Computer Aided Design of Micro- and Nanoelectronic Devices

Chinmay Kumar Maiti
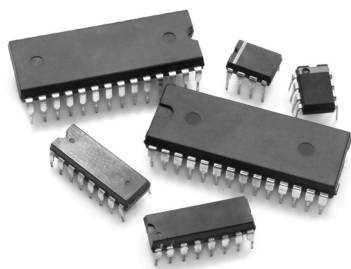
Computer Aided Design of Micro- and Nanoelectronic Devices

C

This page intentionally left blank

Computer Aided Design of Micro- and Nanoelectronic Devices

Chinmay Kumar Maiti
Siksha 'O' Anusandhan University, India

Dedication

In the loving memory of

Srimati Santilata Maiti

and

Srimati Bina Das

c

v

This page intentionally left blank

Preface

The road towards enhancing the performance of integrated circuits comprises fundamental improvement of the devices via new structures, scaling the size and transistor features which increase its speed and reduce power consumption. Combination of new device architectures and use of strain-engineered materials have led towards nanoelectronic devices and pushed the limits of CMOS and BiCMOS scaling. Typical transistor gate lengths have reached 10 nm. The first processor built with the 14 nm technology is the Intel Core-M processor. At the beginning of 2015, the fifth generation Intel Core processor was released, which is also built with the 14 nm technology and with 1.3 billion transistors.

Computer aided design and simulation of transistors for upcoming technology nodes are the main focus of this book. The monograph address mainly the design issues of advanced CMOS and bipolar devices with technical depth and conceptual clarity and presents leading-edge device design solutions to address the new challenges presented by advanced technology nodes. The device design examples covered in this monograph are in use and provide useful technology and device physics insights. The purpose of this monograph is also to bring into one resource presenting a comprehensive perspective of advanced micro- and nanoelectronic device design using advanced simulation tools.

One major limitation of currently available books in related areas is that mostly the design and simulation results presented are using 2D simulators. Due to ultra-small size of the state-of-the-art devices, 3D effects have become dominant. In order to achieve a better understanding of simulated and fabricated device characteristics, 3D process/device simulation is essential. Also 3D integration has become one of the main directions to fit into device performance roadmaps.

vii

After a brief discussion on the basic device physics, the readers are introduced to device design and simulation. Device design examples considered are from various areas, viz., bipolar and CMOS technology, memory and power devices, and solar cells. This monograph also presents design, simulation and analysis of heterostructure silicon and III-V compound semiconductor devices. Detailed and extensive technology CAD simulations have been presented for stress- and strain-engineered 3D FinFET devices. This monograph attempts to fill the gap in the literature in a rapidly evolving field as it blends together a wide ranging description of TCAD activities in process design to device design to compact model generation in Si, SiGe and III-V materials, technology, and their applications.

This monograph is primarily intended for senior undergraduate and graduate students and professors who wish to find a technology CAD teaching reference book and others who are interested in learning about semiconductor device design using simulation. This monograph may be used as a reference for engineers involved in advanced device and process design. Technology CAD specialists may find this monograph useful since it discusses the organization of TCAD simulation systems and also presents many case studies where the reader may apply the CAD tools in different situations. Approaches described in this monograph are expected to boost advanced device design using challenging TCAD simulations and help with characterizing new processes and devices.

I am especially grateful to all my research students whose association has been the source of learning for many new topics covered in this monograph. Special thanks go to Dr. A. Bag, who has kindly contributed a chapter (Chapter 13) on Heterojunction Solar Cells. We would also like to express our deep appreciation to WSPC team (V. Vishnu Mohan and Steven Patt). Finally, I would like to thank my family members (my wife Bhaswati, sons Ananda and Anindya) for their support, patience and understanding during the preparation of the manuscript.

C. K. Maiti
SOA University,
Bhubaneswar
December 2015

List of Tables

ix

This page intentionally left blank

List of Figures

xi

This page intentionally left blank

Contents

C

Chapter 1

Introduction

Since the invention of the transistor more than 60 years ago, the successful downscaling of silicon devices and advances in processing technologies have now enabled us to achieve highly complex electronic systems. The intense downscaling of CMOS transistors has been the major driving force behind the growth of the semiconductor industry for the past 50 years. However, standard bulk transistors severely suffer from short channel effects, although they have been considered down to the 45nm node. In order to understand computer aided device design, it is necessary to discuss the evolution of the devices considered and the issues involved in their design. Metal-Oxide-Semiconductor Field Effect Transistors (MOSFETs) comes first. Since its conception, MOSFET technology has improved steadily and has become the primary technology for ultra-large scale integration primarily because of the simple device structure. VLSI circuit development for greater functional complexity and performance is strongly motivated by the reduced cost per device and has been achieved in part by larger chip areas, but predominantly by smaller device dimensions.

ULSI circuit development for greater functional complexity and performance is strongly motivated by the reduced cost per device and has been achieved in part by larger chip areas, but predominantly by smaller device dimensions. A general guide to the smaller MOSFET devices was proposed by Dennard which was suitable for the chips until around 2005. Planar CMOS technology has reached its scaling limits at the 22nm node, where it is increasingly difficult to design high-performance low-power devices with good yield in the presence of global and local process variations. In continuing the march towards denser circuitry, however, it has become apparent that scaling the classical "bulk" MOSFET below the 22nm node is not practical on account of poor

1

electrostatic behavior. This has triggered research into silicon-on-insulator (SOI) structures like partially-depleted SOI and fully-depleted SOI (FD-SOI) with better short-channel effects (SCE), greater performance, and lower power consumption.

Table 1.1.   Scaling MOSFET device and circuit parameters.

	MOSFET device and circuit parameters	Constant field scaling
Scaling assumption	Device dimensions (t_{ox}, W, L) Doping concentration (N_A, N_D) Voltage (V)	Device dimensions (t_{ox}, W, L) k 1/k
Derived scaling behavior of device parameter	Electric field Carrier velocity Depletion layer width Capacitance Inversion-layer charge density Drift current Channel resistance	1 1 1/k 1/k 1 1/k 1
Derived scaling behavior of device parameter	Circuit delay time Power dissipation per circuit Power-delay product per circuit Circuit density Power density	1/k $1/k^2$ $1/k^3$ k^2 1

Table 1.1 shows the scaling MOSFET and circuit parameters under the constant field scaling and the generalized scaling rules. The generalized

scaling rule assumes that the electric field intensity changes by a factor of α. After, the conventional scaling ended, and the performance enhancement originating from the device scaling, such as microprocessor clock frequency has saturated. This performance saturation has been caused by short channel effects, which appeared in MOSFETs with aggressively shortened distance between the source and drain regions. Unfortunately, it is anticipated that SCE becomes more salient as the length is downscaled. One interesting option to further improve transistor characteristics is to use high mobility materials such as germanium and III-V materials. However, transistors have to be redesigned in order to fully benefit from these alternative materials. Silicon technology has entered the nano-CMOS era with 22nm MOSFETs in mass production in 2011 and sub-10nm transistors are expected to enter mass production in 2020.

As the gate length is scaled down, the channel becomes not only controlled by the gate but also affected by the drain. This behavior causes two-dimensional effects such as short-channel effect and drained-induced barrier lowering (DIBL). The SCE is often represented as a decrease of the MOSFET threshold voltage as the channel length is reduced. The short-channel effect is especially pronounced when the drain is biased at a high voltage equal to the power supply. Downscaled devices should be designed and optimized in order to keep SCE under control and minimized performance erosion. Device scaling causes unintended effects in device performance. These effects can be categorized according to the different sources:

- Electric field strength becomes very high in the channel – leading to

  1. Velocity saturation
  2. Impact ionization near drain
  3. Gate oxide charging
  4. Parasitic bipolar effect

- Electric field profile changes – leading to

  1. Drain induced barrier lowering
  2. Mobility reduction by gate-induced surface field

- Physical separation between the source and the drain decrease – leading to

  1. Punch through
  2. Channel length modulation

Some of the major problems to MOSFET scaling in sub-100nm channel length regime are:
- Short channel effect
- Increased gate leakage
- Poly gate depletion effects
- Source/Drain access resistance reduction
- High field mobility degradation
- Variability

In order to overcome these challenges significant innovations have been made at every new technology node. We saw introduction of new materials and architectures for the MOSFET in the last decade. To mitigate polysilicon depletion effect, polysilicon gate was replaced with metal gate. Alongside, as the gate leakage also became a serious issue due to tunneling current through thin SiO_2 layer, the gate oxide was replaced with high-k dielectric material ($HfSiON$, HfO_2). Stressor layers were introduced to boost mobility of carriers to counter high field mobility degradation.

As the gate length shrank even further, at around 30nm, short channel effects became very serious requiring paradigm shifts in the MOSFET architecture to continue scaling. Effective gate control on the channel had to be increased considerably to reduce detrimental short channel effects below 30nm. It has been demonstrated that this could be achieved reliably through multi-gate FET architecture or through fabrication of planar devices on ultrathin SOI substrate. Now there is a general consensus that sub-22nm nodes will require fully-depleted channel MOSFETs. Various industries have chosen to employ one of these architectures at sub-32nm nodes. Recently, for 22nm node, Intel announced introduction of tri-gate MOSFET on bulk substrate, thereby going from planar to quasi-planar channel architecture. ST Microelectronics plans to roll out its 28nm node devices on FDSOI

architecture. To continue improving the device performance further and to follow the International Technology Roadmap for Semiconductor (ITRS) other device architectures such as fin field effect transistors (FinFETs) are being investigated.

As shown in Figure 1.1 minimum feature size of MOSFETs, which corresponds to the gate length, has been downscaled from 10μm down to 28nm between 1970 and 2011. In parallel, the number of MOSFETs per square millimeter increased from 200 to over 1 million. The gate length in the current MOSFET generation lies between 22 and 17nm. In practice, the distance between the source and the drain is approximately 50% shorter than the gate length, resulting in an effective channel length of only ~10nm. It is predicted that the effective channel length will approach about 5nm in 2020.

For continuous MOSFET downscaling i.e., the performance enhancement, tremendous efforts have been made to solve various challenges. For instance, recent MOSFET structures released by Intel are shown in Figure 1.2. Technology boosters such as gate-stack, composed
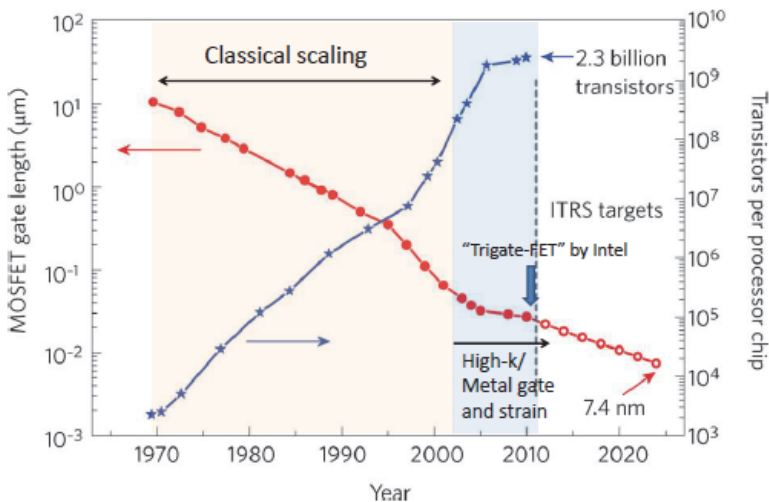
Figure 1.1. Time evolution of MOSFET gate length in microprocessors. ITRS roadmap projections show the gate length to scale down to around 8nm at the end-of-roadmap. After V. Deshpande, Scaling Beyond Moore: Single Electron Transistor and Single Atom Transistor Integration on CMOS, PhD Thesis, University of Grenoble, 2012.
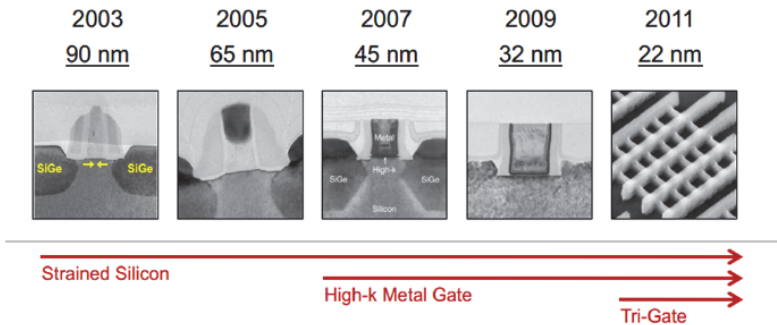
Figure 1.2. Recent MOSFET architectures and process nodes released by Intel Corporation. After M. Koyama, Electrical characterization of interface properties in nano-scaled MOSFET devices based on low-frequency fluctuations, PhD Thesis, Universite Grenoble Alpes, 2015.

of high-k dielectric/metal gate materials instead of SiO_2/Poly-Si, have been introduced from the 45nm process node in 2007 by Intel. SOI and multi-gate (MG) technologies have been proposed to improve electrostatic control by the gate and strongly reduce the SCE. First commercial MG MOSFET was released in 2011 by Intel using the structure of a bulk FinFET.

In an effort to extend CMOS scaling to beyond the conventional bulk device limit of about 10nm gate length, intense research activities are focused on alternative materials and device structures including high-k gate dielectrics, high mobility strained silicon channel, ultrathin silicon-on-insulator (SOI), and double gate (DG) MOSFETs. Some of the alternatives being actively pursued for future technology nodes are the multi-gate device architectures, alternate channel materials with multi-gate FETs (III-V and strained-Ge quantum wells and hetero-FETs), super-steep sub-threshold slope transistors (sub-60mV/dec), graphene based FETs and spin FETs. All of these approaches promise low power operation. Figure 1.2 shows the evolution of the Si process technology after the 90nm node.

ITRS provides specifications typically for two types of applications; one for low power technology (mobile, laptop), and for high performance for logic applications (microprocessors). High performance applications require an increase in the switching speed for logic, which is directly related to the current supplied by the transistor (I_{ON}), and a reduction
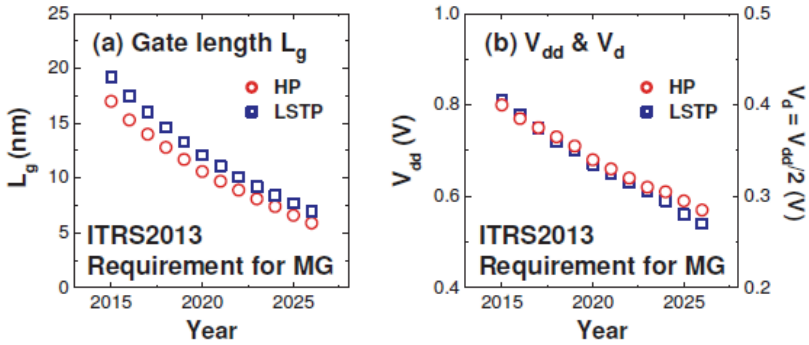
Figure 1.3. ITRS requirements for future high performance (HP) and low standby power (LSTP) logic circuits consisting of MG FETs. (a) MOSFET gate length L_g, and (b) circuit supply voltage V_d and MOSFET drain voltage V_d are plotted as a function of year between 2015 and 2026 (*Source*: ITRS 2013).

leakage current (I_{OFF}). The best architecture for a transistor is where we get the best compromise I_{ON}–I_{OFF} with good control of short channel effects. For the next device generation, multi-gate architecture with extremely shrunk body (nanowires) on bulk or SOI substrates has been proposed. The gate length of MOSFET should become sub-10nm within 10 years, and approach 6–7nm in 2026. It is expected that performance requirements could be achieved by multi-gate devices (Figure 1.3).

First the PDSOI, partially depleted SOI technology with excessive silicon film thickness (greater than 150nm) was employed so that the area of depletion below the conduction channel reaches the buried oxide. This results in a complete isolation of the transistor that is worth to reduce effects of certain parasites transistor, such as latch-up. With a silicon film thickness less than 10–20nm, in FDSOI, SOI is completely depleted, and provides better performance in both static and dynamic. In effect with a junction depth and a depletion thickness limited by the thickness of silicon film, we get better control of short channel effects, increased mobility and a reduction of capacity and junction leakage. Furthermore, the control gate on very thin silicon film leads to a low threshold voltage, enabling one to use a single mid-gap metal gate for both types of transistors. However, SOI devices have their own parasitic effects, for example, related to Joule heating. From the 22nm node and beyond a host of new devices/materials are being investigated, trying to

address this most critical bottleneck to transistor scaling i.e., power dissipation, both static and dynamic. Figure 1.4 summarizes the various types of advanced devices currently receiving special attention.

Strain/stress in the Si MOSFET channel results in higher carrier mobility, thus significantly boosting its ON state performance while HfO_2 with a higher dielectric constant (k ~ 25 compared to 3.9 for SiO_2) and gate dielectric thickness has allowed stronger gate coupling with the channel and an exponentially reduced gate tunneling leakage. The use of metal gate electrode has increased the channel mobile charge concentration by eliminating the polysilicon depletion, thereby resulting in higher drive current. The gate-stack, composed of high-k dielectric/ metal gate materials instead of SiO_2/Poly-Si, has been practically introduced from the 45nm process node in 2007 by Intel. This technology is going to be maintained and the materials will be further optimized to obtain better gate control and lower gate leakage current.

