ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

ARCES - Advanced Research Center on Electronic Systems for Information and Communication Technologies E. De Castro

European doctorate program in Information Technology (EDITH)

Cycle XX - ING-INF/01

TCAD Approaches to Multidimensional Simulation of Advanced Semiconductor Devices

Emanuele Baravelli

Ph.D. Thesis

Tutor

Prof. Guido Masetti

Coordinator

Prof. Riccardo Rovatti

January 2005 - December 2007

Contents

List of Symbols

C	net ionized impurity concentration, defined as $N_D^+ - N_A^-$
D_n, D_p	thermal diffusion coefficients for electrons and holes
$\vec{E}$	electric field
E_C	conduction band energy
E_F	Fermi energy level
E_V	valence band energy
J_n, J_p	current densities for electrons and holes
N_A^-	ionized acceptor concentration
N_C	conduction band density of states
N_D^+	ionized donor concentration
$R(\psi, n, p)$	net carrier generation/recombination
T	lattice temperature
T_n, T_p	electron and hole temperatures
α_n, α_p	carrier ionization coefficients
ϵ	dielectric constant of the considered material
h	Planck constant, 6.626×10^{-34} J·s
$\hbar$	reduced Planck constant, defined as $h/(2\pi)$
k_B	Boltzmann constant, 1.381×10^{-23} J/K
m_e, m_h	electron and hole effective masses
m	density of states (DOS) mass
μ_n, μ_p	electron and hole mobility
n	electron concentration
n_{ieff}	effective intrinsic carrier concentration
p	hole concentration
ψ	electrostatic potential
q	elementary charge, 1.602×10^{-19} C
τ_n, τ_p	carrier lifetimes
v_n, v_p	drift carrier velocities

List of Figures

List of Tables

Summary

Technology Computer-Aided Design (TCAD) is indicated by the International Technology Roadmap for Semiconductors (ITRS) as one of the enabling methodologies that can support advance of technology progress at the remarkable pace of Moore's Law, by reducing development cycle times and costs in semiconductor industry. Several issues classified by the ITRS as difficult TCAD challenges can be seen as different implications of the same general trend, i.e. increasing problem dimensionality. In fact, technology scaling increasingly emphasizes complexity and non-ideality of the electrical behavior of semiconductor devices and boosts interest on alternatives to the conventional planar MOSFET architecture. A three-dimensional representation is mandatory to properly describe such devices: as a result, 3D simulations become a crucial need for everyday tasks. The outlined scenario highlights the need for meshing tools able to represent complex 3D geometries in an accurate yet efficient way, resolving all critical features of the device structure without unacceptable drawbacks in terms of grid size. Automated gridding procedures are also desirable in process and device simulations to provide a suitable mesh adaptation to geometry or solution changes while avoiding artifacts or spurious effects.

Predictive potentialities of TCAD also depend on its contribution to assessment and minimization of the impact of process variations, which get increasingly critical with device shrinking into the deca-nanometer range. Phenomena such as line-edge roughness (LER) and random dopant fluctuations (RD) broaden the device parameter distributions, thus requiring statistical treatment. This results in computationally challenging 4D problems, where the additional dimension is the size of

the considered ensemble.

The aim of this thesis is to present multi-disciplinary approaches to handle this increasing problem dimensionality in a numerical simulation perspective. In particular, the topic of adaptive meshing is tackled in a multiresolution framework which allows for an effective tracking of physical phenomena within two- and three-dimensional domains during quasi-stationary and transient simulations. The further dimensionality increase due to variability in extremely scaled devices is considered with reference to line-edge roughness and random dopant fluctuation issues. Statistical approaches to predict the impact of variability at an affordable computational expense are proposed. Such techniques are then applied to address feasibility of the FinFET architecture as an alternative to conventional CMOS technology for mainstream applications in sub-45 nm nodes.

The thesis is organized in five parts.

- Part I is a brief introduction to the parallel evolution of technology and TCAD simulations, where the role of computer-aided design and the increasing dimension of involved problems are highlighted.

- In Part II, some of the main challenges for TCAD to successfully deal with such problems are described, after illustrating the most common models used for semiconductor device simulation and the increasing complexity needed to describe aggressively scaled technologies (Chapter 1). In particular, problem discretization issues are discussed in Chapter 2, where important mesh requirements for standard TCAD solvers are also described and conventional error detection approaches for mesh adaptation are introduced. The second considered TCAD challenge, i.e. variability estimation, is analyzed in Chapter 3, describing causes as well as modeling and characterization techniques available in literature for LER and RD.

- Approaches proposed in this thesis for tackling the two outlined TCAD issues are presented in Part III. The topic of adaptive

meshing for semiconductor device simulation is addressed in Chapter 4, presenting a new technique, based on mathematical tools and algorithms from the fields of multiresolution analysis and signal processing. After providing the needed theoretical framework, the proposed approach is first introduced within a 2D setting; the extension to three-dimensional domains is then described, highlighting issues and solutions connected to dimensionality increase. A full integration of the developed C++ software into conventional TCAD environments is provided. Chapter 5 describes the adopted approaches for variability estimation. Line-edge roughness is modeled through a Monte Carlo technique: ensembles of microscopically different devices are generated by a Matlab program according to a proper statistical description of LER. Correlation analysis and other techniques to improve efficiency/accuracy of mismatch evaluation are discussed.

- Part IV shows how the proposed approaches help TCAD yielding accurate physical insight and useful predictive results when dealing with multidimensional real-world applications. The Wavelet-based meshing technique is successfully applied in Chapter 6 to automatically generate and dynamically adapt computational grids for 2D and 3D devices including *p*-*n* diodes, MOSFET drivers with complicated geometries and FinFETs. Combining statistical simulations with experimental data, potentialities and shortcomings of the latter architecture are analyzed in Chapter 7. Different process options, such as resist-defined and spacer-defined fin patterning as well as junction doping, are taken into account to evaluate feasibility of FinFET technology for mainstream applications (e.g. SRAM) in future generation integrated circuits (ICs).

- Finally, conclusions and future perspectives of the work are presented in Part V.

Riassunto della tesi

La progressiva contrazione delle dimensioni dei dispositivi a semiconduttore ne rende sempre più complesso e non-ideale il comportamento elettrico, alimentando inoltre l'interesse verso architetture alternative alla tecnologia MOSFET planare. Strumenti TCAD per la simulazione di dispositivi elettronici avanzati sono fondamentali per l'analisi e lo sviluppo di nuove generazioni tecnologiche. D'altronde, la complessità della struttura e del funzionamento di tali dispositivi determina un progressivo aumento di dimensione dei problemi in esame, richiedendo sempre più spesso una modellizzazione tridimensionale di applicazioni del mondo reale. In particolare, il compromesso tra accuratezza e onere computazionale delle simulazioni dipende fortemente dalla discretizzazione del dominio. Inoltre, la dimensione del problema è ulteriormente aumentata dalle variazioni di processo, che diventano sempre più critiche in dispositivi deca-nanometrici. Fenomeni come rugosità geometriche (*line-edge roughness*, LER) e fluttuazioni casuali di drogaggio impongono la rappresentazione del singolo dispositivo come un insieme statistico di istanze microscopicamente differenti, dando luogo a difficili problemi quadri-dimensionali, in cui l'ulteriore dimensione è data dalla cardinalità dell'insieme considerato.

Questa tesi si propone di utilizzare strumenti multidisciplinari per sviluppare approcci che permettano di gestire la crescente dimensionalità dei problemi di simulazione numerica. In particolare, verranno investigati tecniche adattative per la generazione di griglie computazionali e metodi statistici per la stima di variabilità in dispositivi avanzati.

Il primo argomento verrà affrontato proponendo un nuovo metodo (*Wavelet-based Adaptive Method*, WAM) per il raffinamento adattativo ed automatico della discretizzazione di domini 2D e 3D. Il software implementato fa uso di tecniche multirisoluzione basate sulla trasformata Wavelet al fine di ottenere una stima di regolarità della soluzione. Ciò permette di concentrare la risoluzione della griglia nelle regioni del dispositivo dove si manifestano i fenomeni fisici rilevanti, seguendone dinamicamente l'evoluzione al variare delle condizioni al contorno e ga-

rantendo la qualità delle mesh prodotte. In particolare, le principali caratteristiche di WAM possono essere riassunte come segue.

- Il software consente di sollevare l'operatore dal difficoltoso onere di definire manualmente mesh adatte alla simulazione mediante volumi finiti di situazioni applicative del mondo reale: l'input richiesto è infatti una griglia uniforme e molto sparsa.

- Il carattere direzionale delle informazioni fornite dall'analisi Wavelet permette di raffinare in maniera anisotropica le porzioni di dominio che richiedono una risoluzione elevata. Particolari accorgimenti sono stati messi a punto per mantenere una buona selettività dell'algoritmo anche nel caso tridimensionale, garantendo così una notevole efficienza in termini di dimensioni della griglia.

- L'individuazione delle regioni sensibili sfrutta algoritmi di *signal processing* particolarmente efficienti.

- L'adattamento dinamico consente di gestire efficacemente simulazioni quasistazionarie e in regime transitorio, incluse situazioni numericamente delicate come moltiplicazione a valanga dei portatori e breakdown.

- Grazie alla natura semiregolare delle griglie generate da WAM, è stato possibile definire una procedura di controllo della qualità della mesh in grado di identificare e rimuovere automaticamente configurazioni sfavorevoli per il solutore.

L'integrazione di WAM in un ambiente TCAD standard ne consente l'utilizzo per la simulazione di strutture 2D e 3D. Le applicazioni illustrate in questa tesi includono diodi, driver MOSFET con geometrie articolate e dispositivi FinFET. Questi esempi mostrano l'efficacia e l'efficienza dell'algoritmo proposto rispetto a tecniche convenzionali note in letteratura, sia in termini di costo computazionale e proprietà di convergenza della simulazione, sia per l'accuratezza e l'assenza di artefatti numerici nelle caratteristiche I-V prodotte.

Il problema dell'ulteriore aumento di dimensionalità dovuto a variazioni di processo è stato affrontato con riferimento a due fenomeni che stanno acquisendo crescente importanza, quali il line-edge roughness (LER) e le fluttuazioni casuali di drogaggio. Questa attività si inserisce nell'ambito di una collaborazione con il centro di ricerca IMEC (BE), avviata durante un periodo di permanenza di sei mesi presso tale struttura. In particolare, in questa tesi sono descritti alcuni approcci statistici, che consentono di stimare la variabilità ad un costo computazionale accettabile. Con l'ausilio di tali strumenti, viene studiato l'impatto dei fenomeni citati su dispositivi FinFET, che costituiscono una promettente alternativa all'architettura CMOS planare. L'impiego di simulazioni TCAD 2D e 3D, in combinazione con dati sperimentali, ha permesso di valutare le prestazioni di matching della tecnologia FinFET, relativamente a singoli dispositivi e blocchi circuitali di base, come memorie statiche (SRAM), confrontando diverse opzioni di processo legate alla modalità di definizione della fin e ai profili di drogaggio.

In particolare, sono stati analizzati i contributi di mismatch dovuti alle rugosità della fin, del gate superiore e di quelli laterali, valutando la variabilità su insiemi statistici costituiti da numerose realizzazioni microscopicamente differenti. Queste simulazioni evidenziano un forte impatto del line-edge roughness al nodo tecnologico LSTP-32 nm, quando i dispositivi FinFET potrebbero cominciare ad essere impiegati su larga scala. Il contributo più critico risulta quello dovuto alle rugosità della fin, definita mediante il processo di fabbricazione RDF (*resist-defined fin patterning*) comunemente adottato, che non dà luogo ad alcuna correlazione tra la forma dei due bordi. Si mostrerà, infatti, come tali rugosità influenzino il comportamento elettrico del dispositivo prevalentemente variando lo spessore medio della fin nella regione di canale. Similmente, l'impatto delle rugosità dei gate, sebbene di entità minore, è principalmente legato alla variazione della lunghezza media dei rispettivi canali. Queste informazioni, risultanti da un'analisi di correlazione tra variabilità geometrica ed elettrica, possono essere sfruttate sia per ottenere stime di variabilità approssimate ad un costo computazionale estremamente ridotto, sia per la definizione di modelli compatti

utilizzabili ai livelli gerarchici superiori di simulazione TCAD. Diversi modelli statistici sono disponibili in letteratura per la descrizione del line-edge roughness; le simulazioni effettuate mostrano, però, come il contributo più significativo al mismatch sia dovuto alle basse frequenze spaziali della rugosità, ben rappresentate dal modello ad autocorrelazione gaussiana. Utilizzando per i parametri di questo modello i valori tipicamente estratti da misure sperimentali, si prevede che il LER possa condizionare sensibilmente il funzionamento di celle SRAM al nodo tecnologico esaminato. Le fluttuazioni casuali di drogaggio, simulate mediante un approccio perturbativo, appaiono invece meno critiche in corrispondenza dei range di concentrazioni normalmente impiegati per la realizzazione di dispositivi FinFET.

Due possibilità sono state esplorate per minimizzare l'impatto del fin-LER su tali dispositivi. La prima consiste nell'impiego di strutture multi-fin: ciò ha un effetto benefico sul matching dei parametri elettrici, in accordo con la legge di Pelgrom. La seconda opzione consiste nella definizione della fin mediante un processo di tipo *spacer-defined*: oltre ad aumentare la densità di integrazione, tale tecnica dà luogo ad una significativa correlazione tra i bordi della fin. Si prevede che questo possa determinare una notevole riduzione della variabilità elettrica. I dati sperimentali riguardanti celle SRAM composte da FinFET realizzati con tale tecnologia, però, rivelano, allo stato attuale, una marcata instabilità del processo di fabbricazione, che dovrebbe dunque essere perfezionato. I progettisti dovranno prestare, inoltre, particolare attenzione all'ottimizzazione dei profili di drogaggio, poiché le simulazioni effettuate indicano un accentuarsi dei problemi di variabilità in corrispondenza dell'aumento di concentrazione nelle estensioni e della definizione di giunzioni il più possibile brusche.

Combinando strumenti statistici con simulazioni TCAD, il lavoro svolto fornisce dunque indicazioni utili per lo sviluppo di applicazioni basate sull'architettura FinFET nelle prossime generazioni tecnologiche.

Acknowledgments

I would like to gratefully acknowledge all the people who guided, accompanied and supported me during my Ph.D. studies.

Many thanks to Prof. Guido Masetti, who gave me the opportunity to step into the world of research and continuously encouraged my walk.

This walk would not have led anywhere without the patient guide and experienced advise of Nicolò Speciale, who introduced and directed me on the fields of Wavelets and semiconductor device simulation.

Working in a team has been a wonderful and forming experience, not only from a scientific point of view, especially thanks to Luca De Marchi and Francesco Franzè.

Thanks to Marco Messina, Salvatore Caporale and Alessandro Palladini for their friendship and many valuable suggestions.

Words cannot properly express my gratitude to the faithful mate of all my studies, Nicola Testoni.

Finally, I would like to acknowledge Abhisek Dixit, Malgorzata Jurczak, Kristin De Meyer and the whole EMERALD team at the Inter-University Microelectronics Center (IMEC) for launching and supporting me in the exploration of FinFET devices and process variations during my stay-abroad period in Belgium and afterwards.

Part I

Introduction - *TCAD roadmap towards increasing problem size*

"Everything should be made
as simple as possible,
but not simpler."

A. Einstein

Technology progress trends

Modern semiconductor technology has been developed after important inventions and discoveries achieved between 1945 and 1970. Starting from the fabrication of the first bipolar junction transistor in the late 1940s, the technology gradually improved until, in the 1960s, it reached a sufficient level of maturity for the production of good quality gate oxides. This allowed for the metal-oxide-semiconductor field effect transistor (MOSFET) to be introduced and soon inserted into monolithic integrated circuits (ICs), thus giving birth to the CMOS technology era. In 1965, just a few years after the fabrication of the firs IC, Gordon Moore made his famous prediction that the number of transistors in an integrated circuit would double every year [1]. Updated in 1975 with a prospected density doubling rate of two years, the so-called "Moore's law" has been describing the evolution of the semiconductor industry with extraordinary precision so far.

The reason for this exponential increase of chip complexity over time mainly lies in the continuous shrinking of device geometry, known as *scaling*. Since 1992, the Semiconductor Industry Association (SIA) has been providing essential research and development guidelines on the key needs for technology scaling to keep up with the exceptional rate outlined by Moore's law. Initially elaborated on a national basis, such guidelines were periodically updated and gradually extended to include worldwide industry contributions, resulting in a document called the "International Technology Roadmap for Semiconductors" (ITRS), first published in 1998. The document contains a 15-year outlook on the major trends of the semiconductor industry and provides clear research targets as well as possible solutions to emerging requirements

and issues, including forecasts on materials and software.

Role of TCAD

The progress of technology is so fast, that the underlying scientific understanding has frequently proved to be inadequate, leaving wide room to empirical approaches. However, an accurate physical description is essential at various stages of IC design and fabrication as well as to support innovation. In particular, computer simulations turn out to be the only way to investigate physical phenomena which cannot be directly studied through practical measurements. The synergistic combination of modeling and simulation tools, known as technology computer-aided design (TCAD), helps with the critical analysis and detailed understanding at various levels, including

- system and circuit design
- device engineering
- process development
- integration into manufacturing.

In fact, computer simulations allow investigating potentialities and physical limitations of manufacturing processes as well as developing *behavioral models* at the transistor and circuit level of ICs [2]. This is essential to the development of new technology generations, characterized by an increasing design complexity. Beside providing a *deep insight*, especially for aggressively scaled devices, for which complex physical phenomena and small dimensions severely limit the descriptive capabilities of measurements, TCAD simulations exhibit a remarkable *predictive valence* upon calibration to proper experimental data [3, 4]. The generation of predictive models plays a crucial role in reducing development cycle times and costs in semiconductor industry. This role is highlighted by the 2005 edition of the ITRS [5], where TCAD is indicated as a crucial enabling methodology supporting technology progress.

However, several issues are presented in the ITRS as difficult TCAD challenges, that can be read as different symptoms of the same general trend, i.e. the **increasing problem dimensionality**. This comes as a consequence of scaling and has a twofold implication.

On the one hand, more and more *complex device modeling* is needed for computer simulations at the process and physical levels. This is due to geometry shrinking, which enhances the importance of a number of phenomena contributing to the device behavior; moreover, the introduction of new materials and architectures increasingly complicates the transistor structure. In addition, the difficult fabrication of very small features sizes brings about significant parametric variations.

On the other hand, *design complexity* is constantly enhanced by the increasing density of integration, which has led to a huge gap between physical simulation on the nanometer-scale and IC design on a millimeter-scale featuring complexities up to 10^9 components. This problem can only be tackled through a rigorous hierarchical approach to TCAD (see Fig. 1), in which process and device simulations provide informations for the development of compact models, suitable for circuit and system level analysis. These informations include in the first place accurate *SPICE-like parameters* resulting from a realistic investigation of the device electrical behavior. In the second place, *variations of SPICE-like parameters* must be carefully estimated to achieve acceptable model predictivity, including process yield evaluation.

The outlined dimensionality increase is evident in the historical evolution of TCAD, as described in the next Section.

Increasing problem dimensionality in TCAD evolution

The first steps in computer simulations were drawn the late 1960s and 1970s, when one-dimensional (**1D**) approaches were generally sufficient to deal with bipolar technology and early MOSFET devices: 1D charge transport phenomena were predominant in these large, usually n-channel structures characterized by junction depths in the range of

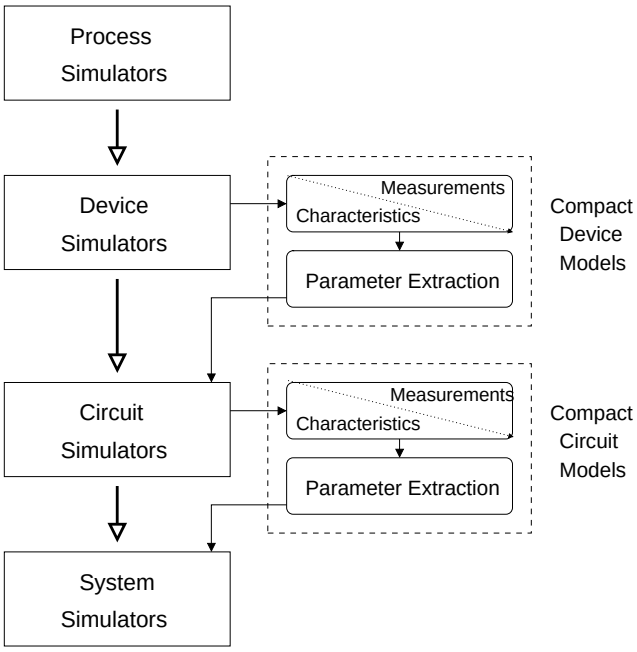


Figure 1: Hierarchical TCAD simulation flow.

fractions of micrometers and channel lengths of several micrometers. Extrapolation of **quasi-2D** doping distributions from sets of 1D profiles helped with process design optimization, although sheet resistances and minority carrier effects could not be predictively evaluated.

Starting from the 1980s, aggressive MOS scaling led to the very-large and ultra-large scale integration (VLSI and ULSI) eras based on CMOS technology. **Fully-2D** simulators soon became indispensable to model increasing process complexity and coupled physical effects, including local oxide isolation (LOCOS), dopant diffusion, subthreshold conduction, parasitic phenomena such as latchup and punchthrough.

The ever-shrinking transistor size led in the 1990s to a growing need for atomic-scale physics to correctly model the device behavior. Short/narrow channel effects and, later on, quantum effects such as gate leakage and carrier confinement required more and more sophisticated transport models, often amounting to several numerically stiff and highly non-linear coupled partial differential equations (PDEs). In addition, physical phenomena, including multi-device interactions,

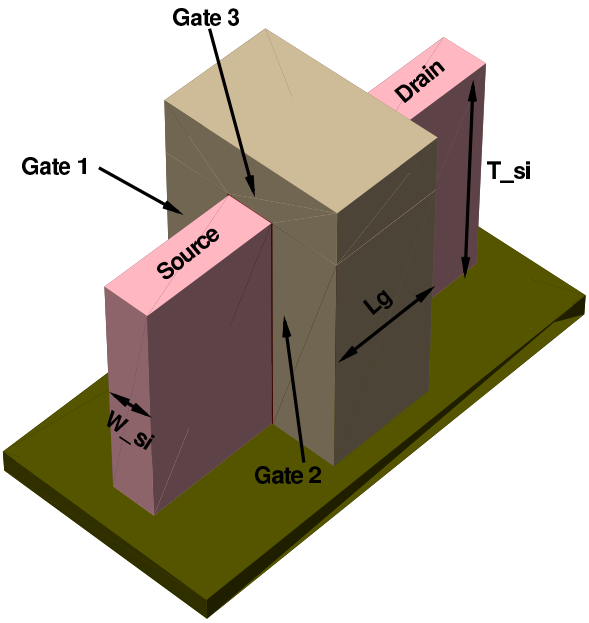


Figure 2: Schematic representation of a FinFET device.

interconnect and substrate parasitics, reliability issues such as electrostatic discharge (ESD), started to become inherently three-dimensional (**3D**).

The need for 3D simulation tools has become indispensable in the last years, when approaching scaling limits of bulk CMOS technology have boosted research on alternative, essentially three-dimensional architectures, e.g. Multiple-Gate devices (MuGFETs) [6, 7]. One of such devices is the FinFET schematically represented in Fig. 2. The silicon fin is surrounded by two sidewall gates and optionally by a top one, thus providing a better short-channel control. Charge transport is therefore a real 3D phenomenon, composed of two current flows parallel to the fin sidewalls and, optionally, an additional third one at the fin top.

The problem size in TCAD simulations is further increased by another major drawback of geometry scaling, i.e. enhanced process fluctuations. Although improved manufacturing tools have reduced absolute variability, relative variability in component geometries is becoming an increasing concern. Polysilicon/metal edge grains, photoresist edge

roughness, gate oxide thickness and permittivity non-uniformities are among the major sources of fluctuations. Moreover, charge transport in nanoscale devices is influenced by random distribution of dopant atoms in the channel. As a result, considerable fluctuations are seen in the device behavior, broadening the electrical parameters distribution and hence limiting IC performance. To take variability into account, each single device has to be represented by an entire distribution of structures with random geometry and doping fluctuations. Not only a 3D description of each device instance is mandatory in most applications, but the full simulation space is transformed into a four-dimensional (**4D**) one: the additional dimension is given by the size of the considered ensemble.

Motivations of this work

The dimensionality increase in TCAD problems and the enhanced complexity of the involved physical models give rise to the fundamental challenge of producing reasonably accurate and predictive results with an acceptable computational effort. In this thesis, two topics are addressed, which have a key role in meeting such a challenge, namely meshing and variability estimation.

The lowest hierarchy levels of TCAD include description of physical characteristics and behavior of the single device. This implies solving coupled PDEs which describe the evolution of either geometry and impurity distribution as a result of manufacturing process steps, or internal physical quantities in response to electrical boundary conditions (BCs). Solutions to such problems can only be sought numerically; thus, a proper discretization procedure is required. Mesh generation is the discrete representation of the considered domain: this operation has a crucial impact on convergence, accuracy and efficiency of the simulation. However, meshing "has become a major issue because device architectures are now essentially three-dimensional" (ITRS 2005 [5]), as also highlighted in the previous Section. Therefore, automatic grid generation and adaptation are highly desirable, both for improving the

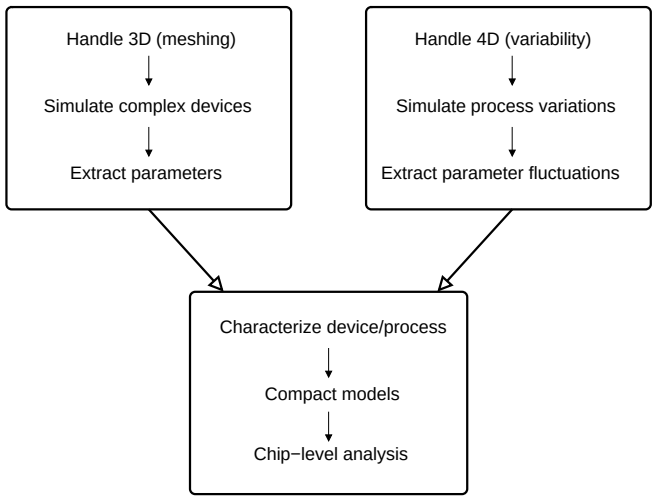


Figure 3: Handling 3D and 4D TCAD simulations enables circuit and system level analysis.

trade-off between computational complexity and solution accuracy, and for relieving TCAD users from a difficult and burdensome task. This motivates the investigation of adaptive meshing techniques for semiconductor device simulation.

In addition to increasing complexity of the device structure and behavior, dimension shrinking collides with the intrinsic discreteness of charge and matter and with difficulties and tolerances in the fabrication process. Random dopant fluctuations and line-edge roughness are two sources of major concern for future technology nodes. Techniques for evaluating variability at an affordable computational cost are sought, which sets the stage for the second analyzed topic.

Approaches described in this thesis can boost feasibility of challenging TCAD simulations and help with characterizing new processes and devices, such as FinFETs. As described in Fig. 3, this allows developing suitable circuit mismatch models, which can be used in predictive simulations of circuit and system-level performance of new technologies.

Part II

Problem setting - *Some TCAD roadblocks*

"That which is static and repetitive is boring.
That which is dynamic and random is confusing.
In between lies art."

J. A. Locke

In this Part, the topic of semiconductor device simulation is introduced, describing modeling and numerical aspects (Chapter 1). Issues related to the discretization of the simulation domain are highlighted, which result in stringent mesh requirements. Consequent difficulties in the mesh generation task represent a challenging TCAD "roadblock" that calls for automatic and adaptive techniques, as discussed in Chapter 2. The main existing approaches in this context are reviewed, which sets the stage for the Wavelet-based adaptive method described in Part III, Chapter 4.

Moreover, the background of parameter variations is outlined in Chapter 3, with particular reference to the impact on circuit mismatch. Line-edge roughness (LER) and random dopant fluctuations (RD) are presented as two major sources of short-range variability in aggressively scaled technologies. Predicting the impact of such effects on device and circuit matching performance is the second TCAD "roadblock" which will be addressed in the thesis, starting from the statistical simulation approach described in Part III, Chapter 5.

Chapter 1

Semiconductor device models

The behavior of real semiconductor devices can be described by partial differential equations which model electrostatic and charge transport phenomena. The simplest PDE system is the drift-diffusion model (DD), widely used in the simulation of conventional devices. However, aggressively scaled and non-conventional transistor structures are poorly described by this model. For example, carrier transport in such devices is strongly conditioned by thermal phenomena, especially in the saturation regime. A more sophisticated physical description which includes similar effects accounting for energy transport of the carriers is provided by the hydrodynamic model (HD). Both DD and HD are derived from a classical representation of the device behavior, but carrier confinement and tunneling phenomena in nanoscale structures can only be accounted for by quantum mechanics.

The quick panoramic provided in this Chapter aims at introducing those models that will be used in device simulations presented in this thesis. The increasing complexity due to technology scaling will be highlighted. An explanation of the adopted symbology in provided in the List of Symbols.

1.1 Drift-diffusion model

In this model, the Poisson equation, which describes the behavior of the electrostatic potential ψ, is directly coupled to the continuity equations

for electrons and holes and to the expression of current densities $\vec{J}_n$, $\vec{J}_p$ as the sum of a drift term, associated to the electric field, and a diffusive one due to concentration gradients:

$$\begin{aligned} \nabla \cdot (\epsilon \nabla \psi) &= q(n - p - C) && (1.1)\\ \nabla \cdot \vec{J}_n - q\frac{\partial n}{\partial t} &= q \cdot R(\psi, n, p) && (1.2)\\ \nabla \cdot \vec{J}_p + q\frac{\partial p}{\partial t} &= -q \cdot R(\psi, n, p) && (1.3)\\ \vec{J}_n &= -q \cdot (\mu_n \cdot n \nabla \psi - D_n \nabla n) && (1.4)\\ \vec{J}_p &= -q \cdot (\mu_p \cdot p \nabla \psi + D_p \nabla p) && (1.5) \end{aligned}$$

The thermal diffusion coefficients in (1.4) and (1.5) are given by Einstein's relations:

$$D_n = \frac{k_B T}{q}\mu_n \quad , \quad D_p = \frac{k_B T}{q}\mu_p \qquad (1.6)$$

The system unknowns are ψ, n and p, even if different rearrangements of the equations were presented (see [8] for a review on this topic). In (1.2) and (1.3), the terms containing time derivatives vanish under quasi-stationary conditions.

1.1.1 Generation/recombination and mobility models

These equations must be combined with suitable models for generation and recombination phenomena as well as carrier mobility.

- *Recombination*
  Contributions to $R(\psi, n, p)$ due to carrier interaction with the lattice are normally modeled through the Shockley-Read-Hall recombination rate:

  $$R^{SRH} = \frac{n \cdot p - n_{ieff}^2}{\tau_p \cdot (n + n_1) + \tau_n \cdot (p + p_1)} \qquad (1.7)$$

  where n_1 and p_1 are approximately equal to the effective intrinsic density if the defect energy level is close to the intrinsic level. n_{ieff} is also influenced by band gap narrowing effects.

- *Avalanche generation*
  Strong electric fields in wide space charge regions give rise to impact ionization phenomena, which can lead to device breakdown. In such conditions, an avalanche generation rate

$$G_{imp} = \alpha_n n v_n + \alpha_p p v_p \tag{1.8}$$

  contributes to the term $R(\psi, n, p)$ in (1.2) and (1.3). Several models are available for the ionization coefficients $\alpha_{n,p}$ (see [9]); one of the most commonly used is the Van Overstraeten - de Man model.

- *Mobility models*
  The main reason why such a simple scheme as the DD is still widely applied in device simulation is because it can be flexibly adapted to the considered problem through mobility calibration. A large variety of models have been developed, which describe mobility dependency on material properties and operating conditions. Different mobility contributions can be combined according to Mathiessen's rule:

$$\frac{1}{\mu} = \sum_i \frac{1}{\mu_i} \tag{1.9}$$

  Here, three models are reported, which have been used in device simulations described in Part IV. The reader is referred to [9] for a detailed explanation of model parameters.

  – The Masetti model [10] accounts for doping dependence of mobility, describing degradation effects due to impurity scattering:

$$\mu_{mas} = \mu_{min1} \cdot e^{-\frac{P_c}{N_i}} + \frac{\mu_{const} - \mu_{min2}}{1 + \left(\frac{N_i}{C_r}\right)^{\alpha}} - \frac{\mu_1}{1 + \left(\frac{C_s}{N_i}\right)^{\beta}} \tag{1.10}$$

  – the Lombardi model [11] describes surface contributions to mobility as affected by acoustic phonon scattering (μ_{ac}) and surface roughness (μ_{sr}):

$$\mu_{ac} = \frac{B}{F_t} + \frac{\mathcal{C} \cdot \left(\frac{N_i}{N_0}\right)^{\lambda}}{F_t^{\frac{1}{3}} \cdot \left(\frac{T}{T_0}\right)^{k}} \tag{1.11}$$

$$\mu_{sr} = \left(\frac{\left(\frac{F_t}{F_{ref}}\right)^{A^*}}{\delta} + \frac{F_t^3}{\eta} \right)^{-1} \tag{1.12}$$

where F_t is the transversal electric field. These contributions are combined with the bulk mobility according to Mathiessen's rule.

– High field mobility degradation due to carrier velocity saturation effects is introduced by the Canali model [12] :

$$\mu_{can}(F) = \frac{\mu_{low}}{\left(1 + \left(\frac{\mu_{low} \cdot F}{v_{sat}}\right)^{\beta}\right)^{\frac{1}{\beta}}} \tag{1.13}$$

where the exponent β and the saturation velocity v_{sat} are temperature-dependent

$$\beta = \beta_0 \left(\frac{T}{T_0}\right)^{\beta_{exp}} \tag{1.14}$$

$$v_{sat} = v_{sat,0} \left(\frac{T_0}{T}\right)^{v_{sat,exp}} \tag{1.15}$$

μ_{low} is the low field mobility, influenced by previously described contributions. The driving force F can be taken as the component of the electric field parallel to the current flow or the gradient of electron/hole quasi-Fermi potentials.

1.1.2 Boundary conditions

Boundary conditions are required to provide unicity to the solution of the PDE system. In particular, Dirichlet BCs are applied at ohmic contacts and homogeneous Neumann conditions at isolating boundaries.

- *Dirichlet boundary conditions*
  The contact potential ψ_c for ideal ohmic contacts is calculated as:

$$\psi_c = \psi_d + \frac{k_B T}{q} \cdot \operatorname{asinh}\left(\frac{C}{2n_{ieff}}\right) \tag{1.16}$$

where ψ_d is the applied external potential. Dirichlet conditions for electrons and holes are obtained by considering vanishing space

charge and thermal equilibrium at ohmic contacts, which leads to:

$$n = \frac{\sqrt{C^2 + 4 \cdot n_{ieff}^2} + C}{2} \tag{1.17}$$

$$p = \frac{\sqrt{C^2 + 4 \cdot n_{ieff}^2} - C}{2} \tag{1.18}$$

- *Neumann boundary conditions*
  Homogeneous Neumann conditions for potential, electrons and holes, respectively, are expressed as follows:

$$\frac{\partial \psi}{\partial \vec{n}} = 0 \tag{1.19}$$

$$\vec{J}_n \cdot \vec{n} = 0 \tag{1.20}$$

$$\vec{J}_p \cdot \vec{n} = 0 \tag{1.21}$$

  Here, $\vec{n}$ denotes the unit vector normal to the considered domain boundary.

- *Interface boundary conditions*
  Application of Gauss's law at interfaces between different materials leads to the following conditions:

$$\epsilon_1 \cdot \left.\frac{\partial \psi}{\partial \vec{n}}\right|_1 - \epsilon_2 \cdot \left.\frac{\partial \psi}{\partial \vec{n}}\right|_2 = Q_{int} \tag{1.22}$$

  where the subscripts 1 and 2 refer to the two considered materials and Q_{int} accounts for possible interface charges.

1.2 Hydrodynamic model

The hydrodynamic model couples the basic semiconductor equations (Poisson equation (1.1) and continuity equations (1.2), (1.3)) with the following energy balance equations for electrons, holes and the lattice:

$$\frac{\partial W_n}{\partial t} + \nabla \cdot \vec{S_n} = \vec{J_n} \cdot \nabla E_C + \left.\frac{dW_n}{dt}\right|_{coll} \tag{1.23}$$

$$\frac{\partial W_p}{\partial t} + \nabla \cdot \vec{S_p} = \vec{J_p} \cdot \nabla E_V + \left.\frac{dW_p}{dt}\right|_{coll} \tag{1.24}$$

$$\frac{\partial W_L}{\partial t} + \nabla \cdot \vec{S_L} = \left.\frac{dW_L}{dt}\right|_{coll} \tag{1.25}$$

Energy fluxes are expressed as:

$$\vec{S_n} = -\frac{5r_n}{2}\left[\frac{k_B T_n}{q}\vec{J_n} + f_n^{hf}\left(\frac{k_B^2}{q} n \mu_n T_n\right)\nabla T_n\right] \tag{1.26}$$

$$\vec{S_p} = -\frac{5r_p}{2}\left[-\frac{k_B T_p}{q}\vec{J_p} + f_p^{hf}\left(\frac{k_B^2}{q} p \mu_p T_p\right)\nabla T_p\right] \tag{1.27}$$

$$\vec{S_L} = -\kappa_L \nabla T \tag{1.28}$$

while energy densities are given by:

$$W_n = n\left(\frac{3}{2}k_B T_n\right) \tag{1.29}$$

$$W_p = p\left(\frac{3}{2}k_B T_p\right) \tag{1.30}$$

$$W_L = c_L T \tag{1.31}$$

In the hydrodynamic case, current densities are defined as a sum of four contributions:

$$\vec{J_n} = q\mu_n\left[n\nabla E_C + k_B T_n \nabla n + f_n^{td} k_B n \nabla T_n - W_n \nabla(\ln m_e)\right] \tag{1.32}$$

$$\vec{J_p} = q\mu_p\left[p\nabla E_V - k_B T_p \nabla p - f_p^{td} k_B p \nabla T_p - W_p \nabla(\ln m_h)\right] \tag{1.33}$$

The first term accounts for spatial variations of electrostatic potential, electron affinity and the band gap. The three remaining terms take into account the contributions due to the gradient of concentrations and carrier temperature, and the spatial variation of the effective masses, respectively. The values of model parameters and the expressions of collision terms (subscript *coll*) in the above equations can be found in [9].

The hydrodynamic model requires the solution of three additional PDEs, i.e. (1.23) ÷ (1.25), with respect to the DD scheme; moreover, more complicated expressions hold for current densities. However, this model allows for a more accurate estimation of velocity overshoot and impact ionization effects in deep submicron (DSM) devices.

1.3 Modeling quantum effects

In aggressively scaled devices, the wave nature of electrons and holes can no longer be neglected. The most rigorous approach to account for quantum effects is to couple previously described device equations with the Schrödinger equation. Assuming a single quantization direction z, the additional 1D PDE to be solved reads:

$$\left[-\frac{\partial}{\partial z}\frac{\hbar^2}{2m_{z,\nu}(z)}\frac{\partial}{\partial z}+E_C(z)\right]\Psi_{j,\nu}(z)=E_{j,\nu}\Psi_{j,\nu}(z) \tag{1.34}$$

where ν labels the considered band valley and $m_{z,\nu}$ is the corresponding (position-dependent) effective mass component in the quantization direction. $\Psi_{j,\nu}$ and $E_{j,\nu}$ are the j-th eigenfunction and eigenenergy in valley ν, respectively. From the solution of equation (1.34), carrier density is computed as:

$$n(z)=\frac{k_BT(z)}{\pi\hbar^2}\sum_{j,\nu}\left|\Psi_{j,\nu}(z)\right|^2 m_{xy,\nu}(z)e^{\frac{E_F(z)-E_{j,\nu}}{k_BT(z)}} \tag{1.35}$$

$m_{xy,\nu}$ being the mass component perpendicular to the quantization direction. The conduction band profile $E_C(z)$ is directly linked to the electrostatic potential ψ provided by the Poisson equation (1.1), so a strong coupling exists between these PDEs. Moreover, a special purpose domain discretization with proper alignment to the quantization direction is required in the region where the Schrödinger equation is solved. Therefore, this approach is extremely expensive from the computational standpoint and prone to convergence problems.

An alternative solution for including quantization effects in a classical device model is to introduce an additional potential-like quantity Λ in the classical density formula:

$$n=N_Ce^{\frac{E_F-E_C-\Lambda}{k_BT}} \tag{1.36}$$

(a similar expression can be adopted for holes). Several models for Λ have been developed. The simplest one is the van Dort quantum correction model [13], which computes Λ as a function of the electric field

E_n normal to the semiconductor-insulator interface, thus accounting for quantization in MOSFET channels:

$$\Lambda = \frac{13}{9} k_{fit} \frac{2e^{-a^2(\vec{r})}}{1 + e^{-2a^2(\vec{r})}} \left(\frac{\epsilon}{4k_B T} \right)^{1/3} \cdot |E_n - E_{crit}|^{2/3} \tag{1.37}$$

(see [9] for model parameters). This model is much simpler and more efficient than the Schrödinger-Poisson scheme; however, it is only suited to MOSFET simulations and it does not give the correct density distribution in the channel, although terminal characteristics are well described.

A good compromise between the two just described approaches is provided by the density gradient approximation (DGA) [14, 15]. This model can be applied to several device structures and gives a reasonable description of both terminal characteristics and internal charge distribution, even in the presence of 2D and 3D quantization effects. In macroscopic terms, the DGA captures the non-locality of quantum mechanics to lowest-order by assuming the electron gas to be energetically sensitive to both the carrier density and its gradient. In this approach, Λ is computed for (1.36) by solving the following PDE:

$$\Lambda = -\frac{\gamma \hbar^2}{12m} \left[\nabla^2 \log n + \frac{1}{2} (\nabla \log n)^2 \right] = -\frac{\gamma \hbar^2}{6m} \frac{\nabla^2 \sqrt{n}}{\sqrt{n}} \tag{1.38}$$

where γ is a fitting parameter. Modified mobility formulas are also available to account for tunneling through semiconductor barriers.

For the sake of completeness, another approach is worth mentioning, which can be considered as a quantum correction. In this model, proposed by Ferry [16, 17], treating electrons and holes as wave packets with a certain space extension results in the definition of a non-local effective potential that replaces the classical one. However, the approach was proven to be equivalent to the DGA formalism by using a first-order expansion wherever the effective potential is a slowly varying function of position.

The outlined hierarchy of device models reflects the increasing challenge posed to TCAD by technology scaling. The discretization of both model equations and the analyzed domain is crucial for finding accurate numerical solutions, as described in Chapter 2.

