron Devices*, vol. 53, no. 10, pp. 2610–2619, Oct. 2006.

[109] C. Miccoli, C. Monzio Compagnoni, S. M. Amoroso, A. Spessot, P. Fantini, A. Visconti, and A. S. Spinelli, "Impact of neutral threshold voltage spread and electron-emission statistics on data retention of nanoscale NAND Flash," *IEEE Electron Device Lett.*, vol. 31, no. 11, pp. 1202–1204, Nov. 2010.

[110] R. Gusmeroli, et al., "Defects spectroscopy in SiO_2 by statistical random telegraph noise analysis," in IEDM Tech. Dig., pp. 483-486, 2006.

[111] A. Calderoni, P. Fantini, A. Ghetti, and A. Marmiroli, "Modelling the V_{th} fluctuations in nanoscale floating gate memories," in *Proc. SISPAD*, 2008, pp. 49–52.

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

[113] A. Ghetti, C. Compagnoni, F. Biancardi, A. Lacaita, S. Beltrami, L. Chiavarone, A. Spinelli, and A. Visconti: *Proc. IEDM Tech. Dig.*, pg. 835, 2008.

[114] S. Shukuri, N. Ajioka, M. Mihara, K. Kobayashi, T. Endoh, and M. Nakashima: *SYMP. on VLSI Technology*., pg. 20, 2006.

[115] Y. H. Song, J. Y. Lee, S. E. Lee, and J. H. Park*: Jpn. J. Appl. Phys*. vol. 46 no.8, pg.5067, 2007.

[116] W. H. Kwon, Y. H. Song, Y. Cai, and S. P. Shim: *Jpn. J. Appl. Phys. v*ol. 47, no.12, 8802, 2008.

[117] H. An, K. Kim, S. Jung, H. Yang, K. Kim, and Y. Song, "The Threshold Voltage Fluctuation of one memory cell for the scaling down Nor Flash," *Proceedings of IC-NIDC*, 2010.

[118] L. Larcher, A. Padovani, P. Pavan, P. Fantini, A. Calderoni, A. Mauri, and A. Benvenuti, "Modeling NAND Flash memories for IC design," *IEEE Electron Dev. Lett.*, vol. 29, pp. 1152–1154, Oct. 2008.

[119] G. Roy, A. Ghetti, A. Benvenuti, A. Erlebach, and A. Asenov, "Comparative simulation study of the different sources of statistical variability in contemporary floating gate non-volatile memory," *IEEE Trans. Electron Devices*, vol. 58, no. 12, pp. 4155–4163, Dec. 2011.

[120] N. Sano, K. Matsuzawa, M. Mukai and N. Nakayama, "On discrete random dopant modelling in drift-diffusion simulations: physical meaning of 'atomistic' dopants," Microelectronics and Reliability 42:22, 189-199, 2/2002.

[121] Ji-Young Lee, Jangho Shin, Hyun-Woo Kim, Sang-Gyun Woo, and Han-Ku Cho, Woo-Sung Han, and Joo-Tae Moon, "Effect of line-edge roughness (LER) and line-width roughness (LWR) on sub-100-nm device performance", Proc. SPIE 5376, 426 (2004)

[122] A. Asenov, S. Kaya and J. H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," IEEE Transactions on Electron Devices, vol. 49, pp. 112–119, 2002.

[123] A. R. Brown, V. Huard and A. Asenov, Statistical simulation of progressive NBTI degradation in a 45 nm technology pMOSFET, IEEE Trans. On Electron Devices (in press).

[124] M. Faiz. Bukhori, S. Roy and A. Asenov, "Simulation of Statistical Aspects of Charge Trapping and Related Degradation in Bulk MOSFETs in the Presence of Random Discrete Dopants," IEEE Trans. Electron Dev. vol. 57, iss. 4, pp. 795–803, Apr. 2010.