For high-speed operating devices, strained-Si channel exhibiting higher carrier mobility has been intensively investigated. Moreover, substitute
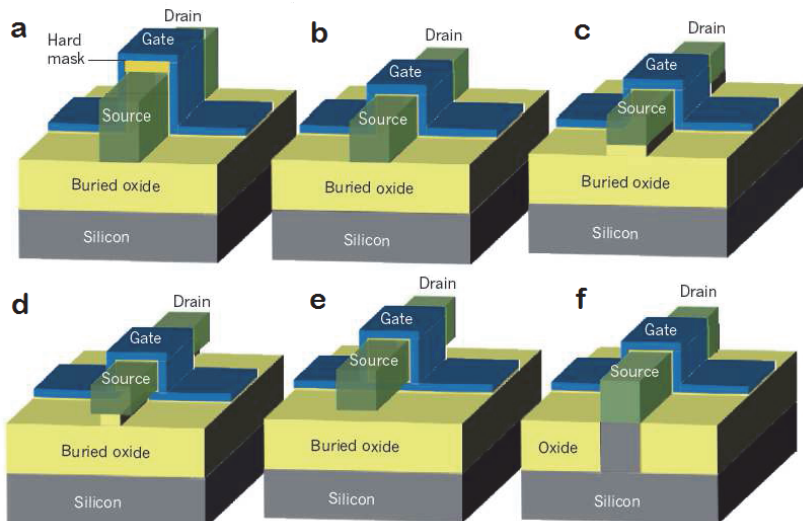
Figure 1.4. Types of multiple gate MOSFETs. (a) SOI DG FinFET, (b) SOI tri-gate FET, (c) SOI π-gate FET, (d) SOI Ω-gate FET, (d) SOI quadruple-GAA FET, and (e) bulk tri-gate FET. After M. Koyama, Electrical characterization of interface properties in nano-scaled MOSFET devices based on low-frequency fluctuations, PhD Thesis, Universite Grenoble Alpes, 2015.

materials, such as Ge and III-V compounds, have been also studied to achieve much higher mobility. SOI and MG technologies have been proposed to improve electrostatic control by the gate and strongly reduce the SCE. First commercial MG MOSFET was released in 2011 by Intel, using the structure of a bulk FinFET. Quantum Well based transistors can be quite suited for this, since they confine the charge carriers to the high-mobility material using a heterostructure. For the next device generation, MG with extremely shrunk body (nanowires) on bulk or SOI substrates has been intensely expected. The gate length of MOSFET should become sub-10nm within 10 years, and approach 6–7nm in 2026.

For the state-of-the-art transistor with 35nm gate length, the gate oxide is 1.2nm which corresponds to 100 A/cm^2 gate tunneling current at supply voltage of 1V. Constant electric field scaling requires the transistor supply voltage to be scaled down as the transistor dimensions are reduced. However, controlling the transistor leakage in the off-state became progressively more difficult as scaling continued. The off-state leakage current would increase by about ten times for every 0.1V reduction of threshold voltage. As the channel length is scaled down, the threshold voltage variation due to channel length variation or short channel effect becomes more prominent. This degrades the noise margin and therefore limits the power supply voltage scaling. For high-performance CMOS, a threshold voltage to supply voltage (V_t/V_d) ratio of <0.3 is desired.

DIBL is qualitatively similar to V_t roll-off in that it is caused by the encroachment of the depletion region from the drain into the channel. Strong DIBL is an indication of poor short channel behavior. However, since the transistor is not operated at low drain bias, DIBL itself is not a parameter that is directly related to circuit operation, but rather it is an indication of the degraded device characteristics such as strong V_t roll-off and high I_{ON}. I_{OFF} (I_d at $V_g = 0$) is one of the most important device characteristics directly related to SCEs. When designing a transistor, not only I_{ON}, but also I_{OFF} should be optimized. I_{OFF} is one of the critical parameters that determine the scalability of a given technology. Other parameters such as V_t roll-off and DIBL indirectly contribute to increasing the off-state leakage current.

During 1960 till early 2000, continuous enhancement of MOSFET performance has been achieved mostly by geometrical scaling as predicted by Gordon Moore in 1965. Device performance enhancement was being achieved by combining gate oxide thickness and gate length scaling. Since 2000s, MOSFET performance enhancements have been achieved mainly via device engineering while keeping the gate dielectric thickness almost constant. Three major innovations viz., channel stress and strain engineering for mobility enhancement (90nm and 65nm nodes), high-k/metal gate technology for reduced gate leakage (45nm and 32nm nodes), and tri-gate device architecture for improved electrostatic control (22nm node) have been introduced. Summary of various types of technology boosters via stress/strain engineering in front end processing is shown in Figure 1.5.

As the area of a MOSFET gets smaller due to scaling, the STI-induced stress in the active region becomes significant. The compressive stress from STI can lead to rapidly increased junction leakage and, depending on layout conditions, can degrade the n-MOSFET drive current. Intentional use of mechanical stress to enhance the MOSFET performance started at the 90nm CMOS technology node. A highly tensile nitride contact-etch-stop layer (CESL) is deposited on the top of
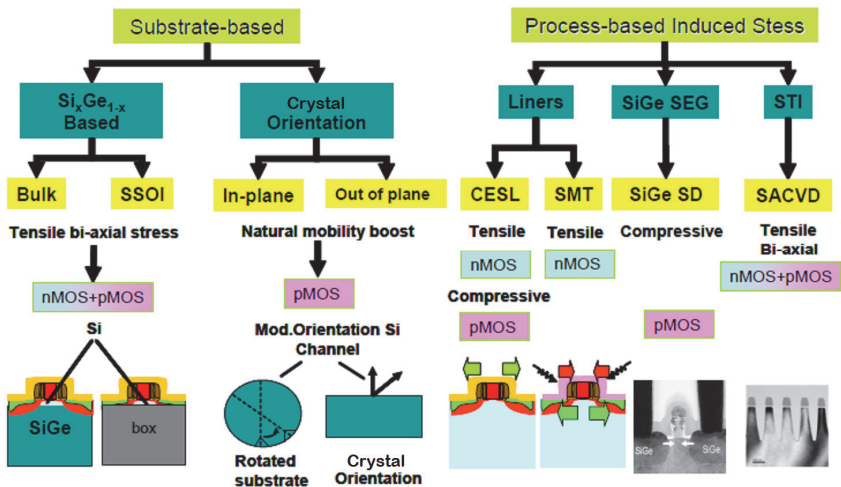
Figure 1.5. Schemes of front end stressors types. After V. Fiori, How do Mechanics and Thermo mechanics affect microelectronic products: Some residual stress and strain effects, investigations and industrial management, PhD Thesis, L'institut national des sciences appliquées de Lyon, 2010.

n-MOSFETs, while SiGe is selectively, epitaxially grown in recessed source/drain regions. A compressive CESL is deposited for p-MOSFETs. These methods introduce longitudinal uniaxial stress along ⟨110⟩ channels in contrast to the biaxial stress in an epitaxially grown Si layer on a SiGe virtual substrate. As schematically illustrated in Figure 1.6, tensile CESL causes tension in n-channel MOSFETs, and compressive CESL and SiGe source/drain areas introduce compressive stress in p-channel MOSFETs. Uniaxial stress techniques include:

- SiGe stressor (e-SiGe)
- Dual Stress Liner (DSL)
- Stress Memorization (SMT)
- Stress Proximity Technique (SPT)
- SiC Stressor
- SiGeSn Stressor
- Strain Transfer Structure (STS)
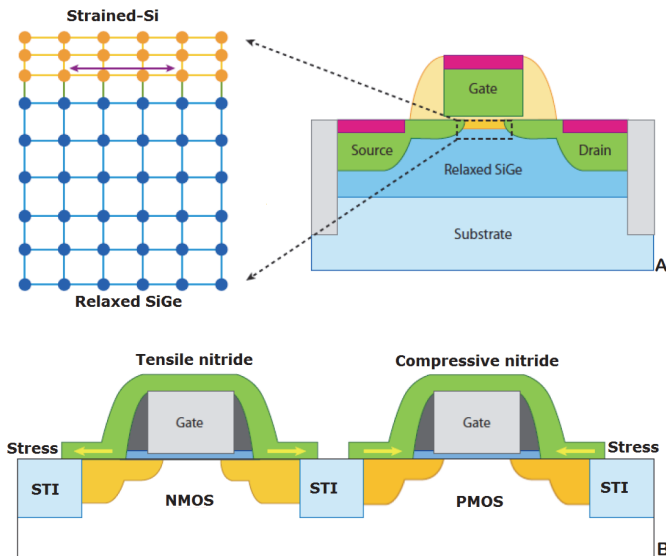
Figure 1.6. Illustration of process-induced stress on Si MOSFETs. (a) Biaxial stress caused from lattice mismatch between the Si channel and the relaxed SiGe substrate. (b) Uniaxial stress induced by the nitride capping layer. After M. Chu, Characterization and modeling of strained Si FET and GaN HEMT devices, PhD Thesis, University of Florida, 2011.

Careful analysis of biaxial and uniaxial strained-Si experimental data suggested that the industry adopt process induced uniaxial strain. The key observations are as follows. First, uniaxial (versus biaxial) stress provides significantly larger hole mobility enhancement at both high strain and high vertical electric field due to differences in the warping of the valence band under strain. Large mobility enhancement at low strain is important since yield loss via dislocations occurs at high strain. Second, uniaxial (as compared to biaxial) stress enhanced mobility provides larger drive current improvement for nanoscale short-channel devices. This results since the uniaxial stress-enhanced electron and hole mobility arises mostly from reduced conductivity effective mass (versus reduced scattering for biaxial stress), since uniaxial shear stress provides significant valence and some conduction band warping.

Even though the predominant focus of the industry in the 1980s and 1990s was on biaxially stressed devices [1.1, 1.2], the current focus has shifted to uniaxial stress. Encouraged by the strain-enhanced planar MOSFETs, researchers recently applied uniaxial stress to multi-gate devices with metal gate and high-k dielectric as a performance booster. Uniaxial stress has several advantages over biaxial stress, such as larger mobility enhancements and smaller shift in threshold voltage [1.3]. Lastly, process-induced uniaxial stress causes approximately five times smaller n-channel threshold voltage shift. Since any threshold voltage shift needs to be retargeted by adjusting channel doping (for industry standard poly-Si gate devices on bulk or partially depleted SOI), the larger threshold voltage shift for wafer substrate-induced biaxial tensile stress causes approximately half of the stress-enhanced electron mobility to be lost.

To predict the effect of strain in advanced technology, the strain-enhanced Si planar MOSFETs need to be understood first, which serves as the foundation of understanding the role of strain in advanced device structures. However, the multiple gates have different surface orientations that respond differently to stress. As such, the overall performance enhancement is not straight forward and needs detailed study.

High-k/metal gates were introduced into mass production in 2007 by Intel in the 45nm CMOS technology generation. This is the first time that

traditional oxides or oxynitrides have been replaced in gate stacks, to enable continuous scaling of the EOT. The replacement of SiO_2 by a high-k dielectric stack must satisfy a series of material constraints and process integration conditions. Although there are many potential high-k materials, based on their permittivity, a strict selection rules out many candidates. First of all, from a gate leakage perspective, a suitable conduction band offset is necessary to provide a sufficient barrier. The narrower band gap of high-k materials cancels the benefit of the high dielectric constant. Thereby, a suitable trade-off between the dielectric constant and the conduction band offset is the first criterion for high-k dielectric candidates. Initially, poly-Si/high-k combination gate stack was considered as a route to improving gate leakage. However theoretical studies and experimental data show mobility degradation compared to the use of metal gates. Metal gate implementation, however, has difficulties. Similar to the high-k dielectrics, the first issue is thermal stability of the work function. Gate-last or replacement-gate processes appear as a remedy for thermal stability problems. The high-k/metal gate stack is realized after all the front end processes, avoiding critical high temperature treatments. This requires the selective removal of a sacrificial poly-Si and oxide layer after planarization of the top of transistor by chemical mechanical polishing. The high-k/metal gate is thereafter subject to back end processes, such as interconnect deposition and packaging, the thermal budgets of which are usually lower than 500°C.

In this monograph we present the outlook for manufacturing advanced transistors for technology node 10nm and beyond. It is particularly intended to address the design and simulation issues associated with the front-end aspects of extending CMOS technology. Areas covered include transistor structures, front-end materials and front-end processes along with their characterization and modeling and simulation infrastructure. As the monograph encompasses broad areas of semiconductor technology, physics, modeling and characterization and this monograph being planned as a design book, original sources have been referred at appropriate places for interested readers for detail. Simulation based scaling study – including device design, performance characterization, and the impact of statistical variability – on decananometer bulk MOSFETs is considered. This monograph is intended both as a self-learning resource or professional reference and as a text for use in

graduate course. A key feature of the monograph is modular organization. The fourteen chapters may be read in almost any order. All simulation results presented throughout this monograph were performed using SILVACO and Synopsys TCAD tools.

Simulation Tools

Over the past decade, transistor structures have evolved a step further from planar, classical single-gate FETs to 3D multi-gate FETs whose behavior can only be fully understood via process and device simulations. Technology computer aided design was originally pioneered at Stanford University. In Chapter 2, overviews on currently available commercial TCAD tools will be presented. The program structures and models used will be described only briefly. Since these programs are in the public domain, the reader is referred to the user manuals for more details.

Simulation Methodology

In order to understand computer aided device design, it is necessary to discuss the evolution of the devices considered and the issues involved in their design. In this context, the Metal-Oxide-Semiconductor Field-Effect Transistors will be described first. Since its conception, MOSFET technology has improved steadily and has become the primary technology for ultra large scale integration primarily because of the simple device structure. Design issues in state-of-the-art device development such as drain-induced barrier lowering, trench isolation, hot electron effects, device scaling and interconnect parasitic are discussed. We shall present the basic concepts and models useful for understanding the device simulation. In Chapter 3, we discuss the effects of process-induced stress on the conduction and valence bands of silicon that will allow us to understand the use of piezoresistivity to model the stress induced to improve performance.

CMOS Technology

We first recall the basics of the CMOS transistor fabrication via process simulation which is the main topic of our studies as we must understand how process affects various parameters that determine the MOSFET

behavior. In Chapter 4, we shall use the SILVACO TCAD simulation tools for CMOS process development and device simulation.

Stress-Engineered CMOS

For several generations of technology, reduced dimensions MOS transistors is no longer sufficient to increase alone performance integrated circuits. The scaling challenges of recent MOSFET technologies demand the introduction of technology inventions and new materials. Remarkable advancements have already been achieved in channel and gate stack engineering. These innovative technology boosters known as equivalent scaling bring performance improvement. Strain engineering is the main technological booster used by semiconductor companies for the 65 and 45nm technology nodes to improve the channel mobility and the electrical performance of devices. The modeling of the stress by TCAD tools is introduced in order to estimate the stress generated and their optimization. Stress evolution modeling for thick layer deposition is presented. The general objective of this chapter is to evaluate the benefits and limitations of a technology that uses high-mobility channel materials and to design transistors that fully benefit from higher mobility. Simulation has been performed in order to explain in details how the CESL transmits stress to the Si channel. In Chapter 5, we explain and quantify the different contributions of the CESL towards the stress transfer in the channel. Finally, guidelines for CESL optimization are given.

Heterojunction Bipolar Transistors

The idea of varying the bandgap in a bipolar transistor structure to increase the emitter injection efficiency is almost as old as the bipolar junction transistor itself. Professor Herbert Kroemer was the first to explain the underlying principle of the heterojunction [1.4], which offers a larger set of device configurations and has become the basis for the so-called field of bandgap engineering. SiGe HBT technologies that exhibit higher operating speed can be leveraged for advanced circuits and systems in different ways; they can open up new applications at very high frequencies (THz). With steady performance gains and continued innovation in process integration, silicon-germanium (SiGe) heterojunction bipolar transistor BiCMOS technology is becoming an
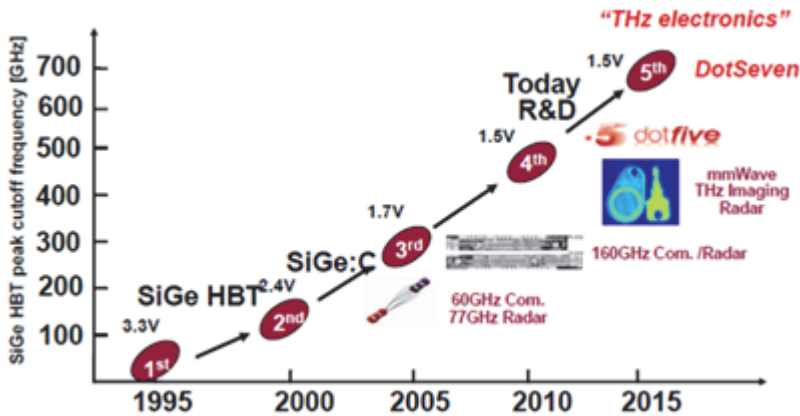
Figure 1.7. Evolution roadmap of SiGe HBTs. *Source*: ITRS 2013.

increasingly viable and affordable solution for highly integrated, high-performance mixed-signal applications. The SiGe-HBT roadmap is shown in Figure 1.7. Because of careful bandgap engineering, lateral scaling, and vertical profile optimization, state-of-the-art SiGe HBTs currently demonstrate peak unity-gain cutoff frequency in excess of 500GHz. In Chapter 6, we describe a systematic approach to investigate the sensitivity of device characteristics to process variations, as well as the trade-offs between different device designs.