Chapter 2

First TCAD issue: problem discretization

Two main approaches are commonly adopted for the discretization of PDE systems, namely the finite element method (FEM) [18] and the finite volume method (FVM) [8]. The first scheme is based on a variational formulation of the problem through the Gauss-Green law and the use of suitable test functions. This approach is mainly implemented in process simulators. Instead, the device equations described in Chapter 1 are discretized through the FVM in nearly all state-of-the-art solvers. One of the main advantages of this scheme is that it imposes the local conservation of fluxes, thus correctly modeling charge conservation inside the device. Prior to the discretization, each PDE is properly normalized for numerical stability [8].

2.1 Finite volume discretization

The FVM, or box integration method (BIM), integrates the PDEs over a set of test volumes covering the simulation domain. Device simulators normally require that these volumes coincide with the Voronoi regions of the points [19] (see Fig. 2.1). First, the Gaussian theorem is applied, resulting in equations with the form:

$$\nabla \cdot \vec{J} + R = 0 \tag{2.1}$$

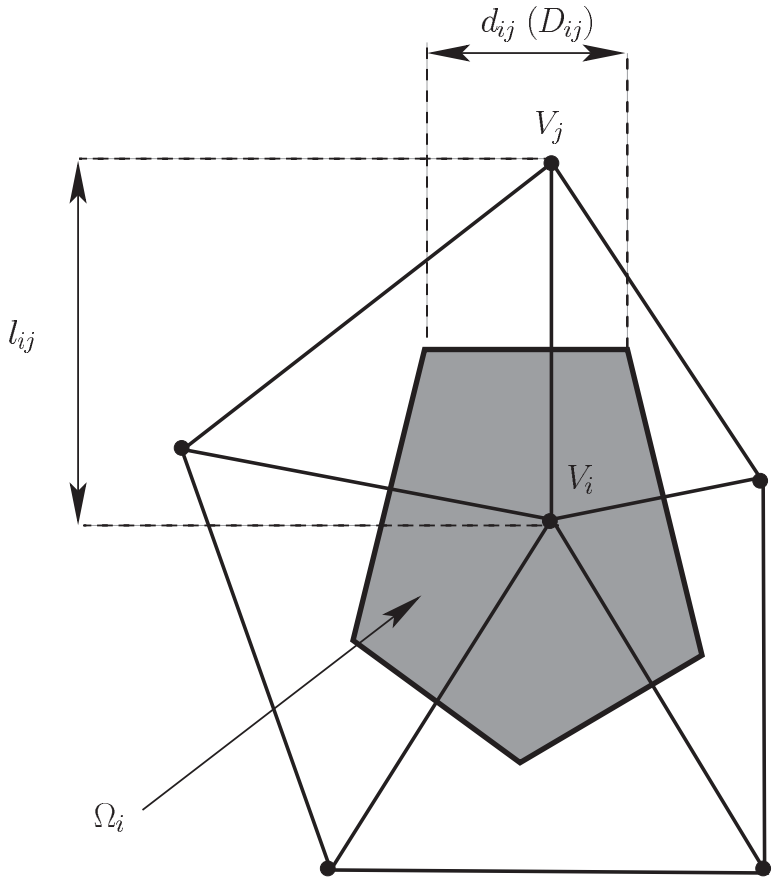


Figure 2.1: Voronoi tessellation of the domain. Ω_i is the Voronoi cell associated to mesh node V_i. l_{ij} is the length of the mesh edge connecting nodes V_i and V_j, while d_{ij} is the length of the Voronoi cell side normal to this edge (in 3D domains, this side is a facet whose area is D_{ij}).

Each PDE is then discretized to a first-order approximation:

$$\sum_{j \neq i} \kappa_{ij} \cdot J_{ij} + \mu(\Omega_i) \cdot R_i = 0 \tag{2.2}$$

In (2.2), κ_{ij} and $\mu(\Omega_i)$ are geometry-related terms whose values are given in Table 2.1 according to the domain dimensionality. Instead, Table 2.2 provides the expression of physical parameters J_{ij} and R_i as resulting from the discretization of Poisson and continuity equations (1.1)-(1.3). The Gummel iterative method [20] is typically adopted to solve the discretized system.

Dimension	κ_{ij}	$\mu(\Omega_i)$
1D	$1/l_{ij}$	box length
2D	d_{ij}/l_{ij}	box area
3D	D_{ij}/l_{ij}	box volume

Table 2.1: Values of geometry-related terms in eq. (2.2).

Equation	J_{ij}	R_i
Poisson (1.1)	$\epsilon(u_i - u_j)$	$-\rho_i$
Electron c. (1.2)	$\mu_n \left[n_i B(u_i - u_j) - n_j B(u_j - u_i)\right]$	$R_i - G_i + \frac{d}{dt} n_i$
Hole c. (1.3)	$\mu_p \left[p_j B(u_j - u_i) - p_i B(u_i - u_j)\right]$	$R_i - G_i + \frac{d}{dt} p_i$

Table 2.2: Expressions of physical parameters in eq. (2.2). $B = x/(e^x - 1)$ is the Bernoulli function, while u and ρ are normalized potential and charge density, respectively.

2.2 Domain discretization

The finite volume discretization of the device equations is based on a subdivision of the simulation domain into a set of control volumes associated to discrete grid nodes. The choice of the mesh (grid points and connectivity), and consequently the domain tessellation, has a crucial impact on convergence, accuracy and efficiency of the simulation. However, there is no general consensus about the definition of a "high quality" mesh. Geometrical features must certainly be taken into account both to comply with the requirements imposed by the discrete solution scheme and to improve convergence. Nevertheless, meshes cannot be designed only based on criteria such as aspect- or volume-ratio of the elements, as this may lead to excessively large mesh sizes or to degraded resolution. Instead, the properties of the problem to be solved need to be considered as well.

2.2.1 Mesh requirements

Some key features in the framework of mesh generation for TCAD simulation can be summarized as follows.

- *Delaunay conformity*
  All major device simulators based on the finite volume method require Delaunay-conform meshes [21]. This is because the Delaunay triangulation corresponds to the dual graph of the Voronoi tessellation defining control volumes. For a given set S of grid points in $\mathbb{R}^n$, the Delaunay triangulation is constructed such that no point of S lies inside the circum-sphere of any simplex (i.e. triangle in 2D, tetrahedron in 3D). Each cell of the dual Voronoi diagram is the region of all points that are closer to the associated grid node than to any other point in S. 2D Delaunay triangulations maximize the minimum angle of all mesh elements. Furthermore, boundary conformity is typically required in 3D, i.e. surface mesh elements should also be Delaunay.

- *Geometrical quality*
  Several quality indicators have been proposed (see for example [21, 22]), most of them related to properties of the single element, such as aspect-ratio measures, which estimate how close each cell is to a regular triangle (in 2D) or tetrahedron (in 3D). These criteria are particularly suited to finite element applications such as process simulations: FEM-based solvers are influenced by the shape of mesh elements, which determine the properties of the resulting discretization matrix. Instead, clear quality criteria for finite volume meshes are still lacking. However, it is well known that obtuse elements can affect FVM convergence. An element is obtuse when it does not contain the associated Voronoi center. Since the method is based on the computation of fluxes, obtuse elements are undesired because the flux between certain nodes is discretized using the area of Voronoi cell faces which are far from the mesh line connecting the two points. A 2D example is shown in Fig. 2.2. Non-obtuse triangulations can be guaranteed in two dimensions [23], while such a guarantee remains an open problem in 3D. Delaunay property and boundary conformity are the only clear requirements for 3D meshes.

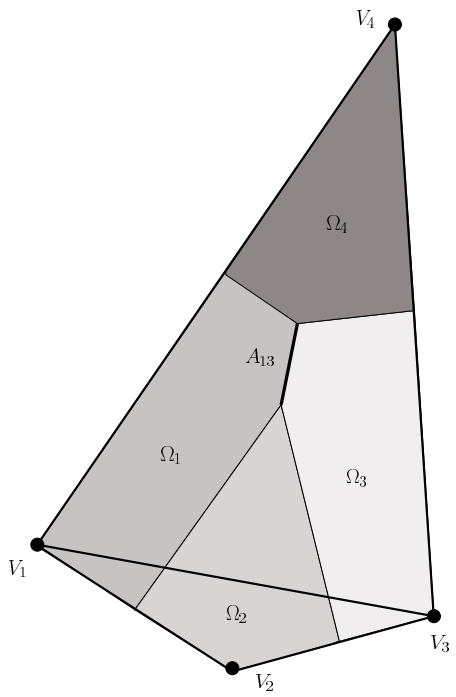


Figure 2.2: Example of adverse 2D Voronoi boxes due to obtuse angles. Fluxes between nodes V_1 an V_3 are discretized using area A_{13}, which is far from the mesh line $V_1 - V_3$.

- *Smoothness*
  In addition to the single element shape, global smoothness of the mesh is also important because rapid volume changes between adjacent cells can translate into large truncation errors.

- *Structural alignment*
  Structural alignment of the grid is essential to accuracy of charge transport computation [24]: since contact currents represent one of the most important informations provided by the simulation, mesh elements should be properly flux-aligned for a correct integration of the PDE system.

- *Non-uniformity*
  Designing uniform meshes would be the simplest solution for domain discretization and it would comply with most of the previ-

ously mentioned requirements. However, an optimal representation of the simulation domain must be sought both in terms of solution accuracy and computational efficiency. Semiconductor device structures generally include very thin layers; layer behaviors are also typical of the solutions produced by physical simulations, due to the singularly perturbed character of the considered PDEs [25]. Grid points should be placed in such a way as to resolve all geometric irregularities and small spatial features as well as accurately approximate any physical quantity of interest, e.g. potential and concentrations. A uniform approach would lead to overwhelmingly large grid sizes. Therefore, suitable non-uniform meshes are required.

- *Anisotropy*
  The strong directional dependency of the problems under investigation calls for anisotropic mesh densities. Such a need has become even more crucial nowadays since new device architectures are essentially three-dimensional. This is in contrast with typical criteria on geometrical quality, thus demanding a difficult trade-off between simulation stability, solution accuracy and computational effort. Although undesired in FEM applications, high aspect-ratio elements that are correctly flux-aligned have been found to provide excellent results in FVM simulations, while keeping the mesh size as small as possible.

- *Unstructured meshes*
  Process simulations usually involve non-planar surfaces and interfaces [26]. The resulting irregular geometries are often the input structures for device simulations. Unstructured meshes are needed to properly handle such situations.

- *Adaptivity*
  TCAD simulations usually involve dynamically changing conditions and hence evolving solutions, including moving geometries and variation of internal quantities. Tackling this problem through a static approach is highly unfavorable. In fact, the operator

should be able to predict interesting evolutions and generate a fixed mesh with the proper resolution in all domain regions where important phenomena could take place during the simulation. Beside requiring extraordinary expertise, such a task would result in large mesh sizes and hence high computational cost. The desirable alternative is a dynamical approach, meaning adaptive meshing of solution changes.

- *Automation*
  It follows from the considerations reported above that a properly-designed mesh requires a deep insight of:
  - the peculiar geometrical features of the structure under investigation;
  - the distribution of internal physical quantities of interest;
  - the bounds on mesh quality for solver stability;
  - the link between resolution in critical domain areas and simulation accuracy;
  - the bounds on computational resources.

  Such a complex set of conflicting requirements makes mesh generation an extremely challenging task, generally resulting in a time-consuming trial-and-error loop accomplished by highly experienced users. Due to the increasing complexity of 3D device structures and physics involved, hand-generation of computational meshes is becoming totally impracticable. Therefore, it is clear how automatic meshing tools represent a key need in a modern TCAD environment. On the other hand, the outlined scenario also provides some hints on the difficulties contrasting the development of such tools.

2.3 Adaptive meshing

The finite element method is the most suitable technique to deal with moving boundaries as typical of process simulations. Grid adaptation

in the FEM arena is broadly studied in literature and quite well assessed (see for example [27, 28]). The development of adaptive techniques for device simulations through finite volumes has proven to be relatively more challenging, mainly due to difficulties emerged in:

1. identifying the most suitable physical quantities to be surveyed;
2. preventing grid changes from producing numerical artifacts and spurious solutions in terminal characteristics;
3. ensuring stability of the FVM.

2.3.1 Review of the most common approaches to error detection

A glimpse at the state of the art concerning the first of the difficulties mentioned above is provided in this Section. The quantities used to detect domain regions to be refined can be roughly distinguished into two groups: *error indicators* and *error estimators*. Error indicators inform on the location of the discretization error. They are generally connected to gradients, curvatures or regularity properties of the physical solution. Usually, the magnitude of error indicators does not provide a direct estimation of the magnitude of the solution error. Instead, the magnitude of error estimators can be used to bind the global accuracy of the solution, therefore providing a stopping criterion for the refinement. It is worth to notice that in order to be useful for mesh adaptation, error estimators should also be able to *indicate the localization* of the error [29], thus blurring the distinction with the former group.

Error indicators/estimators can be either *a-priori* or *a-posteriori*; due to the adaptivity requirement described above, just a-posteriori approaches will be considered in this thesis. Here, by "a-posteriori" it is meant that a preliminary solution must be computed first, which is then analyzed to locate domain areas requiring high spatial resolution. In contrast, "a-priori" approaches to mesh refinement would

Criteria for mesh refinement	**Explicit** *Based on properties of the solution, measuring jumps of relevant quantities*	**Implicit** *Based on the solution of auxiliary problems*
Error estimators	LTE (2.5)-(2.6) estimated through "analytical" formulas [8, 30, 31]	Richardson extrapolation error (2.3) [31]
	ZZ (2.12) [32]	Bank and Weiser formalism (2.4) [31]
	L^2 norm of flux density error vector (2.9) [33]	
	Residual based (discontinuity of gradients across element edges) (2.11) [32]	
Error indicators	Residual local dissipation rate error (2.10) [25]	Dissipation rate error estimation (2.10) based on local Dirichlet problems [25]
	Local curvature of electrostatic or quasi Fermi potentials (2.7) [33]	Hessian matrix of approximation error (2.13) [34]
	Non-uniformity of current density (2.8) [33]	

Table 2.3: Criteria for mesh refinement adopted in semiconductor device simulation.

exploit informations that are available *before* any simulation is performed. Such informations would then imply an a-priori (and hence strictly problem-dependent) knowledge of some solution properties. A further distinction can be made between *explicit* or *implicit* error localization techniques: the explicit ones do not require the solution of additional problems as opposed to implicit localizations, which are consequently more onerous. In Table 2.3 a classification of mesh refinement criteria for device simulation is proposed, based on the above distinctions.

Several criteria listed in Table 2.3 have been reviewed in [31], where

the **Richardson extrapolation error**

$$e_f(x_i, y_i) = f_{h/2}(x_i, y_i) - f_h(x_i, y_i) \tag{2.3}$$

calculated comparing the current solution f_h with one computed on a uniformly refined grid ($f_{h/2}$), is indicated as one of the most accurate implicit estimators, although extremely costly. A still accurate but more efficient alternative is shown to be the implicit estimator based on the formalism proposed by **Bank and Weiser** (BW): in this approach, the error $e = \sum_i \alpha_i v_i$ for each element τ is computed through a FEM formulation with higher order basis functions v_i. The formulation for the 2D Poisson equation has the form:

$$a(e, v_i)_\tau = (-\rho, v_i)_\tau + \langle S, v_i \rangle_{\Omega_B \cap \tau} + \gamma \langle J, v_i \rangle_{\Omega_l \cap \tau} \tag{2.4}$$

where the first two terms are associated to the integration of the left and right hand side of the Poisson equation, respectively, while the remaining two integrals enforce Neumann boundary conditions on each sub-problem. In particular, $\langle S, v_i \rangle_{\Omega_B \cap \tau}$ is associated to surface charges on the device boundary Ω_B and $\gamma \langle J, v_i \rangle_{\Omega_l \cap \tau}$ accounts for jumps in the electric flux along element edges which are internal to the problem domain. Applying the BW estimator to continuity or energy balance equations is more challenging.

Explicit error estimators are also considered in [31], such as the one based on the computation of **local truncation errors** (LTEs) due to discretization schemes. Local truncation errors are also used in [8]. In [30], LTEs associated to the drift-diffusion scheme are re-derived more accurately, resulting in:

$$LTE_\psi = \frac{J_n}{8q\mu_n E^2} h^2 \frac{\partial^2 \psi}{\partial x^2} \tag{2.5}$$

$$\begin{aligned} LTE_n = {} & (h^2 - k^2)\frac{\partial^2 J_{nx}}{\partial x^2} + (p^2 - r^2)\frac{\partial^2 J_{ny}}{\partial y^2} \\ & + (h^3 + k^3)\frac{\partial^3 J_{nx}}{\partial x^3} + (p^3 + r^3)\frac{\partial^3 J_{ny}}{\partial y^3} \end{aligned} \tag{2.6}$$

for equations (1.1) and (1.2), respectively. h, k, p and r are mesh spacings as in Fig. 2.3. A drawback of LTEs is that they are strictly

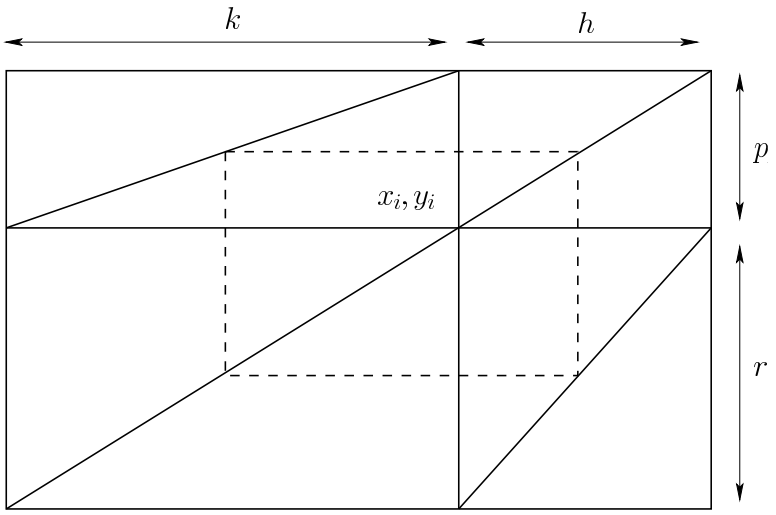


Figure 2.3: Reference mesh structure for the computation of LTEs (2.5), (2.6).

dependent on the particular model used for the simulation.

In [33] other explicit error indicators are compared, based on measuring the **local curvature** of electrostatic potential

$$\beta = \frac{\psi''}{\sqrt{1+(\psi')^2}}\frac{(dx)^2}{2} \tag{2.7}$$

(or quasi Fermi potentials) or **local variations in the current density**

$$\gamma_r = \frac{|J_1 - J_2|}{\max(|J_1|, |J_2|)} \tag{2.8}$$

where J_1 and J_2 are the current densities along two parallel edges of a cell in a box grid discretization. The error estimator already proposed in [28], based on the calculation of **flux densities** F (electric field or current density) is also considered and found to be superior to the previous two. This estimator is computed as

$$\eta_i = \sqrt{\int_\tau |F - F^*|^2 d\Omega} \tag{2.9}$$

where F^* is the expected "true" flux density, approximated by piecewise linear interpolation on each element τ.

In [25] the authors propose an adaptation scheme driven explicitly or implicitly by the variations of a physical quantity known as **dissipation rate** D of the system. This is a weighted sum of the device terminal currents, derived for the drift-diffusion model:

$$D = \int_\Omega \mu_n n |\nabla \phi_n|^2 d\Omega + \int_\Omega \mu_p p |\nabla \phi_p|^2 d\Omega + k_B T \int_\Omega R \ln\left(\frac{n \cdot p}{n_{ieff} \cdot p_{ieff}}\right) d\Omega \tag{2.10}$$

In the implicit approach, the dissipation rate associated to the computed solution at each element is compared to an estimation obtained by solving Dirichlet problems on locally refined grids. The explicit alternative consists in measuring jumps of D across element boundaries and is seen to provide similar accuracy at a lower computational cost.

Two explicit error estimators are considered in [32]. The first one is a residual based estimator, formerly proposed in [35]:

$$\eta_k = h_k \left(\sum_{E_{k,int}} \|J_{E,n}(u_h)\|_E^2 + \sum_{E_k} \|J_{E,t}(u_h)\|_E^2 \right) \tag{2.11}$$

It derives from the observation that a piecewise affine interpolation of the solution function fulfills the Laplace equation in the interior of mesh elements, while local errors arise from discontinuities of the tangential ($J_{E,t}$) and normal ($J_{E,n}$) components of the function gradient at element boundaries E_k. In (2.11), $E_{k,int}$ are interior mesh edges and h_k is a characteristic length of the k-th element. The second considered quantity is the **Zienkiewicz-Zhu** (ZZ) error estimator, which measures the difference between the piecewise constant numerical solution and a smoothed version obtained through a piecewise affine interpolation on each element τ_k:

$$ZZ_k = \sum_i U_i^2 + \sum_{i \neq j} U_i U_j \tag{2.12}$$

The proposed adaption strategies are validated for Laplace equations only: applicability to more sophisticated problems is not discussed.

Finally, in [34] the adaptation is driven by the **Hessian matrix of an error** e_h computed hierarchically enriching the finite element

approximation space V_{ih} to which the solution u_{ih} of

$$a(u_{ih}, v_h) = \langle f, v_h \rangle \ , \quad \forall v_h \in V_{ih} \subset V \tag{2.13}$$

belongs. Though applied to device simulations, this implicit approach is only suitable for FEM solvers.

2.3.2 Refinement-Solver interaction

Once regions with a poor resolution have been identified through some error indicator/estimator, two problems arise: (a) how to perform the refinement and to re-mesh the domain, and (b) how to redistribute the solution on the newly inserted nodes.

Since most of the detection approaches mentioned above are element-based, the refinement is usually performed element-wise in such a way as to equidistribute local errors over the domain. An efficient refinement should follow the anisotropic features of the solution; however, *directional informations* are generally not provided by the discussed error estimators and indicators. To overcome this drawback, some authors [29, 31, 36] introduce auxiliary sources of directional informations, although the computation of these new quantities implies additional overhead; moreover, the effectiveness of directing the refinement through a quantity that may be poorly connected with the discretization error is doubtful. The alternative is to refine each selected element isotropically, with clear drawbacks in terms of mesh size. It is also worth to notice that the equidistribution of local error estimators seems to be inappropriate in semiconductor device problems [37, 38], because of the layer behavior of the solution. In fact, layer regions will show high discretization errors up to extremely small grid resolutions: trying to reduce them by redistribution over the whole domain is likely to result in highly redundant refinements. Domain re-meshing is also crucial: in particular, it should conform to features described in Sec. 2.2.

As for problem (b), the solution can be redistributed either through a naive linear interpolation (which is quite "dangerous" as interpolation errors will be relevant in the most critical domain regions) or by means of more onerous procedures, such as solution recomputation on the

new nodes with local Dirichlet problems or homotopy techniques [25]. A rule of thumb for an effective combination of refinement and solution recomputation is not to insert or move too many nodes (maximum 10% new nodes [8]) at each adaptation step. This approach allows to simplify the remeshing procedure as the Delaunization can be recomputed only locally, and obviously reduces the number of interpolations.

Choices to cope with topics (a) and (b) above are particularly relevant to stability of the FVM and to smoothness of the curves produced by quasi-stationary or time-varying simulations because of the coupling between the FV solver and the adaptation process.

Chapter 3

Second TCAD issue: variability estimation

Designing a proper mesh according to the considerations illustrated in Chapter 2 is instrumental to any TCAD simulation. Today, one of the fundamental roles of computer-aided simulations is to bridge the gap between process development and circuit design by estimating the effects of statistical variations on yield and electrical performance. These statistical variations are inherent to the IC manufacturing process and are usually classified into *global* and *local* components.

The first group includes all parameter fluctuations occurring between different dice (*inter-die*), be them placed on the same wafer or on different wafers, belonging to the same lot or to different lots. Global variations are caused by process gradients across the wafer/batch due to equipment variations and spatial drifts, such as non-idealities in photomasks and optical lenses, or non-uniformities of the photoresist and oxide thickness. This results in *systematic* parameter fluctuations for identically designed groups of devices and hence compensation techniques can be applied to minimize their impact on electrical performance.

Variations affecting two components within the same chip (*intra-die*) belong to the second group. Historically, local variability has been small with respect to the global component; however, in modern technologies this is no longer the case because of geometry scaling [39].

In fact, local fluctuations are related to the discrete nature of charge and matter and hence gain importance as the involved distances are becoming comparable to the device dimensions. Small feature sizes and low supply voltages increase the impact of variations of transistor currents (5-30%) and voltages (10-100mV) on chip- or system-level performance, causing yield loss and delayed time-to-market. Parameter variations produced by local sources cannot be easily compensated because they are totally *random*. This thesis is particularly concerned with short-range variations because of their increasing importance, especially in the development of new technology generations.

3.1 Local variation sources: RD and LER

Among the sources of local variations, two phenomena have attracted considerable interest in recent years because they are predicted to become predominant for technology nodes of immediate interest: random dopant fluctuations (RD) and line-edge roughness (LER). Fluctuations in number and position of impurity atoms result from two contrasting trends. On the one side, increasingly high doping concentrations are required to achieve the target sub-threshold behavior in short channel MOSFETs; on the other, the total number of atoms in MOSFET's channel is decreasing due to scaling. LER is the random deviation of printed device feature edges from the ideal shape, mainly due to granularity of the materials, especially polysilicon and photoresist, and tolerance of optical equipments. TCAD tools are essential in predicting the impact of these phenomena on device performance. The history of TCAD investigation of RD and LER is a significative example of problem dimensionality increase.

The impact of random dopants on bulk MOSFETs with channel lengths down to 100 nm was initially studied by means of 2D device simulations [40–42], by introducing statistical fluctuations of the number of dopants in the volume associated with each discretization node. A similar simplified approach was adopted in the first 3D studies [41, 43]. However, to carry out a more realistic analysis of the phenomenon, in-

cluding random fluctuations in the number and spatial distribution of impurities, ad-hoc "atomistic" simulators have been developed. The first work presenting this kind of approach is [44], where a 3D discrete doping region is defined and atoms are placed according to a rejection technique that produces a Poisson distribution, mimicking the physical process of ion-implantation. Unfortunately, very small ensemble sizes of only 24 devices are considered due to computational resources limitations, which is not sufficient to provide quantitative statistical predictions. A similar technique and similarly small ensembles appear in [45], while extensive statistical analysis of 3D "atomistic" structures is carried out in [46]. In order to simulate hundreds or thousands of device instances, a simplified model is used for the current continuity equation and the extraction of device parameters is performed at low drain voltages. Together with model simplification, the use of special solution techniques, such as multigrid, and hardware parallelism are the main strategies to cope with the large problem dimensionality, while adaptive meshing techniques discussed above cannot be used in the "atomistic" framework because uniform grids are required. A totally different approach is proposed in [47] and [48], where RD fluctuations are treated as a noise source, whose impact on terminal current and threshold voltage is evaluated through a small-signal analysis. This method is very efficient since it does not involve the simulation of statistical ensembles. Perturbation techniques have been applied to ultra-small devices, revealing a reasonable accuracy when compared to direct Monte Carlo evaluation approaches [49].

Line-edge roughness mainly affects planar bulk MOSFETs by varying the channel length across the device width. A deterministic approach was used in early 3D studies on the problem, approximating the roughness with a single step in the gate edge [50, 51]. A 2D statistical treatment of LER was proposed in [52], based on an approximation of the three-dimensional device by means of several 2D slices with different gate lengths. L_g values were generated through a Monte Carlo program producing a Gaussian distribution. An analogous approach is adopted in [53–55] and [56], but in the latter work an actual

spectral distribution from SEM images is used to generate gate length values. A full-scale 3D study of LER using "atomistic" simulations is presented in [57]. Here the roughness is generated based on a Fourier synthesis technique: a power spectrum corresponding to a Gaussian or exponential autocorrelation function is calculated, introducing random phase variations; then, the corresponding height function is obtained by inverse Fourier transform. Decananometer MOSFETs are simulated using a drift-diffusion model with constant mobility: though rude for such small devices, this approximation is necessary to reduce the computational effort and is justified in the paper by the interest in relative parameter variations.

Together with short-channel effects (SCEs) and oxide thickness reduction, the discussed local variation issues represent the main obstacles in further scaling of bulk CMOS technology. Multi-gate architectures such as the FinFET device described in the Introduction are a promising alternative due to a stronger coupling to the channel, which results in an improved SCE control both in the subthreshold and superthreshold regimes. Moreover, the slight doping concentration in the fins of such devices should alleviate the problem of RD fluctuations. However, LER in FinFETs is much more challenging than in planar devices, because the roughness affects several features, including top and sidewall gates as well as the fin edges. In this case 2D approximations cannot provide any realistic evaluation of the whole variability, which would require statistical ensembles of highly complex 3D device structures due to the coupling between different spatial directions. An example of 3D investigation of LER in double-gate MOSFETs is provided by [58], although only the gate roughness is considered. Oxide and body thickness fluctuations are discussed in [59], but here a kind of backward propagation of variance (BPV) [39] procedure is applied to avoid a full Monte Carlo analysis.

3.2 Statistical characterization

The impact of process variations on device and circuit performance cannot be studied by deterministic approaches because the considered fluctuations are essentially random. Statistics is therefore mandatory for this analysis.

Two main approaches are commonly used, as mentioned in the review presented in Sec. 3.1. The direct (*Monte Carlo*) method consists in simulating many microscopically different devices and statistically describing the variability of relevant electrical parameters. Usually, the magnitude of parameter fluctuations is expressed in terms of standard deviations. Of course this technique is extremely burdensome from a computational standpoint. The alternative, more efficient, approach is the *propagation of variance* [39], in which a relationship between standard deviations of physical (Ph) and electrical (El) parameters of device compact models is sought in terms of partial derivatives $\partial El_i/\partial Ph_j$. Although extremely useful for its clear predictive potentialities and low computational cost, this technique has the drawback of being strictly model-dependent. Moreover, the desired relationships may be particularly difficult to determine in cases of highly localized physical variations (e.g. RD and LER) not exhibiting explicit connections to the electrical performance through the compact model.

Short-range variations represent a particularly critical issue for all those circuits whose operation relies on perfectly matched transistor pairs. This is the case of many analog applications, including differential pairs, current mirrors, comparators, reference sources, digital-to-analog converters, but even digital blocks such as SRAM and DRAM cells are becoming more and more sensitive to such a problem. For the above reason, local fluctuations are generally characterized in terms of *stochastic mismatch*, i.e. by studying time-independent random variations in physical quantities of two identically designed devices in terms of *difference parameters* $\Delta P = P_1 - P_2$ (P_1 and P_2 being the values of parameter P for the two considered devices). Techniques to trade-off accuracy and computational cost of variability and mismatch estima-

tion will be presented in Part III, Chapter 5.

Part III

Proposed approaches - *Multidisciplinarity at the aid of TCAD*

"Any intelligent fool can make things
bigger, more complex, and more violent.
It takes a touch of genius
- and a lot of courage -
to move in the opposite direction."

E. F. Schumacher

Some techniques to deal with the two TCAD roadblocks outlined above are proposed in this Part. In Chapter 4, an automatic approach to adaptive meshing is presented, which relies on an estimation of solution regularity based on the Wavelet Transform (WT). This technique is suitable for 2D and 3D domains and exploits efficient algorithms from the field of signal processing. The quality of generated meshes is controlled through a verification routine, which allows for a full integration of the developed refinement module into a standard TCAD environment.

The topic of variability estimation is discussed in Chapter 5, describing statistical approaches to evaluate LER- and RD-induced variability at a reasonable computational cost. In particular, mismatch estimation through a limited number of simulations is considered and the use correlations to further reduce the computational effort is discussed. An advanced statistical model is also introduced, which could become indispensable when dealing with parameter distributions related to extremely scaled devices. Moreover, a perturbation approach is described as an alternative to "atomistic" simulation of random dopant fluctuations.

Chapter 4

Wavelet-based approach to adaptive meshing

A new **Wavelet-based Adaptive Method (WAM)** able to automatically generate meshes for semiconductor device simulation will be presented in this Chapter. The main feature of this approach lies on its ability to create a non-uniform grid starting from an initial coarse and uniform one, and to dynamically adjust it based on the solution behavior. The proposed strategy comes as a natural consequence of the use of a multiresolution representation. The theoretical background of this technique will now be outlined.

4.1 Wavelet analysis

One of the basic purposes of signal processing is to extract from an input signal all the relevant informative content relative to the considered application. To this aim, transformations are often applied to represent the signal in a new domain where such informations are more evident. Whenever the spectral content of the signal is of interest, the most widely used technique of this kind is the Fourier Transform. Since the employed basis functions have perfect frequency localization but infinite time duration, however, a Fourier expansion allows detecting all spectral components of the considered signal, but it does not provide any information on *when* they are present, i.e. time resolu-

tion is lost. Therefore, applications for which such information is also important require more sophisticated tools, able to decompose the analyzed waveform through basis functions characterized by both time and frequency localization.

Among *time-frequency operators*, the Wavelet Transform has recently found successful application in a variety of different disciplines. The basic idea of this technique is to start from a prototype function that is well localized in both time and frequency (compatibly with Heisenberg uncertainty principle): basis elements are obtained as shifted and dilated/contracted versions of this function, thus originating a two-dimensional representation of the input signal. Moreover, several choices are available for the prototype *mother Wavelet*, thus providing an extremely flexible tool. Time shift corresponds to scanning the signal along its duration, while the dilation factor determines the size of the waveform portion that is associated to each basis element, and therefore the range of frequencies which can be detected by that element. This is similar to analyzing the signal through a time window, as in the Short-Time Fourier Transform, but here the window size is not fixed, thus allowing for a variable time and frequency resolution. Convolving the input waveform with each basis function corresponds to calculating the *details* of the signal associated to a specific range of frequencies in a certain time interval. If high-frequency details are removed from the signal, what is left corresponds to the lowest part of the original spectrum, therefore representing a low-pass *approximation* of the given waveform. In addition to its powerful analysis capabilities, the Wavelet Transform can thus be used as a synthesis tool for approximating signals with different degrees of accuracy, depending on which details are left or removed. This *multiresolution representation* can be successfully exploited for compressing data. The following Subsections provide a more formal mathematical description of the key concepts outlined above.

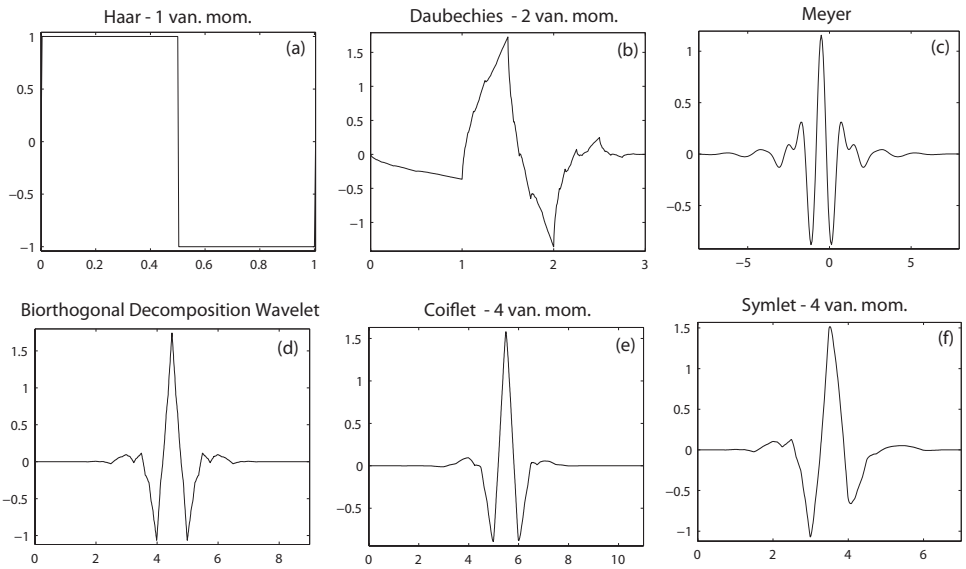


Figure 4.1: Examples of Wavelet functions $\psi(x)$.

4.1.1 Continuous Wavelet Transform

The Continuous Wavelet Transform (CWT) was originally introduced by Goupillaud, Grossman and Morlet [60]. For a function $f \in L^2(\mathbb{R})$, it is defined as:

$$\begin{aligned} CWT_{a,b}[f(x)] &= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\psi^* \left(\frac{x-b}{a} \right) dx \\ &\equiv \langle f, \psi_{a,b} \rangle \quad , \; a \in \mathbb{R}^+, \; b \in \mathbb{R} \end{aligned} \tag{4.1}$$

where a and b are usually called *scale* and *translation* parameters and $\psi(x)$ is a suitable *Wavelet* function[1]. The admissibility condition for Wavelets implies that $\psi(x)$ must have a band-pass like spectrum. Hence

$$\int_{-\infty}^{\infty} \psi(x)dx = 0$$

and therefore ψ must be oscillatory like a *wave*. Fig. 4.1 shows some of the most commonly used Wavelet functions, while examples of basis elements obtained by translation and dilation of the mother Wavelet are displayed in Fig. 4.2.

[1]Note the overloading of the symbol ψ, also used to indicate the electrostatic potential.

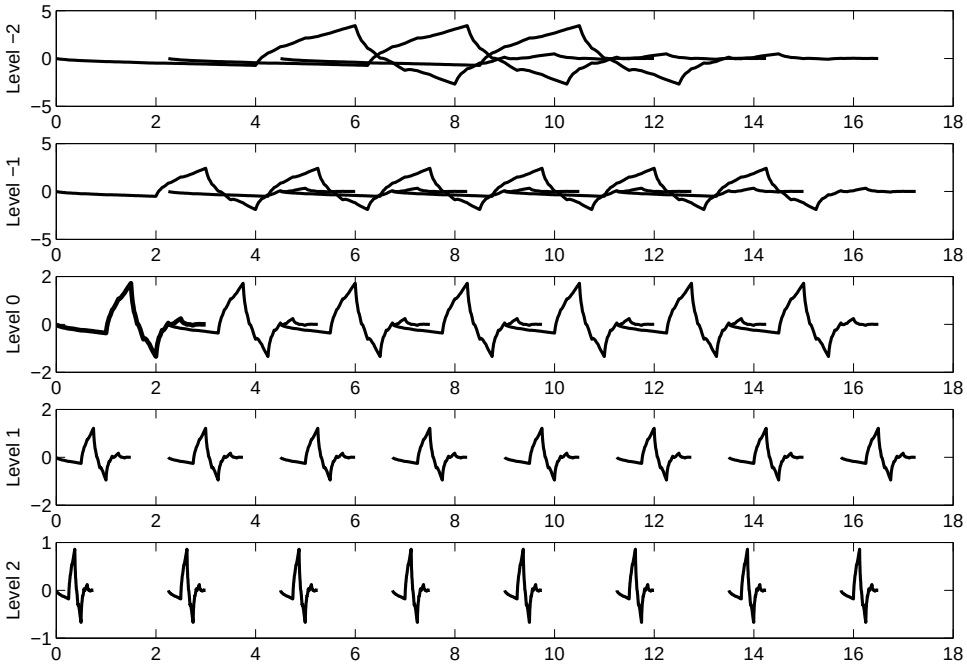


Figure 4.2: Basis functions resulting from translation and dilation of one of the mother Wavelets shown in Fig. 4.1.

4.1.2 Localization property

Both ψ and its Fourier transform Ψ are *window functions*, in space and frequency respectively[2], with centers $\tilde{x}_\psi$, $\tilde{\omega}_\psi$ and radii Δx_ψ^2, $\Delta\omega_\psi^2$. It can be shown [61] that the Wavelet transform provides a rectangular space-frequency window of size:

$$[b + a\tilde{x}_\psi - a\Delta x_\psi, b + a\tilde{x}_\psi + a\Delta x_\psi] \times \left[\frac{\tilde{\omega}_\psi}{a} - \frac{\Delta\omega_\psi}{a}, \frac{\tilde{\omega}_\psi}{a} + \frac{\Delta\omega_\psi}{a}\right]$$

with constant area of $4\Delta x_\psi \Delta\omega_\psi$. Wavelets are chosen so that most of the energy is restricted to a finite interval, i.e. either they are *compactly supported* functions, or a fast decay is imposed away from their center of mass (space localization, see Fig. 4.1). Instead, frequency localization corresponds to the band-pass like spectrum of the Wavelet.

[2]Here the signal domain is assumed to be *space* (rather than time) because this is the case in the applications discussed later on. Therefore, the transformed domain is *spatial frequency*.

The space-frequency resolution is limited by Heisenberg uncertainty principle [62], which, in signal processing terms, states that it is impossible to know both the exact frequency and the exact space position where such frequency occurs within a signal. However, this resolution varies over the two-dimensional domain of the WT, which allows analyzing high frequencies with a good space resolution but poor spectral resolution, and viceversa. Actually, in Wavelet theory, space and frequency correspond to the translation (b) and scale (a) parameters, respectively. In general we can say that for a given feature of the analyzed waveform, located at position $x = v$, there is a *cone of influence* in the scale-translation plane. This is constituted by the set of points (a, b) such that v is included in the support of the scaled version of the Wavelet function

$$\psi_{a,b} = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)$$

If ψ has a compact support $[-C, C]$, the cone of influence is defined by:

$$|b - v| \leq Ca$$

An example is depicted on Fig. 4.3, where WT coefficients of an input signal are represented in the scale-translation plane: the Wavelet Transform gradually *zooms-in* to the singularity with a good localization at small scales, i.e. local maxima of Wavelet coefficients at finer scales allow to locate high-frequency features of the analyzed function.

4.1.3 Characterization property

Roughly speaking, the Wavelet Transform calculates a resemblance index between the analyzed waveform $f(x)$ and the Wavelet located at position b and scale a, that is, the coefficient produced by (4.1) represents how closely correlated the Wavelet is with a certain portion of the function: the larger the coefficient, the stronger the resemblance. In particular, as the Wavelet is an oscillating function, the transform coefficient $CWT_{a_0,b_0}[f(x)]$ measures local variations, at scale a_0, of the function f around point b_0: for example, jumps of f or discontinuities in

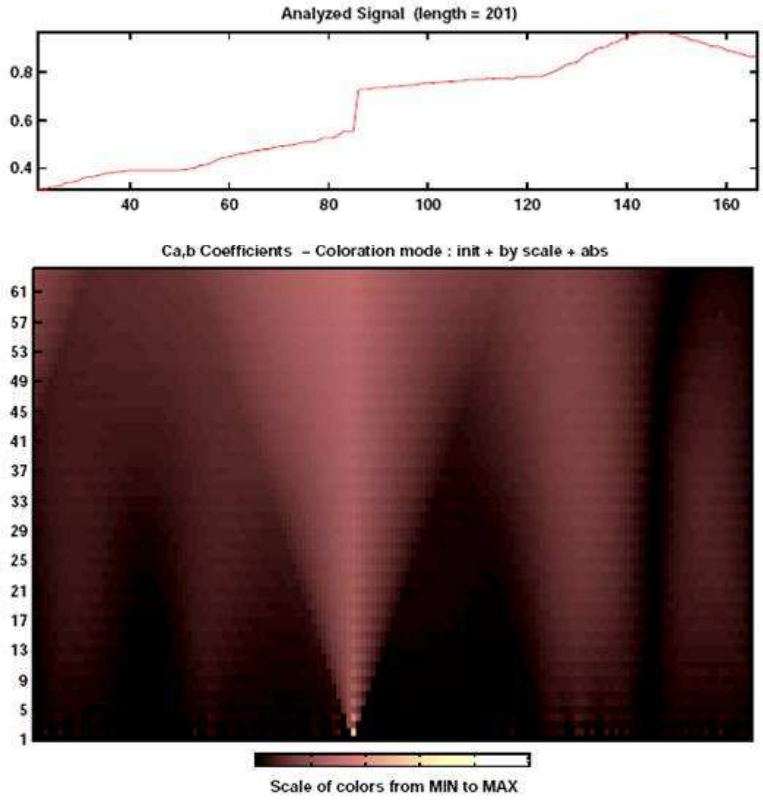


Figure 4.3: CWT of a sample signal. The pixel intensity represents the modulus of Wavelet coefficients for a certain position b (abscissa value) at a given scale a (ordinate). Strong gradients and singularities can be localized following local maxima across the scale-translation plane. The cone of influence of a sharp region occurring around $x = v$ is located in the space-scale plane where $\psi_{a,b}$ intercepts v.

its low-order derivatives generate high Wavelet coefficients. More precisely, the described zooming property of the Wavelet Transform allows to characterize the *local regularity* of signals: regularity at a particular location can be analyzed independently of the behavior elsewhere because the support of $\psi_{a,b}(x)$ becomes arbitrarily small at sufficiently small scales.