Strain-Engineered HBTs

Stress and strain engineering technology has so far been used mostly in CMOS as a technology booster to enhance the carriers transport due to band structure changes and mobility enhancement. Devices with impressive cutoff frequency f_T values have been demonstrated that would have been believed to be reserved for III–V technologies. Due to the continuous demand for devices in areas of technology where high speed and high frequency response are required, it becomes imperative to develop new bipolar device architectures suitable for high frequency and power applications. However, the mobility of charge carriers in bipolar devices can also be enhanced by creating mechanical tensile strain in bipolar devices. SiGe HBTs have proven their capability to support large bandwidth and high data rates for high-speed communication systems.

New bipolar devices concepts and novel device architectures that are based on strain engineering technology may be explored using TCAD modeling. In Chapter 7, we explore the impacts of mechanical strain engineering technology principle on Si bipolar and SiGe heterojunction bipolar devices using TCAD process and device simulations.

FinFETs

Over the past decade, transistor structures have evolved from planar to 3D. Multi-gate 3D structures are expected to replace FDSOI and other scaled planar bulk devices. From the fabrication perspective, the most likely candidate for widespread adoption is the FinFET. Over the past few years, considerable research has been directed towards issues dealing with improving and integrating FinFET technology into the conventional CMOS process. The FinFET device structure consists of a silicon fin surrounded by shorted or independent gates on either side of the fin, typically on an SOI substrate. It is observed that fin shape significantly impacts leakage in bulk tri-gate FinFETs with thin fin widths when the fin body doping is optimized to minimize leakage. Fin shape can be used to implement multi-threshold FinFETs. Optimization of triangular fin with gate-source/drain underlap may lead to the design of ultralow power FinFETs. In Chapter 8, we have explored the fin shape dependence of leakage in bulk tri-gate FinFETs with thin fins when the fin body doping profile is optimized to minimize leakage. Fin shape is evaluated as a technique to construct multi-threshold n-FinFETs without increasing chip area consumption.

Advanced Devices

III-V compound semiconductor materials and Ge, which have superior transport properties, high breakdown electric fields or high thermal and/or high heat capacity and thermal conductivity are suitable for applications that are driven more by performance and less by cost such as high power and high frequency application and where silicon technology cannot meet the performance requirements such as high dynamic range or low noise figure. Figure 1.8 shows the prediction for GaN-based future technologies. In Chapter 9, different advanced devices involving III-V semiconductors and Ge are considered. A novel GaN-based high electron mobility transistor is considered which significantly increases
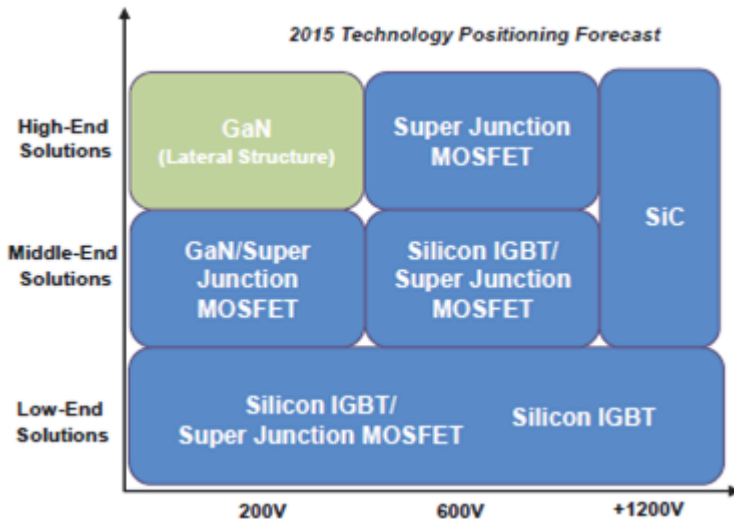
Figure 1.8. Prediction for GaN-based technologies. After D. J. Macfarlane, Design and fabrication of AlGaN/GaN HEMTs with high breakdown voltages, PhD thesis, University of Glasgow, 2014.

the drive current and thereby the power efficiency. In the recent years, International Technology Roadmap for Semiconductors has shifted the focus from More-Moore (Moore's law scaling) to More-than-Moore (MtM). Beyond CMOS, emerging technologies will reach device dimensions below 5nm and will use nanowires transistors, quantum devices, carbon nanotubes, graphene, or molecular electronics.

Memory Devices

As CMOS approaches the 22nm node and below, to address the scaling challenges, new materials and approaches are required towards memory design which use conventional CMOS materials. The flash memory is the dominant technology and offers a very high density on-chip and very low idle power consumption and low cost. The nonvolatile memory has drawn much attention over the past years due to their applications in consumer electronics market where memory devices with a retention time of ~10 years are desired. Lower operating voltages and faster switching can be achieved by using bandgap engineered gate stacks, multiple metal floating gates, thinner oxides, and tunneling as the main

programming mechanism. Floating gate devices are commonly used as pure storage elements but they may find applications as logic circuit elements within CMOS logic. So-called nanocrystal floating gate (NCFG) devices are superior in many aspects to its counterpart. NCFG devices may further be modified to use a high-k dielectric or a dual-layer structure to improve charge transfer and/or retention characteristics. In Chapter 10, we discuss the key challenges in scaling conventional FLASH memories and examine the scaling of nanocrystal floating gate devices.

Power Devices

Over the last decade, there has been a growing research interest in the area of power electronic applications. In general, bipolar transistors are not suitable for high speed switching applications because they saturate when their collector-base junctions is forward-biased. MOS power transistors have several advantages over their bipolar counterparts being a majority carrier device, simpler drive requirements, and lower forward voltages. Conventional MOS structure are also not suitable especially in medium and high voltage power ICs as both small gate length and thin oxide thickness are related to the breakdown. To alleviate the effect of the electrical field on the gate oxide, several power MOS device structures have been developed. The power MOSFET family can be divided into two different categories: lateral and vertical power MOSFETs. Super junctions are used in lateral double diffused MOS transistors (LDMOS) to greatly increase the breakdown voltage of small geometry devices. In Chapter 11, we shall present the results of 3D process/device simulation of buffered Super Junction LDMOS devices.

Solar Cells

Renewable energy sources have become increasingly important because of global environmental concerns. To make the solar cell technology more viable as an energy source, the study of how to increase the solar cell efficiency is important. A tandem solar cell usually has two cells having different band gaps placed one over other with an encapsulation between them. They are designed in such a way that the top cell has a larger band gap and hence can absorb shorter wavelengths. The bottom cell with a relatively smaller band gap can absorb longer wavelengths

that pass through the top cell. In Chapter 12, we describe the fabrication processes as well as the methods for physical and electrical characterization of solar cells. In Chapter 13, iron disilicide based heterojunction solar cell design is considered. Effects of various parameters such as work-function of ITO, electron affinity of $FeSi_2$, thickness of different layers and traps on heterojunction solar cell performance are investigated.

SPICE Parameter Extraction

While device simulators are reasonably accurate for a given transport model framework, they are very slow for large-scale circuit simulation. Here, compact models serve as a crucial link between process technology and circuit simulation, by leveraging inputs from TCAD simulation. SPICE models used in circuit design are traditionally extracted from measurements taken on working transistors fabricated in a process technology. A typical flow diagram (procedure) for SPICE parameter extraction using TCAD is shown in Figure 1.9. In Chapter 14, we provide a discussion on the methods used to characterize, simulate and optimize the performance of devices using TCAD tools and SPICE parameter extraction.
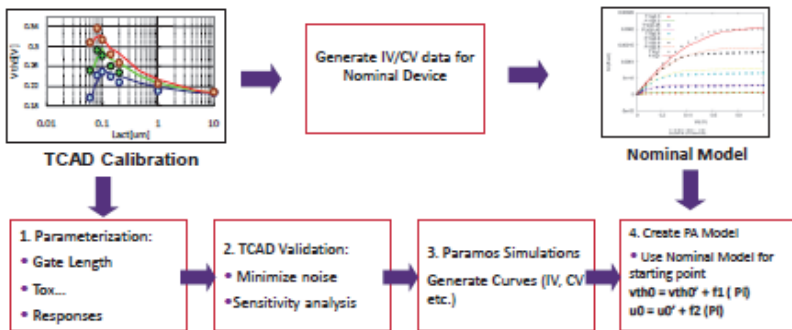
Figure 1.9. SPICE parameter extraction procedure using TCAD. Source. Dr. T. K. Maiti, private communication.

Summary

In this chapter we have presented a comprehensive overview of the monograph. Scaling of key parameters of bulk MOSFETs was considered first. We discussed the scaling rules: constant-field scaling and generalized scaling and we explored the new scaling features beyond the 90nm CMOS technology and the ITRS projections of design and performance over the next generations of devices. The scaling challenges facing CMOS were described in detail, including: the trade-off between power dissipation and performance; the vertical and lateral scaling challenges such as gate direct tunneling and short channel effects; reliability and statistical variability. Finally, the technology boosters, such as stress engineering and high-k/metal gates, all employed to enable continued scaling were presented. This chapter provides the basis of understanding needed for the next chapters presented in the monograph.

Chapter 2

Simulation Tools

In this chapter, currently available (in public domain) common semiconductor process and device simulation tools are introduced. The evolution of semiconductor simulation tools goes back to the late 1960s. Process development for the fabrication of semiconductor devices presents a formidable task and mandates the development of a range of advanced physical models for process and device simulations. Integrated circuits (IC) development for more than sixty years has been dominated by the MOS technology. In the 1970s and 1980s NMOS was favored due its speed, however, with technology limitations and concerns related to isolation, parasitic effects and process complexity. During NMOS dominated VLSI era, the emergence of process simulation took place and reached maturity in terms of realizing robust process design which is now an integral process technology design tool, and is used universally across the industry.

Starting in the late 1960s, the process and device simulation were predominantly in 1- and 2-dimensions. Device simulation, dominantly 2-dimensional owing to the nature of devices, became the workhorse in the design and scaling of devices. The transition from NMOS to CMOS technology resulted in the necessity of tightly coupled and full 2D simulators for process and device simulations. Due to ultra-small size of the state-of-the-art devices, the 3D effects have now been dominant. In order to achieve a better understanding of simulated and fabricated device characteristics, 3D process/device simulation has now been essential. Currently, there are several TCAD software platforms for process/device simulation around the world. Most of them were first developed by the universities worldwide and later integrated in a suite by the TCAD vendors. Since the early 1980s, commercial vendors such as SILVACO, ISE-TCAD, TMA and AVANTi developed their own

graphical user interfaces (GUI) around existing frameworks, which facilitated the integration of process and device simulation tools by non-specialists within a wider engineering environment. The concept of general purpose process and device simulators that allows flexible simulation of different structures in different technologies became a reality.

Since 1990s, there has been an increasing demand for simulators that provide the capability of modeling in a comprehensive manner for as many different types of semiconductor devices and processes as possible [2.1, 2.2]. This goal can either be achieved by incorporating new features in a simulator or by establishing links between simulators with different capabilities, such as coupling of electronic design (ECAD) with technology computer aided design (TCAD). Supporting the process and device engineers in an efficient and intuitive way in exploring the design space calls for a number of capabilities of the process simulation environment, such as free interaction and analysis of all aspects of a device, transparency with respect to real-life data and control to facilitate the management of large-scale and iterative experiments.

TCAD originated from bipolar technology in the late 1960s, solving the 1-D and 2-D process control issues. As the advent of computer modeling of semiconductor in the 1970s and the popularity of MOSFET technologies in the 1980s, tremendous efforts were put in the development of the TCAD. This electronic design automation became an essential toolset and gained broad deployment as a workhorse of integral technology in both semiconductor process and device simulation. The goal of TCAD is to offer interactive physical description and investigation of semiconductor devices by simulation and modeling to support circuit design as well as to reduce R&D time and cost. Using TCAD, conventional semiconductor device characteristics can be predicted with time and economic efficiency based on developed model and simulation algorithm, while novel process, structure and materials can be visualized through phenomenological and semi-empirical models to catch the physical insights and speed up the learning curve [2.2].

Commercially available simulation tools for semiconductor device and process simulations are MEDICI and SUPREM4 from Synopsys Corporation, ATHENA and ATLAS from SILVACO, TCAD studio etc.

For model parameter extractions AURORA from Synopsys and UTMOST from SILVACO are the most commonly used tools. MEDICI is a device simulator, which allows user to create a two dimensional structure of a semiconductor device, including the definition of oxide and silicon regions, doping profiles to simulate the current – voltage characteristics of the device. MEDICI features include transient and AC small-signal analysis, impact ionization, gate current and ionization integrals. There are different transport models available in device simulators for example, drift-diffusion and hydrodynamic models. AURORA is general purpose optimization tool for fitting analytical models to data and extract parameters for circuit simulation.

Tools are being designed to address the aforementioned challenges by combining the best-in-class features together with a wide range of new features arising out of new technology generations. Typical requirements for the TCAD simulation GUI are: 1. Easy to use – new users should be able to pick up the basics within minutes, 2. Real time help for the command syntax, and 3. Integration with other GUIs of the software suite, like the 3D setting up tool and plotting GUI. One highly demanding features of TCAD simulation GUI is the capability to perform batch simulation: this refers to a series of simulations associated with different process parameters and different device simulation parameters. In the semiconductor industry this is useful because optimizing device performance often requires adjusting process parameters. There may be hundreds of possible process step combinations so it is useful to automate the process as much as possible.

2.1 Introduction to Synopsys TCAD Tools

The Sentaurus TCAD toolset, provided by Synopsys, is a comprehensive nanoscale process, device design and parameter extraction simulation tool [2.3–2.13]. It supports industry leading process and device simulation with a powerful GUI-driven simulation environment for managing simulation tasks and analyzing simulation results. In this monograph, Synopsys Sentaurus TCAD tools will be used extensively for the semiconductor device and process design, characterization, modeling and analysis. The design and simulation flow in Sentaurus TCAD framework is shown in Figure 2.1. The device structure with

doping profile is designed by either Sentaurus Process (SProcess) or Sentaurus Structure Editor. The former is a 3D capable process simulator for silicon semiconductor process technology development and optimization. Its comprehensive process models cover implantation, diffusion, annealing, etching, oxidation, epitaxial growth, etc. The latter is not only a 3D-capable device editor but also a 3D-capable process emulator in which geometric and process emulation operations can be mixed freely, adding more flexibility to the generation of 3D structures. The Sentaurus Structure Editor is suitable for design of non-classical devices such as novel silicon transistor and non-silicon device. The device created by either Sentaurus Structure Editor or Sentaurus Process is a "virtual" device whose physical properties are assigned to a finite number of discrete grids of nodes. This grid adaptation procedure is completed by Sentaurus Mesh, which is a mesh generator that incorporates two mesh generation engines: an axis-aligned mesh generation engine and a tensor-product engine that produces rectangular or hexahedral elements.

The choice of the appropriate mesh generator depends largely on the geometry of the device and the essential surfaces within the device. For planar devices such as conventional MOSFET and layer-stack HEMT, axis-aligned mesh generator is recommended. For device where most important surfaces are non-axis-aligned or curved, the tensor-product engine Noffset3D is adopted to produce meshes containing layers to better conform to the curved surfaces. The mesh quality is controlled by the refinement information according to user requirements. Generally, a total node count of 2000 to 4000 is reasonable for most 2D simulations. 3D structures require a considerably larger number of elements. It is noted that the mesh quality will have impact on the simulation accuracy, efficiency and robustness.
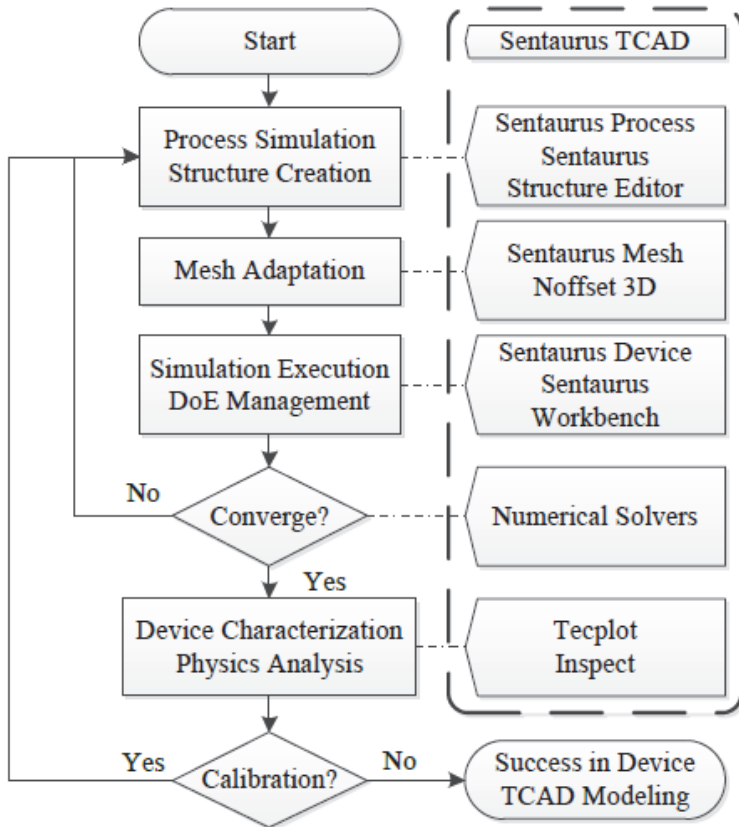
Figure 2.1. Sentaurus TCAD flow. After P. Feng, Design, modeling and analysis of non-classical field effect transistors, PhD Thesis, Syracuse University, 2012.

A most suitable mesh strategy should compromise fine mesh for high current density, high electric fields, high charge generation regions to ensure the accuracy and robustness, and coarse mesh to relatively low physics activity regions such as substrate and most source/drain regions for the improvement of simulation efficiency. After the device passed by the mesh engine, next steps include applications of physics models, bias condition and numerical solution algorithm in Sentaurus Device (SDevice), a comprehensive, semiconductor device simulator capable of electrical, thermal and optical device characteristics simulation. Similar to other TCAD toolsets, Sentaurus TCAD solves fundamental equations

along with its own advanced model in explaining of non-ideal device phenomena during device simulation.

Four different carrier transport models are available for the description of current density in various applications. The Drift-Diffusion (DD) model can be selected for long channel device isothermal simulation. The thermodynamic model accounting for self-heating effect is suitable for device with low thermal exchange. The hydrodynamic model can be applied in small active region device and the Monte Carlo (MC) model is capable of full band structure solution. Generally, one employs the first three carrier transport models and select the appropriate one to device simulation according to the device material, structure and operation, and interested device characteristics.

In addition to basic transport equations, TCAD modeling of semiconductor devices also need to address the band structure and the mobility properties. The band structure models are involved with bandgap, electron-affinity, and effective mass and effective density of states. The mobility models include the impact of the phonon scattering, impurity ion scattering, and carrier-carrier scattering and surface roughness. When one simulates devices other than MOSFET, particular models for material and physical phenomena should be included in the simulation. For instance, the tunneling and trapping must be considered in the simulation of a nonvolatile memory whereas spontaneous and piezoelectric model should be introduced in the investigation of heterostructure devices. Besides general simulation methods, phenomenological and semi-empirical models are often used to account for novel process, material or structure and to compromise the simulation efficiency and robustness. For example, the high-k metal gate stack can be modeled through changes to the equivalent oxide thickness with an additional dipole layer; the Ohmic contact on the nitride materials can be visualized as a Schottky contact with only electron tunneling to improve the simulation convergence rate.

The simulation convergence as well as simulation time depend on the iterative algorithm settings of the Newton solver. Several factors are needed to be addressed for the trade-offs among simulation efficiency, accuracy and robustness: the maximum number of iterations; the desired precision of the solution; the linear solver appropriate for interested

device operation; and the introduction of damping methods in expediting the initial solution search at the cost of resultant accuracy. To achieve good convergence rate of simulations, efforts need to be put on the compromise among the physics model, mesh quality and solution algorithm. It should be pointed out that examination of the simulated device physics properties such as electron density, electric field and electrostatic potential will help identify the cause to the convergence issue.

Mixed-Mode Simulation

TCAD being a powerful and rigorous single device simulator, it can be extended further for the mixed-mode device and circuit simulation to perform multi-device simulation. Similar to SPICE, TCAD needs to define a circuit netlist to virtually connect the active and the passive components and solve the whole system of devices. In addition, TCAD mixed-mode simulation can add more device physics details during the circuit operation and provide results more close to the real integrated circuits.

Design of Experiment

Through the careful design and setting of the simulation, modeled devices with good convergence rate can be further investigated by Design of Experiments (DoE) for calibration, fine tuning or physics properties study. The DoE for semiconductor process and device design refers to variation of fabrication parameters that can be applied for device development. It aids to obtain information for device characterization and optimization, to improve manufacturing processes, and to design and develop new processes and products. Proper DoE and the choice of samples for analysis are important factors that increase the likelihood of success in semiconductor device fabrication. DoE can be performed in Sentaurus Workbench, the primary graphical front end that integrates Sentaurus simulation programs into one environment. Also in the workbench, graphic visualization can be obtained through Tecplot and device characteristics variables can be extracted through Inspect.

Simulation Example for Sentaurus TCAD

The TCAD flow begins with simulating a layout in process simulation to create a structure representing a semiconductor device. The structure is prepared for device simulation by remeshing and adding contacts. After device simulation is complete, the electrical response is captured in plot form. This plot is analyzed to extract electrical parameters. This methodology is implemented into the Synopsys TCAD framework. Figure 2.2 illustrates a general TCAD simulation methodology that begins with process simulation and ends with electrical parameter extraction.
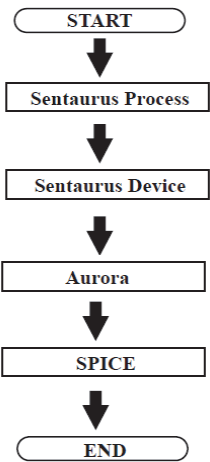
Figure 2.2. TCAD simulation flow begins with process simulation and ends with electrical parameter extraction.

Simulation Setup

Family Tree						Variable Values	
SPROCESS		SDE		INSPECT			
tool1		tool2		tool3			
	param1		param2		param3	VAR1	square
1			[n12]: ABC	[n22]: --	[n32]: lin	2	1
2	[n2]: 1	[n7]: --			[n33]: log	2	1
3			[n13]: DEF	[n23]: --	[n34]: lin	2	1
4					[n35]: log	2	1

Figure 2.3. Simulation setup in Sentaurus Workbench.

Synopsys TCAD tools include the process simulator, structure editor, device simulator, and curve plotting utility are integrated into the Sentaurus Workbench as shown in Figure 2.3. The tool sequence in the workbench starts with the process simulator, Sentaurus Process, to simulate transistors. The Sentaurus Structure Editor creates contacts and optimizes the mesh for device simulation. Sentaurus Device runs a series of DC, AC, and transient mixed-mode simulations. The resulting I–V plots are analyzed by Inspect, the curve plotting and extraction tool. By performing the complete simulation flow within the workbench, each transistor can be fully studied for its charge injection/clock feed through characteristics.

Layout Driven Process Simulation

Process simulation creates a semiconductor device representation using modern semiconductor processing methods, such as diffusion, oxidation, etching, deposition, and implantation. Details and complexities of performing process simulation may be found in reference SProcess manual. Layout driven process simulation is an efficient method for creating various structures within one process flow. In order to utilize the tool's layout driven simulation capability, a layout must be provided along with a defined simulation domain. Figure 2.4 is an example of defining a cross section in a layout to be simulated by Sentaurus Process.
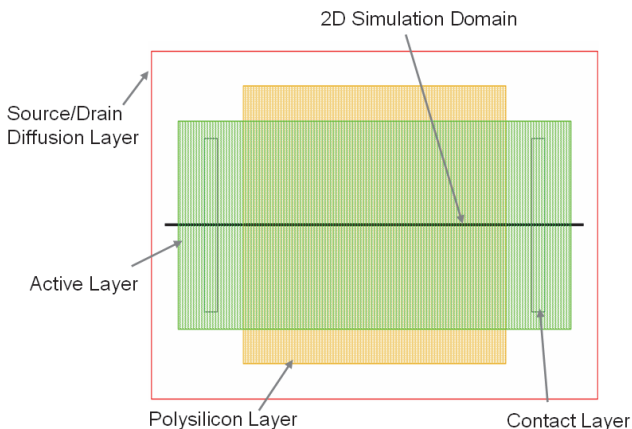
Figure 2.4. Defining the simulation cross section in the NMOS top view.

Model Selection for Calibration

One of the challenges in process simulation is to reproduce an accurate representation of the semiconductor device. Models for a given technology are specified in the process simulator to ensure accurate results. This involves proper selection of diffusion models, implantation models, and boundary conditions. For example, it is important to choose the point defect assisted dopant diffusion model when anneal follows an ion implantation step with a moderate dose.

Meshing

TCAD simulators are based on finite element solutions to solve the non-linear partial differential equations. Therefore, good meshing techniques are required to perform numerically stable and accurate simulations. Optimum meshing is important to maintain accuracy without dramatically increasing process simulation time. In a transistor, an initial mesh is defined by specifying the spacing in the vertical direction. Additional mesh refinements are necessary in critical areas of the device, such as the channel, source, drain, and lightly doped drain (LDD) regions. This technique is illustrated in Figure 2.5.

The key process steps in the NMOS process flow are shallow trench isolation (STI) formation followed by well implantation to set a threshold
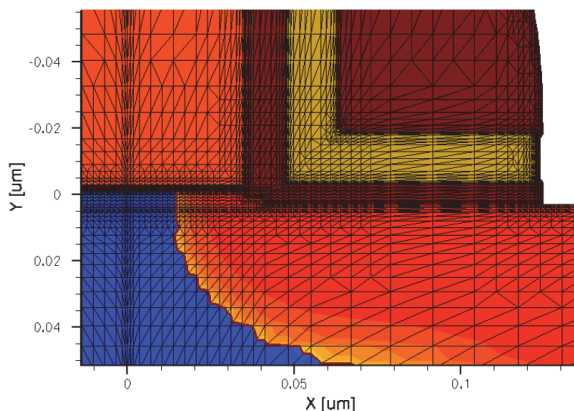
Figure 2.5. Mesh requirements are defined differently for different areas of the device. An initial vertical spacing is specified followed by denser mesh specifications in the channel, source, drain, and LDD regions.
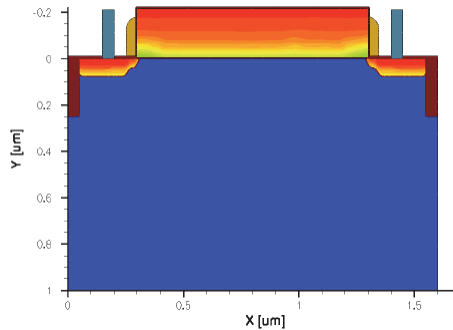
Figure 2.6. 2D process simulation generates a full cross section of an NMOS transistor.

voltage of 0.5V. After 50Å of gate oxide is grown, polysilicon deposition and etch occur to form the transistor's gate. LDD implantation is self-aligned to the gate to reduce electric field, followed by nitride deposition to form the sidewall spacer. After source/drain implantation, the damage is annealed using rapid thermal anneal. Figure 2.6 shows the final structure after the process simulation has completed.

Structure Editing

In order to prepare a structure for device simulation, structure editing is performed using Sentaurus Structure Editor. The NMOS structure from process simulation is reduced in size by removing the lower portion of the substrate. Next, contacts are placed for the substrate, gate, source, and drain. After contact placement, the structure is ready to be remeshed. Since mesh requirements are different for process and device simulations, the structure editor is also utilized to remesh the structure prior to device simulation. Optimum mesh is important to resolve key regions of the transistor pertaining to charge injection and clock feed through; these areas include the inversion layer and overlap regions. To properly resolve the inversion layer, mesh in the gate oxide is 10Å, followed by a fine mesh in the first 300Å of the channel. In the channel's first 10Å the mesh spacing is 2Å, and as the channel depth increases to 300Å the mesh spacing increases to 20Å. The source and drain are remeshed to complete the NMOS device.

Device Simulation

In order to fully understand the electrical response of a semiconductor device, different device simulations are performed. Sentaurus Device is used to perform DC, AC, and transient mixed-mode simulations. DC simulations provide $I_d - V_g$ and $I_d - V_d$ characteristics, AC analysis generates C–V characteristics. Sentaurus Device is utilized to perform a series of DC, AC, and transient mixed-mode simulations and the resulting I–V plots are analyzed by Inspect.

3D Simulation Methodology

Given present day memory and computer processing power, 2D simulation is fast and used as much as possible. However, when it is necessary to capture effects in the third dimension, 3D simulation is utilized. Effects such as width and explicit three-dimensional structures can be studied. Discussion of 3D simulation methodology focuses primarily on the setup of process simulation and the structure editor. Device simulation is nearly identical to the 2D setup and the methodology used to generate 3D structures is discussed briefly. The setup of 3D simulations is based on 2D simulations. For example, process flow, calibration models, and device simulations are nearly identical in 2D and 3D simulations. The main difference in 3D simulations is how the structure is created from process simulation. Challenges with 3D process simulation involve the meshing complexities with moving boundaries. Therefore, present day 3D process simulators cannot stably perform the full processing steps that 2D simulators can. The Synopsys TCAD tools perform 3D process simulation by dividing the tasks between Sentaurus Process and Sentaurus Structure Editor. During process simulation, Sentaurus Process performs all implantation and diffusion steps, and internally calls upon Sentaurus Structure Editor to perform etching and deposition.

In order to reduce simulation time, Sentaurus Process simulates one-quarter of the transistor and Sentaurus Structure Editor performs resizing and reflection to form the full transistor. Simulation of the standard NMOS begins by defining the 3D region to be simulated by Sentaurus Process. By taking advantage of the standard NMOS symmetry, only the gate and drain portions are simulated in 3D process to save

computational resources. The resulting structure is later reflected in Sentaurus Structure Editor to form the full NMOS structure. Advanced TCAD Tools from Synopsys includes:

- Sentaurus Workbench, Framework tool for the co-ordination and analysis of multiple simulation runs
- Sentaurus Workbench Visualization, Extraction and display of physical and electrical characteristics
- Sentaurus Process, Multidimensional semiconductor process simulator
- Sentaurus Process Kinetic MC, Atomistic kinetic Monte Carlo implantation and diffusion simulator
- Sentaurus Device, Multidimensional device simulator. Includes Advanced, 3D, DSM, Compound, Optical, Parallel, EMW, Power and Monte Carlo options
- Sentaurus Structure Editor, Multidimensional process emulator and graphical device structure editor. Includes 3D option
- Sentaurus PCM Studio, Builds process compact models for parametric yield analysis
- Sentaurus PCM Library, Builds process compact models for parametric yield analysis
- DIOS, Process simulator for semiconductor device fabrication
- Taurus Modeling Environment, Framework tool and multidimensional visualization package
- Taurus Medici, Device simulation tool
- Taurus TSUPREM-4, Process simulation tool
- Raphael NXT, 2D/3D field solvers for interconnect
- Aurora, Device characterization and parameter extraction

2.2 Overview of SILVACO TCAD Tools

SILVACO International provides a comprehensive set of electronic design automation software, which allows semiconductor companies to design both analog and mixed-signal integrated circuits. SILVACO delivers products like TCAD tools for process and device simulation [2.14-2.19]. It provides analog semiconductor process, device and design automation solution in CMOS, bipolar transistors, SiGe, strained-Si and

other compound semiconductor technologies. The main process and the device simulation tools are the ATHENA and ATLAS for process, device simulation and characterization, respectively.

SILVACO provides Virtual Wafer Fab (VWF) tool involving process simulation, device simulation, and process and device characterization. A brief overview of SILVACO TCAD software is presented in Figure 2.7. The Virtual Wafer Fab software allows split parameters to be defined for any of the SILVACO process, device and circuit simulators and all simulations are carried out in parallel either on a cluster of workstations or on a single machine. This allows either multiple computers to parallel computer as a project or to use a single computer attached to multiple processors, as for the process simulation software depends on the degree of dimension simulated. In addition to these softwares, SILVACO provides other softwares that can be used as peripheral tools for conducting simulations. These includes TonyPlot, TonyPlot3D, DeckBuild, MaskViews, and DevEdit. Figure 2.7 describes the flow diagram when simulating through SILVACO TCAD. All these core tools are activated together through DeckBuild interface.
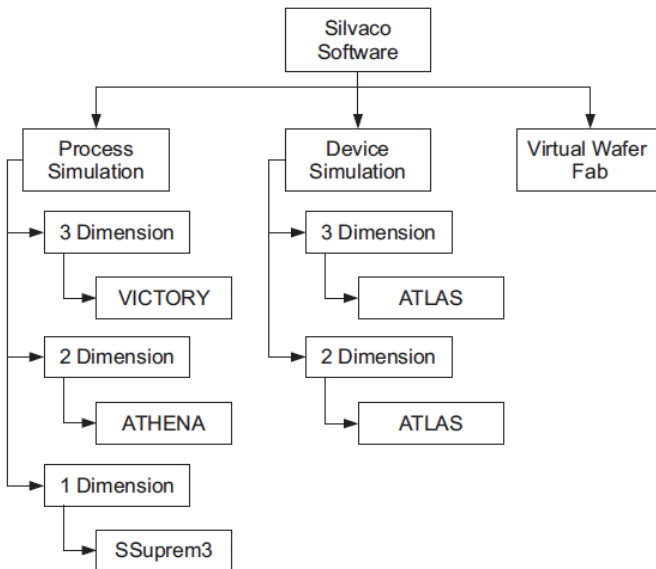
Figure 2.7. A brief overview of SILVACO TCAD software.