Singularities of f (i.e. discontinuities in the signal or its derivatives) at a given point v can be characterized by the Lipschitz exponent, i.e.

a positive real number α such that

$$\forall x \in \mathbb{R},\ |f(x) - p_v(x)| \leq K|x - v|^{\alpha} \tag{4.2}$$

for a certain $K > 0$. In the previous relationship, p_v is the Tailor polynomial expansion of f, with degree $m = \lfloor \alpha \rfloor$. An important theorem due to Hwang and Mallat relates this exponent to the decay of Wavelet coefficients: it is proved [61] that the Lipschitz exponent α at v can be computed as the maximum slope of $log_2 |CWT_{a,b}[f(x)]|$ as a function of $log_2(a)$ along the maxima lines converging to v. According to definition (4.2), large values of α characterize smooth functions: this corresponds to steep slopes, i.e. a fast decay across scales of the associated Wavelet coefficients. On the other hand, singularities or strong local variations of the analyzed signal give rise to large and slowly decaying coefficients: the lower the function regularity, the smaller the Lipschitz exponent and, therefore, the slower the coefficient decay at successive resolution levels.

An example is provided by Fig. 4.4, where a sample function (Fig. 4.4(a)) is analyzed (Fig. 4.4(b)) and maxima lines associated to three irregular features are plotted as specified by Hwang and Mallat's theorem (Fig. 4.4(c)). The signal is discontinuous at $x = 400$, resulting in large and slowly decaying Wavelet coefficients. The two almost parallel maxima lines in Fig. 4.4(c) are steeper than the other one because they are both associated to a jump in the first-order derivative of the signal, at $x = 200$ and $x = 800$, respectively. However, the strongest jump occurs at the latter position, resulting in higher values of the coefficient moduli. Thanks to the described property, the Wavelet analysis allows locating and characterizing singularities of the analyzed signal with a zooming procedure on the space-scale domain.

4.1.4 Wavelet series

Due to redundancy of the CWT, the scale and translation parameters can be discretized without loss of information according to this rule [63]:

$$(a, b) = (a_0^j\ ,\ k \cdot b_0 \cdot a_0^j)\ ,\quad j, k \in \mathbb{Z} \tag{4.3}$$

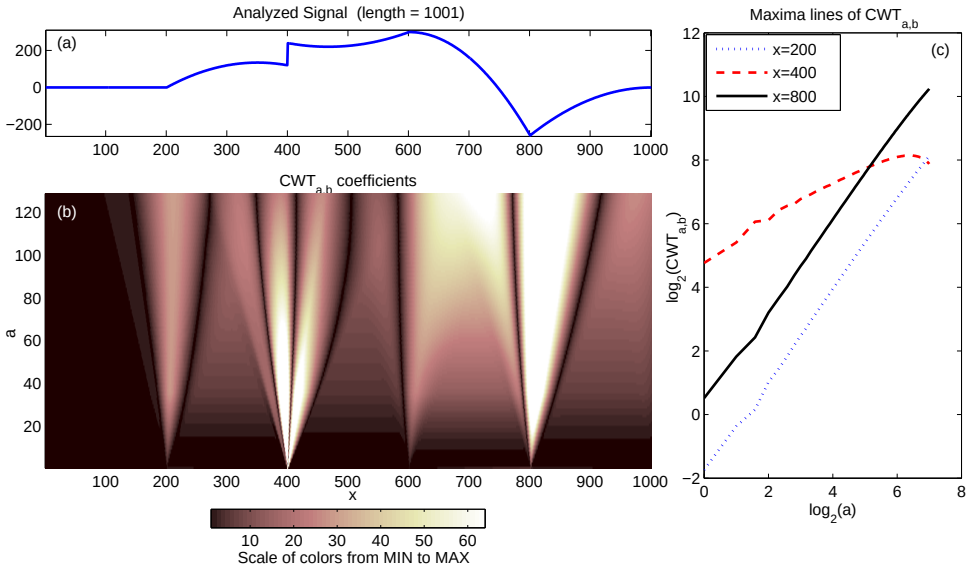


Figure 4.4: (a) Sample signal. (b) Continuous Wavelet Transform. (c) Logarithmic plot of Wavelet coefficient maxima around $x = 200$, 400, 800 as a function of the scale parameter.

where a_0 and b_0 are suitable constants[3]. In some cases it is possible to produce a rigorous orthonormal decomposition: this is the case with the choice [64] $a_0 = 2$, $b_0 = 1$, which gives rise to the following basis:

$$\{\psi_{jk}(x)\}_{j,k\in\mathbb{Z}} = \{\ 2^{-j/2} \cdot \psi(2^{-j}x - k)\ \}_{j,k\in\mathbb{Z}} \tag{4.4}$$

The analyzed function $f(x) \in L^2(\mathbb{R})$ can therefore be expanded into a *Wavelet Series* (WS):

$$f(x) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} d_{j,k}\psi_{jk}(x) \tag{4.5}$$

where the WS coefficients (usually called *details*) are defined as:

$$d_{j,k} = \int f(x)\psi^*_{j,k}(x)dx \tag{4.6}$$

The underlying *dyadic* relationship between basis elements originates a logarithmic subdivision of the scale domain: each function $\psi_{j,k}$

[3]Note that with choice (4.3) smaller scales, and therefore increasing resolution, correspond to *decreasing* values of the index j whenever $a_0 > 1$.

contributes to the fluctuations of f at scale 2^j in a neighborhood $I_{j,k}$ of size $2^j \cdot |\text{Supp}(\psi)|$, around the point $2^j k$. In particular, if $f \in C^N(\mathbb{R})$ and the Wavelet has N *vanishing moments*, that is:

$$\int_{-\infty}^{+\infty} x^k \psi(x) dx = 0 \quad \text{for} \quad 0 \leq k < N \tag{4.7}$$

then the magnitude of Wavelet coefficients is linked to the N-th derivative of f as [65]:

$$|d_{j,k}| \leq C 2^{jN} \max_{x \in I_{j,k}} |f^{(N)}(x)| \tag{4.8}$$

Eq. (4.8) is exemplified by Fig. 4.5, which compares Daubechies2 WS coefficients of a C^2 function with the second-order derivative of the signal itself. The shown coefficients, corresponding to non-overlapping $I_{j,k}$ supports, are seen to approximate the signal derivative f'' when scaled through a factor which behaves as 2^{2j} (see Fig. 4.5(g)).

4.1.5 Multiresolution approximation

In the framework of Multiresolution Analysis (MRA) [64, 66], a function $f \in L^2(\mathbb{R})$ can be represented with different degrees of accuracy by means of projection onto a nested sequence of approximation spaces $\{V_j\}_{j \in \mathbb{Z}}$, $V_j \subset V_{j-1}$. Starting from $f_j \in V_j$ at a given resolution level j, a finer approximation at level $j-1$ is obtained by adding to f_j the details belonging to W_j, the orthogonal complement of V_j in V_{j-1}:

$$V_{j-1} = V_j \oplus W_j$$

This procedure can be iterated to obtain a multiscale decomposition, in which f is expanded into the sum of its coarsest approximation f_0 and additional details g_j ($j \leq 0$):

$$f = f_0 + \sum_{j \leq 0} g_j = f_0 + \sum_{j \leq 0} \sum_{k \in \mathbb{Z}} d_{jk} \psi_{jk} \tag{4.9}$$

as shown in Fig. 4.6. In (4.9), $g_j = f_{j-1} - f_j$ represents the fluctuation of f between two successive resolution levels j and $j-1$ and can be expressed as a linear combination of W_j basis functions. For specific types of multiresolution methods, the basis of band-pass detail spaces W_j is

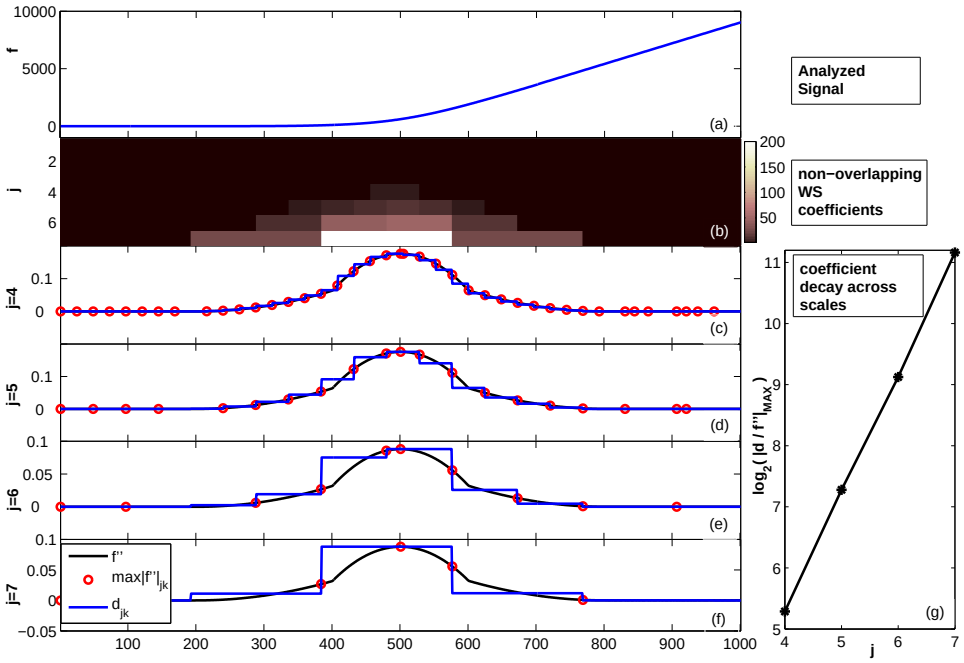


Figure 4.5: (a) Analyzed signal $f \in C^2(\mathbb{R})$. (b) WS coefficients corresponding to non-overlapping $I_{j,k}$ supports. The mother Wavelet is Daubechies2 (2 vanishing moments). (c)-(f) f'' and coefficients at different resolution levels, scaled with factor $K_j = \max_k |d_{j,k}/f''|)$. (g) $\log_2(K_j)$ plotted as a function of j.

the Wavelet basis (4.4). Similarly, the basis of low-pass approximation spaces V_j can be constructed with scaled and translated versions of a unique prototype scaling function $\phi(x)$.

The approximation theory studies how to provide an accurate approximation of a certain function f with a *reduced number of basis vectors*. For example, an approximation f_M could be constructed using M Wavelets (or scaling functions), which must be chosen in order to minimize the error $\|f - f_M\|$ due to the discarded WS projections. The best approximation is found by simply choosing the M largest Wavelet coefficients (in absolute value) $\{c_\lambda\}_{\lambda \in \Lambda_M}$:

$$f_M = \sum_{\lambda \in \Lambda_M} c_\lambda \psi_\lambda \xrightarrow[M \to \infty]{} f$$

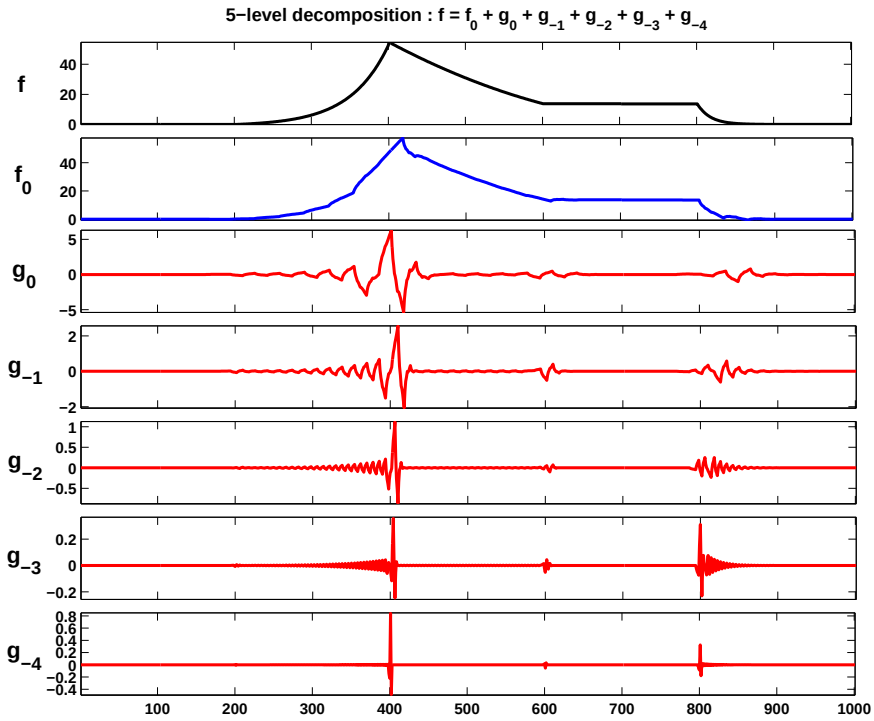


Figure 4.6: Multiscale decomposition of a sample signal f. Approximation f_0 is obtained after subtracting details g_j at five resolution levels.

The approximation properties of the basis can be evaluated through the speed of convergence as more vectors are added. This is given by the largest α for which:

$$\|f - f_M\| = \mathcal{O}(M^{-\alpha}) \tag{4.10}$$

The larger the speed of convergence α, the lower the number M of Wavelet coefficients that are needed to capture the essential information contained in f. In particular, (4.10) holds [61] for functions belonging to *Besov spaces* [67] of smoothness α. This is the case of piecewise smooth functions that model the behavior of many real-life signals.

For these classes of functions, the nonlinear Wavelet-based method described above exhibits a faster convergence with respect to other approaches such as linear Fourier-based methods or adaptive spline approximations [61]. Wavelets are therefore *optimal bases for compress-*

ing, estimating and recovering functions in Besov spaces. In particular, a high number of vanishing moments (4.7) ensures a *sparse* representation of piecewise smooth signals because the Wavelet coefficients will be essentially zero wherever the analyzed signal is well approximated by the first terms of its Taylor series.

4.1.6 Discrete Wavelet Transform

The multiresolution decomposition of a signal $f \in L^2(\mathbb{R})$ can be computed through a fast algorithm in the discrete case, thanks to a recursion relationship between approximation and details at different resolution levels. This relation in expressed by the following *two-scale equations* [66]:

$$\phi(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} \tilde{g}[n]\phi(2x-n) \tag{4.11}$$

$$\psi(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} \tilde{h}[n]\phi(2x-n) \tag{4.12}$$

(4.11) and (4.12) hold because $\phi(x)$ and $\psi(x)$ belong to V_0 and W_0, respectively, which are both subsets of V_{-1}. It follows from these equations that the projections of $f(x) \in V_0$ onto V_1 and W_1 can be computed as:

$$f_1[n] = \sum_k g[2n-k]f_0[k] \tag{4.13}$$

$$d_1[n] = \sum_l h[2n-l]f_0[l] \tag{4.14}$$

where $g[n] = \tilde{g}[-n]$ and $h[n] = \tilde{h}[-n]$. According to (4.13) and (4.14), the approximation (f_1) and detail (d_1) projection coefficients are obtained by filtering f with the low-pass and high-pass filters g and h, respectively, and downsampling by 2. The decomposition can be continued by iterating this procedure on the approximation f_1.

The computational structure outlined above can be implemented through a bank of octave-band FIR filters, which leads to the *Discrete Wavelet Transform* (DWT) [66], schematically depicted in Fig. 4.7. The reconstruction algorithm which synthesized the original signal from

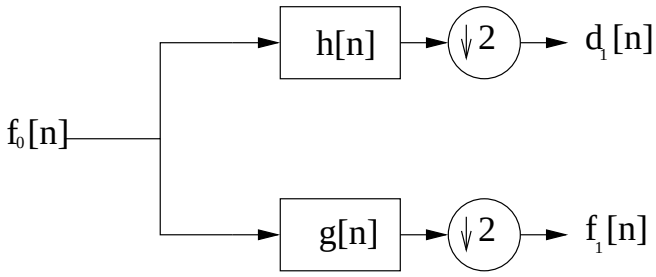


Figure 4.7: Computational structure of the Discrete Wavelet Transform. $g[n]$ and $h[n]$ are the low-pass and high-pass FIR filters used to calculate approximation and details, respectively.

approximation and detail coefficients is simply the reverse of the decomposition process. Basically, f_j and d_j are upsampled by two, passed through the low-pass and high-pass synthesis filters and then added together. To reconstruct the original signal, this process must be iterated on the same number of levels j as in the decomposition.

Obviously, efficiency of the DWT computation is determined by the filter length. The following theorem ([61], p.243) relates the support size of the Wavelet filter h to the supports of ψ and ϕ:

Theorem 1 *The scaling function ϕ has a compact support if and only if h has a compact support and their supports are equal. If the support of h and ϕ is $[N_1, N_2]$ then the support of ψ is $[(N_1 - N_2 + 1)/2, (N_2 - N_1 + 1)/2]$.*

The DWT algorithm has been described in this Section assuming that $f \in V_0$, i.e. $f = f_0$, which is generally not true. However, f_0 can be approximated through a natural sampling procedure if V_0 corresponds to a sufficiently fine resolution, because $\phi(x)$ is a low-pass filter with an integral equal to 1. More generally, in numerical computations the samples are often obtained through a low pass filtering of $f(x)$ followed by uniform sampling. If the original input $f(x)$ is related to the discrete sequence $f[n]$ by means of a suitable interpolating function $\chi(x)$ as

$$f(x) = \sum_n f[n]\chi[x - n]$$

then the relation between the continuous and discrete Wavelet coefficients is [68]:

$$CWT_{2^j,k2^j}[f(x)] = DWT_{j,k}(f_{int}[n]) \tag{4.15}$$

where

$$\begin{aligned} f_{int}[n] &= \sum_m f[m]P_f[n-m] \\ P_f[n] &= \int \chi(x)\phi(x-n)dx \end{aligned} \tag{4.16}$$

However, in many cases the pre-filtering of equation (4.16) is avoidable because almost ineffective. Such a strict link between the continuous and the discrete transform allows for the properties of regularity characterization described in the continuous case to be exploited when analyzing sequences, although in this case we are limited by the resolution of measurements.

4.1.7 Multidimensional DWT

Application of the Wavelet decomposition can be extended to multidimensional signals. A separable 2D transform is obtained by defining two-dimensional Wavelets and scaling functions as tensor products of one-dimensional components. By doing so, a 2D scaling function

$$\phi(x,y) = \phi(x)\phi(y)$$

and three 2D Wavelets

$$\begin{aligned} \psi^{HH}(x,y) &= \psi(x)\cdot\psi(y) \\ \psi^{GH}(x,y) &= \phi(x)\cdot\psi(y) \\ \psi^{HG}(x,y) &= \psi(x)\cdot\phi(y) \end{aligned}$$

are obtained, which allow to calculate a low-pass approximation of the considered signal $f(x,y)$ and *three directional details* corresponding to high-frequency features in the horizontal, vertical and diagonal direction, respectively.

The discrete version of the so called "square Wavelet Transform" is computed in two steps, as depicted on Fig. 4.8. First, a 1D DWT is

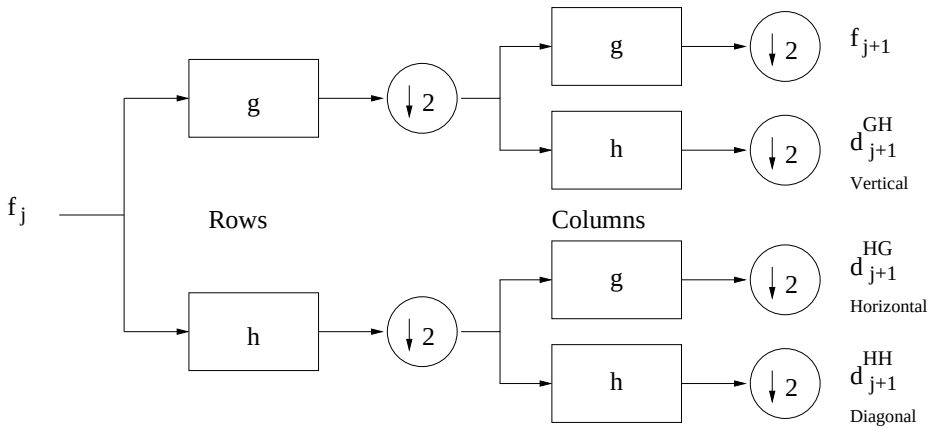


Figure 4.8: 2D DWT decomposition: H, G are the high-pass and low-pass filters, respectively. Starting from approximations at level j, they produce approximation (A) and detail (D) coefficients at level $j+1$.

performed on all rows of the original signal, yielding two matrices which contain down-sampled low-pass and high-pass coefficients of each row, respectively. Then, a similar decomposition is applied to all columns of these two matrices, thus producing four types of coefficients:

- f_{j+1} are the *approximation* coefficients resulting from low-pass filtering in both directions;
- coefficients d_{j+1}^{GH} result from a low-pass filtering (g) of the rows followed by a high-pass filtering (h) of the obtained columns and therefore represent *vertical* details;
- *horizontal* details d_{j+1}^{HG} are calculated by row-wise high-pass filtering followed by column-wise low-pass filtering;
- coefficients d_{j+1}^{HH} resulting from a convolution with $h[n]$ in both directions highlight *diagonal* variations of the two-dimensional signal.

A 2D multiresolution analysis can be obtained by iterating this procedure on the approximation f_{j+1} (see Fig. 4.9(a)-(b)). Alternatively, it is possible to also include transformation of the details at each step through a "rectangular two-dimensional transform", as represented in

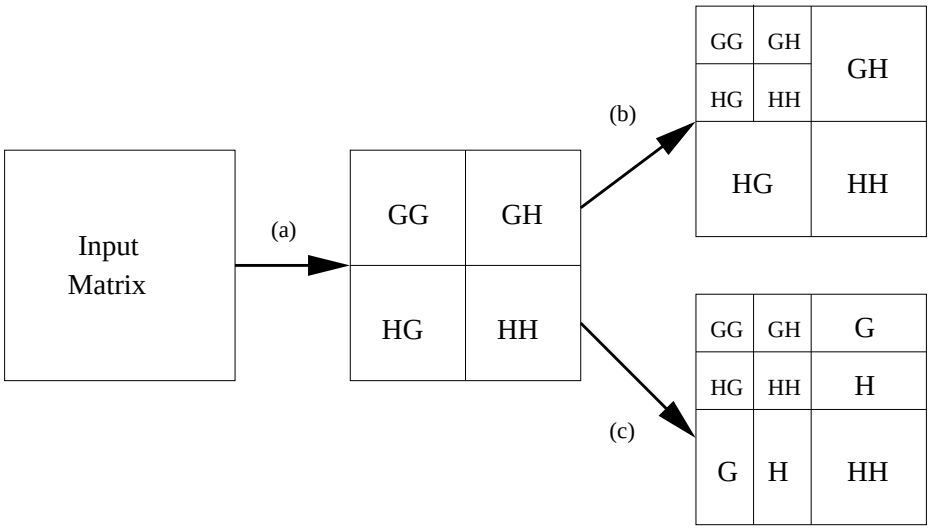


Figure 4.9: Two-dimensional Wavelet Transforms: the input matrix is decomposed into four components (a). Then the algorithm can be iterated just on the low pass component GG (*square two-dimensional transform* - case b); otherwise, the signal may be decomposed with an anisotropic basis (*rectangular two-dimensional transform* - case c)

Fig. 4.9(c). Basis functions for such a decomposition are tensor products of Wavelets at different scales:

$$\psi_{j,k}(x)\psi_{i,l}(y)$$

Such basis elements with variable aspect-ratios produce an anisotropic representation of the analyzed signal. The transform scheme outlined in this Section can be straightforwardly extended to three or more dimensions.

4.2 Wavelet properties applied to mesh refinement

The amplitude of Wavelet coefficients is related to the local regularity of the analyzed signal (Sec. 4.1.3). Therefore, a non-linear approximation that keeps the largest Wavelet inner products (Sec. 4.1.5) is equivalent to constructing an **adaptive approximation grid**, whose resolution is

locally increased where the signal is irregular. If the signal has isolated singularities, this non-linear approximation is much more convenient than a linear scheme that maintains the same resolution over the whole signal support, in terms of both accuracy of the representation and computational cost.

This is the case when the considered signals are the numerical solutions of PDE systems which describe the internal behavior of electronic devices, as explained in Part II. Wavelet properties described above can be applied to the adaptive mesh refinement for device simulation, producing grids that comply with the requirements pointed out in Sec. 2.2.1.

- *Localization* properties (Sec. 4.1.2) can be exploited to *identify sensible domain regions*, where mesh resolution must be increased to capture singularities and layer behaviors of the solution. Since such behaviors produce a *cone of influence* in the space-scale plane, the grid resolution is increased gradually, thus ensuring *smoothness* of the global mesh.

- The *characterization* property of Wavelet coefficients (Sec. 4.1.3) allows for a *regularity-estimation-based mesh refinement*. This is something different than error estimation techniques on which most adaptation strategies reported in literature are based, because:

  - calculation of an error often requires supplementary reference quantities, normally computed on auxiliary grids with increased resolution, or solving additional problems; since regularity is estimated on the solution itself, nothing similar is needed in the Wavelet-based approach, which is therefore *explicit* rather than *implicit*;
  - error indicators are usually defined point-wise or element-wise, while any Wavelet coefficient gives informations on a specific domain region including a certain amount of nodes and elements (the size of such region is determined by the

extension of Wavelet support at the considered resolution level);

– regularity estimation is not solver-sensitive, i.e. it can be performed on the desired physical quantities independently of the particular models used for the simulation.

Theorems on the *decay* of Wavelet coefficients also provide a *stopping criterion* for the refinement. Moreover, the *local regularity characterization* enabled by the compact support of basis functions allows achieving different resolution levels in distinct domain regions, i.e *non-uniform, unstructured meshes* can be created.

- The *dyadic discretization* of the space-scale plane introduced in Sec. 4.1.4 suggests an analogous policy for grid refinement. The *semi-regular* nature of resulting grids is favorable for *mesh quality control* as well as *flux-alignment* whenever axis-aligned structures are simulated.

- The *non-linear approximation* procedure based on selection of the largest Wavelet coefficients (Sec. 4.1.5) can be exploited to control the number of inserted nodes, i.e. to *trade-off accuracy and mesh size*. Furthermore, an *automatic mesh adaptation* to solution changes is obtained by monitoring large Wavelet coefficients.

- Complexity of the DWT computation described in Sec. 4.1.6 is $\mathcal{O}(N)$ for a signal composed of N samples. Such an *efficient discrete algorithm* allows for a *negligible computational overhead* of regularity estimation for mesh refinement.

- *Directional detail* informations provided by the *multidimensional transform* (Sec. 4.1.7) can be exploited to construct *anisotropic meshes* suitable for *2D and 3D simulations*. Anisotropy is intrinsic to the nature of multidimensional Wavelet coefficients, i.e. no additional sources of directional informations need to be introduced, in contrast to other approaches proposed in literature (see Sec. 2.3.2). Complexity of the multidimensional DWT is still

$\mathcal{O}(N)$, i.e. the computational cost has a *linear dependence on the number of mesh nodes* even in 2D and 3D applications.

4.3 Review of Wavelet approaches to device simulation

In recent years, different Wavelet-based methods have been applied to the solution of semiconductor device equations. Some of the most interesting approaches are reviewed in this Section.

- A Wavelet Series expansion was exploited in [69] to obtain the solution for an abrupt junction diode, by projecting the problem on a Wavelet approximating space. However, just a one-dimensional solution compared with the one obtained through a central difference method using the same grid points was presented.

- The Wavelet Transform was used in [70] as a multigrid regression and projection operator, or as a preconditioning operator for the solution of a coupled Schrödinger-Poisson system. However, this method was adopted for electronic structure calculations (atomistic simulations) rather than for the solution of charge transport problems.

- In [71] a MESFET was simulated, finding a time-dependent solution for carrier density, energy and momentum. Computational cost reduction was achieved by compressing the data with the Wavelet Sparse Point Representation (SPR) introduced by Holmström [72]; this technique was also used in [73] to simulate a 2D diode. The derivatives at collocation points were calculated from an interpolated solution on a uniform grid at the finest considered resolution level. A drawback of this approach lies in the additional overhead introduced by the interpolation. Moreover, just grids with a very limited resolution are shown in both [71] and [73].

- In [74] a Wavelet Transform was applied to a partial solution, calculated for a MESFET device on a uniform fine grid, and then the resulting coefficients were used to select and remove redundant points. Once again, the method is just suitable for time-varying problems and requires a very large initial mesh size to capture the layer features of the solution.

- Finally, in [75], spatial and temporal projectors were constructed using a multigrid framework (as in [70]) coupled to a Wavelet-based gridding procedure. This technique was used to solve a quantum-corrected drift-diffusion model on a 1D structure, but six scalar parameters (thresholds) had to be set for detecting domain regions to be re-gridded, and further intelligence was required to decide how the addition of new points should be performed; moreover, no results on multi-dimensional structures have been shown yet.

4.4 The WAM approach

A Wavelet-based Adaptive Method (WAM) for mesh refinement has been developed based on the considerations reported in Sec. 4.2. The basic idea is to use a hierarchy of fixed nested grids at different resolutions, which offers the possibility of locally selecting the appropriate discretization level. In this approach, a partial solution is calculated on a uniform coarse grid, which is then iteratively and automatically refined only in the regions where Wavelet coefficients associated to the preliminary results are greater than a given threshold. For semiconductor applications, the multiresolution analysis is performed on significative internal quantities of the considered device (e.g. electrostatic or quasi-Fermi potentials, carrier concentrations or current densities). At each level of the analysis, these quantities are the result of a previous simulation performed with a finite volume solver. WAM is therefore inserted into a *validation tool* including the solver and a meshing engine. An additional module has been implemented, which improves the quality of the generated meshes. This validation tool is described in

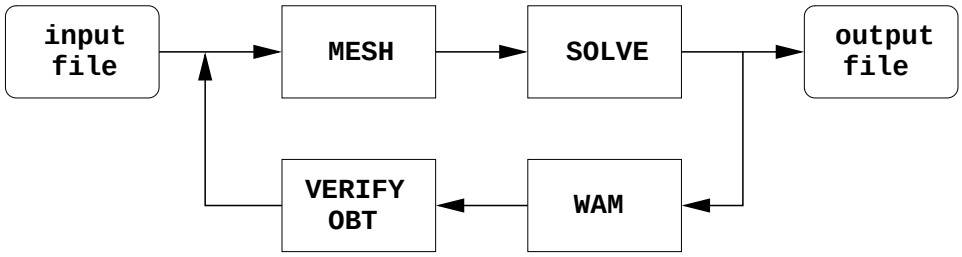


Figure 4.10: Validation tool block diagram for the proposed multiresolution analysis.

the next Section.

4.4.1 Solve-refinement cycle

Since the multiresolution analysis is structured over different levels, the validation tool implements a solve-refine cycle through the four blocks depicted on Fig. 4.10. Before entering the loop, an initialization phase is required, in which a coarse, uniform mesh is generated according to the device geometry and materials. Then the solve-refine cycle is started, which goes on until the desired resolution is reached.

- The `MESH` module must be able to produce a Delaunay triangulation/tetrahedralization of a domain described through a list of nodes, faces and regions, or to build a new mesh from an old one and a list of additional points to be inserted. These features are included in the open-source programs Triangle [76] and TetGen [77], which have been chosen as meshing engines for 2D and 3D domains, respectively. A filter has been implemented for the conversion of the resulting mesh to the specific format required by the solver.

- The `SOLVE` block represents the chosen simulator. Since a boundary conforming Delaunay mesh is produced by both Triangle and TetGen, the proposed approach can be directly carried out within a typical industrial TCAD environment. In our validation tool, Sentaurus Device [9] has been used, but any other *finite volume*

simulator could be employed[4] since the WAM approach is solver-independent (see Sec. 4.2).

- Simulation results are then filtered to extract relevant functions on which the Wavelet analysis is performed by the `WAM` block. The WAM algorithm is invoked at each resolution level after all informations about mesh node coordinates and corresponding values of sensible functions have been stored in a grid object with a hierarchical structure resembling the mesh topology. The handle of such object is one of `WAM` inputs, together with the device dimensionality, the current resolution level and spacing steps in each direction, the number of analyzed variables and the threshold values. The `WAM` module scans the loaded grid to compute Wavelet coefficients. This allows to decide if and how the mesh has to be refined; such information is given in terms of new node coordinates. The whole set of additional nodes are inserted either into the grid of the previous level or into the initial uniform grid. The second option allows for grid *coarsening* when necessary (see Sec. 4.5.5).

- New grid points produced by `WAM` are then meshed: the block `VERIFY OBT` looks for undesired node patterns in the grid and adds Steiner points in order to prevent badly-shaped elements. After the correction procedure has been performed, the final mesh is built by the `MESH` module and a new simulation can be started.

4.5 WAM algorithm description

4.5.1 Choice of the Wavelet functions

A suitable Wavelet function for the transformation can be chosen according to several features, including the number of vanishing moments, minimum support [64] and problem characteristics. In WAM, a function with $N = 2$ vanishing moments (4.7) has been selected: as a con-

[4] Of course in this case different input/output filters should be implemented to integrate the solver into the validation tool.

$g[n]$	$h[n]$
$(1+\sqrt{3})/4$	$(1-\sqrt{3})/4$
$(3+\sqrt{3})/4$	$-(3-\sqrt{3})/4$
$(3-\sqrt{3})/4$	$(3+\sqrt{3})/4$
$(1-\sqrt{3})/4$	$-(1+\sqrt{3})/4$

Table 4.1: Filter bank coefficients $g[n]$ and $h[n]$ for the `db2` scaling function and Wavelet, respectively.

$g[n]$	$h[n]$
1	-1
1	1

Table 4.2: Filter bank coefficients $g[n]$ and $h[n]$ for the Haar scaling function and Wavelet, respectively.

sequence, the magnitude of Wavelet coefficients is particularly related to the second-order derivative of the analyzed quantities according to eq. (4.8). In turn, this provides informations on the behavior of local truncation errors, expressed by (2.5)-(2.6) for the drift-diffusion model.

Filters for the discrete computation have been chosen among the **Daubechies** N family, which is characterized by the shortest possible support for a given number of vanishing moments, i.e. the filter length is $2N$. Our case $N = 2$ corresponds to the `db2` low-pass ($g[n]$) and high-pass ($h[n]$) filters, whose four taps are listed in Tab. 4.1. The corresponding Daubechies2 Wavelet is depicted in Fig. 4.1(b).

An additional transformation step has been introduced in the 3D extension of WAM (see Sec. 4.5.4). A more local basis function was needed for this purpose: therefore, the Daubechies1 Wavelet, better known as **Haar** [78, 79], has been chosen. The Haar waveform is depicted in Fig. 4.1(a) and the associated two-tap filters are described in Tab. 4.2.

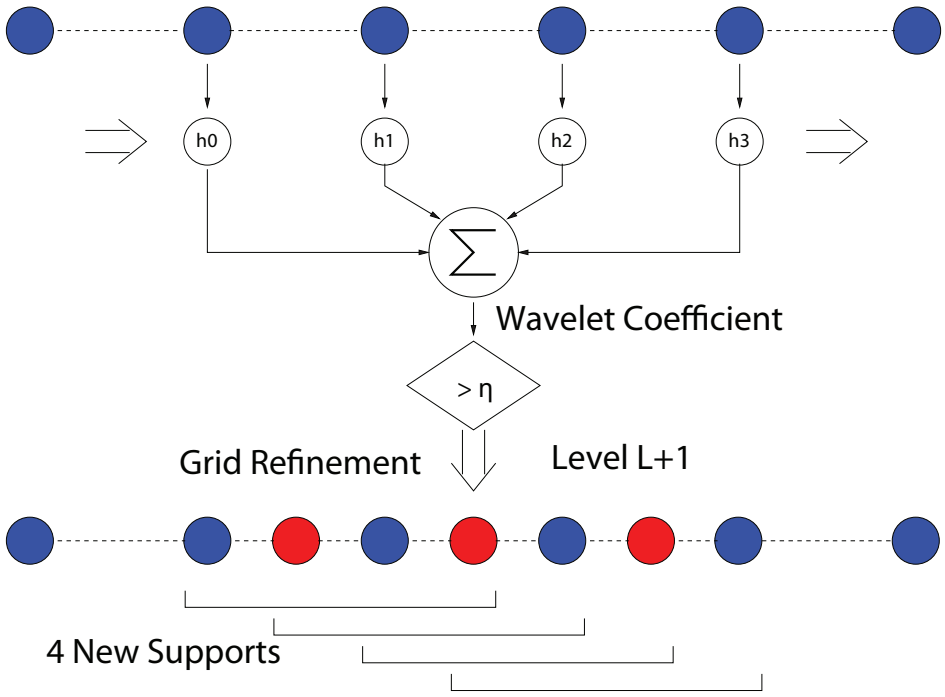


Figure 4.11: The solution on the sparse grid is convolved with the Wavelet filter $h[0-3]$; if the resulting coefficient is greater than threshold η, a dyadic refinement is imposed.

4.5.2 1D WAM computation

The details of the proposed algorithm [essderc05, sse06] will now be discussed, considering at first a simple one-dimensional case. As shown in Fig. 4.11, each Wavelet coefficient is calculated by convolving the analyzed function samples on four equidistant mesh points (the stencil or *support* of the Wavelet computation) with the taps $h[n]$ in Tab. 4.1. If a coefficient is greater than the given threshold η, three new nodes are inserted in the mesh by imposing a dyadic refinement of the support, i.e. new points are added midway between the old ones. Repeating this procedure for all available supports, the computational grid used in the next iteration is obtained. Moreover, through the described strategy four new supports at a finer resolution level are generated for each refined stencil, which allows for a smaller-scale analysis at the next iteration. Therefore, a *dyadic semi-regular mesh* is dynamically created

by this multiresolution approach.

4.5.3 Algorithm for 2D domains

An extension of the proposed approach to multidimensional domains can be obtained by means of tensorial product techniques. A two-dimensional implementation of WAM algorithm is possible through the 2D DWT described in Sec. 4.1.7. Such transform leads to a decomposition of the approximation at level j into four components, namely the approximation at level $j+1$ and details in the horizontal, vertical and diagonal direction (see Figure 4.8). If the order of the corresponding filter is equal to N, the 2D DWT coefficient is calculated convolving N^2 numerical values, i.e. $4^2 = 16$ points in the `db2` case, as depicted in the left part of Fig. 4.12. WAM analysis only requires horizontal and vertical details: if both are greater than the threshold, a uniform refinement is imposed (Fig. 4.12(A)), otherwise the analyzed region can be refined anisotropically, as illustrated in Fig. 4.12(B) and (C). The new rectangular supports produced by this approach allow to iterate the analysis at finer scales, although different resolution levels in different directions are associated to stencils generated by the non-uniform refinement: this leads to a rectangular two-dimensional transform, as described in Sec. 4.1.7. This strategy results in anisotropic grids, which is a very important feature for multidimensional device simulation, though not included in several standard adaptation methods.

Figure 4.13 shows an example of the vertices produced by the automatic 2D refinement applied to a MOSFET structure. Two main features are to be noted: (*i*) the proposed refinement strategy correctly captures the most sensible regions such as the channel and the drain junction, and (*ii*) the dyadic structure of the grid is clearly visible.

4.5.4 Extension to 3D domains

The procedure described above can be straightforwardly applied in three-dimensions as schematically represented in Fig 4.14: Wavelet di-

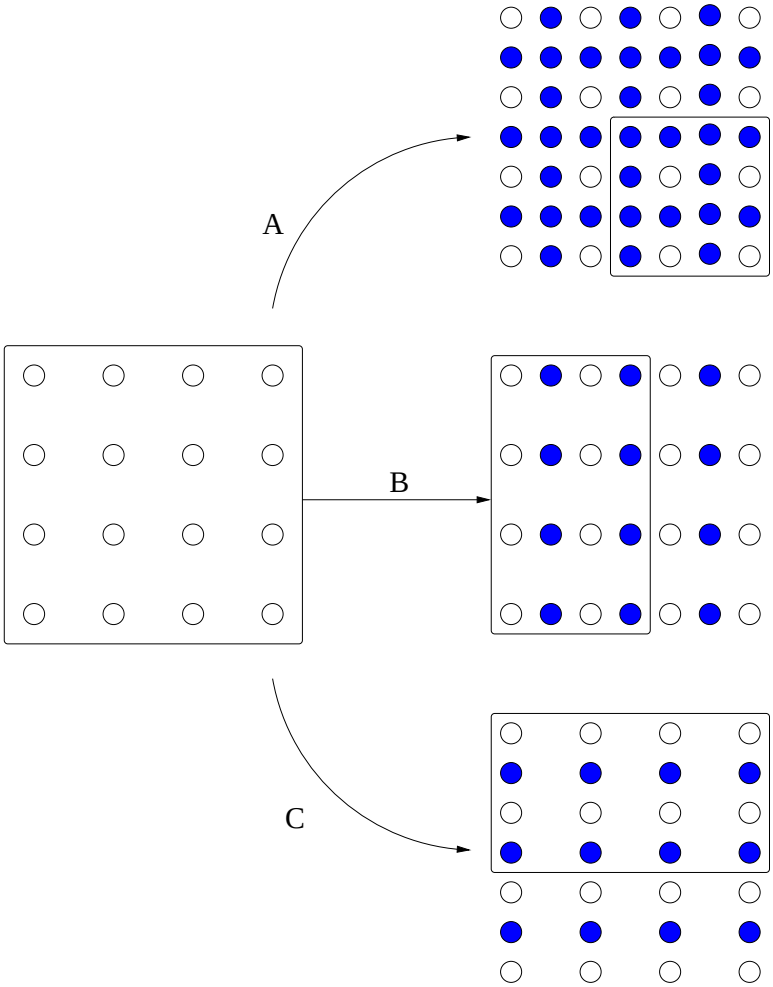


Figure 4.12: Uniform (A) or anisotropic (B, C) refinement of a 2D `db2` support.

rectional details are calculated convolving 4^3 equi-spaced numerical values associated to a 3D Daubechies2 support. A dyadic refinement can be performed in each direction corresponding to a high Wavelet coefficient, in analogy with the approach described in the previous Section. However, since supports partially overlap and lower-level ones often cover large domain areas, such a simple refinement strategy suffers from redundancy problems. While this inconvenient is well tolerable in the 2D case, it can lead to excessively large grid sizes in 3D applications.