DECKBUILD

Interactive Deck Development and Runtime Environment DeckBuild is an interactive runtime and input file development environment within which all SILVACO TCAD and several other SIMUCAD products can run. DeckBuild has numerous simulator specific and general debugger style tools. This includes powerful extract statements, GUI based process file input, and line-by-line runtime execution and intuitive input file syntactical error messages. DeckBuild contains an extensive library of pre-run examples decks which cover many technologies and materials, and also allow the user to rapidly become familiar with simulation. Key features of DeckBuild are:

- Provide an interactive runtime and input file development environment for running several core simulators
- Input deck creation and editing
- View simulator output and controls
- Set popups that provide full language and run-time support for each simulator
- Automatic interface among simulators
- Many input file creation and debug assist features, such as run, kill, pause, stop at, and re-start
- Extracted quantities can be used as targets in DeckBuild internal optimizer, allowing automatic cyclical optimization of any parameter.

TONYPLOT: 2D/3D Interactive Visualization Tools

TONYPLOT is a powerful tool designed to visualize TCAD 2D and 3D structures produced by SILVACO TCAD simulators. TONYPLOT provides visualization and graphic features such as pan, zoom, views, labels and multiple plot support. TONYPLOT also provides many TCAD specific visualization functions, 1D cut lines from 2D structures, animation of markers to show vector flow, integration of log or 1D data files and fully customizable TCAD specific colors and styles. Some of the key features of TONYPLOT are:

- Flexible graphical analysis tool specifically developed for TCAD visualization assists in rapid prototyping and developing of process and device designs

- Common visualization tool across all SILVACO TCAD products
- Plotting engine supports all common 1D and 2D data views including: 1D x–y data, 2D contour data, 2D meshed data, smith charts and polar charts
- Exports data in many common formats for use in reports or by third party tools.
- Supported formats include; jpg, png, bmp, Spice Raw File and CSV
- Flexible Labels allow plots to be annotated to create meaningful figures for reports and presentations
- Integrated suite of probes, rulers, and other measurement tools allows detailed analysis of 1D and 2D structures
- Overlays allow multiple plots to be easily compared
- Overlaying 1D log files enables visualization of how process conditions effect on electrical results
- Cutline tool allows 1D slices to be generated from 2D structures. Slicing can be automated to generate several slices through a structure
- Function and Macro editor allows complex functions and macros to be defined that can be visualized as normal 1D quantity. This feature allows calculation of M-Plots for OLED devices
- Quasi 3D mode allows visualization of multi-dimensional data
- Fully customizable including; colors, materials, legends, toolbars and shortcuts

DevEdit

DevEdit can be used to either create a device from scratch or to remesh or edit an existing device, creating standard SILVACO structures that are easily integrated into SILVACO 2D or 3D simulators and other support tools. This software allows the user to define areas of the device that requires higher resolution and accuracy during simulations.

SILVACO ATLAS

ATLAS enables one to simulate the electrical, optical, and thermal behavior of semiconductor devices. ATLAS provides a physics-based, easy to use, modular, and extensible platform to analyze DC, AC, and time domain responses for semiconductor devices in 2 and 3 dimensions.
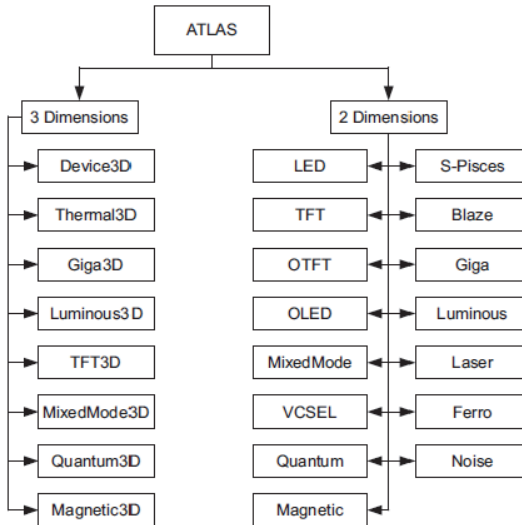
Figure 2.8. SILVACO ATLAS framework architecture.

Accurately characterizes physics-based devices in 2D or 3D for electrical, optical, and thermal performance without costly split-lot experiments. It solves yield and process variation problems for optimal combination of speed, power, density, breakdown, leakage, luminosity, or reliability. Figure 2.8 shows the ATLAS Framework. ATLAS is able to conduct simulation based on quantum theories, magnetic effect, lattice heating, and many more. Collectively, these features allow one to conduct a wide variety simulations related to the modern semiconductor technologies. ATLAS device simulation program for 2 dimension and 3 dimension utilizes the same program, but depending on the degrees of dimensions, different smaller programs are packaged into ATLAS as presented in Figure 2.8. Figure 2.9 shows the flow diagram of integrated ATLAS tool. Simulation flow diagram of SILVACO ATLAS is shown in Figure 2.10.

Blaze – 2D Device Simulator for Advanced Materials

Simulates devices fabricated using advanced materials. Includes a library of physical models and material parameters for binary, ternary and quaternary semiconductors.

S-Pisces – 2D Silicon Device Simulator

Advanced 2D device simulator for silicon based technologies that incorporates both drift-diffusion and energy balance transport equations. A large selection of physical models is available which include surface/bulk mobility, recombination, impact ionization and tunneling models.

Giga – 2D Non-isothermal Device Simulator

Simulates self-heating effects when combined with S-Pisces or Blaze. Models include heat generation, heat flow, lattice heating, heat sinks, and effects of local temperature on physical constants. Thermal and electrical physical effects are coupled through self-consistent calculations.

Device3D – 3D Device Simulator

Device3D is a 3D device simulator for silicon and other material based technologies. The DC, AC and time domain characteristics of a wide variety of silicon, III-V, II-VI and IV-IV devices can be analyzed. Device3D can accurately characterize physics-based devices for electrical, optical, and thermal performance without costly split-lot experiments. Device3D solves yield and process variation problems for optimal combination of speed, power, density, breakdown, leakage, luminosity and reliability.

Thermal3D – Thermal Packaging Simulator

Thermal3D is a general heat flow simulation module that predicts heat flow from any power generating devices (not limited to semiconductor devices), typically through a substrate and into the package and/or heat sink via the bonding medium. Operating temperatures for packaged and heat sinked devices or systems can be predicted for the design and optimization phase or for general system analysis.

MixedMode3D – Circuit Simulation for Advanced 3D Devices

MixedMode3D is a circuit simulator that includes physically-based 3D devices in addition to compact analytical models. Physically-based

devices are used when accurate compact models do not exist, or when devices that play a critical role must be simulated with very high accuracy. Physically-based devices are placed in a SPICE net list circuit description and may be simulated using any combination of ATLAS3D modules.

ATLAS Inputs/Outputs
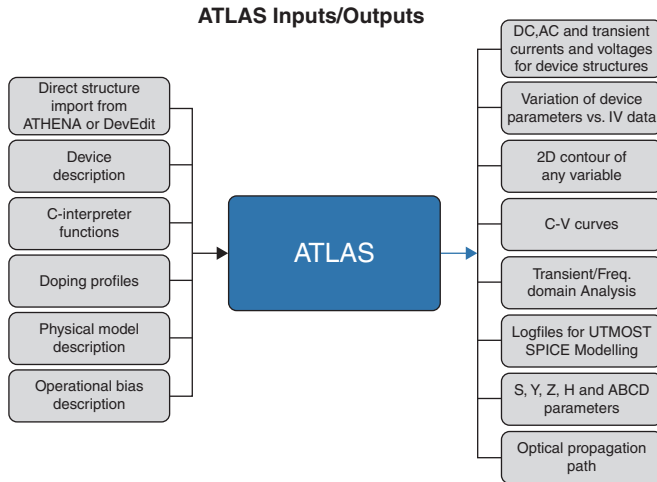
Figure 2.9. SILVACO ATLAS inputs and outputs. After W. Y. Jin, Mobility enhancement of nanoscale biaxial strained silicon metal-oxide semiconductor field effect transistor, Master's Thesis, Universiti Teknologi Malaysia, 2009.
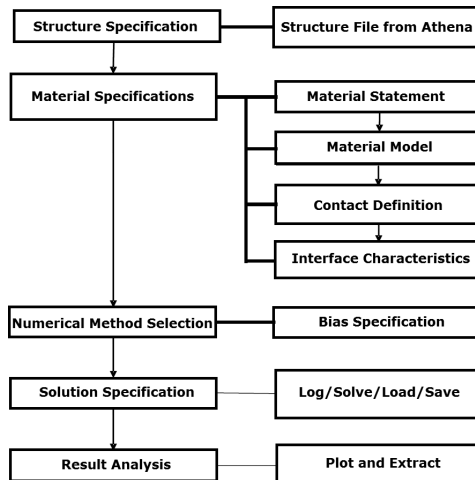
Figure 2.10. Simulation flow diagram of SILVACO ATLAS. After W. Y. Jin, Mobility enhancement of nanoscale biaxial strained silicon metal-oxide semiconductor field effect transistor, Master's Thesis, Universiti Teknologi Malaysia, 2009.

For device simulation using ATLAS, order of statements are as follows:

Structure Specification
- Mesh
- Region
- Electrode
- Doping

Material model Specification
- Material
- Models
- Contact
- Interface

Numerical Models Specification
- Method

Solution Specification
- Log
- Solve
- Load
- Save

Result Analysis
- Extract
- TonyPlot

Silvaco ATHENA

ATHENA software is used for the process simulation and generation of the device structure. ATLAS software is used for investigation of the electrical characterization of the device. ATHENA has evolved from a world-renowned Stanford University simulator SUPREM-IV, with many new capabilities developed in collaboration with dozens of academic and industrial partners. SSuprem3 is used for one dimensional problem solution, while ATHENA employs SSuprem4 and a collection of other smaller programs as presented in Figure 2.11. Fully-integrated ATHENA
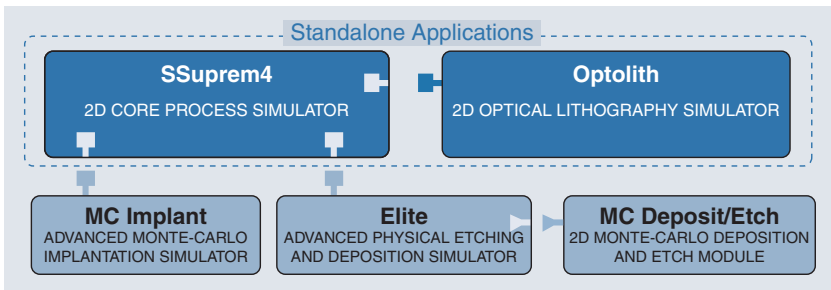
Figure 2.11. SILVACO ATHENA framework architecture.
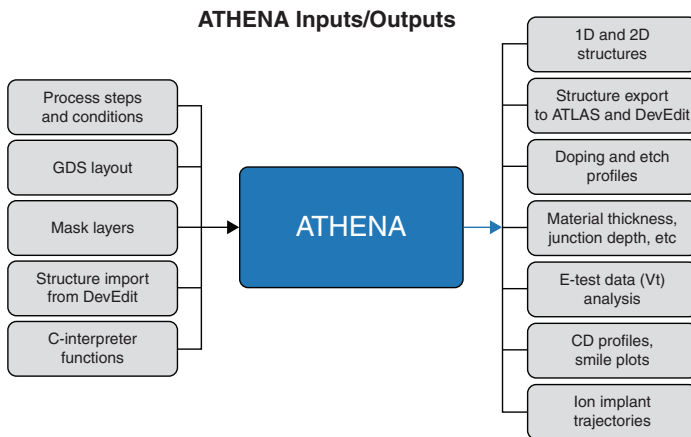
Figure 2.12. SILVACO ATHENA inputs and outputs.

process simulation software includes several comprehensive visualization package and extensive database of example. One can choose from the largest selection of silicon, III–V, II-VI, IV-IV, or polymer/organic technologies including CMOS, bipolar, high voltage power device, VCSEL, TFT, optoelectronic, LASER, LED, CCD, sensor, NVM, ferroelectric, SOI, FinFET, HEMT, and HBT technologies. For process simulation using ATHENA, the user is required to input several parameters in accordance to ATHENA requirements, as presented in Figure 2.12.

VictoryProcess

VICTORY PROCESS (VictoryProcess) is a fully three-dimensional process simulator which allows user to perform a wide range of device

structures of desired shape [2.20]. Most of the supported operations correspond to real processes such as etching or deposition, CMP, epitaxy and others so that the process engineer can establish a direct link between the technological processes and an input statement for VictoryProcess. VictoryProcess can simulate complicated full physics-based etching and deposition processes. Advanced process simulation may be performed very quickly and accurately using either in the 'geometrical' mode or simple physical models. VictoryProcess takes into account; reactor characteristics (particle flux), shading effects and/or, and secondary effects like redeposition of the etched material. Figure 2.13 shows the inputs and outputs flow. The etching and deposition process characteristics in VictoryProcess include:

- Rounded corners
- Tapered sidewall
- Non-conformal epitaxial growth
- Selective geometrical etch
- Selective CMP

VictoryProcess is capable of simulating complex processes within a feature scale for three-dimensional structures. The technological parameters from the reactor scale can be included by means of transport characteristics. For example, different crystal planes of silicon are known to have different oxidation rates, e.g., the silicon plane with Miller indices $\langle 111 \rangle$ is oxidized approximately 1.7 times faster than the $\langle 100 \rangle$ plane. During an oxidation process the geometry of a 3D structure may change significantly and the silicon/oxide interface may pass through various crystal planes with different oxidation rates. Therefore, the orientation and type of the silicon wafer affects the resulting geometry of
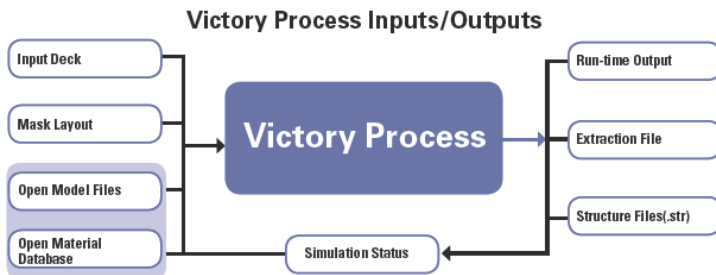
Figure 2.13. SILVACO VictoryProcess inputs and outputs.

an oxidized structure on this wafer. VictoryProcess offers several modes for the simulation/emulation of an oxidation process, namely,

- the analytical oxidation mode
- the empirical oxidation mode
- the full physical oxidation mode

VictoryDevice

VictoryDevice is a general purpose 3D device simulator [2.21]. A tetrahedral meshing engine is used for fast and accurate simulation of complex 3D geometries. VictoryDevice performs DC, AC, and transient analysis for silicon, binary, ternary, and quaternary material-based devices. Figure 2.14 shows the inputs and outputs flow for VictoryDevice.

VictoryDevice features include:

- Tetrahedral mesh for accurate 3D geometry representation
- Advanced physical models with user-customizable material database
- Stress-dependent mobility and bandgap models
- Physical models using the C-Interpreter or dynamically linked libraries
- DC, AC, and transient analysis
- Drift-diffusion and energy balance transport equations
- Self-consistent simulation of self-heating effects including heat generation, heat flow, lattice heating, heat sinks, and temperature-dependent material parameters

Figure 2.14. SILVACO VictoryDevice inputs and outputs.

- Advanced multi-threaded numerical solver library
- Atlas-compatible

VictoryStress

VictoryStress is a generic 3D stress simulator that allows accurate prediction of stresses generated during semiconductor fabrication as well as assist users in various aspects of stress engineering [2.22]. The three main modules in VictoryStress are: input file generation, 3D VictoryStress simulation, and output generation. Figure 2.15 shows the inputs and outputs flow. Accurate simulation of mechanical stress/strain distributions generated during device fabrication is an important part of technology and device design. In many cases, stress effects should be taken into account to predict better manufacturability and to increase reliability of semiconductor devices. In recent years, the stress simulation has become a critical issue in TCAD due to advances in stress engineering. Stress engineering could be described as a collection of device optimization methods based on deliberate introduction of stress in device structure. These methods include Dual Stress Liners, strain SiGe channels, SiGe (or SiC) S/D pockets, hybrid orientation for NMOS and PMOS devices, and etch stop layer. All these methods are used to improve device performance by altering carrier mobility in specific areas of the device structure. VictoryStress provides comprehensive set of models and capabilities covering various aspects of stress simulation and stress engineering:

- Stress analysis of arbitrary 3D device structure
- Import of device structures from VictoryProcess, VictoryCell, and ATLAS3D
- Export of 3D stress distributions to VictoryDevice and ATLAS3D

Figure 2.15. VictoryStress information flow.

Models for various sources of strain and stress include:

• Thermal mismatch between material layers
• Local lattice mismatch due to doping
• Initial intrinsic stress in specified regions
• Hydrostatic stress from capping layers
• Stress/strain generated in previous processing step (e.g., oxidation)
• Stress simulation for various crystalline (e.g., Si, SiGe, GaAs) and isotropic (e.g., silicon nitride and oxide) materials
• Generic 3D anisotropic stress simulation accounts for wafer orientation and arbitrary wafer flat rotation
• Estimation of mobility enhancement factors (p- and n-type) by use of piezoresistivity model
• Can be used with Virtual Wafer Fab for Design of Experiments to analyze stress dependence on process and geometrical parameters of semiconductor

VictoryCell

VictoryCell is a 3D process simulator designed to be a very fast way of creating devices [2.23]. The choice of mesh formats that can be output are optimized to be as device simulator friendly as possible, minimizing device simulation times. This makes VictoryCell an ideal tool for simulation of numerous technologies, such as SiC IGBT and MOSFETs. Figure 2.16 shows the input/output flow information. GDSII Layout Driven Process Simulation has the following advantages:

• Layout-driven simulation allows creation of high aspect ratio 3D structures
• Simulates realistic etch and deposition steps very efficiently with unstructured tetrahedral mesh
• 3D Monte-Carlo ion implantation takes full account of crystal orientation and ion stopping including the effects of damage and amorphization

VictoryCell key features include:
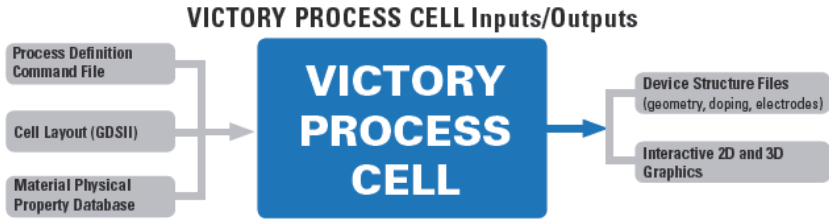
Figure 2.16. VictoryCell input/output flow information.

- Layout-driven device creation, using either GDSII or SILVACO layout formats
- Fast 3D process modeling of etch, deposit, implant, diffusion and photolithography
- Easy to learn and user-friendly SUPREM-like syntax
- Interfaces to both SILVACO 3D device simulator Device3D and VictoryDevice

MASKVIEWS

MASKVIEWS is an IC layout editor from SILVACO. It is designed to interface IC layout or any complicated structured device with SILVACO process simulators. MASKVIEWS can be used to draw and edit device and IC layout, store and then load the complete layout, and import or export the layout details by using the industry standard GDSII and CIF layout formats. Any part of the layout can be simulated – the most interesting process, and without the hassle of structuring the codes in ATHENA; one can obtain the accurate design based on the layout that has been designed. MASKVIEWS also provides features to allow layout experimentation such as misalignment, polygon over sizing or under sizing, global rescales and region definition - depending on combinations of present mask elements.

2.3 Other Simulators

In the following, a list of other simulators (with a brief description) available in the public domain is presented.

Cogenda – VisualTCAD

Cogenda is a vendor of Technology Computer Aided Design software. Their core expertise is the TCAD simulation of semiconductor fabrication processes and devices. VisualTCAD is designed to suit novice TCAD users. VisualTCAD is capable of device simulation of 2D and 3D, SPICE circuit simulation and mixed device/circuit simulation. It consists of the following modules:

- Device structure drawing tool
- Circuit schematic capturing tool
- GUI simulation controller
- Visualization tool of simulation results
- X–Y plotting tool

VisualTCAD combined with the Genius Device Simulator may be used for advanced TCAD solution. Genius is a parallel 2D/3D TCAD device simulator, featuring a wide range of advanced physical models and simulation capabilities. Genius Device Simulator incorporates parallel computation technology.

Crosslight – APSYS

APSYS is a general purpose two-dimensional finite element analysis and modeling software program for compound semiconductor devices (with silicon as a special case). It includes many advanced physical models and flexible modeling and simulation environment for compound semiconductor devices. Advanced features include hot carrier transport, heterojunction and quantum well models.

The simulation software allows the user to develop own physical models. For example, the composition and temperature dependence of all of the physical parameters (band gap, mobility, etc.) are located in a user accessible macro library with formulas written in the syntax of C/FORTRAN. These formulas are parsed and incorporated into the simulation software only at run-time so that the user can modify and fine tune these formulas any time. Such an approach to physical parameters

meets the need of computer aided design (CAD) for a new generation of semiconductor devices when the search for new material and new structures never seem to stop.

APSYS's numerical stability against mesh points regardless the structure of the device is good. For a minimal amount of mesh points used, the simulator runs smoothly for a device with structural variation from a few nanometers in one direction (such as quantum wells) to hundreds of micrometers in another direction. Application areas of the APSYS software:

- Diodes, transistors and various other types of silicon devices
- LEDs and OLEDs
- Solar cells
- Photo detectors
- High electron mobility transistors
- Heterojunction bipolar transistors
- Resonant tunneling diodes
- Quantum well infrared photo detectors
- Small MOS devices with strong quantum mechanical effects (Quantum-MOS)
- Power devices

APSYS includes the following physical models and advanced features:

- Hydrodynamic models for hot carriers
- Heat transfer equations
- Thermionic emission model
- Impact ionization model
- Deep level trap and trap dynamics
- Interface states
- Pool-Frenkel model
- Low temperature simulation model below 77K
- Guided optical modes (multimode)
- k.p theory for strained/unstrained QW/barrier
- Field dependent mobility model
- Large number of material models
- Temperature dependent model

- Flexible material parameter format
- Cylindrical coordinate system

MiniMOS-NT

Global TCAD Solutions (GTS) is a spin-off company of Vienna University of Technology. MiniMOS-NT is a general purpose semiconductor device simulator providing steady state, transient, and small signal analysis of arbitrary two and three dimensional device geometries. In mixed-mode device and circuit simulation, numerically simulated devices can be embedded in circuits consisting of compact device models and passive elements. A comprehensive set of physical models allows simulating various kinds of advanced device structures, such as advance CMOS devices, silicon-on-insulator devices, and heterostructure devices. Taking into account the atomistic nature of traps and dopants, MiniMOS-NT provides reliability and variability modeling of highly scaled transistors such as bulk planar devices and silicon-on-insulator FinFETs having a channel length down to 22 nm and below.

MiniMOS-NT features:
- Two- and three-dimensional device structures
- Drift diffusion and energy-transport models
- dc, ac, transient, and mixed-mode
- Atomistic traps and dopants
- Density-gradient model
- Self-heating simulation
- Heterostructure interfaces
- Mixed mode circuit and device simulation
- Schematic editor for mixed-mode simulation
- Circuit analysis with a thermal circuit
- BSIM4 compact model
- Latest hot-carrier degradation and BTI models
- Mobility degradation models
- Statistical variability and reliability analysis
- Unstructured and structured meshes in two and three dimensions

Nextnano

The Nextnano GmbH is a spin-off from the Walter Schottky Institute (WSI) of the Technische Universitat München, Germany. Nextnano is a simulator for calculating, in a consistent manner, the realistic electronic structure of three-dimensional heterostructure quantum devices under bias and its current density close to equilibrium. The electronic structure is calculated fully quantum mechanically, whereas the current is determined by employing a semi classical concept of local Fermi levels that are calculated self-consistently. Nextnano has been used for simulating quantum dots, quantum wires, RTDs, MOSFETs, HEMTs, etc. and includes group IV materials (Si, Ge, and SiGe), all III-V and II-VI materials and its ternaries as well as lattice-matched quaternaries; the nitrides and the II-VI materials are available in the zinc blende and wurtzite crystal structures.

Nextnano features include:

- flexible structures and geometries (1D, 2D and 3D)
- fully quantum mechanically, based on the 8-band k.p model within a finite differences grid
- includes strain, piezo- and pyroelectric charges
- growth directions along [001], [011], [111], [211], in short, along any crystallographic direction
- equilibrium and non-equilibrium, calculation of current close to equilibrium (semi-classical), ballistic transport
- magnetic field
- optical absorption

NanoFEM

NanoFEM platform is an environment for TCAD simulation of nanoscale devices. The NanoFEM platform is based on the finite element method for technology CAD simulation and visualization of nanoscale devices. The NanoFEM platform addresses the following features:

- Flexibility – simulation user and developers can modify and create physical parameters, boundary conditions or even whole governing equations;
- Interactive pre-processing and post-processing (namely geometry definition, mesh generation, visualization);
- Automatic generation of meshes suitable for FEM;
- Ability to control flow of simulation modules (either with an interactive designer or a scripting language);
- Ability to run simulation modules and methods on remote servers;
- Control of sparse linear algebra solver methods;
- Extendibility and modularity;
- Portability – using virtualization possibilities to be able to use the software on different operating systems.

BIPOLE3

The Bipole software developed by Professor David Roulston as a tool for assisting in the design of advanced bipolar transistors for industry originated in the late 1970s.Current version (BIPOLE3), available from the University of Waterloo, is capable of high precision quasi three dimensional numerical simulation of a range of bipolar devices including discrete and integrated bipolar transistors, silicon-germanium (SiGe) heterojunction bipolar transistors and diodes. BIPOLE3 may be used for predicting the terminal (dc, small signal and hf) characteristics of bipolar transistors including all major three dimensional effects. For the integrated bipolar transistor, BIPOLE3 includes the three-dimension effects such as, base and collector resistances, total junction capacitances with separate identification of plane and sidewall capacitances.

Availability:

Cogenda - VisualTCAD;
www.cogenda.com/

Crosslight Software (APSYS, LASTIP and PICS3D);
http://www.crosslight.com/Product_Overview/prod_overv.html

ENEXSS Integrated 3D TCAD system;

http://www.mizuho-ir.co.jp/english/solution/enexss/memory.html

GTS by Global TCAD Solutions;
http://www.globaltcadsolutions.com/

MiniMOS-NT;
www.globaltcad.com/en/products/minimos-nt.html

SEQUOIA Design Systems - Device Designer;
http://www.sequoiadesignsystems.com/products.html

Siborg - MicroTec: Semiconductor Process and Device Simulator;
http://www.siborg.ca/microtec.html

Silvaco tools (Atlas, DevEdit, DeckBuild, TonyPlot and TonyPlot3D);
http://www.silvaco.com/

Synopsys Sentaurus Device/Process: 1D/2D/3D device/process
simulator;
http://www.synopsys.com/Tools/TCAD/Pages/default.aspx
http://www.synopsys.com/Tools/TCAD/Pages/default.aspx

FLOODS/FLOOPS;
http://www.flooxs.tec.ufl.edu/

General purpose Semiconductor Simulator (2D);
http://gss-tcad.sourceforge.net/

NanoTCAD;
http://monteverdi.iet.unipi.it/~fiori/ViDES/ViDES.html

Nemo 3-D;
http://cobweb.ecn.purdue.edu/~gekco/nemo3D/

Nextnano - Software for the simulation of electronic and optoelectronic
semiconductor nanodevices and materials; http://www.nextnano.de/

TiberCAD;
http://www.tibercad.org

Some open source or freely available tools for device simulation (sometimes only for non-commercial usage or provided without source code) are: Archimedes (2D Quantum Monte Carlo simulator for semiconductor devices); http://www.gnu.org/software/archimedes/

BIPOLE 3:
www.bipole3.com/mainframe.html

Summary

A brief background on the past and present simulation development activities has been provided along with a number of relevant references that the reader may wish to refer to for more detailed information. Also, it is noted that other development groups in academia and industry have developed their own set of simulation programs. Each of these programs has certain advantages over the other regarding specific physical models, the precise mathematical equations solved, the numerical algorithms used, and the pre- and post-processing handling of data.

Chapter 3

Simulation Methodology

In the semiconductor industry, technology computer aided design tools are used to plan, design and test the device structures and their performance without the need of actual wafer fabrication and device characterization. Important tests allow the researchers to gain invaluable insights into determining the relationship between the change in the process conditions or device design and the overall performance of the device. Some examples of these tests may be to determine the energy levels, carrier distribution under biasing conditions, impact ionization, etc. These tests are carried out using the same computer, in contrast to specialized experimental setup. Test results for simple designs are obtainable as fast as 30 minutes while complex mesh structure may take up to 2 hours. The procedures may be summarized as follows:

- Virtual fabrication of the device using a process simulator or a device editor
- Creation of a mesh suitable for device simulation
- Device simulation that solves the equations describing the device behavior
- Post processing i.e., generation of figures and plots
- Parameter extraction

In this chapter, we shall discuss fundamentals of the semiconductor device and process simulations, mathematical model of elasticity, define system of governing equations and corresponding boundary conditions, and describe sources of strain/stress and material parameters used for stress calculations in isotropic and anisotropic (crystalline) materials. Finally, we will briefly discuss the simulation procedure and numerical methods used in VictoryStress.

57

3.1   Device Simulation

Any TCAD framework acts as a virtual factory specifically tailored for device and technology development. TCAD tools model the behavior of semiconductor devices using fundamental physical models like the current continuity (drift-diffusion) and Poisson equations. The ovals on the right of Figure 3.1 describe the processes in the actual factory while the TCAD tools corresponding to the ovals on the left form the virtual factory. It should be noted that each oval in the virtual factory corresponds to a set of tools — all of which may interact with each other at the unit process step level. A robust virtual factory should share inputs
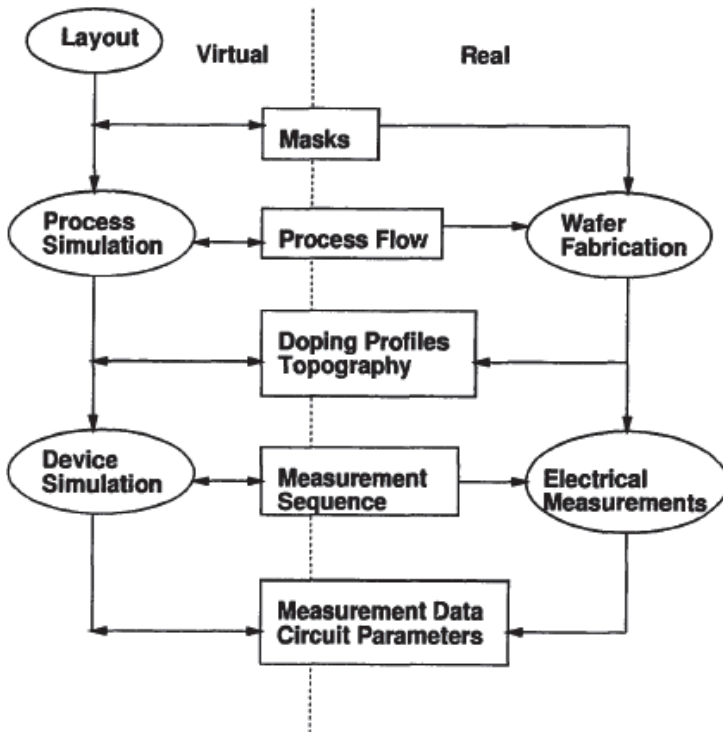
Figure 3.1. Device design methodology. After A. Quiroga, Investigation and development of advanced Si/SiGe and Si/SiGeC Heterojunction Bipolar Transistors by means of Technology Modeling, PhD Thesis, Universite Paris-Sud, 2013.

to the actual factory and allow data from the actual factory to be used as input. For example, given masks and a process flow, the output of process simulation and wafer fabrication is a set of doping profiles and topography that describe the wafer. Optimization of physical model parameters based on the comparison between simulation and the actual fabrication of control wafers provides the tightly-coupled interaction necessary for accurate simulation results. These goals are shown by the rectangles between the virtual and real world. The common inputs are described as representations of the manufacturing process and representations of measurements respectively while the common outputs are descriptions of the wafer state.

A TCAD framework also provides the mechanism to link TCAD and EDA. In many companies TCAD and EDA interact only at the circuit model parameters and design rules level. Transfer of this information between technology and design groups is not very automated and may be as simple as paper documents delivered between the groups. A framework supporting both circuit and device simulation provides the impetus for automated updating of circuit simulation models based on changing conditions in the factory and for mixed-level circuit/device simulation. Mixed-level circuit/device simulation is necessary to predict performance of critical blocks whose behavior is not well characterized using the analytic model formulations characteristic of circuit simulators.