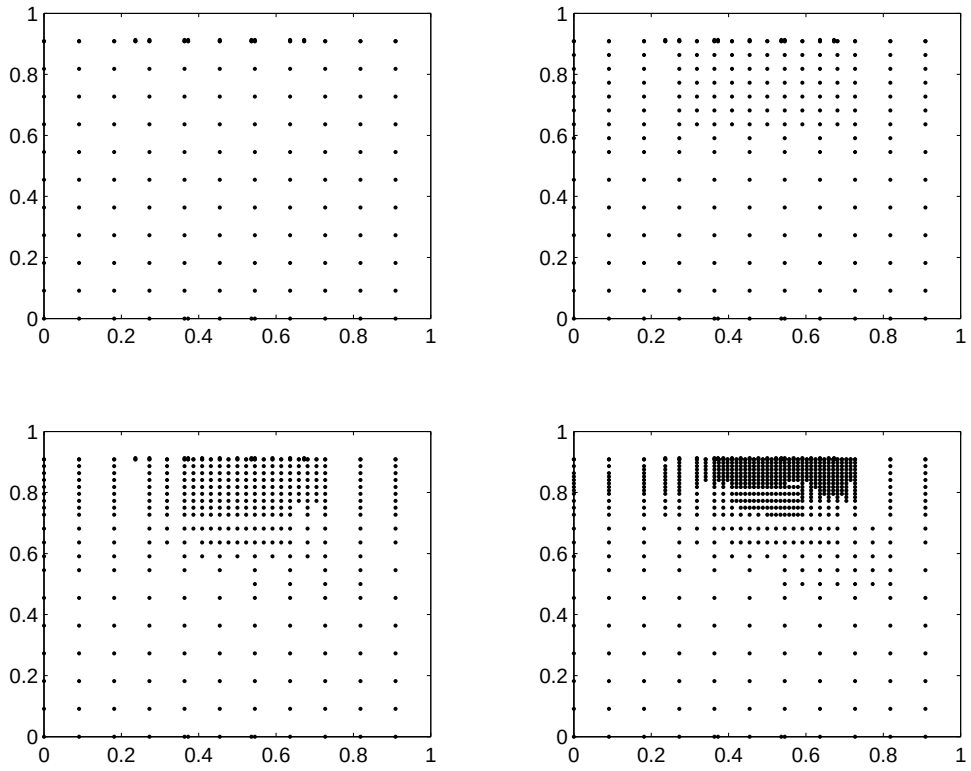


Figure 4.13: Anisotropic refinement of a prototype MOSFET device. Grid density is progressively increased under the gate and in the drain junction region.

In order to deal with three-dimensional domains, more sophisticated strategies [sispad06, tcad07] have been introduced with the following goals:

- *decoupling* the effects of a singularity on different directions while detecting the zones where an *anisotropic refinement* is desired;
- allowing for partial refinement of each support in case of highly localized singularities (*two-step refinement*).

Decoupled anisotropic refinement

In case of anisotropic refinement, new grid points can be used for higher level coefficient calculation in two ways: while a straightforward extension of the method proposed in the 2D case [sse06] consists in defining

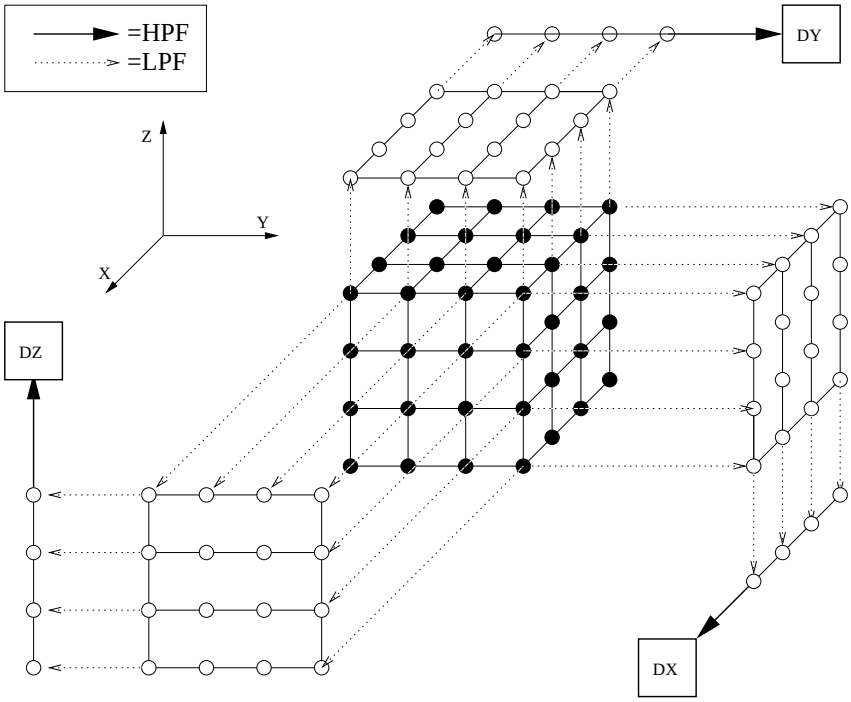


Figure 4.14: 3D Wavelet coefficients calculation. LPF and HPF are the averaging and high pass 4-taps Daubechies filters [64], respectively. Directional details `DX`, `DY` and `DZ` can be calculated by alternated application of these filters in different directions.

prismatic supports with different node spacing in the different directions, an alternative approach based on the creation of new supports with reduced dimensionality has been adopted for 3D domains. The two possibilities are compared in Fig. 4.15(b) with reference to a bidirectional refinement of a 3D stencil. More generally, the new strategy leads to generation of 1D, 2D or 3D supports (see Fig. 4.16) according to the number of directions in which the resolution must be increased.

The main advantage of this approach is that multidimensional supports have the same resolution level in all directions. This avoids Wavelet directional detail information to be corrupted by averaging operations performed at different resolution scales in other directions, which is a possible drawback of the rectangular transform described

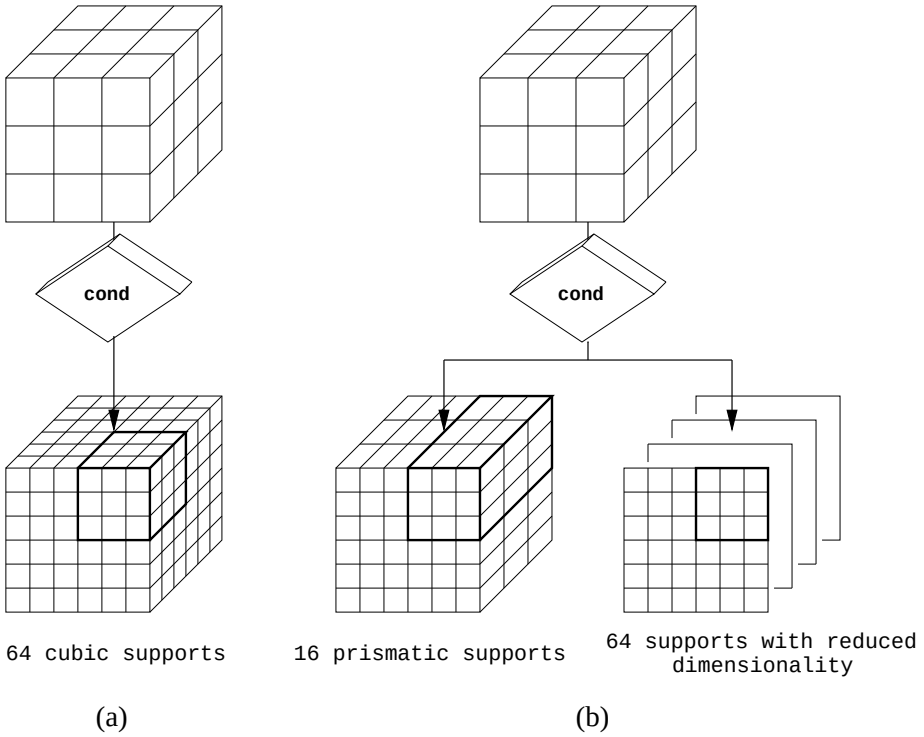


Figure 4.15: (a) 3D uniform dyadic refinement. (b) Anisotropic refinement: while the strategy in [sse06] introduces new prismatic stencils, the alternative approach [tcad07] adds smaller 2-dimensional supports.

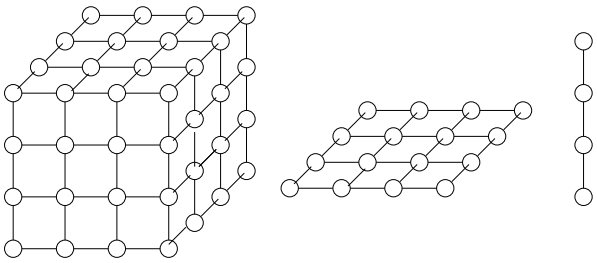


Figure 4.16: Examples of 3D, 2D and 1D `db2` supports introduced by the decoupled anisotropic refinement.

in Sec. 4.1.7. Moreover, the effect of local singularities in the solution is prevented from propagating the refinement to regions larger than necessary, thus relieving redundancy issues. Computational cost of the

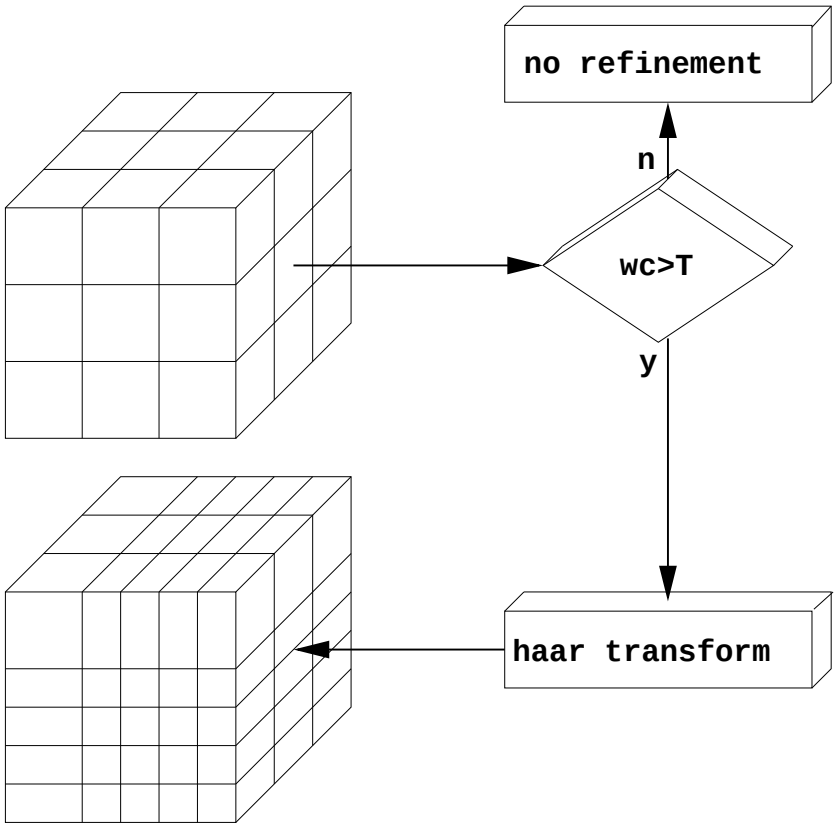


Figure 4.17: Details of two-step Wavelet refinement. The Wavelet coefficient is calculated convolving 4^3 samples of the computational grid. A further step based on the Haar Transform is added to the algorithm to keep the number of inserted nodes as small as possible.

Wavelet analysis is also reduced in the case of 2D and 1D coefficient calculations.

Two-step refinement

The 3D Wavelet analysis has been implemented as a two-step algorithm which allows to refine each Daubechies2 Wavelet support only partially, according to additional informations provided by the Haar Transform (see Sec. 4.5.1). The procedure is schematically represented in Fig. 4.17. Since the `db2` support is made of 4 grid samples in each direction, while the Haar support only includes two samples, a 3D `db2` stencil can be split into 3^3 Haar supports: together with the `db2` direc-

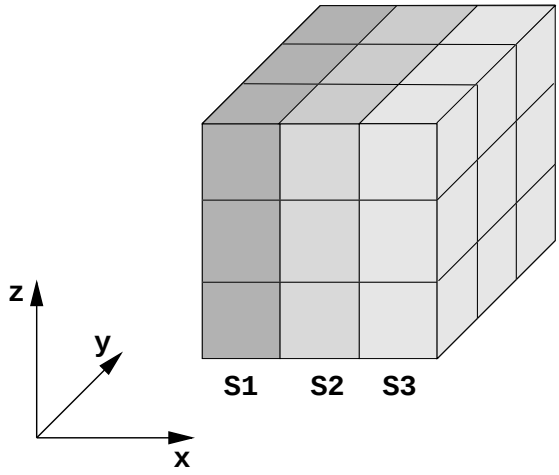


Figure 4.18: Haar analysis of a 3D db2 support in the x direction: the stencil is split into three portions $S1$, $S2$, $S3$ and the average Haar coefficient is calculated for each of them. Ratios between the resulting values discriminate if $S1$ or $S3$ can be excluded from the refinement.

tional coefficient, the associated Haar coefficients are also computed for each analyzed function. If at least one of the db2 coefficients is greater than the corresponding threshold η, the considered direction is refined. Haar coefficients are used to decide whether the refinement has to be performed on the whole support or some portions can be excluded. In other words, the basic idea is to use Daubechies2 Wavelets, characterized by moderately large stencils, to detect sensible regions and directions, and more local Haar Wavelets to further locate singularities inside db2 supports.

Fig. 4.18 shows the case of a 3D stencil analyzed in the x direction: in this example, the support can be split into three slices $S1$, $S2$ and $S3$ orthogonal to the x axis, each one including 9 Haar supports. For each slice, the x-directional average Haar coefficient is computed, resulting in three values h_{coe1}, h_{coe2} and h_{coe3}. If the condition

$$(h_{coe1} > M \cdot h_{coe2}) \text{ AND } (h_{coe1} > M \cdot h_{coe3}), \quad M > 1$$

is satisfied for all analyzed functions, then slice $S3$ is not refined; otherwise, $S1$ is excluded from the refinement if

$$(h_{coe3} > M \cdot h_{coe1}) \text{ AND } (h_{coe3} > M \cdot h_{coe2}), \quad M > 1$$

By adopting this strategy, no exclusion is allowed if the central slice exhibits large Haar coefficients. Such stringent criteria have been chosen to guarantee a smooth grading of the mesh outside the regions where singularities occur. Fig. 4.17 shows an example in which a bidirectional refinement is imposed excluding the upper and left parts of the stencil. The two-step strategy combined with the anisotropic adaptation allows for flexible refinements.

4.5.5 Dynamic mesh adaptation

One of the most powerful features of WAM is that it can be directly applied during a quasistationary simulation sweep: in such a case, a fully dynamical grid adaptation to the solution changes is produced by bias variations. When the desired accuracy has been reached at the first bias point, the simulation proceeds to the next one, as illustrated in Fig. 4.23, Sec. 4.7.2: new vertices can be added where they are needed, but it is also possible to coarsen the grid in regions which are losing influence on the solution, by dropping out the points inserted in the previous iterations. This is achieved by combining the two following expedients:

1. Wavelet supports are stored in a file, which is updated at each iteration with the new stencils produced by the refinement;

2. new nodes calculated by WAM are added to the initial uniform grid.

Thanks to the first expedient, Wavelet supports corresponding to different resolution levels can be analyzed at each step: this allows understanding when previously inserted points have become unnecessary. Such points can be removed through the second expedient. Suppose a Wavelet coefficient was greater than the threshold in the previous step, thus inducing a refinement of the corresponding stencil, but a small coefficient is associated to the same support in the current iteration (for example due to a solution change produced by updated bias conditions): the support is not refined now, i.e. nodes of the current

grid that are associated to a refinement of the considered stencil will not be included in the list of additional points produced by WAM. Since these points are added to the initial grid, the effect is a removal of the unnecessary nodes. Note that this would not be possible if only highest-resolution supports were considered at each iteration. The computational overhead introduced by the described multi-level analysis is negligible because of the fast DWT algorithm and efficient encoding of stored support informations.

Moreover, after the resolution has been increased up to some final level $levMax$ by multiple iterations at the first bias step, only one refinement cycle with the same resolution limit $levMax$ is performed at each successive bias point, usually followed by a solution recomputation on the adapted mesh. Such an approach keeps the final scale level constant through the whole simulation, thus fixing the degree of accuracy. This is beneficial to smoothness of resulting I-V curves, as shown by the results reported in Chapter 6. Finally, it is worth to notice that the described strategy can be straightforwardly applied to transient simulations as well.

4.6 Mesh quality check procedure

To provide the possibility of integrating the adaptive method into the framework of conventional TCAD tools, some requirements have to be fulfilled, as described in Sec. 2.2.1. Despite obvious intrinsic limitations in terms of geometrical flexibility, the semi-regular nature of Wavelet-based grids exhibits some advantages in this context. First of all, it guarantees mesh alignment to current flux whenever axis-aligned structures are simulated, as pointed out in Sec. 4.2. Moreover, in 2D domains the number of grid patterns generating undesired obtuse elements (i.e. obtuse triangles) is small, and for each one a stable correction strategy has been defined [prime06], based on either edge swapping or the insertion of Steiner points. On 3D domains, an extension of the correction algorithm has been implemented [tcad07], which eliminates all obtuse element faces parallel to coordinate planes.

4.6.1 2D obtuse correction algorithm

The 2D verification routine identifies and corrects a finite set of grid patterns that are responsible for all obtuse angles inside the mesh. The correction is performed by the `VERIFY OBT` module just after new points have been added by the `WAM` block (see Fig. 4.10) and consists of the following steps:

1. Delaunay triangulation of the convex hull defined for each subdomain.
2. Check for triangle patterns that are not valid and add Steiner points.
3. Repeat steps 1. and 2. up to the complete removal of obtuse angles.

Wrong patterns can be subdivided into two categories, instanced by Fig. 4.19(a) and (b), respectively. Pattern (a) consists of a missed node at specific mesh line intersections (we named this configuration a *hole*): for similar situations, a correction is performed even when the point locations do not create any obtuse angle, because such a pattern could affect simulation convergence and accuracy. Each hole can be eliminated by simply adding the point marked by a square in Fig. 4.19(a1).

All patterns (b) in Fig. 4.19 and (c) in Fig. 4.20 include an obtuse triangle. These configurations can be modified in two different ways according to node positions. If two triangle vertices have the same x or y coordinate, then a rectangle is built around the largest non-axis-aligned edge of the triangle and one of the rectangle vertices is added, in particular the first one that does not exist in the mesh yet. The technique is depicted on Fig. 4.19(b1), where the added point is represented by a square. If the wrong triangle has no axis-aligned edges, then two segments are considered, as shown in Fig. 4.20(c1). The vertical segment ($\overline{V3V4}$) is built by using the abscissa of the obtuse angle vertex ($P3$) and the ordinates of the other two triangle vertices ($P1$, $P2$), while for the horizontal one ($\overline{V1V2}$) the y of $P3$ and x's

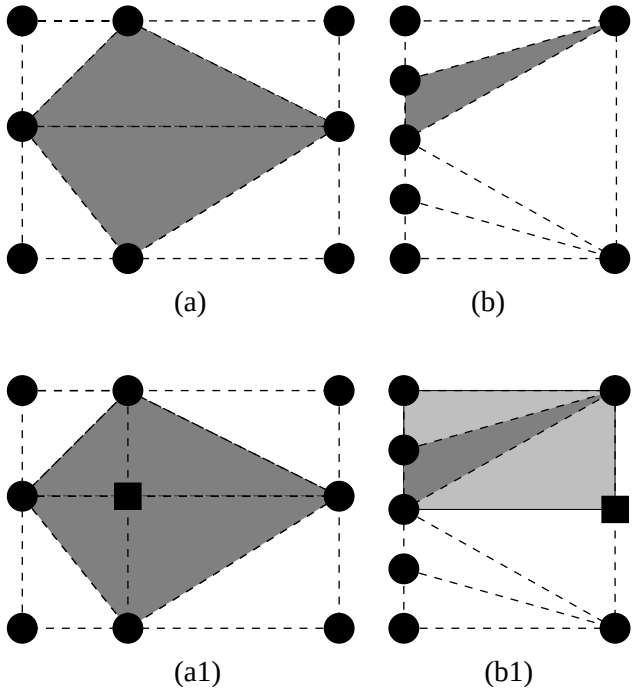


Figure 4.19: Possible undesired patterns after triangulation of the refined grid. In particular (a) is simply a hole in the mesh (not necessarily including angles greater than 90 degrees), while (b) is an obtuse triangle. (a1) and (b1) show the correction procedure for these patterns.

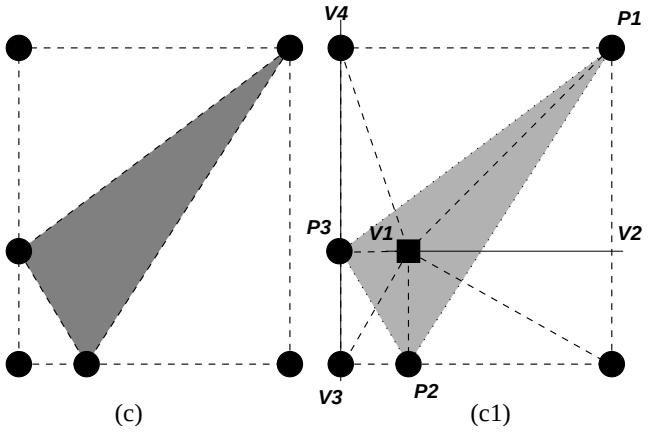


Figure 4.20: Obtuse triangle with no axis-aligned edges (c), and corresponding correction strategy (c1).

of the other nodes are taken. Between these two segments, the one that intersects the triangle edge opposite to the obtuse angle is selected and one of its end points is added to the mesh, (in Fig. 4.20(c1), the one nearest to $P3$). This technique either directly eliminates the wrong pattern or transforms it into another one belonging to one of the former cases, that will be removed in the next iteration.

Termination of the algorithm after a limited number of iterations has been observed in all considered test cases, with a limited increase in the grid size. Fig. 4.21 shows an example of mesh correction during a MOSFET device simulation.

4.6.2 Correction procedure in three dimensions

As explained in Sec. 2.2.1, mesh quality is a challenging issue in 3D applications and non-obtuse tetrahedralizations are still an open problem. In `VERIFY OBT`, the choice is to apply the procedure described above to any mesh element face parallel to coordinate planes. Inputs to the block are informations about the original mesh, the list of new nodes produced by `WAM` and, optionally, a box enclosing the region in which the correction has to be performed. The algorithm is very similar to its 2D counterpart:

1. Delaunay tetrahedralization of the domain, performed by calling TetGen as a library function;

2. identify the set of triangles to be checked through a loop on mesh elements that selects all faces parallel to each of the coordinate planes xy, xz and yz;

3. check for triangle patterns that are not valid and add correction points;

4. loop on steps 1. to 3. up to the complete removal of undesired configurations.

Obtuse faces are corrected exactly as illustrated in Figs. 4.19 and 4.20.

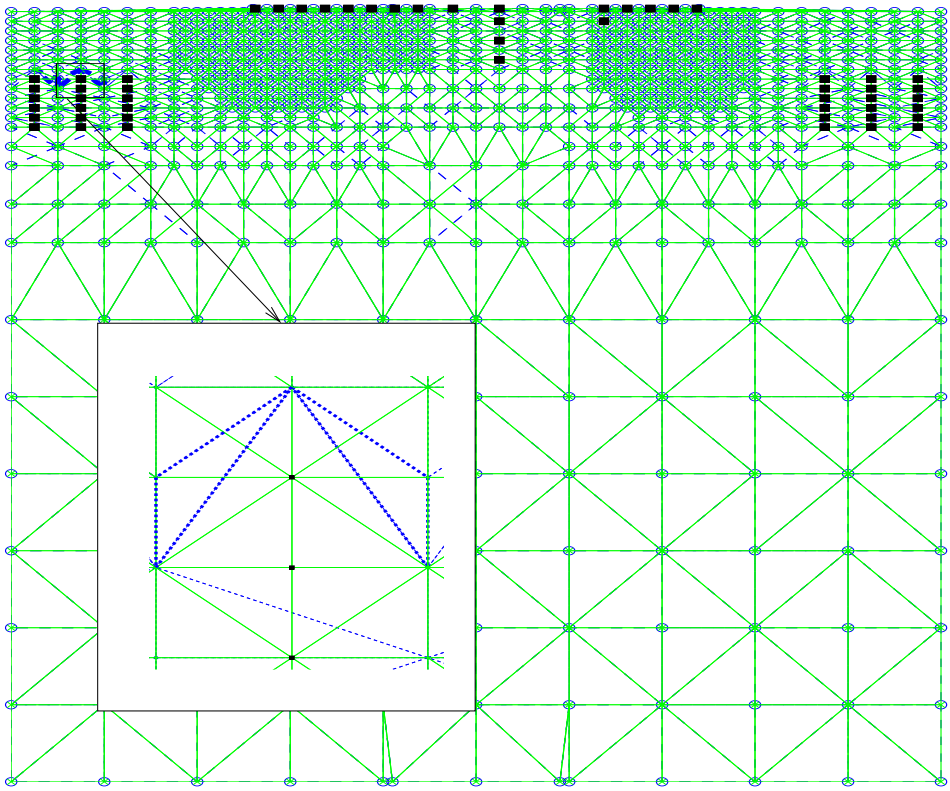


Figure 4.21: Mesh changes produced by the obtuse triangle correction. The inset shows identification and correction of one of the wrong patterns. The dashed blue segments are mesh edges before the correction, Steiner points are marked with squares and solid green lines represent the mesh after the verification step.

This approach has proven to be beneficial to mesh quality. A significative example is provided by Figs. 4.22(a) and (b), which show identification and correction of some undesired node patterns within a power MOSFET driver. Poor quality mesh configurations marked in the left part of the figure affect element faces parallel/orthogonal to the device channel and cause inaccurate discretization of internal quantities (doping concentration in the shown example). Beside relieving these inconveniences, the proposed correction algorithm also improves global mesh smoothness, as will be shown in Chapter 6.

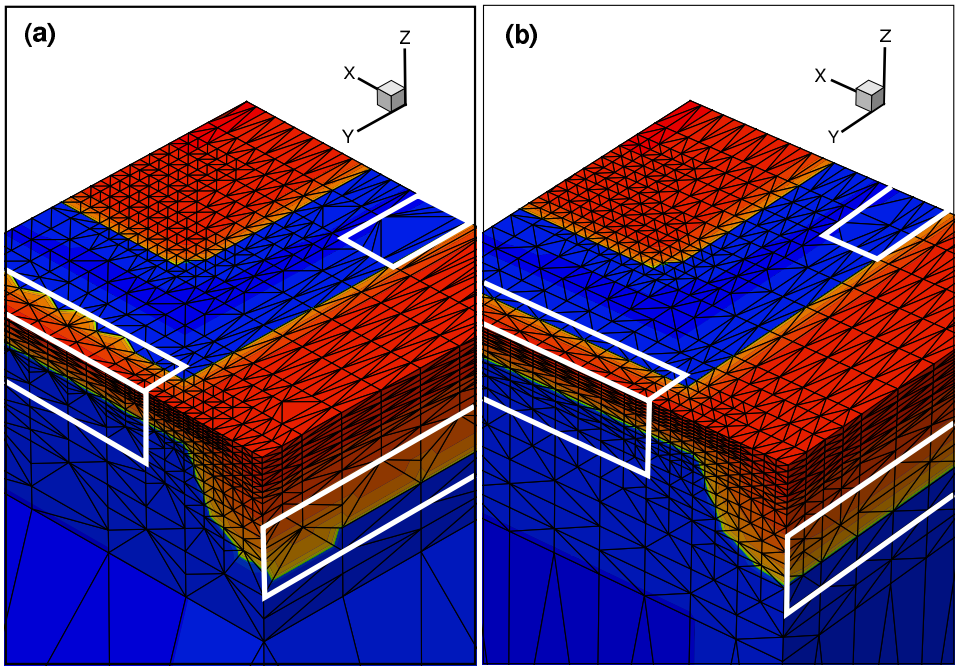


Figure 4.22: Examples of undesired mesh patterns (a) and quality improvement through the 3D quality check procedure (b) during mesh refinement of a MOSFET driver.

4.7 Implementation details

A few relevant details concerning software implementation of the 3D WAM algorithm and validation tool structure are described in this Section.

4.7.1 WAM internals

The `WAM` module is implemented in C++. It is able to refine a domain region enclosed by a given box, on which a virtual tensor grid is defined, with a minimum spacing in each direction as determined by the current resolution level and tensorial subdivision of the first-level grid.

Internal grid representation

Only points belonging to the virtual tensor grid mentioned above (i.e.

dyadic points) are retained by WAM when loading the actual mesh to be refined at each step. In this phase, a grid object is progressively created, that implements a linked multiple list: the grid is described through a list of x locations, each one linked to a list of y locations, each of which further linked to a list of z locations. Accepted grid nodes are stored in the leaves of this structure. However, since topology is defined in terms of previous and next element for each list location, memory occupancy is only determined by *actual* grid points and the structure is dynamically expanded as new nodes are loaded. Moreover, dyadic cartesian coordinates (x, y, z) are converted into *even* (i, j, k) indices, i.e. a virtually doubled grid is considered, in which only even positions are occupied, while odd locations will be filled by new nodes during the refinement phase. Optionally, a second grid object is created, containing the initial coarse mesh, which allows for grid coarsening in addition to refinement.

Wavelet support description

Informations associated to each node include (i, j, k) position, corresponding values of the analyzed quantities and a list of *supports* for which this node is a "head". In fact, each Wavelet support is identified through its *support head*, i.e. one of its corners: specifically, the node of the stencil with smallest (i, j, k) indices. Informations on the supports are given in terms of resolution levels in each direction (including zeros for lower-dimensional stencils), that are stored in a byte-wise format, so that memory occupancy for the description of each support is only one `int`. At the first analysis level, a loop on grid points is performed to identify all available 3D supports. The list of grid nodes and associated stencils, including the new ones produced by the refinement, is saved at the end of WAM task. At each successive iteration this list is re-loaded and updated.

Refinement loop

The refinement algorithm is implemented through a loop on mesh nodes. Wavelet directional details are calculated on `db2` supports as-

sociated to each node and compared with the threshold. Supports on which large coefficients have been detected are refined according to further information provided by the Haar Transform. The refinement is described in terms of new `db2` supports, whose head nodes are added to a temporary grid object. At the end of the loop, the input grid is filled with all additional nodes corresponding to the new supports.

Computational cost
Due to the use of expandable lists for grid representation and compressed support description, memory occupancy required by WAM is about 160 bytes per node. The efficient ($\mathcal{O}(N)$) DWT computation and simple node insertion mechanism allow for negligible refinement time (typically ranging from some milliseconds to a few seconds) with respect to the simulation time.

4.7.2 Validation cycle and user interface

A C++ system integration software has been developed, which connects the four blocks of the validation tool in Fig. 4.10, implementing the automatic solve-refinement cycle and providing a simple user interfacing. The program is supported by a set of auxiliary filters, written in C++ or Python language, which provide file format conversion between the different tool blocks and flow control. In particular, these filters include:

- generation of an initial tensor grid, given a boundary domain description, bounding box (or multiple boxes for separate analysis of different domain regions) and number of mesh lines in each direction;
- mutual conversion of the mesh description between the formats required by the `MESH` and `SOLVE` modules;
- extraction of analyzed physical quantities as well as terminal quantities from simulation results;
- bias condition update during sweep simulations.

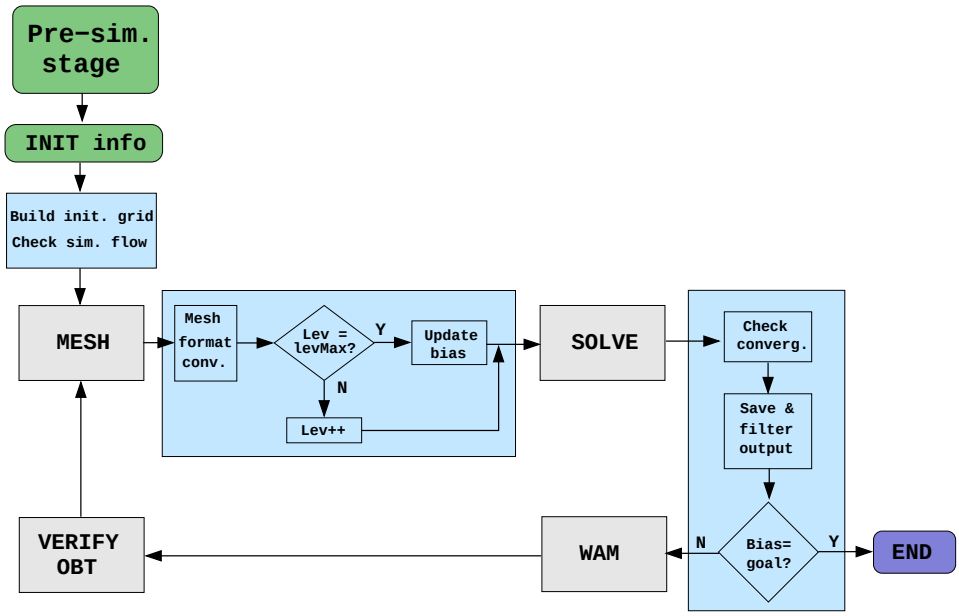


Figure 4.23: Block diagram of the system integration software. The first two blocks are the only steps requiring user interaction. Light-blue modules represent the filters that control the solve-refine cycle and allow interfacing of the heterogeneous blocks `MESH`, `SOLVE`, `WAM` and `VERIFY OBT`.

A more detailed block diagram, including these filters, is depicted in Fig. 4.23. An additional interpolation step is usually required in sweep simulations to provide a reasonable initial guess when moving between two successive bias points. Interpolation capabilities of Tecplot [80] have been used for this purpose.

The described filters have been implemented assuming TetGen and Sentaurus Device as the `MESH` and `SOLVE` modules, respectively. As a result, the developed software provides a full integration of the validation tool into the Synopsys TCAD environment, i.e. most of the tasks, that are required to start the simulation flow, can be performed by the user exactly in the same way and with the same tools as in the mentioned environment. These tasks, which are represented by the `Pre-sim. stage` block in Fig. 4.23, include:

- device structure and doping definition,

- simulation flow description.

The basic difference is that the user does not need to define a proper mesh for the device anymore, thus avoiding one of the most critical and onerous steps. Instead, the operator must provide the following informations.

- Structure name and number of desired refinement cycles.
- Device bounding box, or multiple boxes in case different refinement regions are desired. Optionally, the refinement can be enabled only for a specified material in each box. Non-refinement boxes can also be defined, where a fixed tensor grid will be generated.
- Initial tensorial subdivision of each box.
- Device quantities to be surveyed by the Wavelet analysis and corresponding thresholds in each direction. Thresholds are given in relative terms, i.e. as a fraction of the maximum value assumed by the analyzed quantity.
- An optional flag which disables the mesh quality check procedure.

These informations are interactively required by the program during the initialization phase (`INIT info` in Fig. 4.23); alternatively, they can be specified in an input file. During quasistationary and transient simulations, the marching step is automatically adjusted according to the convergence trend.

The modular structure of the described code allows for interchangeable solvers and meshing engines: to substitute these modules one has to write different implementations of the filters described above, while `WAM`, `VERIFY OBT` and the structure of the system integration software remain unchanged.

Chapter 5

Statistical approaches to variability estimation

The Wavelet-base Adaptive Method described in Chapter 4 is an example of the increasing multidisciplinarity which is a general trend in the evolution of scientific research. WAM borrows a variety of concepts from different scientific fields, including multiresolution analysis, signal processing, numerical analysis and computer science, and conveys them to the development of an auxiliary tool for TCAD to deal with complex 2D and 3D simulation problems. However, the dimension of real-world TCAD problems is further increased by enhanced non-idealities in the manufacturing process of ultra-small devices, resulting in non-deterministic effects. Accounting for process variations is fundamental in the design of integrated circuits: again, knowledge from different disciplines, especially statistics in this case, must be exploited for this purpose. In this Chapter, approaches to estimate the impact of parameter variations due to line-edge roughness and random dopant fluctuations will be presented, with particular emphasis on strategies to deal with the high computational cost of the involved simulations. The proposed techniques will be applied to study variability issues for FinFET technology in Part IV, Chapter 7.

5.1 Monte Carlo approach for LER impact evaluation

The most accurate statistical approach to include process variations in TCAD simulations is the Monte Carlo method, as explained in Sec. 3.2. This involves evaluating the impact of short-range fluctuations on device electrical performance through the following general steps:

1. statistical characterization of the considered source of variations;
2. generation of ensembles of device structures preserving the statistical features determined at the previous step;
3. device simulation and extraction of representative electrical parameters for variability estimation;
4. statistical characterization of parameter distributions extracted from simulation results.

In particular, application of this approach to evaluate the impact of LER on FinFET performance will be considered in this thesis.

5.1.1 Statistical models for LER

Step 1 in the above list involves physical insight on the causes of analyzed variations and often relies on measurements, whose reliability depends on accuracy of metrological tools and difficulties in measuring the physical phenomena. To characterize LER, recurring statistical features of the roughness in the considered technology must be determined. This is typically done by extracting and analyzing line-edge waveforms from micrographs obtained with a scanning electron microscope (SEM) or atomic force microscope (AFM), though the first method is preferred because it is faster, easier and does not damage the wafer. LER is often described through a two-parameter model obtained from the power spectrum of detected edges. The involved parameters are rms amplitude Δ and correlation length Λ: the first one represents the standard

deviation of line-edge fluctuations from a best straight fit, while the second one is the largest distance beyond which two points along the edge can be considered as statistically independent. The two most common models of this kind assume a Gaussian or exponential autocorrelation function, resulting in the following power spectra, respectively [57]:

$$S_G(k) = \sqrt{\pi}\Delta^2\Lambda e^{-(k^2\Lambda^2/4)} \quad (5.1)$$

$$S_E(k) = \frac{2\Delta^2\Lambda}{1+k^2\Lambda^2} \quad (5.2)$$

where

$$k = i\frac{2\pi}{Ndx}\,,\ i = 0, 1, .., N/2 \quad (5.3)$$

and dx is the discrete spacing between the N edge point samples. Spectral densities associated to these models are shown in Fig. 5.1.

With values for Δ and Λ as extracted from measurement data, equations (5.1) and (5.2) can be exploited to generate random rough sequences with the aim of modeling LER effects in TCAD simulations. As explained in [57], this involves introducing random element phases and creating a symmetric power spectrum array with respect to $N/2$, in order to obtain a real-valued LER sequence after inverse Fourier transform. Sequences resulting from the Gaussian model are smoother than those associated to the exponential one, whose spectrum includes a wider range of spatial frequencies (see Fig. 5.1).

5.1.2 Generation of the statistical ensemble

In general, device generation (*Step 2* in Sec. 5.1) is the result of either process simulation or direct structure definition by the operator. Two important choices are required at this stage, regarding the dimensionality of the simulation domain (1D/2D/3D) and the size of the statistical ensembles. These two choices determine the final problem dimension and influence, respectively, accuracy in modeling the physical effect and confidence on statistical results.

The application described in this thesis does not involve process simulations because of their high computational cost. Instead, a Matlab program able to automatically generate a geometrical description of

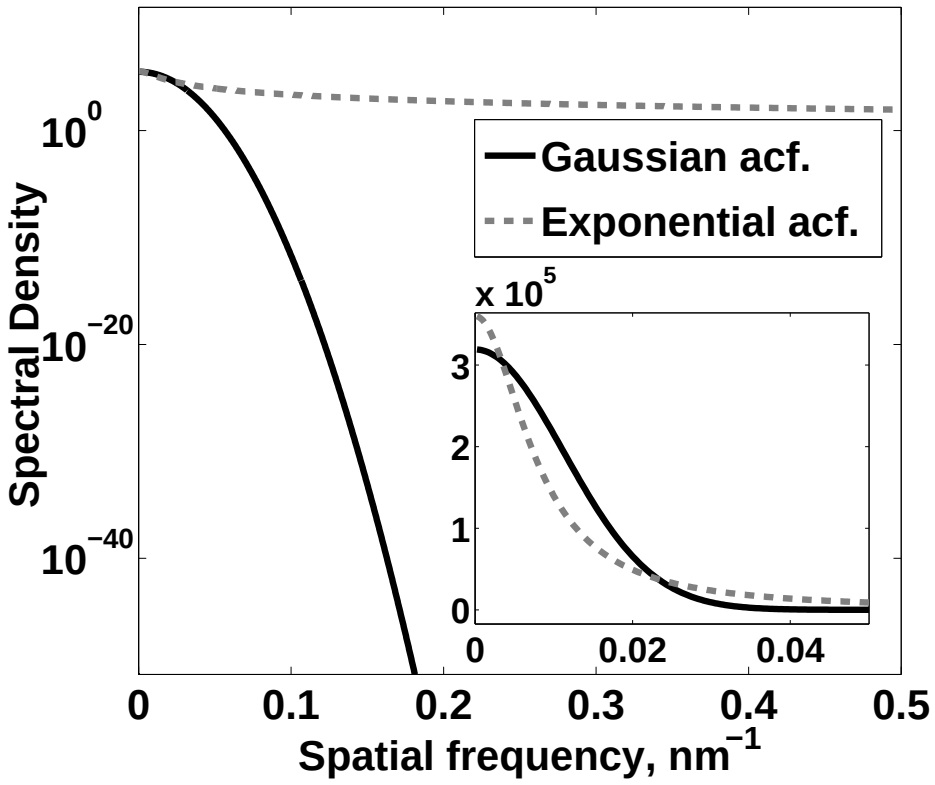


Figure 5.1: Spectral densities corresponding to the Gaussian and exponential autocorrelation functions (Δ = 1.5 nm, Λ = 20 nm) typically used to model LER statistics. The Gaussian model only accounts for low spatial frequency components, while the exponential includes a wider spectrum. A zoomed view of low-frequency spectral components is provided in the inset.