3.2 Overview of Simulation Procedure

Traditional device such as diodes, bipolar transistors and long channel MOS transistors can be numerically analyzed by drift diffusion solver. For this solver, Poisson's equation and both the electron and hole current continuity equations are solved self-consistently by a full Newton scheme. For a power transistor Joule heating cannot be neglected. In order to take this effect into account, the level two DD solver considers an extra lattice temperature equation. Simulation of the behavior of deep submicron devices such as, advanced bipolar and CMOS transistors need to be done by solving the electron and hole energy balance equations self-consistently with the other device equations in a level three energy balance solver. Up to six equations are solved in a fully coupled model

by Newton scheme. Effects such as carrier heating and velocity overshoot are accounted for and their influence on device behavior can be analyzed. For simulation of deep submicron and nanometer MOS devices, the density gradient model (which based on the lowest moments of the Wigner Function) is necessary and in this model, three basic DD equations plus two quantum potential equations are solved consistently.

In a modern device simulator, the typical equations that describe the motion of charge carriers in a semiconductor device are the Poisson equation and carrier continuity equations for electrons and holes. Poisson equation relates the electrostatic potential to the space charge density:

$$\text{div}\left(\varepsilon \, \nabla \, \psi \right) = -\rho$$

where ψ is the electrostatic potential, ε is the local permittivity, and ρ is the local space charge density. The electric field is obtained from the gradient of the potential

$$\vec{E} = -\nabla \Psi$$

The continuity equations for electrons and holes are defined by equations:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\text{div}\,\vec{J_n} + G_n - R_n$$

$$\frac{\partial p}{\partial t} = \frac{1}{q}\text{div}\,\vec{J_p} + G_p - R_p$$

where n and p are the electron and hole concentration, J_n and J_p are the electron and hole current densities, G_n and G_p are the generation rates for electrons and holes, R_n and R_p are the recombination rates for electrons and holes, and q is the magnitude of the charge on an electron.

The current density equations, or charge transport models, are usually obtained by applying approximations and simplifications to the Boltzmann transport equation (BTE). These assumptions can result in a

number of different transport models such as the drift-diffusion model, the energy balance transport model or the hydrodynamic model. The choice of the charge transport model will then have a major influence on the choice of generation and recombination models. The simplest model of charge transport that is useful is the Drift-Diffusion Model. Until recently, the drift-diffusion model was adequate for nearly all devices that were technologically feasible. The drift-diffusion approximation, however, becomes less accurate for small feature sizes. More advanced energy balance and hydrodynamic models are therefore becoming popular for simulating deep submicron devices.

Drift-Diffusion Transport Model

The drift-diffusion formalism is the simplest of transport models, and is derived from the BTE under the relaxation-time approximation. It has been the workhorse of most device simulators up to the beginning of the deep-submicron regime. The drift-diffusion current relations are derived under the assumption that the carriers are in thermal equilibrium with the lattice. Derivations based upon the Boltzmann transport theory have shown that the current densities in the continuity equations may be approximated by a drift-diffusion model. In this case, the current densities are expressed in terms of the quasi-Fermi levels ϕ_n and ϕ_p as:

$$\overrightarrow{J_n} = -q\mu_n n\nabla\phi_n$$

$$\overrightarrow{J_p} = -q\mu_p p\nabla\phi_p$$

where μ_n and μ_p are the electron and hole mobilities. The quasi-Fermi levels are then linked to the carrier concentrations and the potential through the two Boltzmann approximations:

$$n = n_{ie}\exp\left[\frac{q(\Psi - \phi_n)}{kT_L}\right]$$

$$p = n_{ie}\exp\left[\frac{-q(\Psi - \phi_p)}{kT_L}\right]$$

where n_{ie} is the effective intrinsic concentration and T_L is the lattice temperature. These two equations may then be re-written to define the quasi-Fermi potentials:

$$\phi_n = \Psi - \frac{kT_L}{q} \ln\left[\frac{n}{n_{ie}} \right]$$

$$\phi_p = \Psi + \frac{kT_L}{q} \ln\left[\frac{p}{n_{ie}} \right]$$

By substituting these equations into the current density expressions, the following current relationships are obtained:

$$\overrightarrow{J_n} = qD_n \nabla n - qn\mu_n \nabla \Psi - \mu_n n(kT_L \nabla(\ln n_{ie}))$$

$$\overrightarrow{J_p} = -qD_p \nabla p - qp\mu_p \nabla \Psi + \mu_p p(kT_L \nabla(\ln n_{ie}))$$

The final term accounts for the gradient in the effective intrinsic carrier concentration, which takes account of bandgap narrowing effects. Effective electric fields are normally defined whereby

$$\overrightarrow{E_n} = -\nabla\left(\psi + \frac{kT_L}{q} \ln n_{ie} \right)$$

$$\overrightarrow{E_p} = -\nabla\left(\psi - \frac{kT_L}{q} \ln n_{ie} \right)$$

which then allows the more conventional formulation of drift-diffusion equations to be written as

$$\overrightarrow{J_n} = qn\mu_n \overrightarrow{E_n} + qD_n \nabla n$$

$$\overrightarrow{J_p} = qp\mu_p \overrightarrow{E_p} - qD_p \nabla p$$

It should be noted that this derivation of the drift-diffusion model has tacitly assumed that the Einstein relationship holds. In the case of Boltzmann statistics this corresponds to:

$$D_n = \frac{kT_L}{q}\mu_n$$

$$D_p = \frac{kT_L}{q}\mu_p$$

The ever-shrinking transistor size led to a growing need for atomic-scale physics to correctly model the device behavior. Various models are provided in the simulation software package to model the device physics that will determine the behavior of the device. Short and 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. The selection of the model to be used requires a very good understanding of the device, the various physics involved in its working and its working environment. Taking these into consideration the models are selected.

Hierarchically, transport models can be arranged on one side in microscopic models and macroscopic models and on the other side as semi-classical models and quantum models. Transport models can vary considerably in ease and accuracy of solution: these can be physically very sophisticated but difficult to solve and time expensive or easy and fast to solve but with inaccurate results. Figure 3.2 shows different levels of modeling from the most accurate with few physical assumptions but the most difficult to solve, to the less accurate with many physical assumptions but the most practical to solve.
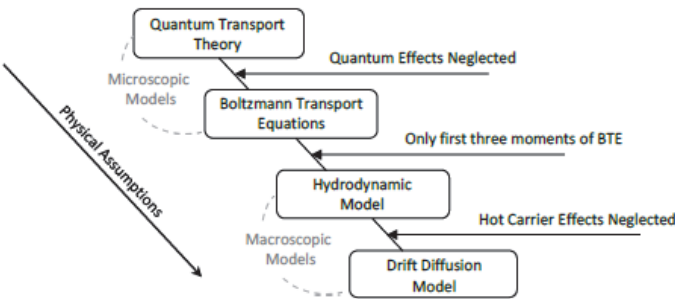
Figure 3.2. Hierarchy overview of the semiconductor transport theories. After A. Quiroga, Investigation and development of advanced Si/SiGe and Si/SiGeC Heterojunction Bipolar Transistors by means of Technology Modeling, PhD Thesis, Universite Paris-Sud, 2013.

Figure 3.3. Transport models used in device simulation. After A. N. Bhoj, Device-circuit co-design approaches for multi-gate FET technologies, PhD Thesis, Princeton University, 2013.

Figure 3.3 shows the typical transport models used in modeling nanoscale semiconductor devices as well as the scope of traditional

continuum TCAD device simulation which is based on semi-classical approaches with quantum corrections. The starting point for the semi-classical approach to modeling transport is the Boltzmann transport equation.

For ultra-small devices simulation, one needs to include the following features:

- Electronic band structure with stress effects
- Analytic bands for carrier trajectory and scattering calculations
    - Two-band model for electrons
    - Six-band model for holes (Luttinger–Kohn Hamiltonian augmented by the Bir–Pikus strain terms)

- For electrostatics update
    - Deformation Potential model for band gap and band offset Scattering mechanisms

- Phonon scattering
    - For electrons, three g-type and three f-type intervalley processes, and inelastic intravalley scattering
    - For holes, optical phonon scattering and inelastic acoustic phonon scattering

- Impurity scattering
    - Brook–Herring model with calibration to the Masetti formula

- Surface roughness scattering
    - Phenomenological model by selecting appropriate probabilities of specular or diffusive scattering events at the interface to the insulator

- Soft optical phonon scattering and remote Coulomb scattering
    - High-k degradation impact

- Impact ionization

Energy Balance Transport Model

At the next level, hydrodynamic/energy balance transport formalisms increase modeling complexity, as they are derived from higher moments of the BTE and can account for effects like velocity overshoot. Non-local effects could lead to higher currents (velocity overshoot) and increased breakdown voltage compared with predictions of the drift-diffusion model. On the other hand, self-heating effects lead to decreasing mobility and even negative output conductance for high gate biases, and decreased impact ionization rates for elevated lattice temperature. Therefore, a need for a device simulator that can account self-consistently non-local and self-heating effects for accurate simulation and optimization of SOI devices is important. The non-isothermal energy balance (NEB) model is a set of six partial differential equations for electrostatic potential, electron and hole concentrations, electron and hole carrier temperatures, and lattice temperature. The current density expressions from the drift-diffusion model are modified to include this additional physical relationship. The electron current and energy flux densities are then expressed as:

$$\overrightarrow{J_n} = qD_n\nabla n - qn\mu_n\nabla\psi + qnD_n^T\nabla T_n$$

$$\overrightarrow{S_n} = -K_n\nabla T_n - \left(\frac{k\delta_n}{q}\right)\overrightarrow{J_n}T_n$$

$$\overrightarrow{J_p} = -qD_p\nabla p - qp\mu_p\nabla\psi - qpD_p^T\nabla T_p$$

$$\overrightarrow{S_p} = -K_p\nabla T_p - \left(\frac{k\delta_p}{q}\right)\overrightarrow{J_p}T_p$$

where T_n and T_p represent the electron and hole carrier temperatures and S_n and S_p are the flux of energy (or heat) from the carrier to the lattice.

While modeling nanoscale devices like multi-gate FETs, it is essential to account for the effect of structural and electrical quantum confinement on threshold voltage. In semi-classical transport approaches, quantum

effects are typically included as a potential-like correction to the quasi-fermi level based calculations for carrier concentrations. Overall, in order to simulate multi-gate devices accurately using a commercial device simulator like SDevice or ATLAS, quantum hydrodynamic models are the best option in terms of the trade-off between simulation accuracy vs. computation time.

Quantum Confinement

Modern semiconductor devices most likely have parts that are comparable in size to the quantum mechanical wavelength of the carriers and therefore quantization effects must be taken into account to achieve better accuracy. While modeling nanoscale devices like multi-gate FETs, it is essential to account for the effect of structural and electrical quantum confinement on threshold voltage. In semi-classical transport approaches, quantum effects are typically included as a potential-like correction. The common form of quantization is to utilize quantum wells to form a two-dimensional (2D) gas of carriers with 2D density of states (DOS) and quantized levels. To model the quantum confinement of carriers in MOSFET or heterostructure, one needs the Schrödinger-Poisson solver and Density Gradient model [3.1].

3.3 Stress Model in Front-End-of-Line Structures

For TCAD simulations one needs to use simulation tool that includes mathematical and physical models of stress and appropriate numerical methods to solve the 3D stress equations. Stress modeling in Front-End-of-Line Structures includes:

- Coordinate System for Stress
- Loading the Stress
- Impact of Stress on Band Structure
- Impact of Stress on Quantization Models
- Stress-dependent Mobility Models
- Impact of Stress on Saturation Velocity
- Discretization Options for Tensor Mobility Models

In a typical TCAD flow, the stress tensor throughout a device is computed during the process simulation and is saved to a file along with the device mesh and other quantities such as the doping. Then, the device simulator loads the mesh and doping profile by specifying the input file and the stress tensor is loaded. The stress tensor is a 3×3 symmetric tensor and is defined by default within the simulation coordinate system. In a typical TCAD flow, the stress tensor is computed during the process simulation for a device with a particular orientation with respect to the crystallographic co-ordinate system. Typically, one needs to specify the directions of the x- and y-axes of the simulation coordinate system relative to the crystallographic axes. Then, the direction of the z-axis is computed automatically. For example, FinFET with a (110) sidewall and ⟨110⟩ channel in which the x-axis is along the channel direction, the y-axis is perpendicular to the sidewalls, and the z-axis is along the height of the fin. Alternatively, FinFET with a (100) sidewall and ⟨100⟩ channel in which the x-axis is along the channel direction, the y-axis is perpendicular to the sidewalls, and the z-axis is along the height of the fin.

The silicon crystal has a cubic lattice structure whose axis directions are specified in the Miller notation. The transport properties depend upon the
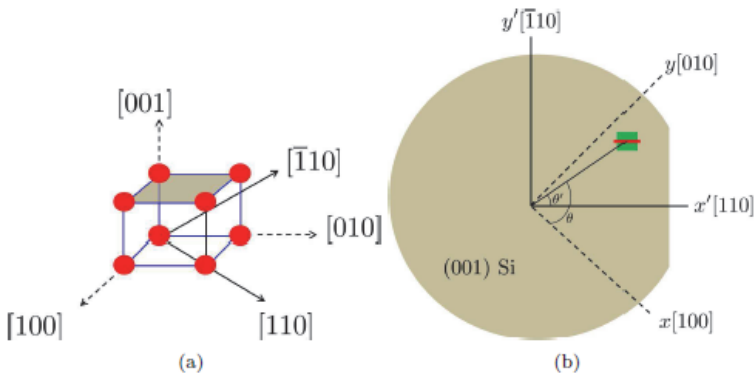
Figure 3.4. (a) Miller indices (b) Coordinate axes in (100) Si with a wafer flat orthogonal to the [110] orientation. The transistor channel here is perpendicular to the [110] axis i.e., $\varphi' = \pi/2$. After S. K. Marella, Performance variations due to layout-dependent stress in VLSI circuits, PhD Thesis, University of Minnesota, 2015.

silicon crystal orientation. While band structures are typically defined in the Cartesian coordinate system, actual electronic transport may physically take place along a direction that maximizes the mobility of the charge carriers. Hence, the stress state need to be appropriately defined in the orientation along which electrical transport takes place.

Figure 3.4a shows the Miller index directions. The shaded plane is the (001) plane perpendicular to the [001] direction. The crystal orientation refers to the Miller index of the silicon crystal. The principal crystallographic axes create a coordinate system that corresponds to the [100], [010], and [001] directions, which is identical to the Cartesian coordinate system. Within this system, the orientation of a wafer is defined as the direction normal to the plane of the silicon wafer. Most integrated circuits are manufactured on wafers which are perpendicular to the [001] axis direction or along the (001) plane. Due to symmetry, the (100), (010), and (001) orientations are equivalent. In CMOS integrated circuits, hole transport is superior along the $\langle 110 \rangle$ set of directions, while electron transport is best along $\langle 100 \rangle$ directions. However, the magnitude of electron mobility is always greater than hole mobility. Since CMOS integrated circuits prefer a single orientation for both NMOS and PMOS transistors for ease of manufacturing and for compact layouts, the transistors are oriented along the $\langle 110 \rangle$ directions so that the relatively weaker hole transport is maximized. The orientation of transistors on a wafer is determined relative to the wafer flat, as shown in Figure 3.4b: transistors may be parallel or perpendicular to this feature. Therefore, a rotated coordinate space with a new x′-axis that is perpendicular to the wafer flat is a convenient frame of reference. This x′-axis is in the [110] direction, and therefore, the [100]– [010] axes must be rotated by 45°.

In the following section, we provide the basic equations of elasticity and the solution approaches to determine the stress state. Apart from process induced stress, layout-dependent stress effects contribute to performance variations in integrated circuits. In earlier technologies, transistor sizes were large enough that their electrical behavior was independent of the final layout. However, in highly scaled technologies with smaller geometries, the electrical performance of a transistor has become increasingly dependent on its location in the layout.

Stress–Strain Relationships

When a material is stretched or compressed along a certain direction, there is a corresponding contraction or expansion of the material in the orthogonal direction so that the volume of the body remains the same. The negative ratio of the strain in the orthogonal direction to the strain along the direction of the applied force is known as the Poisson's ratio of the material. It should be noted that Poisson's ratio is always a positive fraction less than one. Beyond a stress level known as the yield point or yielding stress, the stress and strain are no longer linear and the material ceases to be elastic. The stress (σ) at a point may be determined by considering a small element of the body enclosed by area (ΔA) on which forces act (ΔF) and its unit is Pascal (Pa). By making the element infinitesimally small, the stress (σ) vector is defined as the limit:

$$\sigma = \lim_{\Delta A \to 0} \frac{\Delta F}{\Delta A} = \frac{dF}{dA}$$

Any stress on an isotropic solid body can be expressed as a stress matrix. Generally, the stress tensor σ is a symmetric 3×3 matrix. Therefore, it only has six independent components. To compute the strain tensor (which is needed for the deformation potential model development), the generalized Hooke's law for anisotropic materials is applied as:

$$\varepsilon_{ij} = \sum_{j=1}^{6} S_{ij} \sigma_j$$

where, S_{ij} is the elasticity modulus. In crystals with cubic symmetry such as silicon, the number of independent coefficients of the elasticity tensor (as other material property tensors) reduces to three by rotating the coordinate system parallel to the high-symmetric axes of the crystal. Mechanical stress causes change in the band edges in silicon. These band edge shifts in terms of the strain – the relative distortion of the crystal lattice produced by an applied stress are given by the deformation potential theory. VictoryDevice has several methods for determining the strain. The most accurate and general approach is to read the strain data

directly from a structure file that has been produced by a tool like VictoryStress. This approach is commonly adopted in simulation. However, if the structure file contains stress data but not strain data, VictoryDevice computes the strain tensor from the stress tensor and the material compliance coefficients.