FinFET instances has been implemented. Rough features of individual devices in an ensemble are obtained in this approach by splitting long LER sequences produced through the Fourier synthesis technique described in Sec. 5.1.1. LER components from the fin, top- and sidewall-gates have been decoupled in order to compare the impact of individual contributions. As shown in Fig. 5.2, this also allows for a reduction of the domain dimensionality in the first two cases, since an approximate evaluation of fin- and top-gate LER can be carried out through 2D simulations, whereas fully-3D device structures are mandatory to account

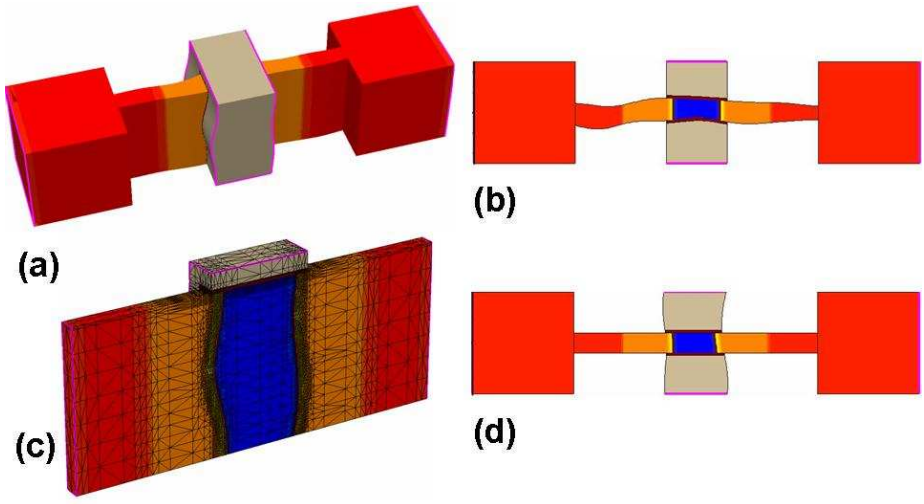


Figure 5.2: 3D FinFET instance (a) and generated structures with fin-LER (b), top-gate LER (d) and sidewall-gate LER (c).

for roughness of the sidewall-gates. However, symmetry of the structure can be exploited to reduce the computational cost in this case, by simulating only half the domain, as shown in Fig. 5.2(c). This assumes equal shapes for the two sidewall gates, but simulation results can also be exploited to predict variability for the full structure in the opposite situation of totally uncorrelated rough features on the two sides. Doping profiles have been defined following the gate shape in each device instance. Adaptive meshing techniques proposed in Chapter 4 could be exploited to further reduce the computational effort of 3D simulations by optimizing the mesh size. However, variability estimation could be affected by numerical noise arising from differences in the meshes of individual instances. Therefore, a fixed mesh definition has been preferred to alleviate this inconvenience, assigning the same resolution to corresponding regions (e.g. channels, source/drain, extensions) of all FinFET instances.

Finally, the ensemble size has been chosen both based on theoretical considerations reported in Sec. 5.1.3 and experimentally monitoring the dependence of statistical results on the number of simulated instances, as described in Part IV, Chapter 6.

5.1.3 Choice of representative parameters

Step 3 in Sec. 5.1 involves selecting a minimum set of electrical parameters able to characterize the overall device performance over a wide range of operating conditions. For MOSFETs, focus is generally on evaluating the mismatch in drain current of nominally identical devices. To this aim, the most useful parameters are threshold voltage V_T and current factor β [81]; additional parameters can be used to describe the body effect, subthreshold (I_{OFF}) and moderate inversion behavior. V_T mismatch accounts for fluctuations in several charge quantities, including fixed oxide charges, the depletion charge density (depending on dopant atoms' distribution) and threshold adjust implant dose. Variations in device dimensions and channel mobility are reflected in the current factor mismatch. Both V_T and β also depend on the gate oxide capacitance per unit area and are therefore correlated to each other. These parameters have been used to evaluate the impact of LER on FinFET electrical performance. The maximum transconductance ($g_{m,max}$) method has been used for the extraction of linear threshold voltage and current factor as $\beta = g_{m,max}/V_{ds,lin}$ (where $V_{ds,lin} = 50$ mV); then, saturation threshold voltage $V_{T,sat}$ has been calculated through a constant current method. On- and off-state currents (I_{ON}, I_{OFF}) have been extracted in the saturation regime ($V_{ds} = 1$V) at $V_{gs} = 1$V and $V_{gs} = 0$V, respectively.

In addition to DC performance, impact of local fluctuations on the device transient behavior is often significant. This can be evaluated by taking into account the Power-Delay Product. The PDP of a single device, be it a bulk MOSFET or a FinFET, can be calculated through its equivalent input capacitance C_{in}, which in turn is estimated by mimicking the device with the equivalent circuit of Fig. 5.3, thus computing the charge flowing through the gate [82]:

$$Q = C_{ref} V_{max} \;\Rightarrow\; C_{in} = \frac{Q}{V_{dd} W} \;\Rightarrow\; PDP = C_{in} V_{dd}^2 \qquad (5.4)$$

In fact, when the gate voltage V_g is raised from 0 V to V_{dd}, a reference capacitance C_{ref} is charged up to a voltage V_{max} through current $I = I_g$

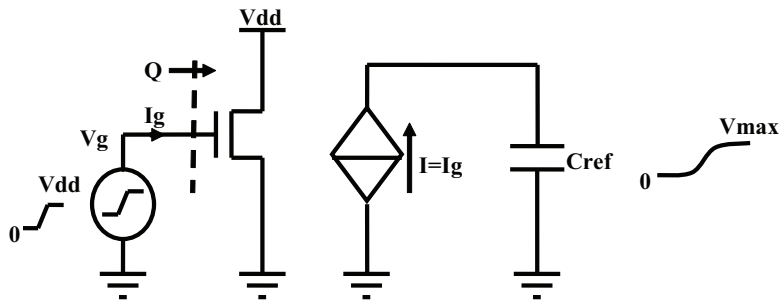


Figure 5.3: Simulated circuit for the estimation of MOSFET/FinFET PDP through relations (5.4), assuming $C_{ref} = 1$ fF.

provided by the unity gain current-controlled-current-source. Mixed-mode simulations yield V_{max} and hence the PDP through (5.4).

Further parameters can be used to characterize the performance of basic circuit blocks. For example, SRAM is considered as a highly process- and mismatch-sensitive building-block in CMOS circuits, as well as a useful test vehicle for advanced technologies and device architectures [83]. Static-Noise-Margin (SNM), extracted from butterfly curves as shown in Fig. 5.4, provides a measure of stability and functionality of SRAM cells. Moreover, ΔSNM also defined in Fig. 5.4 is especially suitable for characterizing short-range variations occurring within a single cell. In majority of the reported work, SNM refers to the read operation of six-transistor (6T) SRAM bit-cells [83–85]. However, LER-induced fluctuations in FinFET-based SRAMs have been studied here by considering the stand-by mode of cell operation: this approach makes SNM analysis independent of cell-sizing, i.e. all the four transistors highlighted in Fig. 5.5 are kept minimum sized. Moreover, for SRAM bit-cell mixed-mode simulations, computational time and complexity are considerably reduced with four transistors instead of six.

The choice of models for device simulation also contributes to determining the overall problem size. In FinFETs, significant quantum confinement of the carriers is expected because of the small fin width. The density gradient approach described in Sec. 1.3 has been used to account for this phenomenon. This approximation was shown to provide

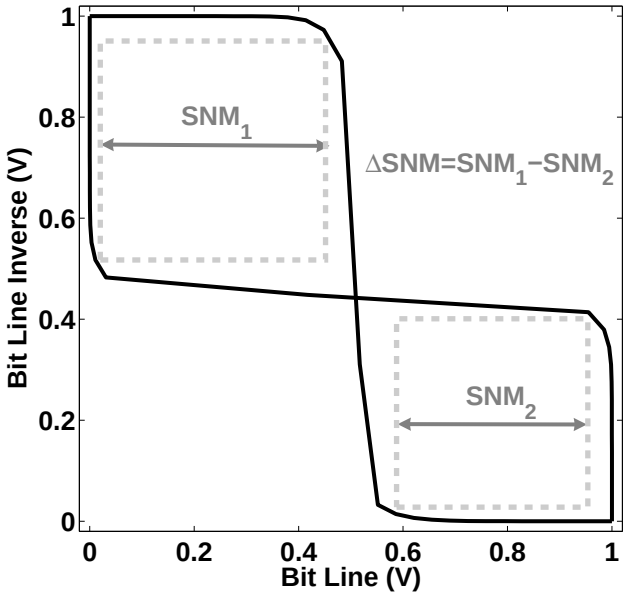


Figure 5.4: Butterfly curves in stand-by mode at $V_{dd} = 1$ V. SNM=min(SNM_1, SNM_2), ΔSNM=SNM_1-SNM_2.

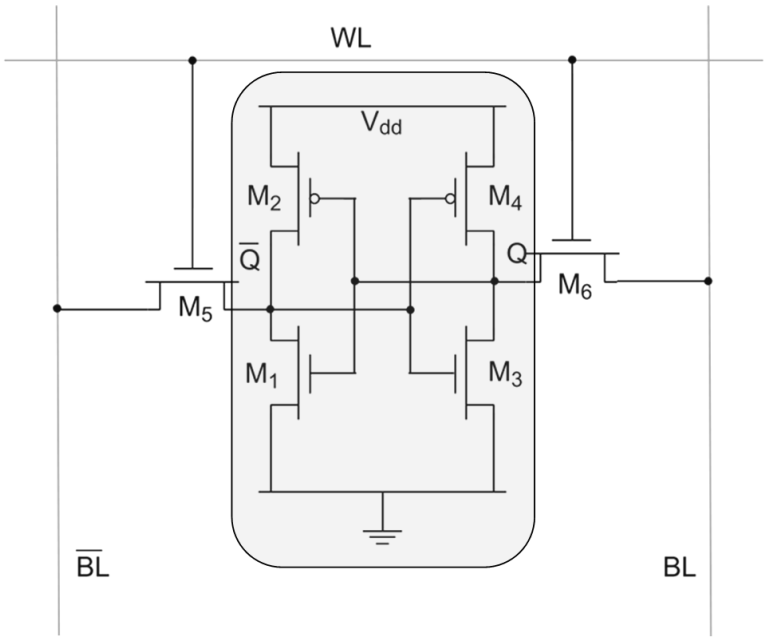


Figure 5.5: Schematic of a 6T SRAM cell. The highlighted zone corresponds to the relevant circuit in stand-by mode.

reasonable accuracy when compared to a more rigorous Schrödinger-Poisson self-consistent solution [59]. Many analog applications rely on matching pairs operating in the saturation regime: therefore, mismatch estimation should be particularly accurate at high drain bias conditions. However, in nanoscale devices the saturation regime is not properly described by the drift-diffusion model. To improve accuracy, simulations have been performed with the hydrodynamic model introduced in Sec. 1.2. Mobility degradation due to normal electric field, high-field velocity saturation and carrier tunneling through the potential barrier at the source have also been considered. The gate work function has been calibrated for threshold voltage adjustment.

5.1.4 Statistical analysis of simulation results

The simulation and extraction phases produce distributions of key electrical parameters, whose statistical behavior must be properly modeled in order to get a meaningful variability estimation at *Step 4* of the approach described in Sec. 5.1. It is reasonable to assume that each considered parameter P follows a Normal distribution and to express variability in terms of the average value $\langle P \rangle$ and standard deviation σ_P. Accuracy of these estimates depends on the sample size N according to the following relationships [46]:

$$\sigma_{\langle\rangle} = \sigma_P^{asy}/\sqrt{N}\,,\ \sigma_\sigma = \sigma_P^{asy}/\sqrt{2N} \tag{5.5}$$

where $\sigma_{\langle\rangle}$ and σ_σ are the standard deviations of $\langle P \rangle$ and σ_P, respectively, and σ_P^{asy} is the "true", asymptotic value of σ_P. This means that as many as 200 samples are needed, for instance, to bound the uncertainty σ_σ within 5% of σ_P^{asy}. However, some strategies described in Sec. 5.2 can be exploited at this stage to improve efficiency, accuracy or informative content of the variability estimation approach.

Finally, it is worth highlighting that informations from *Step 4* can be used at higher levels of the simulation hierarchy to evaluate the impact of local fluctuations on the overall IC performance.

5.2 Techniques to improve the efficiency-accuracy trade-off

Direct statistical estimation of variability involves a further dimension increase represented by the ensemble size. This is often extremely expensive from a computational standpoint and hence techniques are needed, which can provide as accurate as possible statistical results from a limited number of simulations. Reported in this Section are some observations that can help tackling the outlined problem.

5.2.1 Mismatch Evaluation

As stated above, the impact of local variations on device performance is usually estimated in terms of *mismatch* between two nominally identical devices. Typical quantities of interest are the mismatch in threshold voltage and current factor, the latter often normalized to its average value to measure relative variations:

$$\Delta V_T = V_{T1} - V_{T2} \ , \quad \frac{\Delta\beta}{\langle\beta\rangle} = 2\frac{\beta_1 - \beta_2}{\beta_1 + \beta_2} \tag{5.6}$$

where 1 and 2 are the indices of the two considered devices. To evaluate statistics $\sigma[\Delta V_T]$ and $\sigma[\Delta\beta/\langle\beta\rangle]$, two distributions of devices should be simulated, each one generated according to the statistical features of the considered source of mismatch. Following (5.5), this implies doubling the number of simulations to achieve the same amount of confidence on variability estimation for the difference parameter. However, if the two distributions are independent from each other, they will be characterized by the same statistical average $\mu[\cdot]$ and standard deviation $\sigma[\cdot]$, which allows to estimate mismatch by simulating one distribution only [ted07, nova]:

$$\sigma[\Delta V_T] \simeq \sqrt{2}\sigma[V_T] \ , \quad \sigma\left[\frac{\Delta\beta}{\langle\beta\rangle}\right] \simeq \frac{\sqrt{2}\sigma[\beta]}{\mu[\beta]} \tag{5.7}$$

When characterizing stochastic mismatch, it is important to take into account area dependence. Intuitively, local fluctuations must become larger as the involved area decreases, since they are related to

discreteness of charge and matter. The most widely used model to quantitatively describe this area dependence is the one proposed by Pelgrom *et al.* [86], although several corrections have been proposed for deep submicron (DSM) technologies (see for example [87]). In this model, the following relationship describes the mismatch in parameter P between two identically drawn devices with nominal dimensions W, L and whose centers' distance on the wafer is d:

$$\sigma^2[\Delta P] = \frac{A_P^2}{WL} + S_P^2 d \tag{5.8}$$

where A_P and S_P are the fitting parameters for the area- and distance-dependent terms, respectively. Usually, statistical simulations do not allow accounting for the second term in (5.8), which models long-range and often systematic variations. Instead, mismatch for a certain technology is characterized by estimating A_P through linear regression of $\sigma[\Delta P]$ vs. $1/\sqrt{WL}$ data extracted from simulations of several device geometries (which implies a big computational effort). However, since the linear regression must be forced to intercept the origin for physical reasons (stochastic mismatch of two paired transistors converges to zero as they become infinitely large), a reasonable estimation of A_P can be provided by two additional points only, i.e. simulating two device geometries should be sufficient.

5.2.2 The Half-Normal Statistics

As scaling proceeds, the absolute value of device geometric and electrical parameters tends to become smaller and smaller, while local variations are enhanced. Since certain parameters must be strictly positive for physical reasons, the corresponding distributions are expected to exhibit asymmetries between the left and right tail, because the first one is bounded by zero. In other words, deviations from the Gaussian behavior are expected, resulting in "deformed bell-shaped" histograms with different decays at the two sides of the peak value, which now corresponds to the statistical mode rather than the average.

In such cases, a more appropriate fit than the Normal one can be provided by Half-Normal statistics [ted07, nova]. In this approach, μ is

calculated as the mode and two separate Gaussian fits are provided for the left and right parts of the distribution, ensuring a smooth joint at μ. The left and right Half-Normals are characterized by two different standard deviation values σ_L and σ_R. These can be combined to remap the asymmetric Half-Normal fitting into an equivalent Normal distribution with standard deviation:

$$\sigma = \sqrt{\sigma_L \sigma_R + \left(1 - \frac{2}{\pi}\right)(\sigma_L - \sigma_R)^2} \tag{5.9}$$

The described approach allows for a more accurate modeling of asymmetric distributions, ending up with an equivalent estimation of mismatch in terms of the conventional standard deviation parameter. As an example, the Half-Normal fitting has been applied to a distribution of current factor values resulting from the simulation of 85 3D structures of the kind shown in Fig. 5.2(b), used to investigate the impact of sidewall-gate LER on FinFET matching. Although the zero-bound is not yet a severe limitation for the considered technology, asymmetries in the distribution are observed, which are correctly captured by the Half-Normal model, as shown in Fig. 5.6. This approach could become indispensable for analyzing future technologies.

5.2.3 Exploiting Correlations

Finally, correlations are another fundamental topic in statistical analysis. Local variations affect the device structure (e.g. geometry, doping, etc.). In turn, this determines fluctuations of the electrical performance: therefore it is reasonable to expect correlations between structural (x) and electrical (p) parameters. For example, x could be the average size of a printed device feature subject to line-edge roughness, or the number/position of some dopant atoms in the channel, while p could be the corresponding threshold voltage or current factor. Investigation of such correlations can lead to three main achievements:

1. a better physical insight of how variations affect the device behavior;

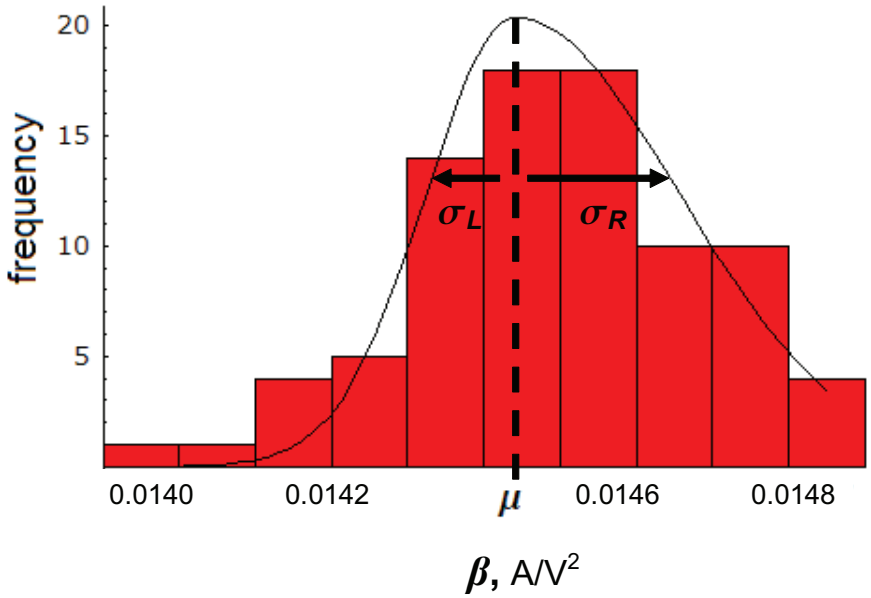


Figure 5.6: Histogram of current factor distribution for 85 3D FinFET structures affected by sidewall-gate LER (see Fig. 5.2(b)). The Half-Normal fitting is also shown; peak position μ as well as left and right standard deviations (σ_L, σ_R) are indicated.

2. inclusion of variation effects into compact models, thus allowing to evaluate their impact at higher complexity levels of IC design and to increase the predictive power of TCAD;

3. reduce computational cost of further statistical simulations.

Only the latter point will be considered here since the focus of this Section is on general methodologies to improve the efficiency-accuracy trade-off. The other advantages of correlation analysis are strictly application-dependent and will be discussed while illustrating simulation results in Part IV, Chapter 7.

Suppose a linear correlation is observed between some structural parameter x and a certain electrical parameter p for a given ensemble, as shown in Fig. 5.7. This suggests a simple way to achieve a faster estimation of variability from very few simulation data [tnano, nova]. The general approach consists of three steps:

a) statistical analysis of *structural* distribution x;

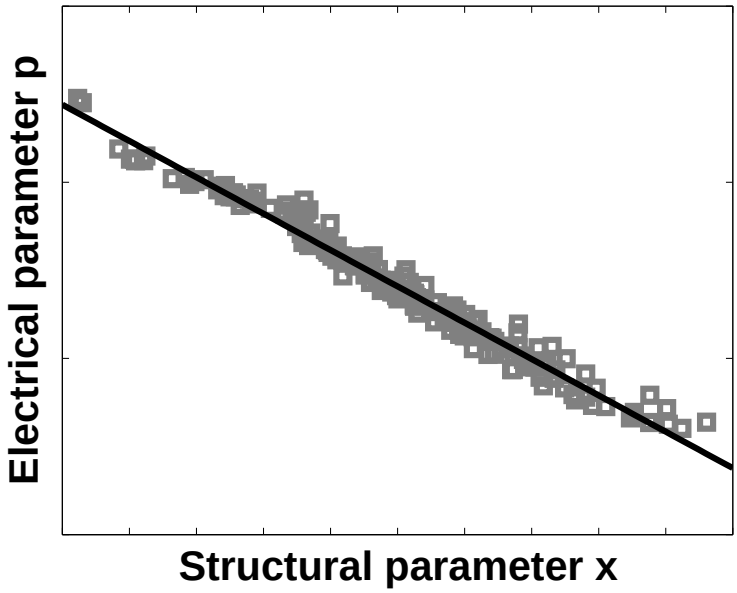


Figure 5.7: Example of linear correlation between structural and electrical parameters in a statistical ensemble of microscopically different devices.

b) select a small number of significative instances for device simulation;

c) use data from previous steps to estimate statistics of *electrical* parameter p.

If device instances do not result from process simulation but are generated by direct structure definition, this procedure can be usually automatized and hence the generation phase is very quick. In this case a large number of samples can be created and a simple Normal fit can be used to describe the structural distribution. Corresponding standard deviation $\sigma[x]$ as well as extremes ("corners") $x_m = \min(x)$ and $x_M = \max(x)$ are calculated at step a), as depicted in Fig. 5.8. Simulation of "corner" devices at step b) provides electrical parameters p_L and p_R and electrical variability (step c)) can be approximately estimated as:

$$\sigma[p] \simeq \left| \frac{p_R - p_L}{x_M - x_m} \right| \sigma[x] \tag{5.10}$$

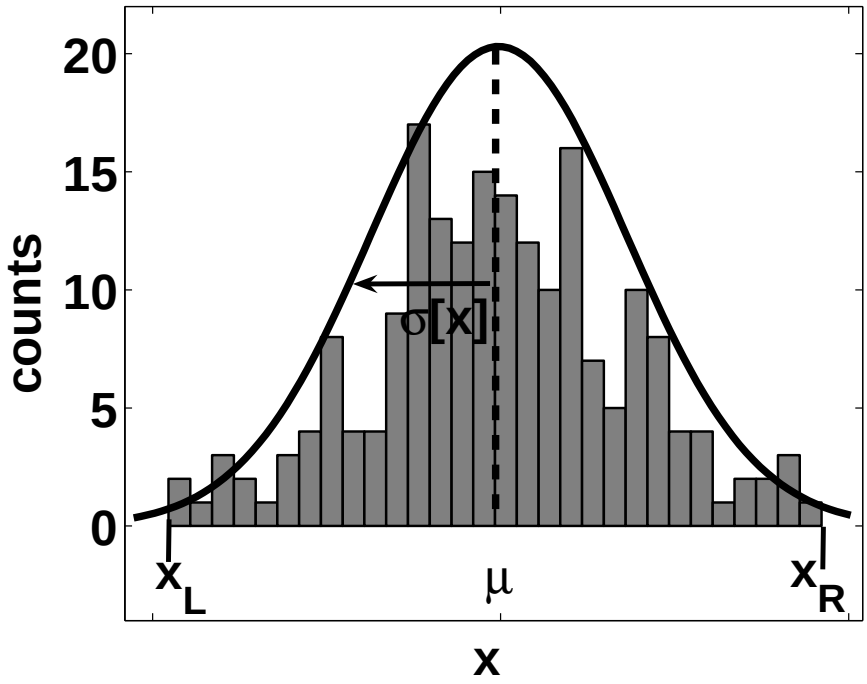


Figure 5.8: Normal fitting of structural distribution x.

This approach is extremely simple and efficient: although the whole ensemble of devices is needed to calculate $\sigma[x]$, only two samples have to be simulated.

However, eq. (5.10) provides quite a rough estimation in cases when the actual distribution of electrical parameter p exhibits strong outliers, i.e. simulation data strongly deviating from the linear trend. Estimation through (5.10) is affected by these outliers when they are located at the distribution tails. Accuracy can be improved by taking as "corners" the devices next to the extremes when convenient, as illustrated in Fig. 5.9. Naming these samples as 2, $N-1$ and the extremes as 1, N, the problem is how to select the suitable couple of instances *a-priori*, i.e. without simulating the full ensemble. A simple algorithm has been developed to automatically perform the best choice for a given test case:

- the four candidate "corner" devices 1, 2, $N-1$, N are simulated;
- two estimates $\sigma_1[p]$ and $\sigma_2[p]$ are calculated through (5.10), using

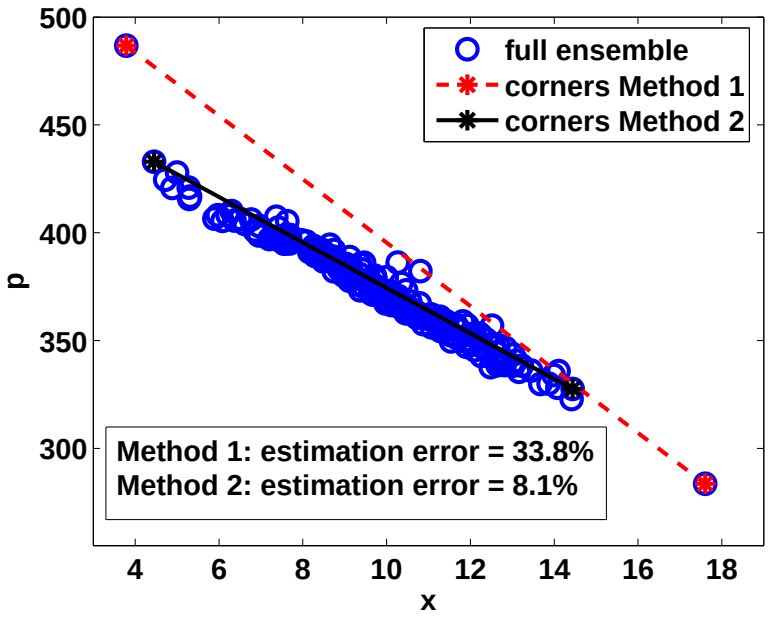


Figure 5.9: Variability estimation of p exploiting correlation to x. Errors in σ values estimated through samples 1, N ("Method 1") and 2, $N-1$ ("Method 2") w.r.t. the value extracted from the full ensemble are also reported.

samples 1, N ("Method 1") and 2, $N-1$ ("Method 2"), respectively;

- the best estimation σ_{opt} is selected as the maximum or minimum between the two computed values according to a threshold T.

If $\sigma_1[p]$ and $\sigma_2[p]$ are "not too different" from each other (i.e. relative difference less than T), then the actual distribution should not be affected by strong outliers. In this case, the maximum between the two standard deviation values is chosen as it provides a worst-case estimate. On the other hand, a large relative difference between the two results is a symptom of outliers, which will probably affect the largest σ. Therefore, the minimum between $\sigma_1[p]$ and $\sigma_2[p]$ is retained in this case. The threshold sets an upper bound to the tolerable difference in the estimations provided by "Method 1" and "Method 2" for distributions without significant outliers. Once $\sigma[p]$ has been calculated,

mismatch can be estimated through relations (5.7).

Exploiting correlations is very convenient from the computational standpoint since it allows estimating variability from only 4 simulations instead of N, with an efficiency improvement of two orders of magnitude for $N = \mathcal{O}(10^2)$. On the other hand, this method involves some inevitable loss of accuracy due to the accumulation of several approximation errors, so the choice between the proposed algorithm and the full statistical approach depends on the acceptable trade-off between accuracy and computational cost.

5.3 Noise analysis for RD investigation

The most accurate technique to evaluate the impact of random dopant fluctuations is again the direct Monte Carlo approach. This involves simulating several device instances with different dopant distributions: the number and position of impurity atoms in each structure should be determined based on accurate statistical models of the ion-implantation process, as mentioned in Sec. 3.1. The classical doping description in terms of continuous profiles is therefore substituted by a more sophisticated representation accounting for charge discreteness. However, simulation of such devices cannot be tackled with standard solvers. Instead, ad hoc "atomistic" simulators are required for this purpose [44–46].

Due to unavailability of such tools, an alternative technique has been adopted here, based on noise analysis [9, 47, 48]. In this approach, fluctuations of contact voltages are computed as a response to a small perturbation of the doping concentration. The impedance field method [48] is applied for this purpose, using Green's functions to calculate the circuit response. This involves a linearization of device equations under the assumptions of small enough doping fluctuations and statistical independence of discrete dopant atoms. The noise analysis is performed in the frequency domain and the simulator [9] computes noise voltage spectral densities at selected circuit nodes, assuming the current flowing through them to be fixed. However, since RD is actually a static phenomenon, these outputs correspond to variances and

correlation coefficients.

5.3.1 Variability estimation technique

To evaluate RD-induced variability in FinFETs, the following linearized system is considered:

$$\begin{cases} V_d = V_d^{(0)} + v_d \\ V_g = V_g^{(0)} + v_g \end{cases} , \begin{cases} I_d = I_d^{(0)} + i_d \\ I_g = I_g^{(0)} + i_g \end{cases} \tag{5.11}$$

In (5.11), the superscript (0) indicates the fixed operating point, d and g are the drain and gate nodes, and the small voltage and current signals (v_x, i_x) are the sum of perturbations $(\delta v_x, \delta i_x)$ induced by dopant fluctuations and corresponding circuit responses $(v_x^{(s)}, i_x^{(s)})$ which allow satisfying boundary conditions:

$$\begin{cases} v_x = \delta v_x + v_x^{(s)} \\ i_x = \delta i_x + i_x^{(s)} \end{cases} , \; x = \{d, g\} \tag{5.12}$$

Under the assumption of linearity, the admittance matrix formalism can be used for these responses:

$$\begin{pmatrix} i_d^{(s)} \\ i_g^{(s)} \end{pmatrix} = \begin{pmatrix} Y_{dd} & Y_{dg} \\ Y_{gd} & Y_{gg} \end{pmatrix} \begin{pmatrix} v_d^{(s)} \\ v_g^{(s)} \end{pmatrix} \tag{5.13}$$

Gate voltage fluctuations are evaluated by prescribing a fixed voltage and current at the drain port ($v_d = 0$, $i_d = 0$). Moreover, half the RD-induced fluctuations must be prescribed in order to get a unique solution: the choice is to set all current fluctuations to zero ($\delta i_d = 0$). This yields:

$$\begin{cases} v_d^{(s)} &= -\delta v_d \\ v_g^{(s)} &= \frac{Y_{dd}}{Y_{dg}} \delta v_d \end{cases} \tag{5.14}$$

Statistics is then applied, with the assumption of zero-mean fluctuations of the doping concentration, i.e. $\langle C \rangle = 0$, which also implies $\langle v_x \rangle = 0$ in the linearized regime. Therefore, the standard deviation of the gate voltage is $\sigma[V_g] = \sqrt{\langle v_g^2 \rangle}$. Exploiting (5.12) and (5.14), this results in:

$$\sigma[V_g] = \sqrt{S_V^{gg} + 2\frac{Y_{dd}}{Y_{dg}} S_V^{gd} + \frac{Y_{dd}^2}{Y_{dg}^2} S_V^{dd}} \tag{5.15}$$

Eq. (5.15) allows computing random-dopant-induced gate voltage fluctuations directly in terms of variances and correlation coefficients provided by the solver:

$$S_V^{xx} = \langle (\delta v_x)^2 \rangle \ , \ S_V^{xy} = \langle \delta v_x \delta v_y \rangle \tag{5.16}$$

By assuming drain as the only port instead, I_d fluctuations can be computed by simply converting the noise voltage spectrum provided by the solver into a noise current spectrum through the admittance matrix:

$$\sigma[I_d] = \sqrt{Y_{dd} S_V^{dd} Y_{dd}^*} \tag{5.17}$$

Equations (5.15) and (5.17) [snw07, tnano] will be used to compute fluctuations of FinFET threshold voltage and on-current, respectively, including models that describe the impact of RD on mobility [9].

Although not as rigorous as an atomistic approach, the impedance field method was shown to provide meaningful results: in literature, this method was tested down to 100 nm gate lengths [48], but similar perturbation approaches were applied to calculate V_T fluctuations of ultra-small devices featuring 50 nm [88, 89] and down to 25 nm channel lengths [49]. Validation versus Monte Carlo simulations was performed in these papers. Remarkable advantages of the adopted technique are computational efficiency and applicability to the same mesh and device models used for other simulations. Therefore, it allows for a direct comparison of random dopant fluctuations and line-edge roughness contributions to FinFET variability, as will be shown in Part IV, Chapter 7.

Part IV

Applications - *TCAD magnifying glass*

"It is the mark of an educated mind
to rest satisfied with the degree of precision
which the nature of the subject admits
and not to seek exactness
where only an approximation is possible."

Aristotle

Adaptive meshing approaches as well as variability estimation techniques help with tackling multidimensional TCAD problems which model complex real-world applications. An accurate yet efficient discrete representation of the internal device behavior is essential to achieving the desired physical insight in TCAD simulations. This allows analyzing and improving current technology as well as designing new device generations and alternative architectures. A statistical approach to TCAD is especially needed when predictive simulations of ultra-small devices subject to significant process variations are performed.

Representative applications of the techniques described in Part III will now be illustrated. In Chapter 6, the Wavelet-based Adaptive Method for mesh refinement will be tested on a set of device structures including challenging geometries, in order to evaluate effectiveness and performance of the proposed approach. Variability estimation techniques will be applied in Chapter 7 to assess feasibility of FinFET as an alternative device architecture for technology nodes of immediate interest.

Chapter 6

Accurate physical insight through adaptive meshing

A set of significative 2D and 3D device geometries has been chosen to validate the WAM algorithm described in Chapter. 4. The test set includes both simple structures used to monitor the behavior of the proposed approach and complicated geometries that challenge its capabilities and effectiveness when dealing with more realistic situations. In all cases, drift-diffusion simulations have been performed including SRH recombination, Masetti, Canali and Lombardi models for mobility as well as avalanche generation when appropriate (see Sec. 1.1). Impact of the mesh quality check, threshold choice and other numerical aspects have been studied. To evaluate the accuracy of WAM-based simulation results, reference meshes have been manually constructed, imposing very small grid spacings in all potentially relevant domain regions for the considered applications.

6.1 2D simulations

Several simulations have been performed to validate the 2D refinement algorithm presented in Sec. 4.5.3 [essderc05, sse06]. Reported here are results related to a power diode and to a 0.18μm n-channel MOS transistor, represented in Figs. 6.1(a) and (b), respectively. The geometry and doping of these devices are described in Table 6.1. In all test cases,

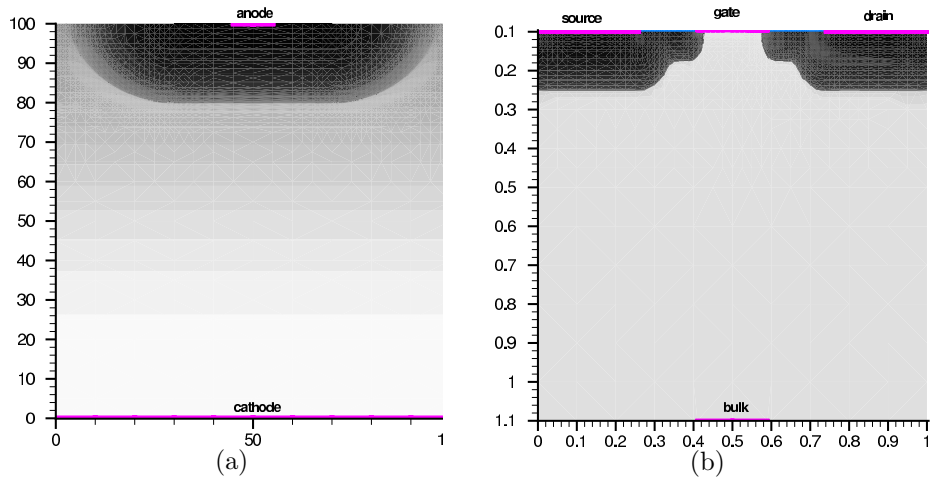


Figure 6.1: Simulated 2D diode (a) and MOSFET (b).

***p-n* junction**			**MOSFET**		
Sim. area	=	$100\mu\text{m}\times100\mu\text{m}$	Sim. area	=	$1\mu\text{m}\times1\mu\text{m}$
			L_g	=	$0.18\mu\text{m}$
			t_{ox}	=	4nm
$N_{D,peak}$	=	$5\times10^{19}\text{cm}^{-3}$	$N_{D,peak}$	=	$5\times10^{18}\text{cm}^{-3}$
$N_{A,peak}$	=	$5\times10^{19}\text{cm}^{-3}$	N_A	=	$3\times10^{15}\text{cm}^{-3}$

Table 6.1: Simulated 2D diode and MOSFET: device description.

"blind" input grids, i.e. coarse uniform grids of about 50 points, are provided by the user and up to 7 refinement cycles are performed to obtain the final grids. Wavelet analysis is performed on the electrostatic potential ψ and carrier concentrations n, p.

Fig. 6.2 shows the simulated IV characteristics of the power diode in both forward and reverse bias. A doping profile with curved junction has been chosen to evaluate the anisotropic capabilities of the refinement algorithm. WAM data in Fig. 6.2 have been obtained through a dynamical mesh adaptation at each bias step according to the scheme in Fig. 4.23 , Sec. 4.7.2, resulting in grid sizes of $1,000$ to $2,300$ points from forward to reverse bias. Simulation results show a good accuracy of the adaptive scheme when compared to a reference fixed mesh with

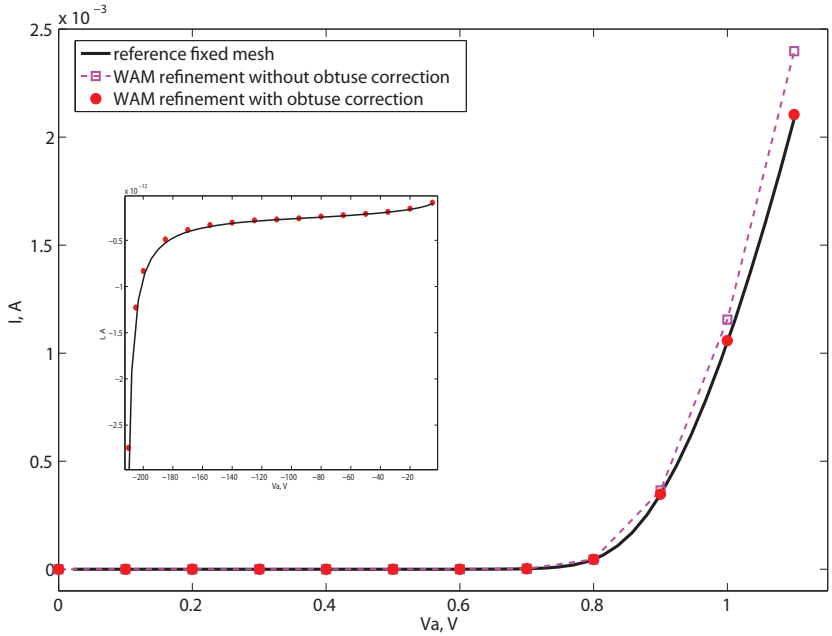
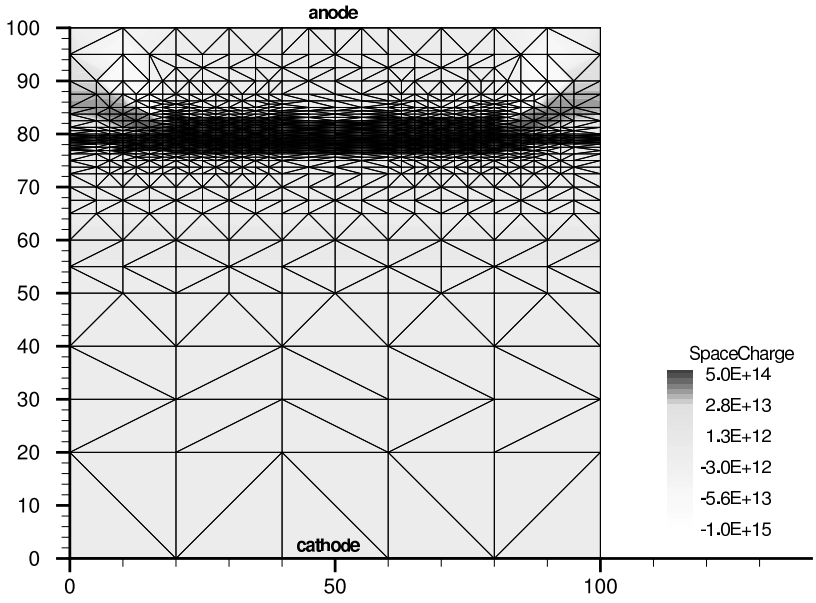
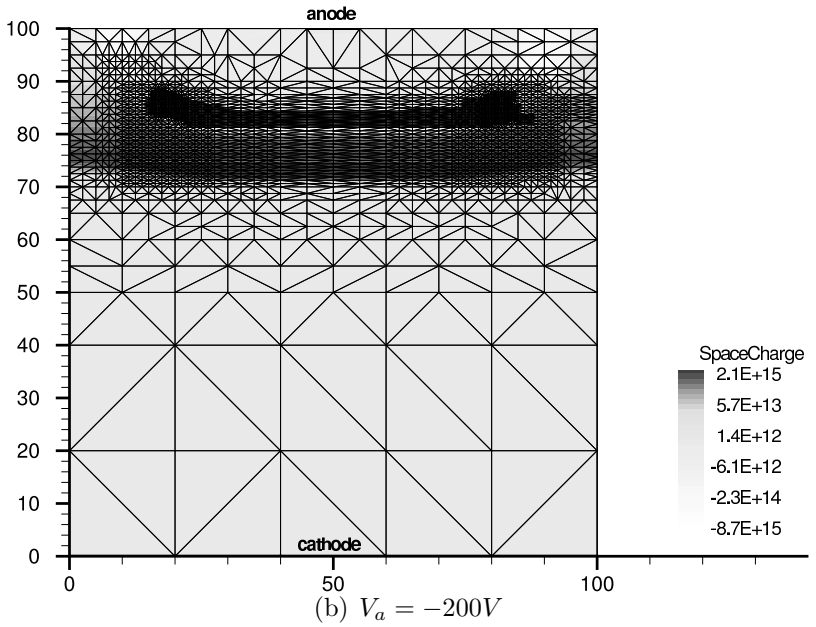


Figure 6.2: Comparison of I-V curves for the simulated 2D silicon p-n diode with curved junction. The WAM refinement provides a good match with reference characteristics when combined with the obtuse triangle correction: this step is essential to ensure accuracy and even to achieve convergence in the reverse bias.

10,000 vertices, while providing a considerable saving in the number of points and preserving the characteristics smoothness within the whole interval of considered anode voltages V_a. It is worth noticing that the obtuse triangle correction described in Sec. 4.6.1 is essential to simulation accuracy. A poor match with reference results is observed in the forward bias when the `VERIFY OBT` module is disabled (see Fig. 6.2); in the reverse bias, even convergence is definitely compromised in such case. Two of the meshes generated by the automatic algorithm at different bias steps of the breakdown simulation are reported in Figs. 6.3(a) and (b). Different refinement levels are clearly visible within the domain. Resolution is especially increased in space charge regions, i.e. where interesting phenomena occur. Spatial variations of these regions due to bias changes are correctly tracked, as can be seen by comparing





(b) $V_a = -200V$

Figure 6.3: WAM meshes for a 2D *p*-*n* junction breakdown simulation.

the considered meshes.

The Wavelet-based refinement strategy has been also successfully

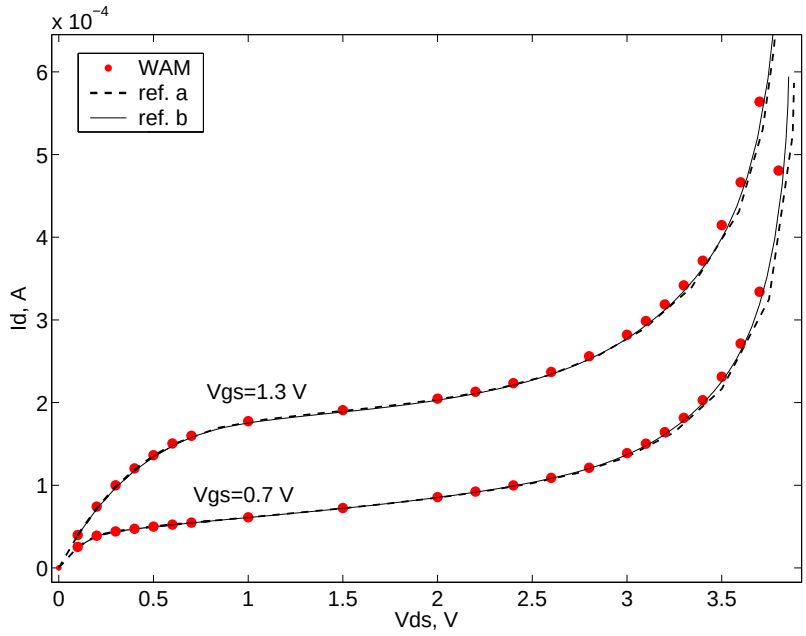
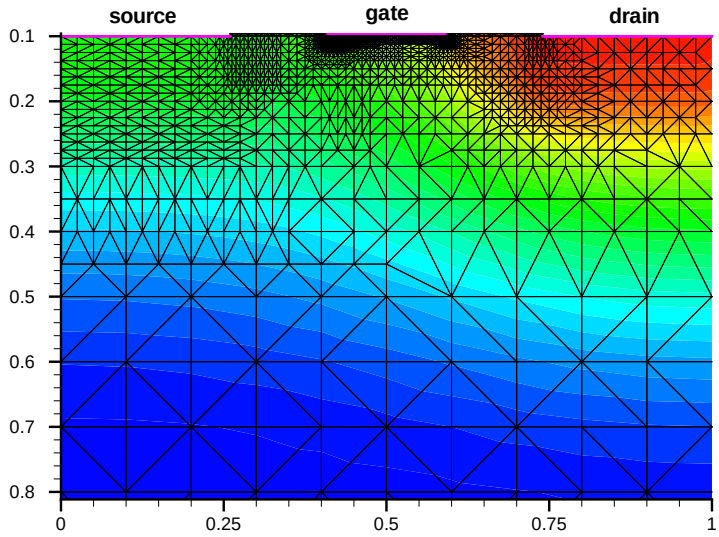


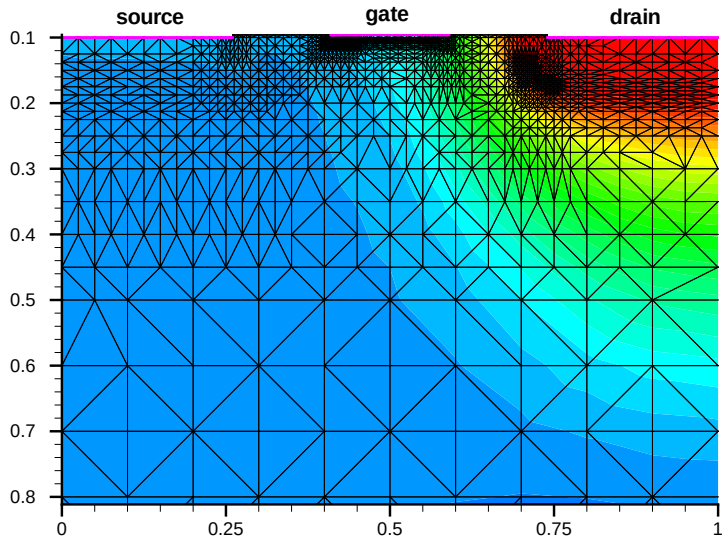
Figure 6.4: nMOSFET $I_d(V_{ds})$ characteristics ($V_{gs} = 0.7V$, $V_{gs} = 1.3V$). "ref. a" and "ref b" are the results obtained with two reference fixed meshes (5,000 and 10,000 points, respectively), while WAM data have been produced by the dynamical mesh adaptation (about 1,600 to 1,900 nodes).

tested on MOSFET structures, both in case of transcharacteristic and output characteristic simulations. An example is provided by the $I_d(V_{ds})$ curves shown in Fig. 6.4 for an n-channel device. Even in this case the automatic grid adaptation provides a reasonable accuracy when compared to reference static meshes, while strongly reducing the grid size (about 1,600 to 1,900 nodes vs. 5,000 and 10,000 points for "ref. a" and "ref b", respectively). Finally, Figs. 6.5(a) and (b) report two different meshes generated with WAM during one of the sweep simulations described above, in the linear and avalanche regime, respectively. It is important to notice:

- how the adaptive strategy accurately meshes the regions which have stringent requirements for impact ionization current calculation,



(a) $V_{ds} = 0.7V$, 1689 grid points



(b) $V_{ds} = 3.4V$, 1875 grid points

Figure 6.5: nMOSFET $I_d(V_{ds})$ simulation with $V_{gs} = 1.3V$.

- the total absence of obtuse triangles,
- the smooth grading of mesh elements.

These examples highlight the efficiency of the Wavelet-based algorithm in terms of both grid size (about 20% lower than data reported in [30] for MOSFET avalanche simulation) and time saving, particularly with respect to a non-automated simulation flow, in which a skilled user has to define a fixed grid with a considerably larger number of nodes in order to ensure accuracy throughout sweep simulations. This is even more relevant in three dimensions, as shown in the next Section.

6.2 3D simulations

Three-dimensional test devices including diodes, power nMOS drivers with different geometries and a FinFET structure have been simulated using the modified 3D WAM algorithm described in Sec. 4.5.4 [sispad06, tcad07]. Sample meshes produced by this algorithm for the considered structures are reported in Fig. 6.6.

The p-n diode in Fig. 6.6(a) provides a simple but useful test case to highlight the 3D anisotropic capabilities of WAM, evaluating effectiveness of strategy improvements described in Sec. 4.5.4. To this aim, Fig. 6.7 shows details of the meshes produced by (a) an isotropic refinement of Wavelet supports, (b) the naive 3D extension of the algorithm presented in Sec. 4.5.3, and (c) the modified 3D approach, using the same threshold on Wavelet coefficients. The isotropic refinement is clearly impractical, especially for 3D domains, while both anisotropic strategies correctly follow the junction shape, increasing the resolution in the required directions, with smooth transitions along the profile corners. However, the improved algorithm ensures better selectivity properties, accurately capturing domain regions where the relevant physical phenomena take place and allowing for a considerable saving in the number of mesh points.

Fig. 6.8 shows impact of the three refinement strategies on mesh size for devices (a), (b) and (d) in Fig. 6.6 as a function of the Wavelet anal-

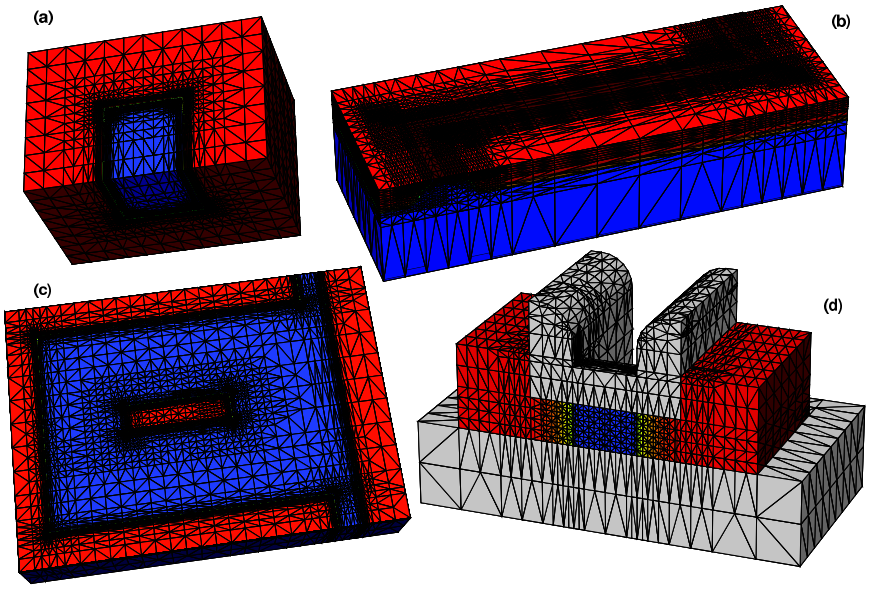


Figure 6.6: 3D WAM anisotropic refinement of four different devices: (a) a 3D p-n diode, (b) and (c) power nMOS drivers, and (d) a FinFET device.

ysis level. Infeasibility of the isotropic approach is confirmed by these graphs, which also show an average reduction in the number of grid points by about 40% with the improved strategy, while no significant loss of accuracy in contact current was observed.

Figs. 6.6(b) and (c) show the meshes for two different n-channel MOSFET drivers; these device structures were selected to test the refiner behavior when dealing with complex geometries and several-tens-of-microns-large domains. Such figures illustrate the good selectivity and quality features of WAM meshes, which are also evident from the magnified view of Fig. 6.9.

Even in the 3D case, WAM allows for a dynamical grid adaptation at each bias step during sweep simulations, including both refinement and coarsening. As illustrated in Fig. 6.10, WAM-based calculated values show a good match with those obtained by using a reference fixed mesh, while the computational cost of adaptive simulations is significantly reduced by the smaller grid size. As in the 2D case, no numerical artifacts are seen in the I-V curves, i.e. smoothness is ensured. Exam-

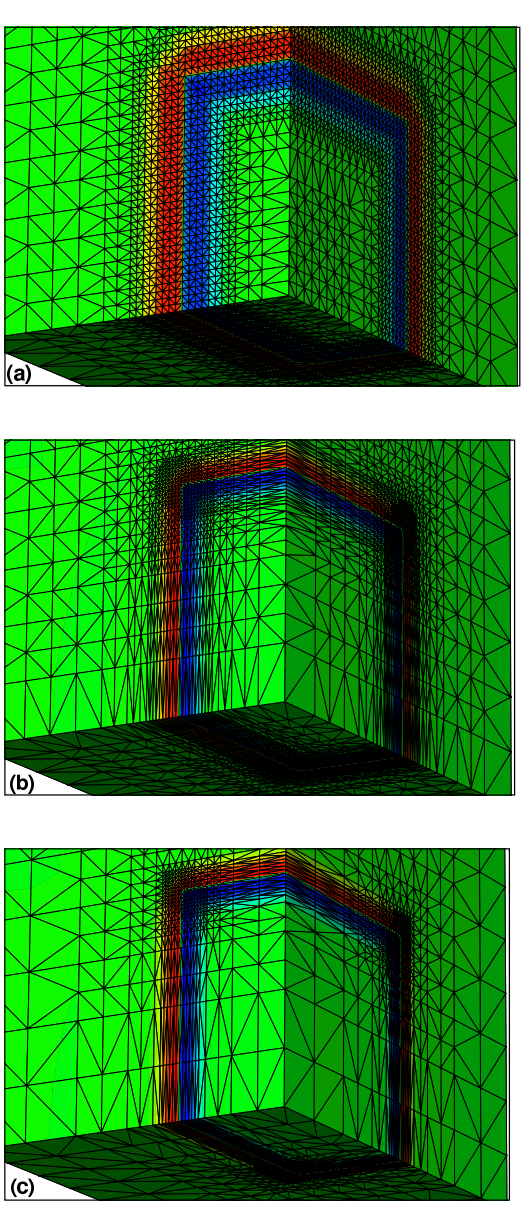


Figure 6.7: Mesh refinement of the p-n junction shown in Fig. 6.6(a) through (a) an isotropic approach, (b) the naive 3D extension of the WAM technique described in Sec. 4.5.3, and (c) the modified 3D WAM approach presented in Sec. 4.5.4. The same value of threshold η on Wavelet coefficients has been used in all three cases.

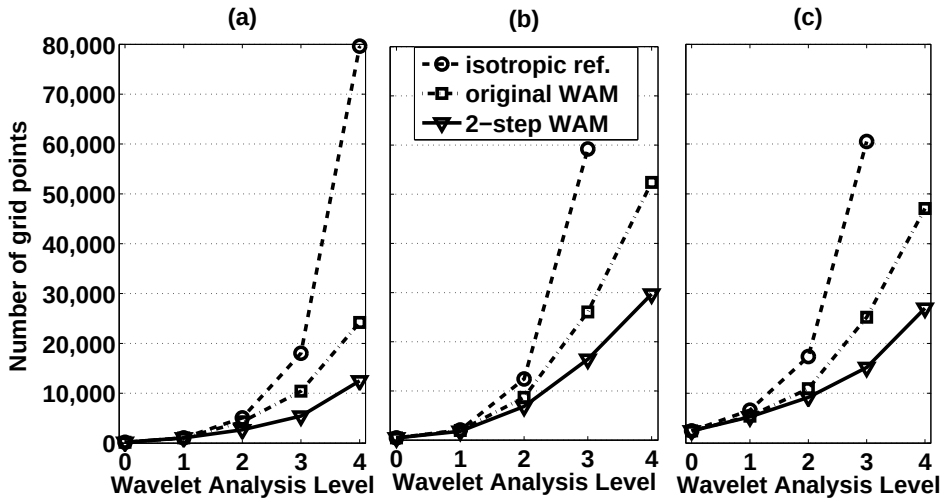


Figure 6.8: Impact of different refinement strategies on mesh size at various levels of Wavelet analysis for (a) the *pn* junction, (b) the MOSFET driver of Fig. 6.6(b), and (c) the FinFET device.

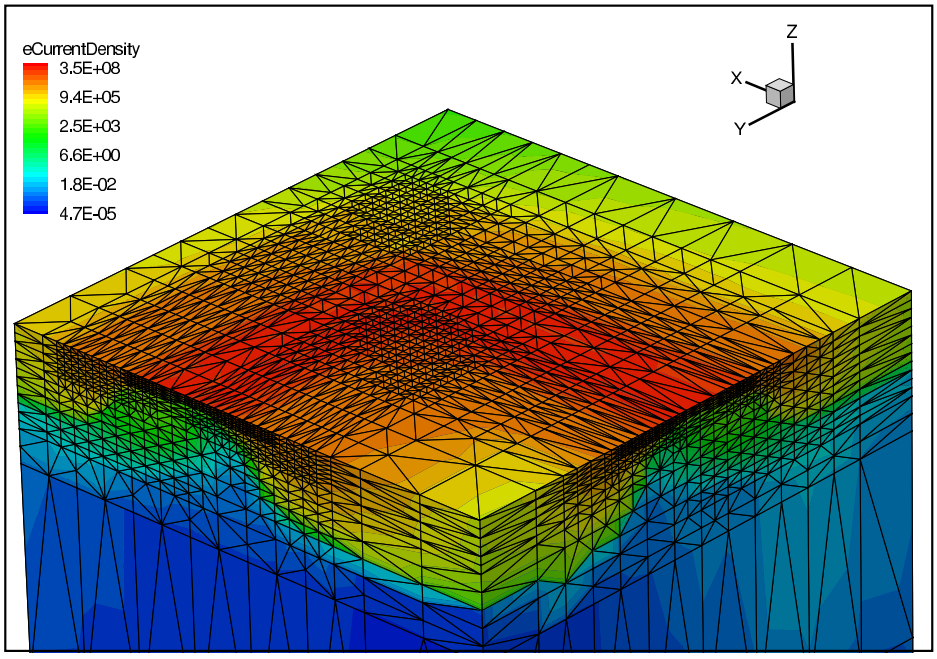


Figure 6.9: Magnified view of mesh details for the MOSFET driver in Fig. 6.6(b). Here, electron current density resulting from a simulation step at $Vgs = 1.3V$, $Vds = 1.78V$ is displayed.

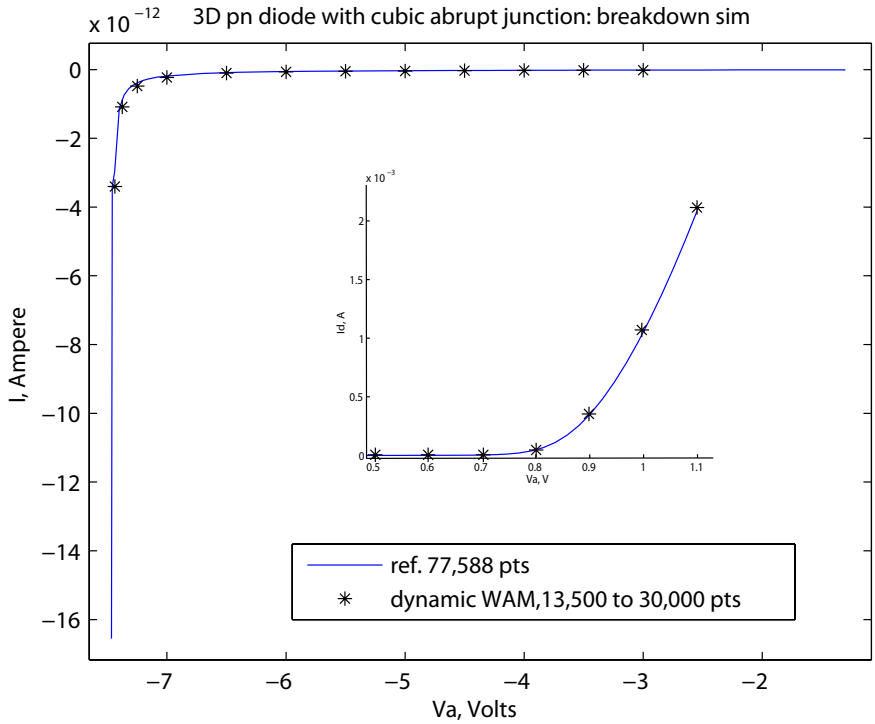


Figure 6.10: Comparison of IV simulations with WAM (stars) and a reference (solid line) fixed mesh for a the 3D p-n diode. WAM is launched with a fully adaptive mesh strategy i.e. adapting the mesh at each bias step.

ples of dynamical mesh adaptation are shown in Fig. 6.11.

Finally, some 3D simulation results related to the FinFET device in Fig. 6.6(d) are discussed. The device geometry is as follows: fin length $L_{FIN} = 70$ nm, fin height $H_{FIN} = 40$ nm, fin width $W_{FIN} = 20$ nm, gate length $L_G = 35$ nm, oxide thickness $T_{OX} = 2$ nm. The doping profiles are assumed to be abrupt. No process variations are considered here despite the small device size: the impact of non-idealities on similar architectures will be analyzed in Chapter 7. Due to the device symmetry along the current-flow direction, only the halved structure was simulated. The test example starts with an initial coarse grid of about $2,400$ points. Due to domain complexity, here three refinement boxes, corresponding to the fin, source and drain regions, are defined

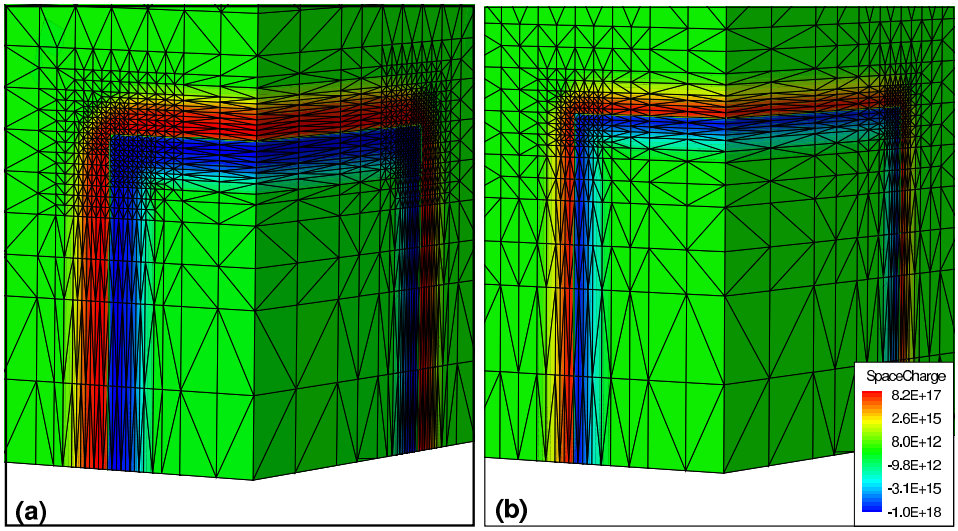


Figure 6.11: Meshes produced by WAM during the sweep simulation reported in Fig. 6.10: (a) $Va = -7.375V$, (b) $Va = 0.1V$.

at the `INIT info` step of the simulation flow (see Fig. 4.23). After the initialization phase, the solve-refinement loop goes on without any further control from the operator. In the considered test case, up to 6 refinement cycles have been performed, resulting in meshes similar to those of Fig. 6.12 and Fig. 6.13: as expected, most grid points are located where relevant physical quantities undergo sharp variations, i.e. in the channel regions. In particular, the refinement strategy correctly captures the anisotropic nature of the device, imposing a finer spacing in the direction perpendicular to the channel. The accuracy of WAM-generated meshes in comparison with reference discretizations is further confirmed by the *Id-Vgs* curve reported in Fig. 6.14.

6.3 Mesh quality

The quality check module plays an essential role in simulation results presented so far. Complete absence of obtuse angles and smooth grading of element sizes are clear benefits produced by the `VERIFY OBT` module in all 2D meshes reported in Sec. 6.1. Nicely shaped and well graded elements are also qualitatively visible in 3D meshes (see

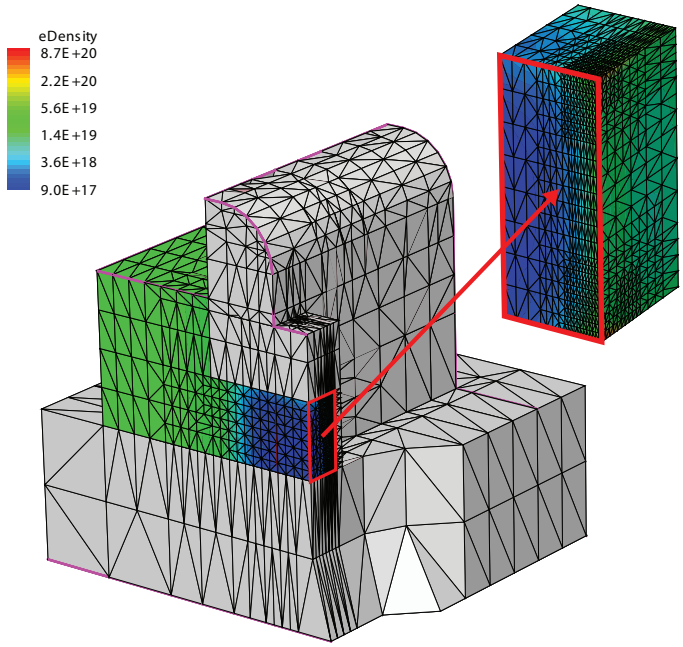


Figure 6.12: Details of the mesh generated by WAM for the 3D FinFET test structure in Fig. 6.6(d).

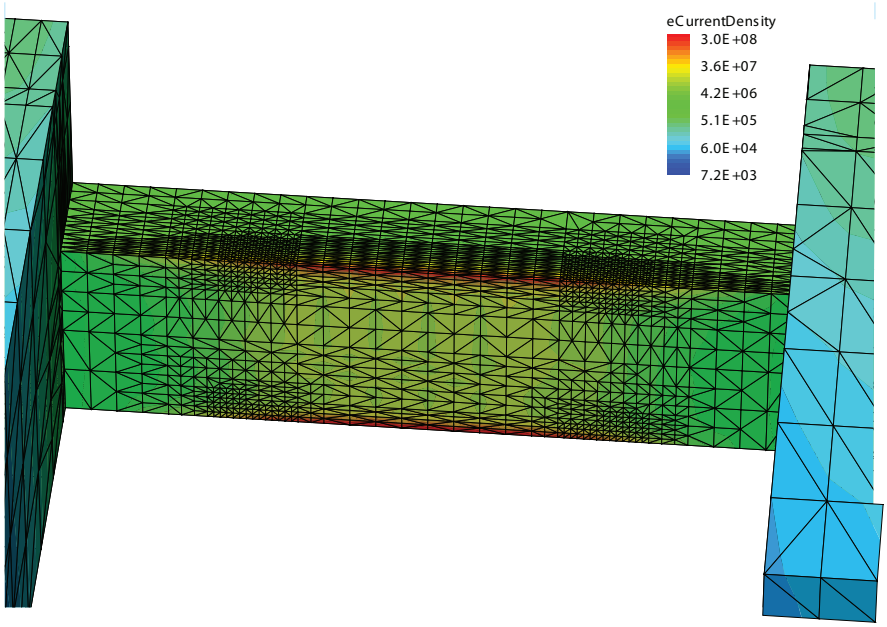


Figure 6.13: Mesh zoom in the FinFET channel region.

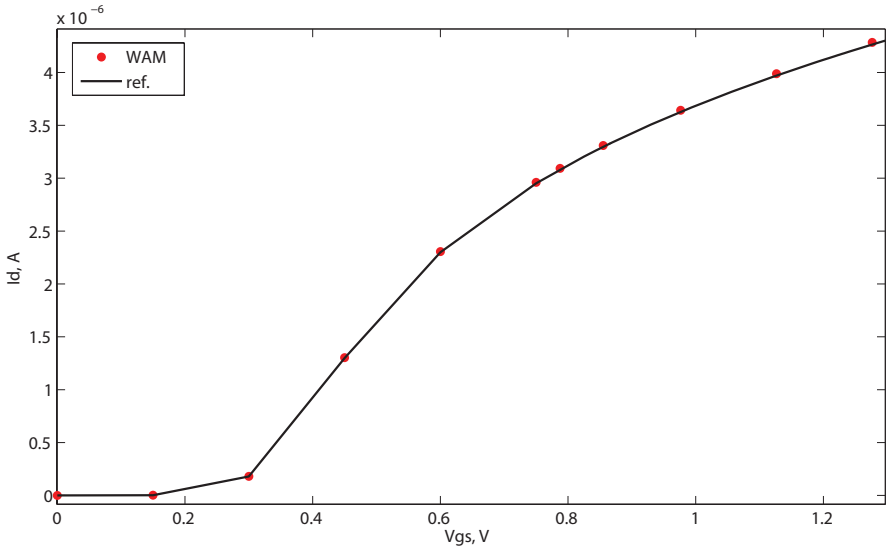


Figure 6.14: Comparison of Id-Vgs curves for the test structure at $Vds = 0.05V$. The WAM-generated mesh (about $17,700$ points) provides a good match with the results obtained with the reference mesh (about $47,500$ points).

Sec. 6.2). In addition, improvement of the mesh smoothness due to the 3D correction algorithm has been *quantitatively* evaluated in terms of volume ratio of adjacent elements and vertex connectivity (number of elements with a given common vertex). The maximum value of these figures of merit is significantly reduced by the correction procedure, as illustrated in Fig. 6.15. In turn, mesh quality improvement is largely beneficial to convergence and accuracy of the solver: masking of the verification routine was seen to cause large errors in contact current and even convergence failure at critical bias conditions in both 2D (see Fig. 6.2) and 3D simulations.

Moreover, in complicated geometries such as the FinFET in Fig. 6.6(d), correction of the badly-shaped elements also provides coherence between the refined silicon domain and surrounding oxide and nitride regions, ensuring *global* quality of the mesh: again, this has a relevant impact on solver convergence.

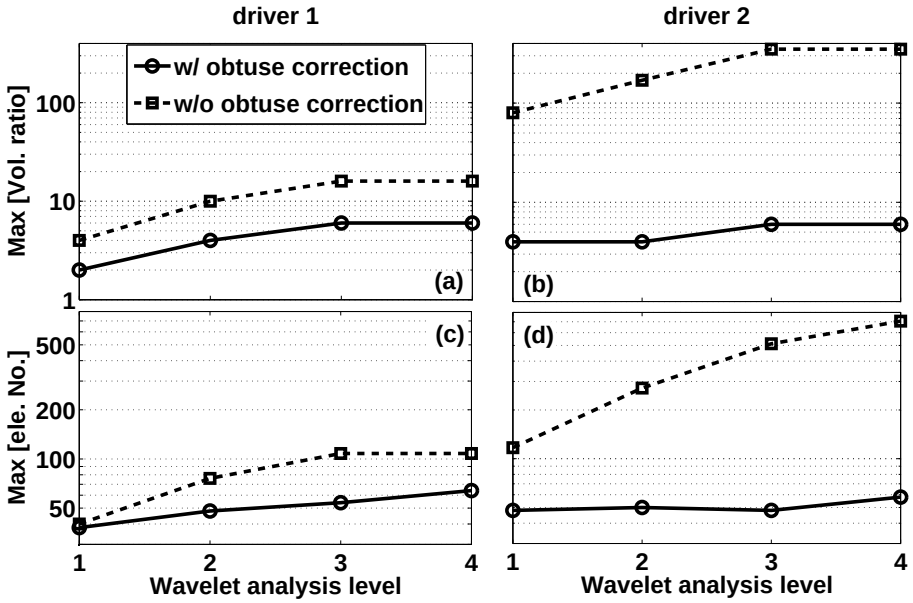


Figure 6.15: Mesh quality in terms of maximum volume ratio of adjacent elements (a), (b) and maximum number of elements with a common node (c), (d) for the two drivers in Figs. 6.6(b) and (c) (here indicated as "driver 1" and "driver 2", respectively).

6.4 Numerical considerations

The effectiveness of WAM approach is influenced by the choice of the threshold η which discriminates relevant Wavelet coefficients. The impact of such a choice can be evaluated by comparing WAM-based results with a reference solution through a suitable relative error, defined on a considered quantity Q (e.g. drain current or electric field) as

$$e_r = \frac{\|Q_\eta - Q_{ref}\|_2}{\|Q_{ref}\|_2}$$

where Q_{ref} and Q_η are the quantities computed on the reference grid and on the adaptive grid obtained with threshold η, respectively. As an example, Fig. 6.16 shows the relative error e_r on drain current I_d, calculated at various refinement levels, versus the number of grid points, for different thresholds but fixed bias conditions. The error was calculated with respect to the solution obtained with an extremely refined reference mesh for the 2D MOSFET considered in Sec. 6.1. The higher

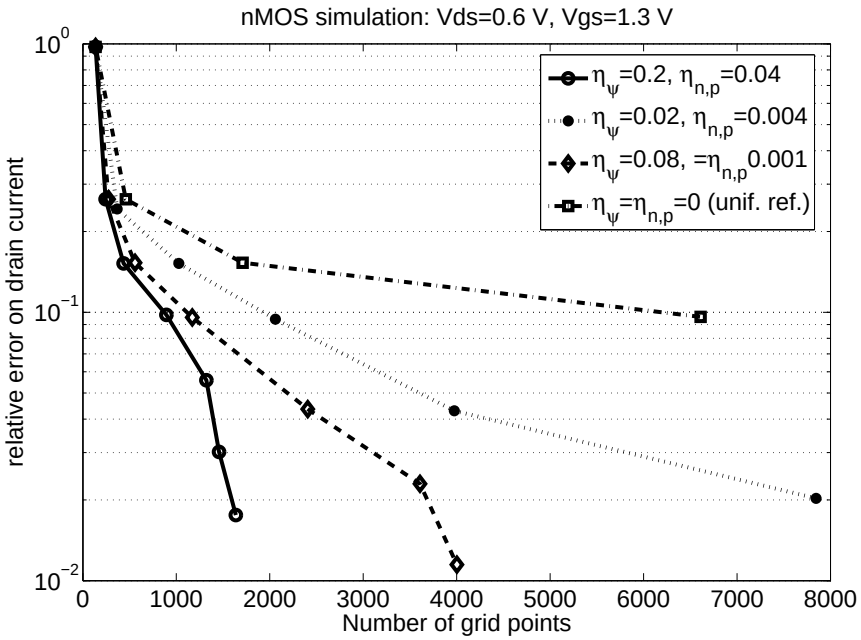


Figure 6.16: Example of threshold influence on accuracy versus number of nodes for drain current in a 2D n-channel MOS. η_ψ and $\eta_{n,p}$ are thresholds on electrostatic potential and carrier concentrations, respectively. Threshold values are given in relative terms (see Sec. 4.7.2).

the threshold, the smaller the mesh size, and therefore the lower the computational cost. However, very high thresholds lead to an unbearable accuracy degradation. Similar results have been obtained in the 3D case, as shown in Fig. 6.17(a) for the p-n diode discussed in Sec. 6.2. Again, a good choice allows for a great saving in the number of nodes, while providing the same degree of accuracy, as clearly visible from Fig. 6.17(b). The same trend is observed at all the successive levels of the Wavelet analysis, thus allowing to *perform the threshold selection at the first level*, when the mesh size is still very small. Fig. 6.17(b) also provides an estimate of decrease in the discretization error as an effect of the regularity-estimation-based resolution increase. As far as the dependence on bias conditions is concerned, it has been verified that keeping the threshold value tied to the applied voltage allows controlling the accuracy-efficiency trade-off throughout the simulation sweep.

Moreover, a *gradual* evolution of the mesh is produced by the dy-

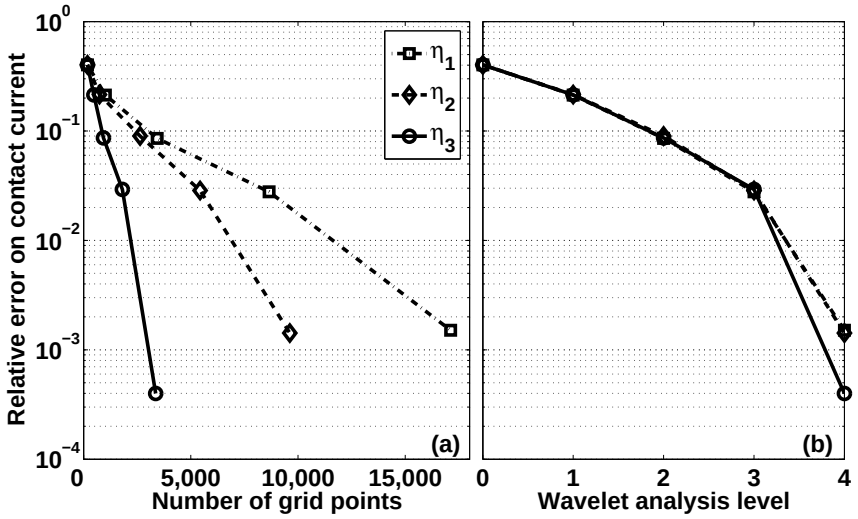


Figure 6.17: Influence of the threshold value on number of nodes (a) and accuracy (b) for a 3D p-n diode simulation ($\eta_1 < \eta_2 < \eta_3$). An extremely refined reference mesh was used to compute errors.

namical adaptation strategy in combination with the mesh quality control, i.e. no strong changes in node number/location are seen between the computational grids associated to two consecutive bias steps. The maximum resolution is also fixed throughout the simulation, as explained in Sec. 4.5.5. The combination of these features leads to the following advantages.

a) Solution recomputation on the adapted grid is not a very challenging issue.

b) Simulations exhibit good convergence trends: extensive numerical tests suggest that convergence is generally reached without the need to implement onerous tasks (such as the nonlinear node block Jacobi iteration or homotopy techniques [9]), even for numerically challenging problems such as breakdown or snap-back simulations.

c) Smooth and physically meaningful I-V curves are obtained, as shown in the previous Sections, whereas a common drawback

of several adaptive strategies is the presence of numerical artifacts such as abrupt variations between successive bias points (see Sec. 2.3).

Finally, due to the fast algorithms which calculate the discrete transform and the limited memory occupancy required by the refinement routine (see Sec. 4.7.1), the use of Wavelet coefficients for grid adaptation is computationally favorable: in all performed tests, the amount of time consumed by the `WAM` module (see Fig. 4.23) was at least two orders of magnitude lower than simulation time. The Wavelet-based Adaptive Method is therefore an effective technique to deal with the increasing complexity and dimensionality of TCAD problems: it exhibits a remarkable efficiency when compared to other automatic adaptation strategies proposed in literature [25, 34, 90] and it relieves the user from a difficult and time-consuming task.

Chapter 7

Impact of variability on future technology generations

The fundamental importance of TCAD in the development of new technology generations is related to its predictive capabilities. These are essential to evaluate effectiveness of innovative materials and process options and performance of new device structures. However, feasibility of these innovations is increasingly opposed by enhanced process variations, which must be properly accounted for. Statistical approaches to handle variability at a reasonable computational cost in TCAD simulations have been discussed in Chapter. 5. These techniques will now be applied to evaluate potentialities, matching performance and scalability of a promising alternative to bulk CMOS technology, i.e. the FinFET architecture illustrated in the Introduction. Line-edge roughness and random dopant fluctuation issues will be investigated for this device, taking into account the impact of several process options, such as fin patterning and doping profiles [iedm06, ted07, snw07, tnano]. FinFET-based SRAM circuits will also be studied to assess variability requirements for mainstream applications of the considered technology. Simulations illustrated in this Chapter have been performed with a hydrodynamic model including density-gradient approximation for quantum confinement, as explained in Sec. 5.1.3, where the extraction procedure for representative electrical parameters is also described. Mobility

W_{fin}	L_{gate}	t_{ox}
25 nm	60 nm	1.5 nm
19 nm	45 nm	1.5 nm
17 nm	40 nm	1.5 nm
12 nm	30 nm	1.5 nm

Table 7.1: Simulated device geometries ($W_{fin} \simeq 0.42 \times L_{gate}$)

degradation due to normal electric field, high-field velocity saturation and carrier tunneling through the potential barrier at the source have also been considered.

7.1 Impact of LER on FinFET scaling in RDF and SDF technologies

Four different FinFET geometries have been considered to study how line-edge roughness affects scalability of this architecture. Dimensions of the simulated n-channel devices are reported in Table 7.1. These FinFETs are ideal except for the random fin-/gate-LER; in particular, ideal box-shaped doping profiles have been used. Simulations have been performed on lightly doped fins (10^{15} cm^{-3}) by adjusting the threshold voltage with a gate work function of 4.62 eV.

In FinFET flows, Si fins are commonly patterned using conventional resist-based processes: this results in random uncorrelated roughness on the fin sidewalls. However, an alternative to resist-defined fin (RDF) patterning has been proposed [91], which is based on the use of dummy spacers and achieves higher fin density as compared to RDF. From the LER standpoint, this patterning process results in an ideally in-phase correlation between the edges of spacer-defined fins (SDF). Fig. 7.1 [iedm06] shows top-down SEM images of resist- and spacer-defined fins.

In this study, LER has been generated through a Gaussian autocorrelation function with rms amplitude $\Delta = 1.5$ nm and correlation length $\Lambda = 20$ nm (see Secs. 5.1.1 and 5.1.2). Values for the statisti-

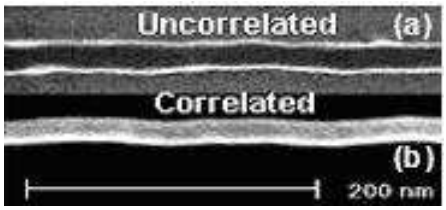


Figure 7.1: SEM image of a Si-fin with (a) uncorrelated and (b) correlated LERs, corresponding to resist- and spacer-defined fin patterning, respectively (IMEC data).

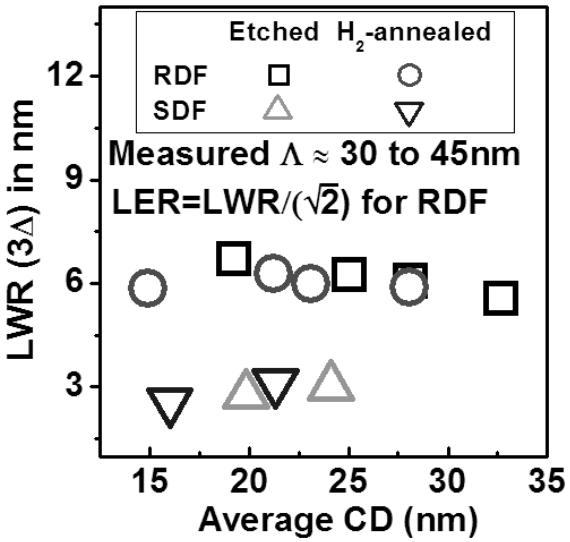


Figure 7.2: Line-width roughness (LWR) measurements for resist- and spacer-defined fins (IMEC data).

cal parameters have been chosen based on experimental measurements reported in Fig. 7.2 [iedm06][1].

Impact of LER contributions from the fin and top-gate of devices in Table 7.1 has been estimated by simulating each component separately. A 2D approximation was judged to be sufficient for the purpose of this Section. RDF and SDF technologies have been compared by assuming an ideal in-phase correlation for the latter case, whereas the

[1]Correlation lengths suggested by these measurements are higher than 20 nm. However, extraction of such parameter is a difficult task: the chosen value is within the typical range of 10-50 nm reported in literature [52].

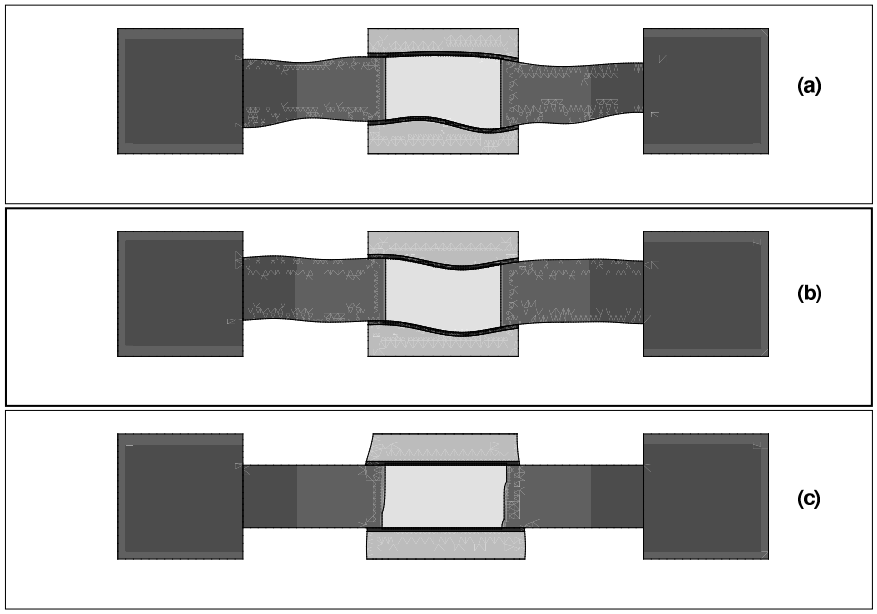


Figure 7.3: Instances of simulated FinFETs affected by fin-LER without (a) and with (b) phase correlation and by gate-LER (c). Nominal device dimensions are $W_{fin} = 25$ nm, $L_{gate} = 60$ nm.

fin edges are completely uncorrelated when resist patterning is considered: Figs. 7.3(a) and (b) show instances of the simulated devices in these two cases, respectively. Instead of fin-LER, an uncorrelated random gate-LER is applied to the device structure shown in Fig. 7.3(c). A statistical analysis of threshold voltage and current factor mismatch has been performed, using eq. (5.7), on ensembles of 200 devices to assess relative importance of the three cases considered in Fig. 7.3. Normal and Half-Normal statistics (see Sec. 5.2.2) have been compared and found to provide similar results. Therefore, a simple Normal fit has been applied to all distributions analyzed in this Chapter. Results of the considered case study are plotted in Fig. 7.4. The mismatch introduced by fin-LER with in-phase correlation is found to be much lower with respect to other contributions, which indicates SDF patterning as a promising technique to alleviate LER issues in FinFETs. When RDF technology is considered instead, the impact of fin-LER on matching performance is as significant as that of top-gate-LER. The overall mismatch due to line-edge roughness tends to increase rapidly at and below

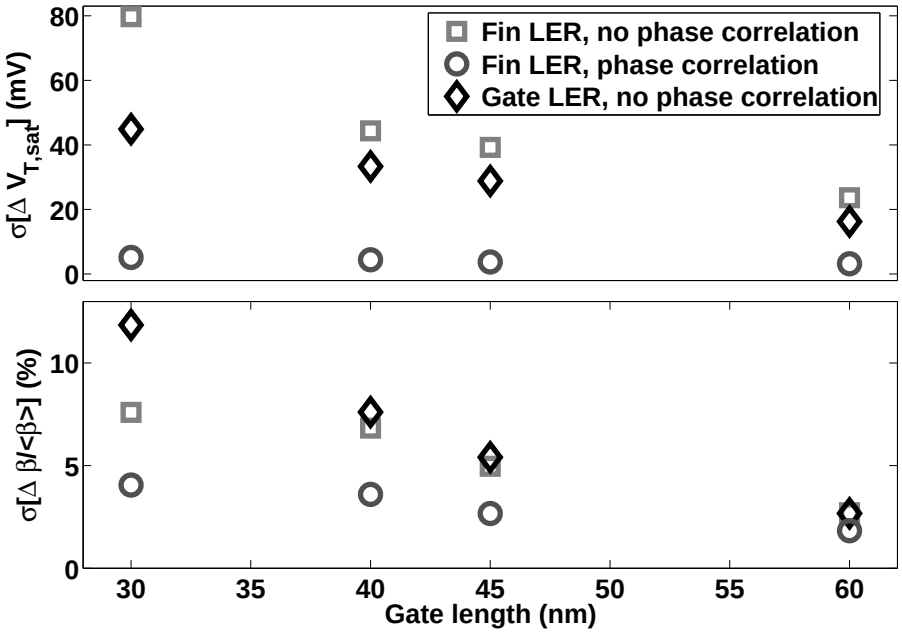


Figure 7.4: Independent contributions to mismatch in threshold voltage (top) and current factor (bottom) for the FinFET geometries shown in Table 7.1. Ensembles including about 200 devices were simulated. LER model: Gaussian autocorrelation function, $\Delta = 1.5$ nm, $\Lambda = 20$ nm.

45 nm gate length, especially for the current factor in Fig. 7.4, unless SDF technology is used to alleviate the problem. Intrinsic Transistor Performance (ITP) data resulting from the same simulations are plotted in Fig. 7.5: RDF-LER is seen to cause spread mainly in off-current, while top-gate-LER affects on-current more.

The FinFET matching performance discussed so far included only low spatial frequency components of LER through the Gaussian autocorrelation function, as explained in Sec. 5.1.1. Results obtained through this model are compared with those provided by the exponential autocorrelation LER model in Fig. 7.6. In addition to low frequency, exponential LER sequences include high frequency components too. However, while using this model for gate-LER, junction profile smoothening due to dopant diffusion must be considered. It can be seen from Fig. 7.6 that the two models mostly provide very similar results,

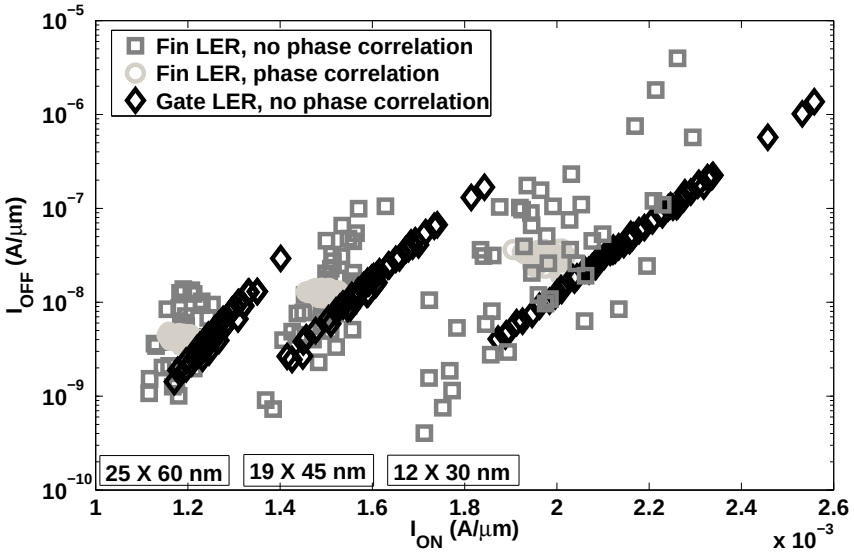


Figure 7.5: I_{OFF} vs. I_{ON} distributions for three of the four simulated geometries. Off-current was extracted at $V_{gs} = 0\ V$, $V_{ds} = 1\ V$ and on-current at $V_{gs} = V_{ds} = 1\ V$.

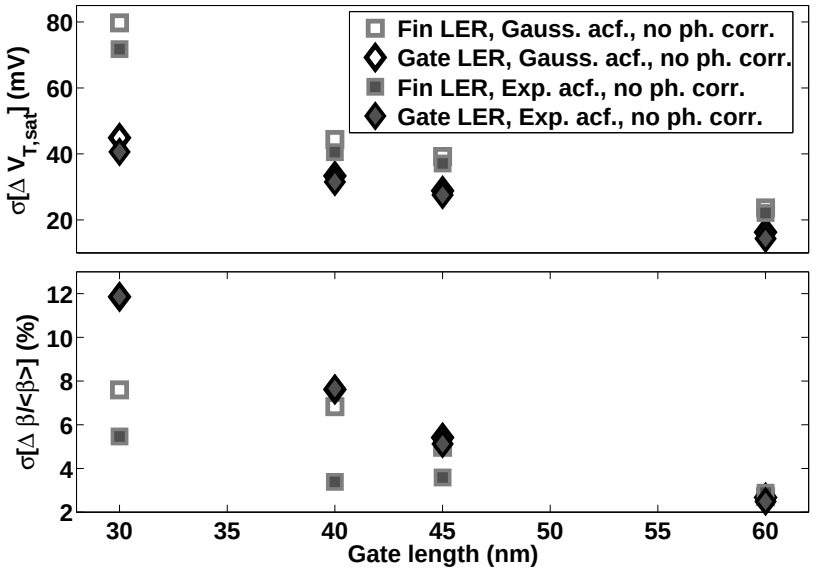


Figure 7.6: Comparison of mismatch contributions from LER generated through the Gaussian and the exponential models ($\Delta = 1.5$ nm, $\Lambda = 20$ nm, ensemble size=200).

with the only exception of the current factor mismatch when fin-LER is considered. However, in practice, surface smoothening processes using H_2 anneal in FinFET flows generally tend to suppress high frequency components of the fin-LER [92]. The absence of a visible degradation in the case of the exponential model in Fig. 7.6 indicates that the main contribution to mismatch comes from low frequency roughness. This fact can be explained in the following way: high frequency roughness components correspond to many peaks and valleys within the single device, resulting in a larger statistical ensemble of LER noise and hence in smaller deviations from the average electrical parameters with respect to low frequency LER, as also reported in [56] for planar bulk technology. The Gaussian model therefore provides sufficient accuracy for investigating the impact of LER on FinFET matching performance and will be used for all further simulations.

7.2 Impact of LER on LSTP-32 nm FinFET technology

As the influence of line-edge roughness has been predicted to become particularly severe for resist-defined fin devices with gate lengths smaller than 45 nm, a more accurate analysis has been performed on an aggressively scaled FinFET geometry. In order to conform to ITRS-ITP specifications for the LSTP-32 nm node, these devices are designed with nominal parameters as shown in Table 7.2 [5, 93], where N_{ch} and N_{hdd} are the channel and source/drain doping concentrations, while $N_{ext,n}$ and $N_{ext,p}$ represent the extension concentrations for n- and p-type devices, respectively. These doping parameters correspond to more realistic profiles with respect to those simulated in Sec. 7.1. In particulare, lower peak concentrations and less abrupt junctions are considered here, resulting in increased S/D resistances. Previously used and new S/D profiles are shown in Fig. 7.7 for n-channel devices.

Geometry	Doping	Electrical parameters
$L_{gate} = 30$ nm	$N_{ch} = 1 \times 10^{15}\text{cm}^{-3}$	$I_{ON} = 750\ \mu\text{A}/\mu\text{m}$
$W_{fin} = 10$ nm	$N_{hdd} = 1 \times 10^{20}\text{cm}^{-3}$	$I_{OFF} = 10\ \text{pA}/\mu\text{m}$
$H_{fin} = 50$ nm	$N_{ext,n} = 5 \times 10^{18}\text{cm}^{-3}$	$V_{T,sat} = 0.36$ V
$t_{ox} = 1.2$ nm	$N_{ext,p} = 2 \times 10^{19}\text{cm}^{-3}$	$SSlope = 69$ mV/dec.

Table 7.2: LSTP-32 nm FinFET specifications

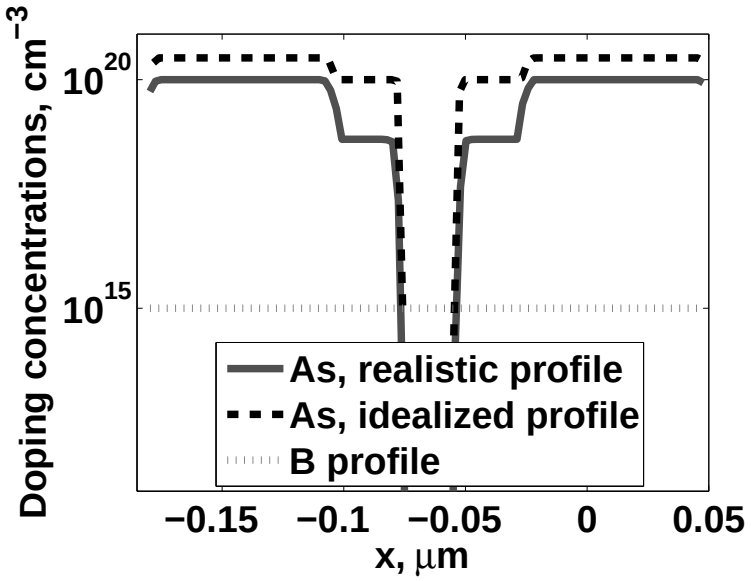


Figure 7.7: Doping profiles of the simulated n-type device (solid lines) compared with those considered in Sec.7.1 (dashed lines).

7.2.1 Mismatch contributions from the fin-, top- and sidewall-gate-LER

Impact of the fin- and top-gate-LER have been estimated separately in Sec. 7.1. However, by considering these contributions to mismatch as uncorrelated, the total net mismatch can also be estimated to a first order. To demonstrate this assertion, ensembles of 200 n-type FinFETs described in Table 7.2 have been simulated in presence of (i) fin-LER only (see Fig. 5.2(b)), (ii) top-gate-LER only (see Fig. 5.2(d)) and (iii) both the top-gate- and fin-LER. In Table 7.3, mismatch parameters obtained in the latter case are compared with those calculated by

Parameter	$\boldsymbol{\sigma_f}$	$\boldsymbol{\sigma_g}$	$\boldsymbol{\sigma_{f+g}}$	$\boldsymbol{\sqrt{\sigma_f^2 + \sigma_g^2}}$
$\Delta V_{T,sat}$ (mV)	28.70	8.15	28.71	29.83
$\Delta\beta/\langle\beta\rangle$ (%)	11.25	4.00	12.06	11.94

Table 7.3: Statistical dependencies of LER contributions to mismatch *(σ_f: rough fin, σ_g: rough top-gate, σ_{f+g}: combined fin- and top-gate-LER)*

combining the standard deviations of individual contributions (i) and (ii): resulting values of the net mismatch are in good agreement. This confirms that treating different LER components as independent contributions to the net stochastic mismatch provides a reasonably good approximation, while substantially reducing the simulation complexity.

In addition to fin- and top-gate-LER, the impact of sidewall-gate roughness has also been investigated for LSTP-32 nm FinFETs. A fully-3D device representation was needed to account for this effect, as explained in Sec. 5.1.2, involving simulation of many structures like the one depicted in Fig. 5.2(c). The combination of three spatial dimensions and an additional dimension provided by the statistical ensemble size thus resulted in a real *four-dimensional* problem, as discussed in the Introduction to this thesis. Due to the large computational effort, ensemble size was reduced to 100 devices. However, this was sufficient to achieve a clear trend, as shown in Figs. 7.8(a) and (c). In these figures, results of the sidewall-gate-LER simulations are compared to fin- and top-gate-LER data discussed above: individual contributions to mismatch in saturation threshold voltage and normalized current factor are plotted as a function of the number of simulated samples. In Figs. 7.8(b) and (d), results extracted from the full ensembles are compared, showing a similar impact of top- ("*G_top*") and sidewall-gate-LER ("*G_sw*"). Assuming these contributions as uncorrelated, in analogy with results reported in Table 7.3, the total impact of gate-LER can be estimated to a worst-case as:

$$G\_tot = \sqrt{G\_top^2 + 2G\_sw^2} \tag{7.1}$$

However, the fin-LER ("*F*" in Figs. 7.8(b) and (d)) is seen to be the

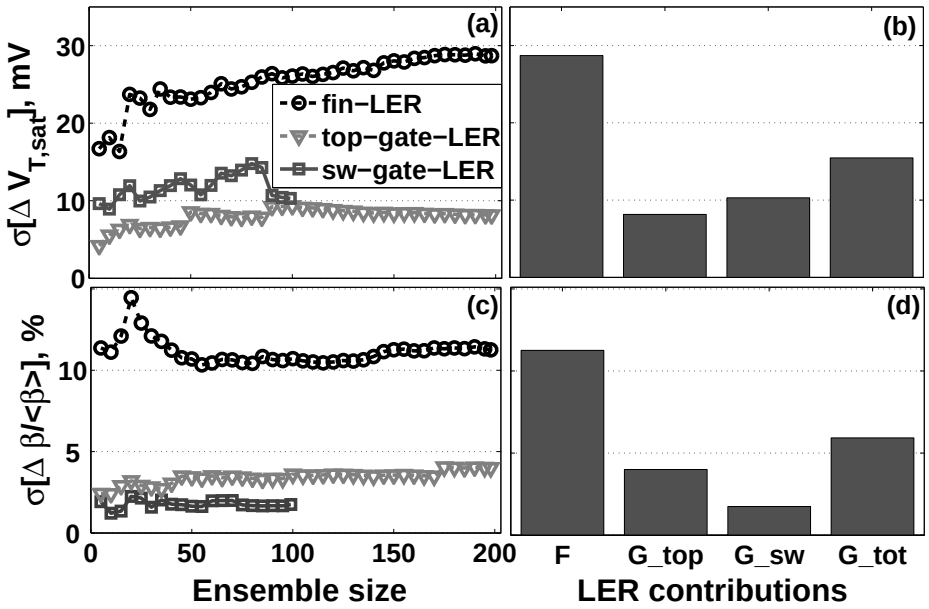


Figure 7.8: Mismatch in threshold voltage and current factor plotted as a function of the ensemble size ((a), (c)) and extracted from the full ensembles ((b), (d)). "*F*", "*G_top*" and "*G_sw*" are contributions to LER from the fin, top-gate and a single sidewall-gate, respectively; "*G_tot*" is the total contribution to gate-LER estimated through (7.1) assuming statistical independence of individual components.

most critical issue for both V_T- and β-mismatch, for identical values of the roughness model parameters.

Choice Δ=1.5 nm, Λ=20 nm is within typically measured ranges for the fin- and top-gate-LER. In order to allow for a direct comparison, the same values have been utilized for the sidewall-gate-LER too, but metrological characterization of the sidewall-gate-roughness is still an issue, i.e. no accurate estimations of rms amplitude and correlation length are available yet. Extraction of such parameters (see Sec. 5.1.1) has been attempted on a limited number of available SEM cross sections like the one in Fig. 7.9 [tnano]. This preliminary investigation yielded rms amplitudes slightly lower than 1 nm. Therefore, results in Fig. 7.8 should provide a worst-case estimation of the impact of sidewall-gate-LER on mismatch.

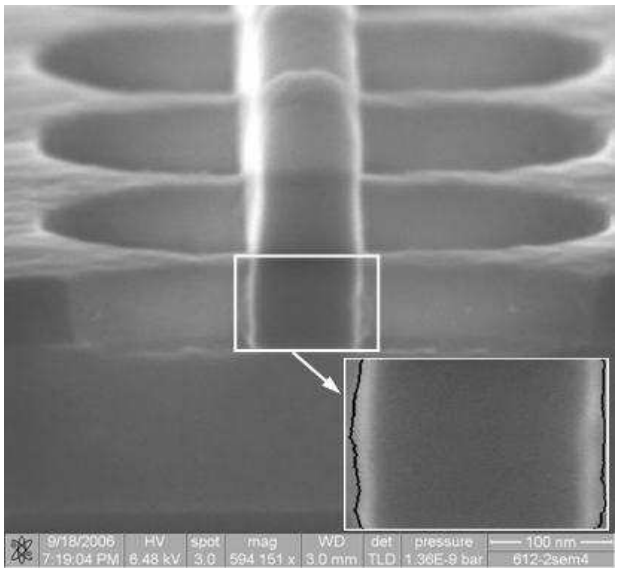


Figure 7.9: SEM cross-section of a multiple-fin FinFET (IMEC data). One of the sidewall gates is highlighted and results of the edge detection are shown in the inset.

7.2.2 Influence of doping profiles and number of fins

According to the above results, the net mismatch of FinFETs described in Table 7.2 is dominated by the contribution from fin-LER. Instead, a similar importance of the fin- and top-gate-roughness was observed in simulations reported in Sec. 7.1. This is due to the difference in doping profiles, i.e. the impact of line-edge roughness is sensibly dependent on FinFET doping. In particular, extension doping is a critical process step as low-energy, high-dose implants are desired at high tilt angles, while avoiding dopant penetration underneath the gate and incurring short channel effects [94]. Advanced techniques as Solid-Phase Epitaxial Regrowth (SPER) are being explored to achieve abrupt junctions [95] and improve device performance [96]. Although a comprehensive investigation of this topic would require sophisticated process simulation, LER dependence on extension profiles has been estimated by varying extension levels and decay rates. To this aim, several n-type profiles have been considered, with extension concentrations N_{ext} rang-

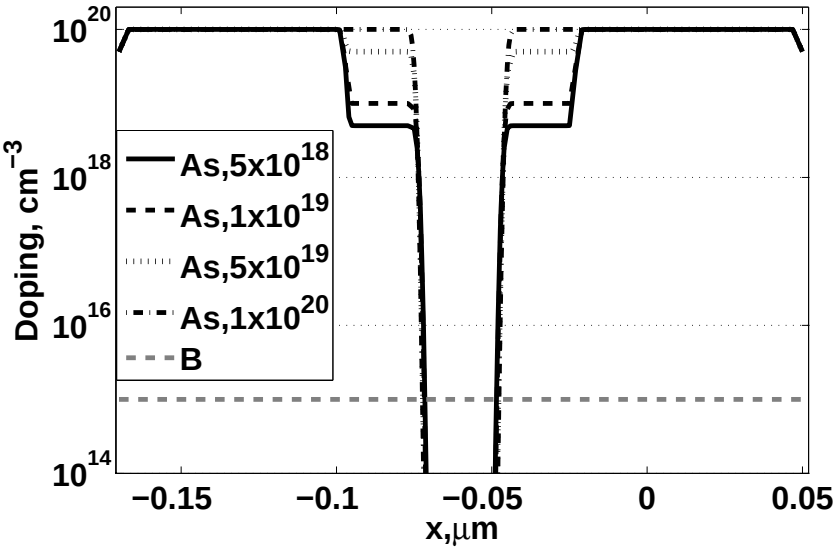


Figure 7.10: Impact of doping profiles on LER-induced mismatch: extension concentrations N_{ext} ranging from 5×10^{18} to 1×10^{20} cm^{-3} have been considered.

ing from 5×10^{18} to 1×10^{20} cm^{-3}, as shown in Fig. 7.10. Simulation results in Fig. 7.11(a) indicate that the impact of LER on threshold voltage mismatch is enhanced with higher extension levels. As for the current factor mismatch, impact of the fin roughness tends to decrease as N_{ext} is increased, whereas the gate-LER contribution rises rapidly, as shown in Fig. 7.11(c). This can be explained as follows. The more ideal, box-shaped *As* extension profiles with increased concentration may be required to boost saturation drain current through reduction in S/D resistance [96] (see Fig. 7.12). Such reduction is believed to be responsible for the diminishing impact of the fin-LER in Fig. 7.11(c). However, this causes saturation current to be more sensitive to the effective channel resistance. In turn, such parameter is particularly sensitive to the gate-LER due to changes in the metallurgical channel length: hence, importance of this roughness component increases in Fig. 7.11(c). Overall impact of the FinFET performance (I_{ON}) optimization on its matching performance can be visualized in Fig. 7.13, where the relative importance of gate-LER is seen to increase with increasing extension concentration. In particular, percentage contribu-

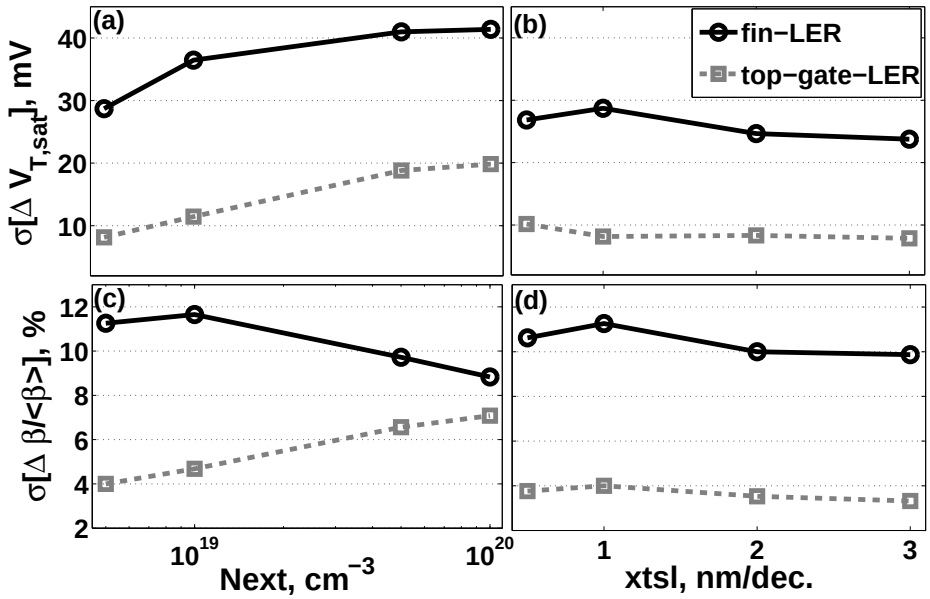


Figure 7.11: Comparison between contributions to mismatch from the fin an top-gate roughness. (a), (c): impact of different extension concentrations N_{ext}. (b), (d): impact of extension slope $xtsl$.

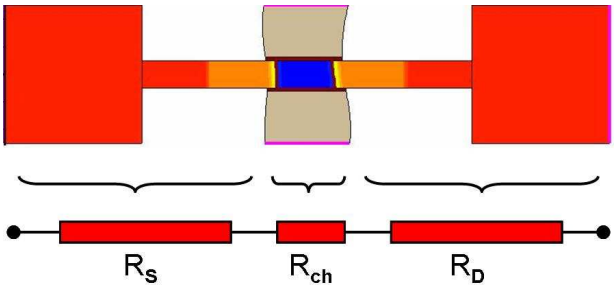


Figure 7.12: Parasitic resistance model of a FinFET. Gate-LER gives rise to gate line-width-roughness, i.e. fluctuations in physical gate length and hence changes in channel resistance (R_{ch}). Increasing extension profile concentration and slope (junction engineering - see Fig. 7.10) reduces S/D resistances (R_S, R_D), thus enhancing the relative importance of R_{ch}.

tion of the top-gate roughness to the current factor mismatch is more than doubled with the "idealized" profile ($N_{ext} = 1 \times 10^{20}$ cm^{-3}) with respect to the "realistic" profile ($N_{ext} = 5 \times 10^{18}$ cm^{-3}).

The impact of profile decay rate has also been studied for the sake

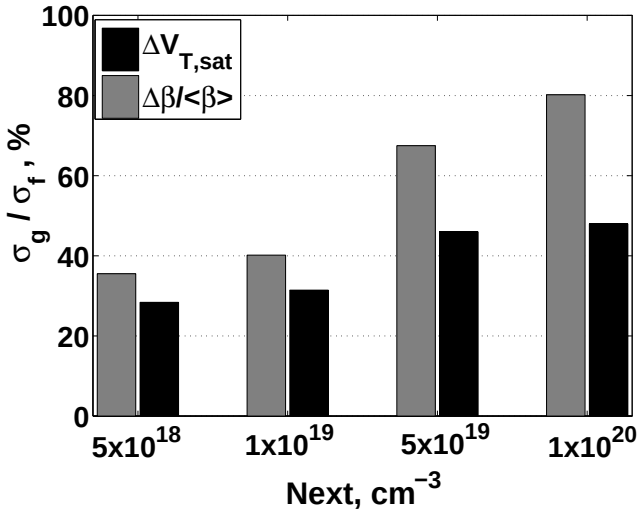


Figure 7.13: Impact of extension concentration on relative importance of the top-gate-LER (σ_g) with respect to the fin-LER (σ_f).

of completeness. V_T- and β-mismatch are plotted as a function of the extension slope $xtsl$ in Figs. 7.11(b) and (d). It can be seen from these figures that steeper junctions correspond to slightly higher mismatch, but comparison with Figs. 7.11(a) and (c) show that the impact of $xtsl$ is less critical than that of the extension concentration. Overall, these simulations show that doping profile engineering might lead to enhanced current factor mismatch in FinFETs: such devices could be unusable without a substantial reduction in gate-LER.

Simulations discussed so far refer to n-channel, single-fin devices. In Figs. 7.14(a) and (c), n- and p-type FinFETs are compared and seen to have similar sensitivity to line-edge roughness. The fin-LER is clearly more critical than the top-gate-LER for both device types. However, in order to keep planar bulk-competitive drain current per unit layout area, FinFETs are essentially designed with multiple fins [97]. Since these narrow fins determine both the sub- and super-threshold behavior, FinFET matching performance is likely to be influenced by scaling the number of fins. This is confirmed by Figs. 7.14(b) and (d), where the mismatch is seen to decrease as the number of fins is increased. Since

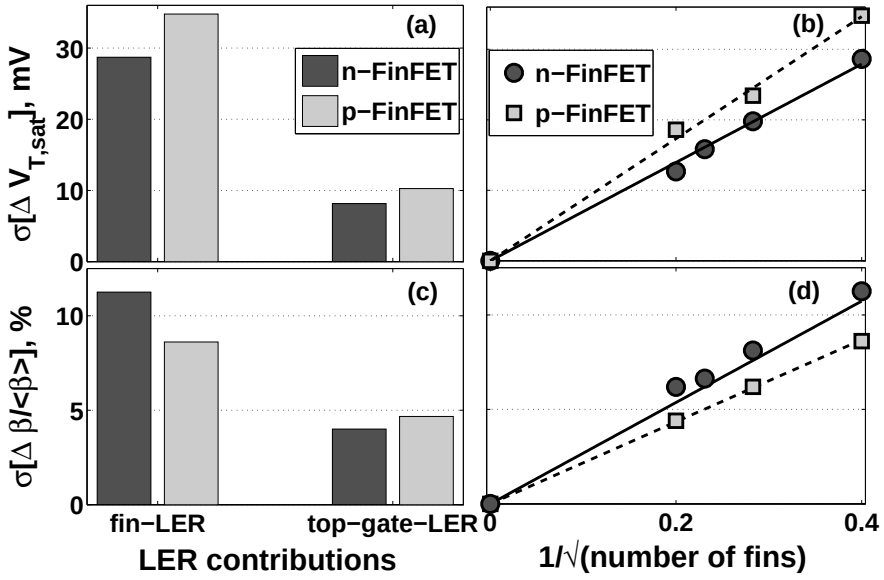


Figure 7.14: (a), (c): impact of fin- and top-gate-LER on mismatch of n- and p-type FinFETs. (b), (d): impact of fin-LER on mismatch of multi-fin devices.

each fin contributes to the total device width, trends in these figures are in accordance with Pelgrom's model of mismatch being proportional to the inverse square root of area [86] (see Sec. 5.2.1). The behavior of n- and p-channel FinFETs is similar even in this case, which allows to conclude that neither device type is more robust than the other to LER issues.

7.2.3 Correlation study

In Sec. 5.2.3, the importance of analyzing correlations between structural and electrical fluctuations has been highlighted. Therefore, a correlation study is presented here, which helps understanding how line-edge roughness affects FinFET performance and can be exploited to improve efficiency of first-order variability estimation with respect to the full Monte Carlo approach adopted so far. The study consists in checking the relationship between the average width of rough device features and resulting electrical parameters for each simulated FinFET instance. Two situations have been considered as for fin-LER simula-

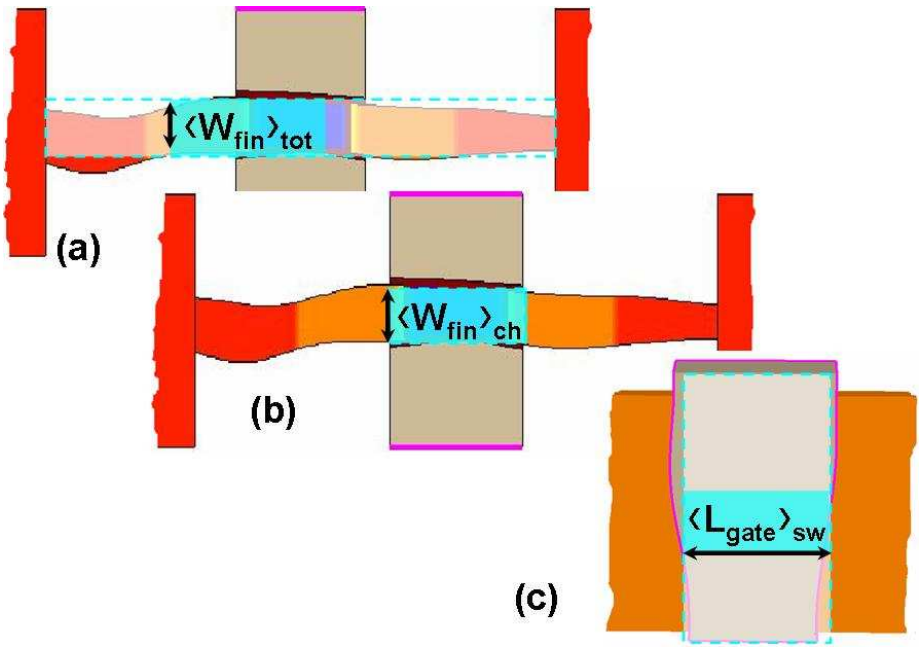


Figure 7.15: Averaging operation for correlation analysis. (a): fin width averaged over the whole fin length. (b): fin width averaged over the channel region. (c): sidewall-gate length averaged over the fin height.

tions, in which the fin width has been averaged over the whole fin length ($\langle W_{fin} \rangle_{tot}$) and over the channel region only ($\langle W_{fin} \rangle_{ch}$), as illustrated in Fig. 7.15(a) and (b), respectively. In the case of sidewall-gate-LER, the average sidewall gate length $\langle L_{gate,sw} \rangle$ has been calculated over the fin height (see Fig. 7.15(c)).

Parameter distributions resulting from simulations of fin- and sidewall-gate-LER discussed in Sec. 7.2.1 (see Fig. 7.8) are plotted in Fig. 7.16 as a function of the respective average feature width. The strong correlation in Fig. 7.16(b) clearly indicates that line-edge roughness mainly impacts the device threshold by changing the average fin width in the channel region. Since the correlation is weak for both V_T and β when the total average fin width is considered instead (see Fig. 7.16(a) and (d)), it is argued that the roughness of extension regions does not severely impact FinFET matching performance. A weak correlation is also seen in Fig. 7.16(e): this suggests the current factor being sensitive

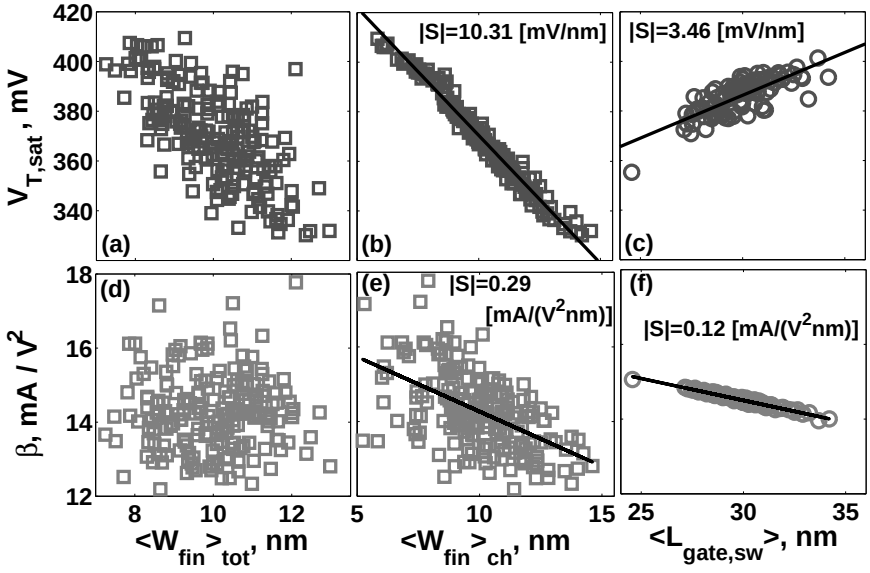


Figure 7.16: Dependence of threshold voltage and current factor on the fin width averaged over the whole fin length((a), (d)), fin width averaged over the channel region ((b), (e)) and sidewall-gate length averaged over the fin height ((c), (f)). Slopes of the linear fits (S) are indicated in the figure.

to the particular shape of the roughness, which probably affects mobility of the different device instances. As for the sidewall-gate roughness, Figs. 7.16(c) and (f) show that the device performance is mainly determined by the resulting average gate length. However, slopes of the corresponding linear fits are smaller than those in Figs. 7.16(b) and (e). This indicates that the device parameters are more sensitive to changes in the average fin width in the channel than to changes in the average gate length, which agrees with the results of mismatch estimation shown in Fig. 7.8.

Similar correlation trends have been observed for all simulation cases described in Secs. 7.2.1 and 7.2.2, including multi-fin devices, n- and p-type channels and different extension profiles. This allows applying the procedure proposed in Sec. 5.2.3 to get a faster estimation of variability for those parameters exhibiting strong correlation properties. Accuracy of such approach has been tested on more than 30 statistical ensem-

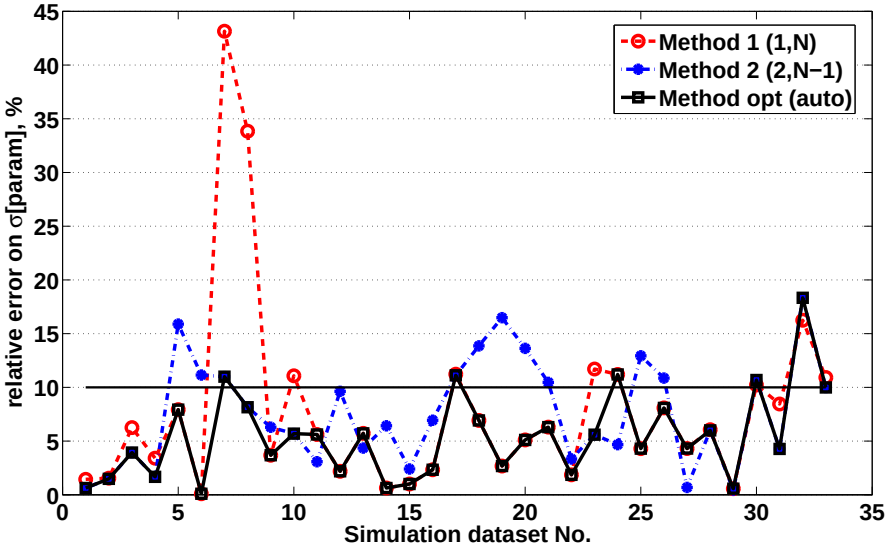


Figure 7.17: Percentage error of correlation-based variability estimation with respect to results extracted from full ensembles, calculated for several datasets.

bles. For each of them, Fig. 7.17 plots the relative error in variability estimation through a small number of selected samples with respect to statistics extracted from the full ensembles. The meaning of "Method 1" and "Method 2" in this figure has been clarified in Sec. 5.2.3: here it is shown that the selection algorithm proposed in that Section automatically chooses the most accurate estimate in 85% of considered test cases. Moreover, relative errors are generally within 10%. Larger errors correspond to particularly unfavorable situations, such as high roughness rms amplitude or correlation length (see Fig. 7.19 in Sec. 7.3). To better visualize the effectiveness of the proposed approach, data of Fig. 7.14(b) are compared to correlation-based estimations in Fig. 7.18.

Although more sophisticated techniques might be investigated, variability estimation through this simple algorithm provides a reasonable accuracy, while allowing for two orders of magnitude improvement in computational efficiency. Assuming correlation properties shown in Fig. 7.16 to be generally valid, the proposed method could be exploited instead of Monte Carlo techniques in future LER investigation of different process options and technology generations.

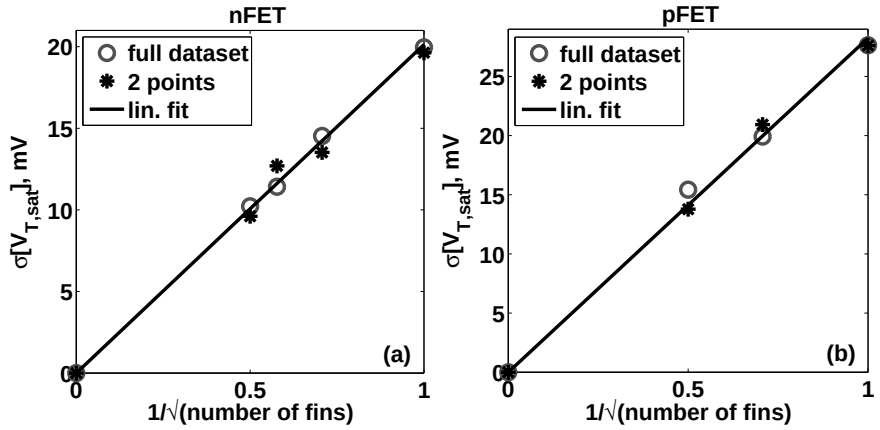


Figure 7.18: Comparison between V_T-mismatch extracted from full simulated distributions and exploiting correlation to $\langle W_{fin} \rangle_{ch}$, for multi-fin n-channel (a) and p-channel (b) devices.

7.3 LER requirements for circuit applications of FinFET: simulations and measurements

In order to address LER requirements for future FinFET technology nodes, the impact of LER on n-type devices in Table 7.2 is analyzed as a function of its rms amplitude (Δ) and correlation length (Λ) in Fig. 7.19. Only contribution from the fin-roughness is considered in simulations reported in this Section as it was shown to be the most relevant component. It can be seen from Figs. 7.19(a) and (c) that the mismatch varies linearly with the rms amplitude of LER. This linear trend is observed for both RDF and SDF technologies. However, slopes of the fitted lines differ and more than 90% reduction in mismatch can be seen for spacer-defined fin patterning at identical rms amplitudes. Standard deviation values corresponding to the maximum allowed V_T-mismatch (approximately 58% of target V_T [5]) are also shown in Figs. 7.19(a) and (b). Considering these values, it can be inferred that the mismatch resulting from current LER parameters (Δ = 1.5 nm, Λ = 20 nm) is critical.

In addition to DC performance, the impact of LER on the transient

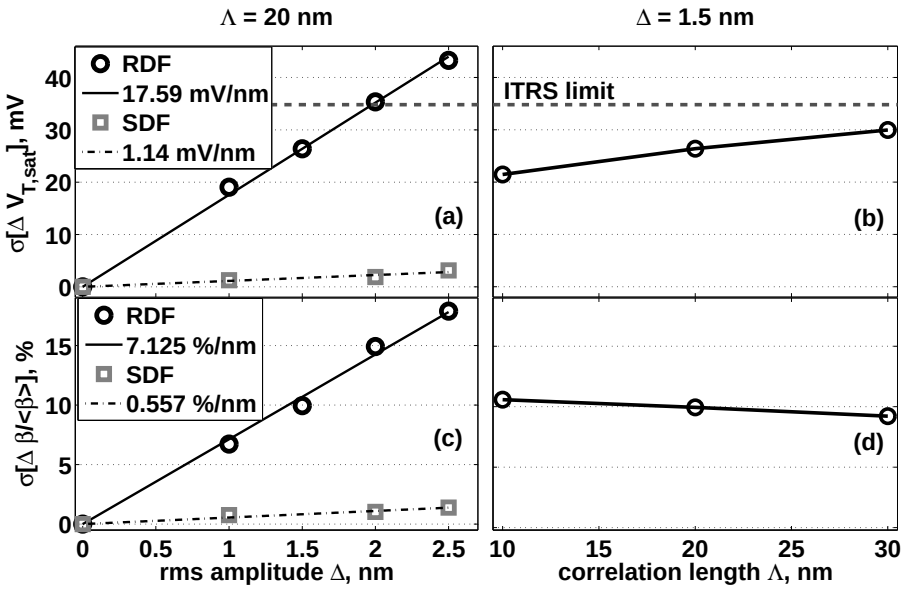


Figure 7.19: Mismatch in threshold voltage and current factor as a function of LER rms amplitude Δ (a), (c) and correlation length Λ (b), (d), for typical ranges of measured values of these parameters respectively. Legends in the left plots show the slopes of linear fits. The maximum threshold voltage variability set by ITRS specifications is also plotted (dashed line).

behavior of FinFETs has also been considered, taking into account the Power-Delay Product, as described in Sec. 5.1.3. Mixed-mode simulations of equivalent circuits like the one in Fig. 5.3 have been performed, where the device is a LSTP-32 nm FinFET affected by uncorrelated fin roughness. Resulting 6σ relative intervals of mismatch in Power Delay Product are shown in Fig. 7.20. It can be seen that in presence of fin-LER, the maximum tolerance to circuit performance variability [5] is exceeded by single fin devices and designing FinFETs with higher number of fins may be useful.

In order to evaluate the impact of fin-LER on LSTP-32 nm compatible SRAMs, mixed-mode DC simulations have been performed on ensembles of four-transistor circuits, as show in Fig. 5.5 (see Sec. 5.1.3). Standard deviations of SNM and ΔSNM distributions resulting from butterfly curves are plotted as a function of the ensemble size in Fig. 7.21, for FinFETs with different numbers of fins. The statistical trend is seen

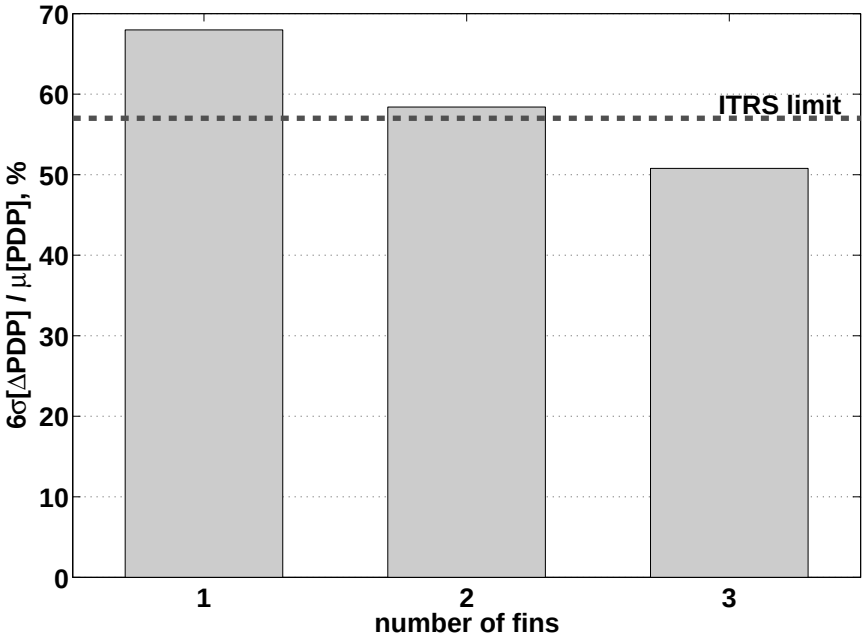


Figure 7.20: 6σ relative interval of ΔPDP versus number of fins for stand-alone FinFETs (bars) and maximum circuit performance variability specifications for the ITRS 32 nm node (dashed line).

to be well-stabilized with 100 simulations. The different statistical behavior of SNM and ΔSNM with respect to the number of fins is probably due to the fact that the latter parameter results from the difference of two correlated random variables. It can be seen from Fig. 7.21 that designing FinFETs with higher number of fins is beneficial for SRAM stability. However, in the case of RDF technology, this choice contrasts with strict area constraints in SRAM design. On the other hand, fin doubling due to spacer-defined patterning provides an opportunity to reduce the intra-bit-cell mismatch, as predicted by these simulations, without increasing the bit-cell layout area [98].

Spacer-defined fin patterning has the potential to increase the number of fins by 2^n, where n is the number of times the spacer patterning is performed. However, as the fin pitch gets smaller, patterning issues related to profiles and mechanical stability of the narrow fins arise and have been reported recently [99]. Since all of the etch and implantation steps utilized for the device fabrication are qualified only

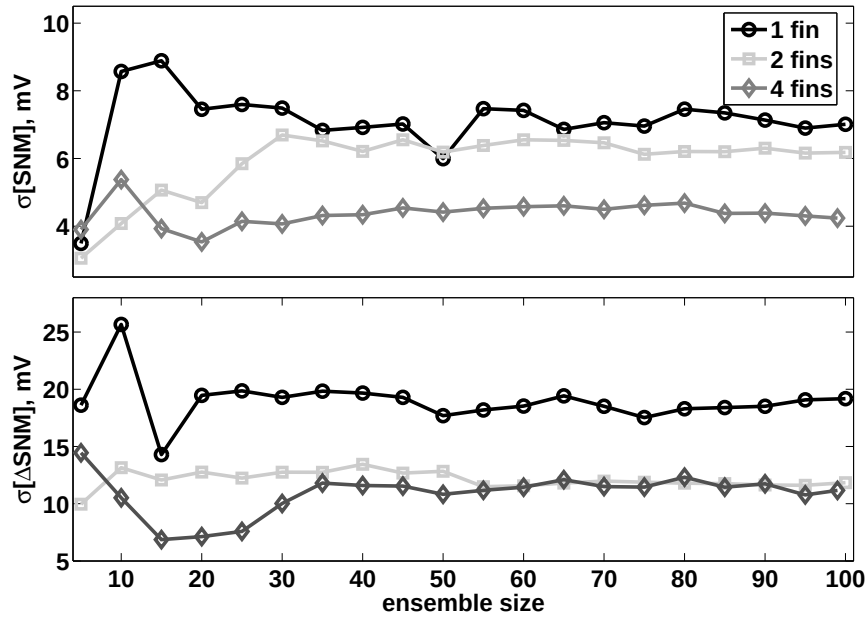


Figure 7.21: SNM and ΔSNM variability extracted from butterfly curves ("Sim.") in Fig. 7.22 and plotted as a function of the number of simulated SRAMs.

at the relaxed fin pitch (RDF-like), subsequent processes in the SDF patterned devices are currently affected in an untraceable manner. To regain the process control in SDF technology, critical steps should be qualified for the target fin pitch. This can be verified from butterfly curves measured on fabricated SRAM cells, as shown in Fig. 7.22, where RDF SRAMs are seen to have lower variability than the SDF counterparts [iedm06]. Statistical parameters are extracted from butterfly curves for the three cases in Fig. 7.22. Resulting standard deviations in SNM and ΔSNM are shown in Figs. 7.23(a) and (b), respectively. As for measured cells, it can be seen from this figure that the additional fluctuations in SDF SRAMs mainly come from short-range process variations affecting intra-bit-cell mismatch, i.e. $\sigma[\Delta\text{SNM}]$, whereas long-range variability represented by $\sigma[\text{SNM}]$ is almost identical for the two technologies. Simulation results in Fig. 7.23 show the *predicted* impact of fin-LER at the LSTP-32 nm node, in an ideal case where this is the only contribution to variability (i.e. real fluctuations including all sources will certainly be larger than these predictions). It can be

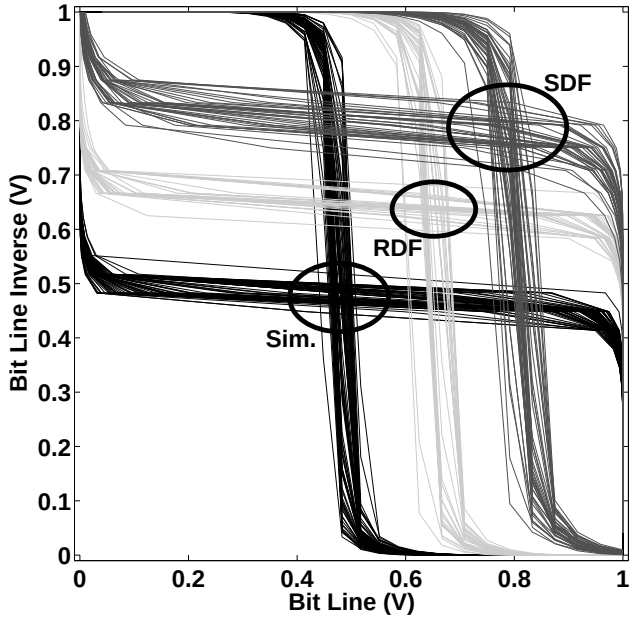


Figure 7.22: Measured ("RDF", "SDF") butterfly curves in stand-by mode at $V_{dd} = 1$ V. Measured SRAM cell devices (fabricated at IMEC) have $W_{fin} = 30$ nm, $L_{gate} = 55$ nm and fin doubling for SDF; the total cell area is 6 μm^2. Simulations ("Sim.") used to extract data for single-fin devices in Fig. 7.21 are also reported to provide an indication of the predicted spread at the LSTP-32 nm node due to the fin-LER contribution alone to variability.

seen from Fig. 7.23(b) that at this node the LER-induced component of short-range variations alone will become approximately as large as the total amount of intra-cell mismatch measured in present-days RDF technology. Thus, in order for future technology nodes to meet variability criteria when other sources of fluctuations are present besides LER, an improved spacer-defined fin-patterning process needs to be developed to contain the contribution of fin-LER to FinFET mismatch.

7.4 Impact of RD fluctuations on FinFET matching

In addition to line-edge roughness, random dopant fluctuations are another major sources of concern for future technology nodes. To provide

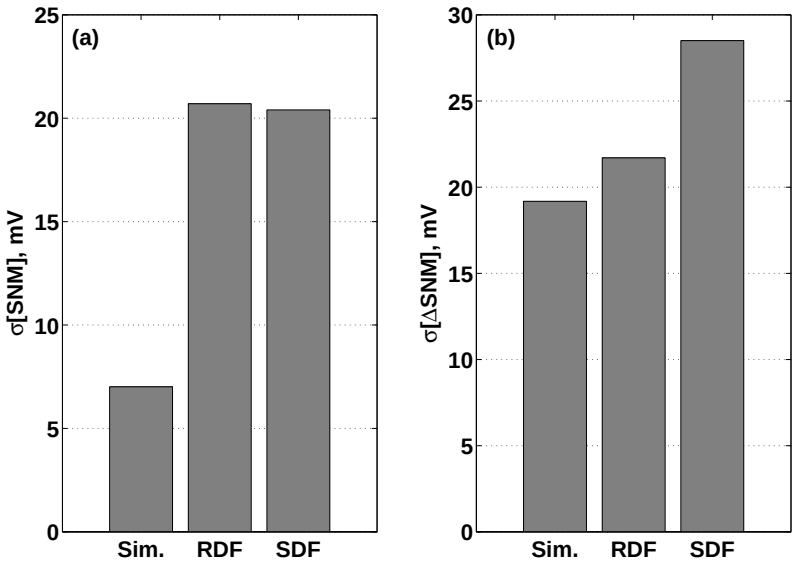


Figure 7.23: SNM (a) and ΔSNM (b) standard deviations extracted from simulated and measured butterfly curves shown in Fig. 7.22.

a first-order estimation of the impact of this issue at the LSTP-32 nm node, two-dimensional simulations of n-channel FinFETs have been performed, including noise analysis, as described in Sec. 5.3. Nominal doping concentrations in the channel, source/drain pads and extension regions have been varied over several orders of magnitude and the impact of RD fluctuations on percentage variation of saturation threshold voltage and on-current has been computed within considered ranges. Results of this analysis are shown in Fig. 7.24, where contributions from the fin-LER are also reported: impact of the two sources of fluctuations can be directly compared because the same models have been used in the simulations. It is evident from Figs. 7.24(b) and (c) that RD-induced fluctuations of the device threshold are almost insensitive to the nominal concentration in the source/drain and extension regions, within typical ranges. Moreover, the impact of line-edge roughness is much more critical. The same considerations hold for doping concentrations up to 10^{18} cm^{-3} in the device channel, as illustrated in Fig. 7.24(a). Convergence issues arise for higher values of N_{ch}, when the threshold is expected to become more sensitive to RD fluctuations. However, such high doping levels can be avoided in FinFETs with suit-

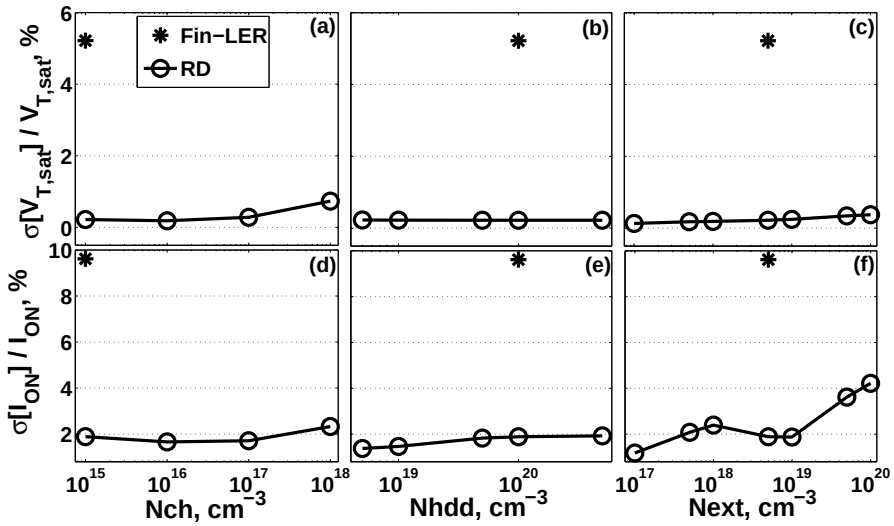


Figure 7.24: RD-induced threshold voltage and on-current percentage variation as a function of the doping concentration in the channel ((a), (d)), S/D ((b), (e)) and extension regions ((c), (f)) of a n-type FinFET. Fin-LER contribution is also shown for comparison.

able metal gates. As for on-current variability, no severe impact of dopant fluctuations is seen in Figs. 7.24(d) and (e), while the RD contribution increases rapidly for concentrations higher than 10^{19} cm^{-3} in the extension regions (Fig. 7.24(f)), although LER is still predominant.

In Sec. 7.2.2, roughness issues were shown to also become more severe with increasing extension doping: the impact of both LER and RD on FinFET matching should be carefully taken into account while engineering the S/D junctions. This is expected to become particularly true for FinFETs fabricated through improved manufacturing processes featuring spacer-defined fin patterning, which has potential to lower the impact of fin-LER on FinFET matching. Therefore, gate roughness and random dopant fluctuations are expected to be the major sources of parameter variations in FinFETs resulting from a mature SDF process.

Part V

Conclusions

"The important thing
is not to stop questioning."

A. Einstein

An effective support from Technology Computer Aided Design is more and more vital to the evolution of semiconductor industry. The technology scaling trend leads to increasing complexity of integrated circuits. Moreover, new materials and architectures are being introduced in device fabrication. As a consequence, fully-3D modeling approaches and an advanced, often non-classical, physical description are needed to represent the complicated structure and behavior of aggressively scaled devices. The highlighted issues are worsened as dimension shrinking collides with the intrinsic discreteness of charge and matter and with difficulties and tolerances in the fabrication process. Consequently, non-deterministic deviations of real devices from the ideal design become more and more critical. While indispensable to achieve a manufacturable technology, accounting for variability implies representing each device through a distribution of microscopically different instances. From the TCAD point of view, the outlined scaling trend is reflected in the increasing dimensionality of the problems which model real-world applications. In this thesis, some approaches have been proposed to address the outlined issues.

Adaptive mesh refinement

First, the meshing stage has been considered because of its crucial role in the device and process simulation flow. Adaptive meshing techniques have been indicated as the main road toward the optimal representation of the simulated domain, both in terms of computational efficiency and solution accuracy. Moreover, automatic adaptation is highly desirable to simplify user-interaction with TCAD tools, which typically requires exceptional expertise.

Potentialities of the multiresolution analysis in this context have

been investigated, leading to the development of a *Wavelet-based Adaptive Method (WAM)* for mesh refinement in two- and three-dimensional settings. In this approach, the adaptation procedure is driven by an *estimation of solution regularity* through the Wavelet Transform, resulting in the following features:

- the possibility of relieving the user from the problem of generating suitable meshes for standard finite-volume solvers to deal with real world tasks;
- the anisotropic refinement of regions which have stringent meshing requirements with a smooth grading of element size;
- the good convergence properties of the scheme, which starts with a uniform coarse mesh;
- the possibility of *dynamically* adapting the mesh at each bias step in sweep simulations;
- the use of fast and numerically efficient algorithms from signal processing for detecting sensible regions.

Good selectivity properties of the algorithm have been obtained even in 3D applications through a two-step Wavelet analysis combined with an effective anisotropy handling. Moreover, the semiregular nature of WAM grids allowed for the implementation of a quality check procedure able to remove undesired mesh patterns affecting simulation convergence and accuracy.

The refiner has been fully integrated into a standard TCAD environment through a modular system integration software. The resulting validation tool has been tested on several 2D and 3D structures, including p-n diodes, nMOS power drivers and FinFETs. Such tests demonstrate the effectiveness of Wavelets as a means to guide the automatic refinement of discretization grids for the simulation of electronic devices, preserving the geometrical and physical features of the problem to be solved.

Extension of the proposed approach to the refinement of completely arbitrary geometries is being investigating by combining the Wavelet analysis with a compatible error indicator for irregular domain portions that cannot be covered by Wavelet supports.

LER and RD fluctuations in FinFETs

Besides accurately modeling the device operation, understanding the impact of process variations is essential to evaluate effectiveness of innovative materials and process options as well as performance of new device architectures through predictive computer simulations. In particular, FinFETs may replace conventional CMOS devices in the future technology generations due to their intrinsically better scalability. Therefore, techniques to deal with variability estimation in these devices have been explored, with particular emphasis on two sources of major concern for future technology nodes, i.e. random dopant fluctuations and line-edge roughness.

The inherently three-dimensional carrier transport in FinFETs makes them sensitive to roughness of several printed features. Contributions to LER from the fin, top- and sidewall-gates have been decoupled and compared by means of 2D and 3D TCAD simulations performed on large statistical ensembles. The mismatch induced by low spatial frequency components of the roughness is shown to become significant below 45 nm gate length geometries. Moreover, results of an in-depth analysis of FinFETs conforming to ITRS LSTP-32 nm specifications indicate random uncorrelated roughness of the fin edges, typically introduced by a resist-defined patterning process, as the main contribution to mismatch in threshold voltage and current factor of nominally identical devices. Top- and sidewall-gate-LER are predicted to have a similar impact, but the total contribution to mismatch from the gate roughness is found to be approximately 50% lower than the fin-related component for both n- and p-channel devices.

Deeper insight on the way line-edge roughness affects FinFET performance is provided by a correlation study. Results show that the

gate-LER mainly impacts the device matching by changing the average gate length. As for the fin-LER, threshold voltage is strongly correlated to the average fin width in the channel region, while the particular shape of the roughness is relevant to the current factor. Correlations can be fruitfully exploited to reduce computational cost of variability estimation by orders of magnitude: inaccuracy of this approach is found to be within 10%.

Simulations reveal that with the current LER parameters, i.e. rms amplitude = 1.5 nm and correlation length = 20 nm, both the DC and transient matching performance of FinFET devices and SRAM cells are in the critical zone. Instead, random dopant fluctuations, simulated through a noise analysis approach, are predicted to be negligible with respect to the LER contribution over wide ranges of doping concentrations in the channel and source/drain regions.

To minimize the impact of fin-LER on FinFET matching, two possibilities have been explored, namely the use of higher number of fins and an in-phase correlation between LERs on the fin sides. It is found that a doubling in the number of fins can reduce the impact of LER on V_T and β matching by 30% and 15%, respectively. Furthermore, spacer-defined fin patterning has been shown as a potential solution to realize in-phase correlated fin-LERs, thus reducing both V_T and β mismatch by 90% with respect to RDF technology for identical LER parameters. However, measured SRAM performance is seen to be significantly affected by process instabilities for SDF technology. Therefore, to meet sub-45 nm variability specifications, *more stable spacer-defined patterning processes are desired*. These processes should pay special attention to *doping profile design* since the importance of both gate-LER and RD contributions to mismatch is expected to increase as S/D profiles are designed with high extension concentrations and box-shaped junctions to improve current drivability.

Overcoming TCAD roadblocks

The physics-based approach adopted in this thesis could be further exploited to extract typical values and fluctuations of parameters suitable for compact models (e.g. BSIM4 [100], MM11 [101], ACM [102], EKV3 [103]). This would allow for predictive simulations of circuit and system-level performance of new technologies, thus bridging the gap between process development and circuit design, which is indicated by the ITRS as a difficult TCAD issue. Moreover, several concepts and algorithms presented in this work are borrowed from a wide range of different scientific areas, including multiresolution analysis, signal processing and statistics. The synergistic interaction of various research fields is shown to result in effective *multidisciplinary* approaches to overcome many modeling and simulation requirements highlighted by the ITRS as critical roadblocks to the assessment of future technology nodes.

Bibliography

[1] G. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, 1965.

[2] R. Minixhofer. TCAD as an integral part of the semiconductor manufacturing environment. In *Proc. SISPAD*, Sep. 2006 (invited).

[3] C. C. McAndrew. Predictive technology characterization, missing links between tcad and compact modeling. In *Proc. SISPAD*, pages 12–17, Sep. 2000.

[4] R. W. Dutton and A. J. Strojwas. Perspectives on technology and technology-driven CAD. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 19(12):1544–1560, Dec. 2000.

[5] *ITRS 2005 Edition*. http://public.itrs.net.

[6] T. Sekigawa and Y. Hayashi. Calculated threshold-voltage characteristics of an XMOS transistor having an additional bottom gate. *Solid-State Electronics*, 27:827–828, 1984.

[7] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, and D. Kyser. FinFET scaling to 10 nm gate length. In *IEDM Tech. Dig.*, pages 251–254, 2002.

[8] S. Selberherr. *Analysis and Simulation of Semiconductor Devices*. Springer, 1984.

[9] *Sentaurus Device Manual. Version X-2005.10*. Synopsys, 2005.

[10] G. Masetti, M. Severi, and S. Solmi. Modeling of carrier mobility against carrier concentration. *IEEE Trans. Electron Devices*, 30(7): 764–769, Jul. 1983.

[11] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi. A physically based mobility model for numerical simulation of nonplanar devices. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 7(11):1164–1171, Nov. 1988.

[12] C. Canali, G. Majni, R. Minder, and G. Ottaviani. Electron and hole drift velocity measurement in silicon and their empirical relation to electric field and temperature. *IEEE Trans. Electron Devices*, 22(11):1045–1047, Nov. 1975.

[13] M. J. van Dort, P. H. Woerlee, and A. J. Walker. A simple model for quantization effects in heavily-doped silicon MOSFETs at inversion conditions. *Solid-State Electronics*, 37(3):411–414, 1994.

[14] M. G. Ancona and G. J. Iafrate. Quantum correction to the equation of state of an electron gas in a semiconductor. *Phys. Rev. B*, 39(13):9536–9540, 1989.

[15] M. G. Ancona and H. F. Tiersten. Macroscopic physics of the silicon inversion layer. *Phys. Rev. B*, 35(15):7959–7965, 1987.

[16] D. K. Ferry and J.-R. Zhou. Form of the quantum potential for use in hydrodynamic equations for semiconductor device modeling. *Phys. Rev. B*, 48(11):7944–7950, 1993.

[17] L. Shifren, R. Akis, and D. K. Ferry. Correspondence between quantum and classical motion: comparing Bohmian mechanics with a smoothed effective potential approach. *Phys. Lett. A*, 274(1–2): 15–83, 2000.

[18] O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method.* McGraw-Hill, fourth edition, 1994.

[19] A. Okabe, B. Boots, and K. Sugihara. *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams.* Wiley, 1992.

[20] D. L. Scharfetter and H. K. Gummel. Large-signal analysis of a silicon read diode oscillator. *IEEE Trans. Electron Devices*, 16(1): 64–77, 1969.

[21] P. Fleischmann. *Mesh generation for technology CAD in three dimensions.* Ph.D. dissertation, Technischen Universität Wien, 2000.

[22] O. Hassan and E. J. Probert. Grid control and adaptation. In J. F. Thompson, B. K. Soni, and N. P. Weatherill, editors, *Handbook of Grid Generation*. CRC Press, 1999.

[23] B. S. Baker, E. Grosse, and C. S. Rafferty. Nonobtuse triangulation of polygons. *Discrete & Computational Geometry*, 3(2):147–168, 1988.

[24] C. Heitzinger, A. Sheikholeslami, J. M. Park, and S. Selberherr. A method for generating structurally aligned grids for semiconductor device simulation. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 24(10):1485–1491, Oct. 2005.

[25] B. Schmithüsen, K. Gärtner, and W. Fichtner. A novel grid adaptation procedure for stationary 2D device simulation. In *Technical Proceedings of the 2003 Nanotechnolgy Conference and Trade Show*, Feb. 2003.

[26] R. Heinzl and T. Grasser. Generalized comprehensive approach for robust three-dimensional mesh generation for TCAD. In *Proc. SISPAD*, pages 211–214, Sep. 2005.

[27] I. Babuška and A. Miller. *A posteriori error estimates and adaptive techniques for the finite element method*. Tech. Rep., Instit. Phys. Sci. Technol., Univ. of Maryland, 1981.

[28] I. Babuška and W. C. Rheinbolt. A posteriori error estimates for the finite element method. *International Journal for Numerical Methods in Engineering*, 12:1597–1615, 1978.

[29] O. Jacquotte. Grid optimization methods for quality improvement and adaption. In J. F. Thompson, B. K. Soni, and N. P. Weatherill, editors, *Handbook of Grid Generation*. CRC Press, 1999.

[30] M. C. Chang, J. H. Chern, and P. Yang. An Accurate Grid Local Truncation Error for Device Simulation. In *Proc. ICCAD-93*, Nov. 1993.

[31] W.M.Coughran, M.R.Pinto, and R.K.Smith. Adaptive grid generation for vlsi device simulation. *IEEE Trans. Computer Aided Design*, 10(10):1259–1275, Oct. 1991.

[32] R. Heinzl, M. Spevak, P. Schwaha, and T. Grasser. A novel technique for coupling three dimensional mesh adaptation with an a posteriori error estimator. In *Proc. PRIME2005*, Jul. 2005.

[33] K. Tanaka, H. Kato, P. Ciampolini, A. Pierantoni, and G. Baccarani. Adaptive mesh generation in three dimensional device simulation. In *International Workshop on Numerical Modeling of Processes and Devices for Integrated Circuits*, Jun. 1994.

[34] G. Yang, R. Wang, and S. Wang. An Adaptive Remeshing Technique based on Hierarchical Error Estimates for Simulation of Semiconductor Devices. *Int. J. Numer. Model.*, 17:17–28, Jan. 2004.

[35] S. Nicaise. A-posteriori error estimations of some cell-centered finite volume methods. *SIAM J. Numer. Anal.*, 43(4):1481–1503, 2005.

[36] W. Wessner, C. Heitzinger, A. Hössinger, and S. Selberherr. Error estimated driven anisotropic mesh refinement for three-dimensional diffusion simulation. In *Proc. SISPAD*, Sep. 2003.

[37] P. A. Markowich, C. A. Ringhofer, S. Selberherr, and M. Lentini. A singular perturbation approach for the analysis of the fundamental semiconductor equations. *IEEE Trans. Electron Devices*, 30(9): 1165–1180, 1983.

[38] B. Schmithüsen, K. Gärtner, and W. Fichtner. A grid adaptation procedure for stationary 2D drift-diffusion model based on local dissipation rate error estimation: Part I background. Technical report, Integrated System Laboratory, ETH Zürich, 2001.

[39] C. C. McAndrew. Statistical modeling for circuit simulation. In *Proc. ISQED*, pages 357–362, Mar. 2003.

[40] T. Hagivaga, K. Yamaguchi, and S. Asai. Threshold voltage variation in very small MOS transistors due to local dopant fluctuations. In *Proc. Symp. VLSI Technol., Dig. Tech. Papers*, page 4647, 1982.

[41] K. Nishiohara, N. Shiguo, and T. Wada. Effects of mesoscopic fluctuations in dopant distributions on MOSFET threshold voltage. *IEEE Trans. Electron Devices*, 39:634639, 1992.

[42] P.A. Stolk and D.B.M. Klaassen. The effect of statistical dopant fluctuations on MOS device performance. In *IEDM Tech. Dig.*, pages 627–630, 1996.

[43] P.A. Stolk, F.P. Widdershoven, and D.B.M. Klaassen. Device modeling of statistical dopant fluctuations in MOS transistors. In *Proc. SISPAD*, pages 153–156, Sep. 1997.

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

[45] D. Vasileska, W. J. Gross, and D. K. Ferry. Modeling of deep-submicrometer MOSFETs: random impurity effects, threshold voltage shifts and gate capacitance attenuation. In *IEEE Cat. no. 98EX116, Extended Abstracts IWEC-6*, pages 259–262, 1998.

[46] A. Asenov. Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 micron MOSFETs: a 3D 'atomistic' simulation study. *IEEE Trans. Electron Devices*, 45:2505–2513, Dec. 1998.

[47] I. D. Mayergoyz and P. Andrei. Statistical analysis of semiconductor devices. *J. App. Physics*, 90:3019–3029, Sep. 2001.

[48] A. Wettstein, O. Penzin, E. Lyumkis, and W. Fichtner. Random dopant fluctuation modelling with the impedance field method. In *Proc. SISPAD*, pages 91–94, Sep. 2003.

[49] P. Andrei and I. Mayergoyz. Analysis of fluctuations in semiconductor devices through self-consistent Poisson-Schrödinger computations. *J. App. Phys.*, 96(4):2071–2079, Aug. 2004.

[50] T. Linton, M. Giles, and P. Packan. The impact of line edge roughness on 100 nm device performance. In *Ext. Abs. Silicon Nanoelectronics Workshop*, pages 82–83, 1999.

[51] Jr. T. D. Linton, S. Yu, and R. Shaheed. 3d modelling of fluctuation effects in highly scaled vlsi devices. *VLSI Design*, 13:103–110, 2001.

[52] P. Oldiges, Q. Lint, K. Petrillot, M. Sanchez, M. Ieong, and M. Hargrove. Modeling line edge roughness effects in sub 100

nanometer gate length devices. In *Proc. SISPAD*, pages 131–134, Sep. 2000.

[53] S.-D. Kim, S. Hong, J.-K. Park, and J. C. S. Woo. Modeling and analysis of gate line edge roughness effect on CMOS scaling toward deep nanoscale gate length. In *Ext. Abs. Int. Conf. Solid State Devices Mater.*, pages 20–21, 2002.

[54] S.-D. Kim, H. Wada, and J. C. S. Woo. TCAD-Based statistical analysis and modeling of gate line-edge roughness effect on nanoscale MOS transistor performance and scaling. *IEEE Trans. Semicinductor Manufacturing*, 17(2):192200, 2004.

[55] J. Wu, J. Chen, and K. Liu. Transistor width dependence of LER degradation to CMOS device characteristics. In *Proc. SISPAD*, pages 95–98, Sep. 2002.

[56] T. Linton, M. Chandhok, B. J. Rice, and G. Schrom. Determination of the line edge roughness specification for 34 nm devices. In *IEDM Tech. Dig.*, pages 303–306, 2002.

[57] A. Asenov, S. Kaya, and A. R. Brown. Intrinsic parameter fluctuations in decananometer MOSFETs induced by gate line edge roughness. *IEEE Trans. Electron Devices*, 50(5):1254–1260, May 2003.

[58] A. R. Brown, A. Asenov, and J. R. Watling. Intrinsic fluctuations in sub 10-nm double-gate MOSFETs introduced by discreteness of charge and matter. *IEEE Trans. Nanotechnology*, 1(4):195–200, Dec. 2002.

[59] S. Xiong and J. Bokor. Sensitivity of double-gate and FinFET devices to process variations. *IEEE Trans. Electron Devices*, 50 (11):2255–2261, Nov. 2003.

[60] P. Goupillaud, A. Grossmann, and J. Morlet. Cycle-Octave and related transforms in seismic signal analysis. *Geoexploration*, 23: 85–102, 1984.

[61] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[62] F. Hlawatsch and G. F. Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Sig. Proc. Magazine*, Apr. 1992.

[63] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inform. Theory*, 36:961–1005, Sep. 1990.

[64] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Commun. on Pure and Appl. Math.*, 41:909–996, Nov. 1988.

[65] O.M. Nielsen. *Wavelets in Scientific Computing.* Technical University of Denmark, 1998.

[66] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pat. Anal. and Mach. Intel.*, 11(7):674–693, Jul. 1989.

[67] R. A. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *Americ. J. of Math*, 114:737–785, 1992.

[68] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Trans. Inform. Theory*, 38(2), Mar. 1992.

[69] Y. Chang F. and P. Pun K. Discrete B-Spline wavelet method for semiconductor device simulation. In *IEEE International Symposium on Circuits and Systems*, 1997.

[70] S. Goedecker and C. Chauvin. Combining multigrid and wavelet ideas to construct more efficient multiscale algorithms. *J. of Theor. and Computat. Chem.*, 2(4):483–495, 2003.

[71] S.Goasguen, M.M.Tomeh, and S.M.El-Ghazaly. Electromagnetic and semiconductor device simulation using interpolating wavelets. *IEEE Trans. Microwave Theory Tech.*, 49(12), Dec. 2001.

[72] M. Holmström. Solving hyperbolic PED's using interpolating wavelets. *Report No. 189/1996*, 1996.

[73] M.Toupikov, G.Pan, and B.K.Gilbert. On nonlinear modeling of microwave devices using interpolating wavelets. *IEEE Trans. Microwave Theory Tech.*, 48(4), Apr. 2000.

[74] Y. A. Hussein and S. M. El-Ghazali. Extending multiresolution time domain (MRTD) technique to the simulation of high-frequency active devices. *IEEE Trans. Microwave Theory Tech.*, 51(7):1842–1851, Jul. 2003.

[75] A.Limon and H.Morris. A multilevel adaptive solver based on second-generation wavelet thresholding techniques. *Numer. Linear Algebra Appl.*, 13:251–273, Feb. 2006.

[76] Jonathan Richard Shewchuk. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In Ming C. Lin and Dinesh Manocha, editors, *Applied Computational Geometry: Towards Geometric Engineering*, volume 1148 of *Lecture Notes in Computer Science*, pages 203–222. Springer-Verlag, May 1996. From the First ACM Workshop on Applied Computational Geometry.

[77] H. Si. *A Quality Tetrahedral Mesh Generator and Three Dimensional Delaunay Triangulator.* WIAS-Berlin, Version 1.4 User's Manual, Jan. 2006.

[78] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 1992.

[79] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Math. Annal.*, 69:331–371, 1910.

[80] *Tecplot User's Manual. Version 10.* Tecplot Inc. Bellevue, Washington, 2005.

[81] K. Lakshmikumar, R. Hadaway, and M. Copeland. Characterization and modeling of mismatch in MOS transistors for precision analog design. *IEEE Journal of Solid-State Circuits*, 21(6):10571066, Dec. 1986.

[82] A. Dixit and V. R. Rao. A novel dynamic threshold operation using electrically induced junction MOSFET in the deep sub-micrometer CMOS regime. In *Proc. 16th IEEE Int. Conf. on VLSI Design*, pages 499–503, Jan. 2003.

[83] E. Seevinck, F. J. List, and J. Lohstorn. Static-noise margin analysis of MOS SRAM cells. *IEEE Journal of Solid-State Circuits*, SC-22(5):748–754, Oct. 1987.

[84] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl. The impact of intrinsic device fluctuations on CMOS SRAM cell stability. *IEEE Journal of Solid-State Circuits*, 36(4):658–665, Apr. 2001.

[85] B. Cheng, S. Roy, G. Roy, A. Brown, and A. Asenov. Impact of random dopant fluctuation on bulk CMOS 6-T SRAM scaling. In *Proc. ESSDERC*, pages 258–261, 2006.

[86] A. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1440, 1989.

[87] P. G. Drennan and C. C. McAndrew. A comprehensive MOSFET mismatch model. In *IEDM Tech. Dig.*, pages 167–170, 1999.

[88] Y. Li and S.-M. Yu. Quantum correction simulation of random dopant-induced threshold voltage fluctuations in nanoscale metal-oxide-semiconductor structures. In *Proc. of 2005 5th IEEE Conf. on Nanotech.*, pages 527–530, Jul. 2005.

[89] I. D. Mayergoyz and P. Andrei. Statistical analysis of semiconductor devices. *J. App. Phys.*, 90(6):3019–3029, Sep. 2001.

[90] K. Tanaka, P. Ciampolini, A. Pierantoni, and G. Baccarani. Comparison between a posteriori error indicators for adaptive mesh generation in semiconductor device simulation. In *International Workshop on VLSI Process and Device Modeling*, May 1993.

[91] Y. K. Choi, N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T. J. King, J. Bokor, and C. Hu. Sub-20nm CMOS FinFET technologies. In *IEDM Tech. Dig.*, pages 421–424, 2001.

[92] Y.-K. Choi, L. Chang, P. Ranade, J.-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T.-J. King, and J. Bokor. FinFET process refinements for improved mobility and gate work function engineering. In *IEDM Tech. Dig.*, pages 259–262, 2002.

[93] J. A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, Q. Sansen, and H. E. Maes. Line edge roughness: Characterization, modeling and impact on device behavior. In *IEDM Tech. Dig.*, pages 307–310, 2002.

[94] H.-J. L. Gossmann, A. Agarwal, T. Parrill, L. M. Rubin, and J. M. Poate. On the FinFET extension implant energy. *IEEE Trans. Nanotechnology*, 2(4):285–290, Dec. 2003.

[95] R. Lindsay, B. Pawlak, J. Kittl, K. Henson, C. Torregiani, S. Giangrandi, R. Surdeanu, W. Vandervorst, A. Mayur, J. Ross, S. McCoy, J. Gelpey, K. Elliott, X. Pages, A. Satta, A. Lauwers, P. Stolk, and K. Maex1. A comparison of spike, flash, SPER and laser annealing for 45nm CMOS. In *2003 MRS Spring Meeting*, Apr. 2003.

[96] A. Dixit, K. G. Anil, N. Collaert, M. Goodwin, M. Jurczak, and K. De Meyer. Analysis of the parasitic S/D resistance in multiple-gate FETs. *IEEE Trans. Electron Devices*, 52(6):1132–1140, Jun. 2005.

[97] K. G. Anil, K. Henson, S. Biesemans, and N. Collaert. Layout density analysis of FinFETs. In *Proc. ESSDERC*, pages 139–142, 2003.

[98] R. Rooyackers, E. Augendre, B. Degroote, N. Collaert, A. Nackaerts, A. Dixit, T. Vandeweyer, B. Pawlak, M. Ercken, G. Dilliway, F. Leys, R. Loo, M. Jurczak, and S. Biesemans. Doubling or quadrupling MuGFET fin integration scheme with higher pattern fidelity, lower CD variation and higher layout efficiency. In *IEDM Tech. Dig.*, 2006.

[99] B. Degroote, R. Rooyackers, T. Vandeweyer, N. Collaert, W. Boullart, E. Kunnen, D. Shamiryan, J. Wouters, J. Van Puymbroeck, A. Dixit, and M. Jurczak. Spacer-defined FinFET: Active area patterning of sub-20nm fins with high density. *Microelectronic Engineering*, 84(4):609–618, 2007.

[100] W. Liu and C. Hu. *BSIM4: Theory and Engineering of MOSFET Modelling for IC Simulation.* World Scientific Publishing, 2007 (in press).

[101] *MOS Model 11.* http://www.nxp.com/Philips_Models/mos_models/model11.

[102] A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro. An MOS transistor model for analog circuit design. *IEEE Journal of Solid-State Circuits*, 33(10):1510–1519, Oct. 1998.

[103] M. Bucher, J.-M. Sallese, F. Krummenacher, D. Kazazis, C. Lallement, W. Grabinski, and C. Enz. EKV3.0: An analog design-oriented MOS transistor model. In *9th Int. Conf. on Mixed Design (MIXDES 2002)*, Jun. 2002.

Author's Publications

[essderc05] L. De Marchi, F. Franzè, E. Baravelli, and N. Speciale. Wavelet-based adaptive mesh generation for device simulation. In *Proc. ESSDERC*, pages 501–504, Sep. 2005.

[iedm06] A. Dixit, K. G. Anil, E. Baravelli, P. Roussel, A. Mercha, C. Gustin, M. Bamal, E. Grossar, R. Rooyackers, E. Augendre, M. Jurczak, S. Biesemans, and K. De Meyer. Impact of stochastic mismatch on measured SRAM performance of FinFETs with resist/spacer-defined fins: Role of line-edge-roughness. In *IEDM Tech. Dig.*, 2006.

[nova] L. De Marchi, E. Baravelli, and N. Speciale. *Progress in Solid State Electronics Research*, chapter TCAD solutions for increasing dimensionality in solid-state device and process simulations. Nova Science Publishers, in press.

[prime06] E. Baravelli, L. De Marchi, F. Franzè, and N. Speciale. Automatic wavelet localization and adaptive meshing of physical relevances in device simulation. In *Proc. of IEEE PhD Research in Microelectronics and Electronics, PRIME2006*, pages 189–192, Jun. 2006.

[sispad06] L. De Marchi, E. Baravelli, F. Franzè, and N. Speciale. 3D mesh generation with wavelet-driven adaptivity. In *Proc. SISPAD*, pages 212–215, Sep. 2006.

[snw07] E. Baravelli, M. Jurczak, N. Speciale, K. De Meyer, and A. Dixit. Impact of LER and random dopant fluctuations on FinFET matching performance. In *Ext. Abs. Silicon Nanoelectronics Workshop*, pages 23–24, June 2007.

[sse06] L. De Marchi, F. Franzè, E. Baravelli, and N. Speciale.

Wavelet-based adaptive mesh generation for device simulation. *Solid-State Electronics*, 50(4):650–659, Apr. 2006.

[tcad07] L. De Marchi, E. Baravelli, F. Franzè, and N. Speciale. Wavelet adaptivity for 3-D device simulation. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 26(11):1967–1977, Nov. 2007.

[ted07] E. Baravelli, A. Dixit, R. Rooyackers, M. Jurczak, N. Speciale, and K. De Meyer. Impact of line-edge roughness on FinFET matching performance. *IEEE Trans. Electron Devices*, 54(9):2466–2474, Sep. 2007.

[tnano] E. Baravelli, M. Jurczak, N. Speciale, K. De Meyer, and A. Dixit. Impact of LER and random dopant fluctuations on FinFET matching performance. *to appear in IEEE Trans. Nanotechnology.*