The governing system of 3D linear elastic equations consists of three partial differential equations for the balance of linear momentum and six strain-displacement relations. In case of isotropic materials the elasticity matrix could be written using only two material parameters, Young modulus (E) and Poisson ratio (*v*). For anisotropic crystals with cubic symmetry (such as Si, Ge, GaAs) there are three independent elasticity constants: C_{11}, C_{12}, and C_{44}. For isotropic materials these constants could be calculated by use of Young modulus and Poisson ratio.

Stress Effects on Bandgap in Silicon

Stress modifies the band structure of silicon by splitting the edges of the valleys and bands that make up the conduction and valence bands. Stress also can modify the shape, or dispersion, of the bands resulting in a change in the density-of-states. Together, these effects modify the threshold voltage in a MOSFET. In simulators, options are available for considering the impact of stress on band structure. The most physical of these options treats the impact of stress on both the band edges and the dispersion. The set of models provided computes the level shifts to the conduction valleys and valence bands using the k.p models, taking the bottom-most conduction valley and the topmost valence band to define the band gap. The stress dependency of the DOS mass is considered as well. User needs to select the models to be used in simulation. Mechanical stress causes changes in the band edges in silicon. The deformation potential theory describes the band-edge shifts in terms of strain – the relative distortion of the crystal lattice produced by an applied stress. VictoryDevice has several methods for determining the strain. VictoryDevice can calculate the strain tensor from the stress tensor and the material compliance coefficients. For an isotropic material,

$$\begin{bmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ \varepsilon_{yz} \\ \varepsilon_{xz} \\ \varepsilon_{xy} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & s_{12} & 0 & 0 & 0 \\ s_{12} & s_{11} & s_{12} & 0 & 0 & 0 \\ s_{12} & s_{12} & s_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2s_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 2s_{44} \end{bmatrix} \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{xz} \\ \sigma_{xy} \end{bmatrix}$$

where

- the ε terms are the components of the strain tensor,
- the σ terms are the components of the stress tensor,
- the s terms are the components of the fourth-order compliance tensor, and the tensors are represented here using Voigt notation.

The non-zero compliance coefficients are given by

$$s_{11} = \frac{c_{11} + c_{12}}{c_{11}^2 + c_{11}c_{12} - 2c_{12}^2}$$

$$s_{12} = \frac{-c_{12}}{c_{11}^2 + c_{11}c_{12} - 2c_{12}^2}$$

and

$$s_{44} = \frac{1}{c_{44}}$$

where c_{11}, c_{12} and c_{44} are the elastic stiffness coefficients.

For silicon,

$$c_{11} = 163.8 \text{ GPa} - T_L \cdot (0.0128 \text{ GPa/K})$$
$$c_{12} = 59.2 \text{ GPa} - T_L \cdot (0.0048 \text{ GPa/K})$$

$$c_{44} = 81.7 \text{ GPa} - T_L \cdot (0.0059 \text{ GPa/K})$$

For germanium,

$$c_{11} = 126.0 \text{ GPa}$$
$$c_{12} = 44.0 \text{ GPa}$$
$$c_{44} = 67.7 \text{ GPa}$$

where T_L is the lattice temperature. The stiffness coefficients for SiGe at composition are obtained by linear interpolation. For SiGe, if no stress or strain tensor is available, VictoryDevice will calculate an isotropic strain from the lattice constants. In three dimensions:

$$\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz} = \left(\frac{1 + v_{\text{SiGe}}}{1 - v_{\text{SiGe}}} \right) \left(\frac{a_{\text{SiGe}} - a_{\text{Si}}}{a_{\text{SiGe}}} \right)$$

where

- v_{SiGe} is Poisson's ratio for SiGe alloy,

- a_{SiGe} is the lattice constant for SiGe, and

- a_{Si} is the lattice constant for silicon.

The lattice constant and Poisson's ratio for SiGe are obtained by linear interpolation between the values for silicon and germanium:

$$a_{\text{Si}} = 5.43102\text{Å} + (1.41 \times 10^{-5} \text{ Å/K})(T_L - 300 \text{ K})$$
$$v_{\text{Si}} = 0.28$$
$$a_{\text{Ge}} = 5.6579\text{Å} + (3.34 \times 10^{-5} \text{ Å/K})(T_L - 300 \text{ K})$$
$$v_{\text{Ge}} = 0.273$$

Change in Band Edge Energy

Assuming Maxwell–Boltzmann statistics, the strain-induced change in the conduction band edge is given by:

$$\Delta E_C = -kT_L \ln\left[\frac{1}{3}\sum_{i=1}^{3}\exp\left(-\frac{\Delta E_{C,i}}{kT_L}\right)\right]$$

while the change in the valence band edge is

$$\Delta E_v = kT_L \ln\left[\frac{r_m}{1+r_m}\exp\left(\frac{\Delta E_{V,l}}{kT_L}\right) + \frac{1}{1+r_m}\exp\left(\frac{\Delta E_{V,h}}{kT_L}\right)\right]$$

With

$$r_m = \frac{MLH}{MHH}$$

where

• $\Delta E_{C,i}$ is the shift in the edge of the ith ellipsoidal conduction band minimum,
• $\Delta E_{V,i}$ is the shift in the edge of the valence band maximum for light holes,
• $\Delta E_{V,h}$ is the shift in the edge of the valence band maximum for heavy holes,
• k is Boltzmann's constant, and
• T_L is the lattice temperature.

The parameters MLH and MHH correspond to the masses of light and heavy holes, respectively. From the deformation potential theory, the shift in the ith ellipsoidal conduction band minimum is:

$$\Delta E_{C,i} = \text{D.DEFPOT} \cdot (\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) + \text{U.DEFPOT} \cdot \varepsilon_{ii}$$

where ε_{xx}, ε_{yy}, and ε_{zz} are the diagonal components of the strain tensor and

ε_{ii} is the component corresponding to the ith conduction band minimum.

D.DEFPOT and U.DEFPOT are parameters representing the dilation and shear deformation potentials for the conduction band. The shifts in the valence band edges are

$$\Delta E_{V,h} = \text{A.DEFPOT}(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) + \sqrt{\xi}$$

$$\Delta E_{V,l} = \text{A.DEFPOT}(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) - \sqrt{\xi}$$

$$\xi = \frac{1}{2}\text{BDEFPOT}^2\left[(\varepsilon_{xx} - \varepsilon_{yy})^2 + (\varepsilon_{yy} - \varepsilon_{zz})^2 + (\varepsilon_{zz} - \varepsilon_{xx})^2\right] + \text{CDEFPOT}^2(\varepsilon_{xy}^2 + \varepsilon_{yz}^2 + \varepsilon_{zx}^2)$$

where

ε_{xy}, ε_{yz}, and ε_{zx} are the off-diagonal components of the strain tensor.

A.DEFPOT, B.DEFPOT, and C.DEFPOT are deformation potentials for the valence band. One can specify values for the various deformation potentials on a MATERIAL statement. Some simulation tools use models that have options for considering additional impacts of stress on quantization. Quantization models, such as density gradient, consider the impact of stress primarily through changes in the band edges. Density gradient model generally use the strained DOS mass in the density-gradient equation.

Stress-Dependent Mobility Models

The mobility gain due to stress in a MOSFET channel depends strongly on the surface and channel orientations. For example, in a FinFET device with multiple surface orientations for the top and sidewalls, it is convenient to use models that treat this orientation dependency automatically. In general, following mobility models are available:

- Sub-band Stress-Mobility Models for Electrons and Holes
- Piezoresistance Model for Electrons and Holes

In general, the sub-band stress-mobility models work as corrections to the relaxed low-field mobility, including automatically accounting for different surface orientations. For example, if in a FinFET, the Lombardi model is selected to compute the low-field mobility, the sub-band stress-mobility models should by default automatically compute the stress-induced mobility gain for the different surface orientations in the fin. For some surface orientations, most notably (110), the relaxed mobility is anisotropic and depends on the channel direction. It is necessary, therefore, to indicate to the sub-band models which channel direction to use. To enable the sub-band stress-mobility models to work correctly for a MOSFET channel, one needs to activate the multi-valley MLDA quantization model [3.2, 3.3].

The piezoresistance model is a straight forward model in which the mobility gain is computed as either a linear or quadratic function of the stress. The appropriate piezoresistance coefficients are specified in the crystallographic coordinate system and are transformed by the tool to the simulation coordinate system. When this model is applied to a FinFET channel, the surface orientation dependency of the stress induced mobility gain and, therefore, the piezoresistance coefficients should be considered. These coefficients can be specified for different surface orientations.

Unlike in a highly doped source/drain region in which the stress-induced mobility gain is suppressed due to high doping, the mobility gain (or reduction) in a MOSFET channel can become relatively large. At large stress, this gain typically saturates at a value that depends on the carrier type and surface/channel orientation. The piezoresistance model is based

on a bulk model, which for silicon means the resulting mobility reflects the cubic symmetry of relaxed silicon. This means that some surface/channel orientations in a MOSFET cannot be completely characterized by this model. In particular, for a $(100)/\langle 100 \rangle$ channel, one cannot model the mobility gain due to both transverse and vertical stress with this model. Nonlinear piezoresistance model considers the impact of the three main types of uniaxial stress on mobility, namely, longitudinal, transverse, and vertical stress. This model is calibrated against reference data generated by Monte Carlo simulation or by the mobility calculation from band structure for a select set of surface/channel orientations.

In general, stress can impact both the low- and high-field mobilities. The impact of stress on the high field mobility is primarily through the effective saturation velocity. Detailed Monte Carlo studies of bulk silicon have shown that the stress dependency of the saturation velocity is rather weak. Thus one may choose to suppress any change in saturation velocity due to stress for long-channel devices. For short-channel devices, quasi-ballistic transport is expected to play a larger role in device behavior. Device Monte Carlo studies have shown that quasi-ballistic transport is affected by stress. While drift-diffusion simulations cannot physically model quasi-ballistic transport, the effect of quasi ballistic transport on device behavior can be approximated by changing the saturation velocity. The effect of stress on quasi-ballistic transport can, therefore, be modeled as a stress dependency of the saturation velocity. Also, saturation velocity itself can be tuned to account for the combined impact of gate length and stress on quasi-ballistic transport. Some of the stress-dependent mobility models produce a tensor mobility. There are several options for selecting how such models should be treated numerically and, in particular, how they are discretized onto the device mesh.

Device simulation is used to analyze the electrical and thermal behavior of the device structure obtained from process simulations. Its main elements are the structure, material parameters, boundary conditions, list of physical effects to be captured, numerical constraints on the solver, carrier transport model, and the modes of simulation, i.e., DC, AC or transient, with specific external biasing conditions. Single device simulation is used to investigate transport phenomena in a single device. Mixed-mode simulation is used to study the behavior of small circuits

constructed out of individual device instances and is generally less rigorous in terms of physical models, owing to the increase in simulation complexity.

Device Simulation Considerations are:

- Material parameters
- Mechanisms

    - Transport (Drift-diffusion or Hydrodynamic)
    - Quantum (Schrödinger solver or Density Gradient)
    - Recombination
    - Trap models
    - Stress

- Coupled equations are solved using Newton or Gummel approach

    - The more effects included, the more equations, the more memory and more time required
    - Convergence

- DC
- AC
- Temperature
- Mixed mode

3.4 Physically Based Simulation

Physically based device simulators predict the electrical characteristics that are associated with specified physical structures and bias conditions. In order to do this, such a simulator first represents the structure and state of a device on a two or three dimensional grid, consisting of a number of grid points called nodes. It next solves a set of differential equations on this grid. These differential equations are derived from Maxwell's laws, and describe the transport of charged carriers (electrons and holes) through the structure. By solving these equations under specific bias

conditions, the electrical performance of a device can be modeled in DC, AC, or transient modes of operation. Before one can do a physically-based simulation, all relevant physics must be incorporated into the simulator and numerical procedures to solve the associated equations must be implemented. To use a physically-based device simulator one must specify the particular problem to simulate. One needs to define:

- The physical structure of the device
- The physical models to be used
- The bias conditions for which electrical characteristics are to be simulated

Physically-based simulation has three notable advantages:

- It is predictive
- It provides insight
- It conveniently captures and visualizes theoretical knowledge
- Doing physically-based simulations is almost always much quicker and cheaper than performing experiments
- It provides information that is difficult or impossible to measure

3.5 Process Simulation

The primary objective of process simulation is to accurately predict the structural geometry of devices at the end of a process run as well as the active dopant and stress distributions. The input to process simulation is a process flow guided by process assumptions and layout/layer masks. Process simulation basically models the semiconductor manufacturing processes. The simulation starts with the bare wafer and finishes with device structures. Processes such as implantation, diffusion, etching, growth, and deposition processes are simulated on a microscopic level.

Process Simulation Features are:

- Process simulation emulates actual fabrication
- Commands for deposit layers, define doping, create implantation profiles, and etch layers

- Each of these steps has a thermal budget where defined diffusion mechanisms are used
- Movement of dopant species is described by the diffusion steps
- In heterostructure simulations elements diffuse across material boundaries creating new mixes of materials but this is not taken into account in commercial simulators

Process Simulation: Model Selection

Implantation Models

- Default is Pearson (or double Pearson), range parameters can be set in the MPLANT or MOMENT statement
- Amorphous Monte Carlo could be useful for multilayered structures, high angled implants, shadowing effects
- Crystalline Monte Carlo is for implants with high channeling probability or to predict a dose dependency
- Depth dependent lateral standard deviation, separate for random scattering and channeling
- Implant damage and amorphization and shadowing effect due to wafer tilt and rotation

Diffusion Models

- High concentration effects (diffusivity enhancement and precipitation)
- Full impurity interaction with point defects
- TED (Transient Enhanced Diffusion) and OED (Oxidation Enhanced Diffusion)

Oxidation Models

- COMPRESS is default which is good for almost all cases
- Stress-dependent VISCOUS is recommended for LOCOS with thick nitride layer and trench corner effects
- Viscoelastic model
- Parallel oxidation by multiple oxidizing species
- Surface orientation dependence for complicated 3D structures

For the simulation of memory devices, if one needs to investigate, verify and optimize designs for nonvolatile memories, such as EPROMs, EEPROMs, and Flash EEPROMs then one needs a device simulator to account for a variety of physical effects such as:

- band-to-band tunneling
- Fowler-Nordheim tunneling
- impact ionization, and floating gates
- Self-consistent descriptions of these effects
- Both steady-state and time-domain transient simulation capabilities
- Time-domain capabilities provide self-consistent calculations of programming and erasing operations

3.6 Numerical Methods

Different numerical methods can be used to solve the semiconductor equations. In general there are three approaches – decoupled (Gummel method), fully coupled (Newton method) or a combination method (BLOCK). The de-coupled technique like the Gummel method solves for each unknown in an equation while keeping the other variables constant and repeats the process until a stable solution is achieved. Gummel's method solves the coupled set of semiconductor equations together with Poisson's equation via a decoupled procedure. First, we solve the nonlinear Poisson's equation. The potential obtained from this solution is substituted into the continuity equations, which are now linear, and are solved directly to conclude the iteration step. The result is then substituted back into Poisson's equation and the process is repeated until convergence is reached. In fully coupled techniques such as the Newton method, the total system of unknowns are solved together. Newton's method is a coupled procedure that solves the equations simultaneously through a generalization of the Newton–Raphson method for determining the roots of a general nonlinear equation. In general, Gummel's method is preferred at a low bias because of its faster convergence and low cost per iteration. At a medium or high bias, Newton's method becomes more convenient. Finally, in the combined or block method, the solution is obtained by solving some equations by the fully coupled method, while others are solved by the de-coupled method.

3.7 TCAD Calibration

The international technology roadmap for semiconductors has identified the importance of device/process simulation. The use of numerical simulation has become more important in the semiconductor industry because it is a tool that has proven to be able to provide accurate prediction for future devices. We start with device generation using process simulation, device simulation and finally an analysis of the results obtained by using TCAD tools. TCAD can be used on a wide class of problems:

- Predictive TCAD
- Advanced Process Control
- Process Centering
- Inverse Modeling
- Early Exploration
- Failure Analysis
- Learning/Insight

To be "predictive" TCAD needs to have a very high level of accuracy. However, proper calibration needed for predictive TCAD is time-consuming and expensive to calibrate the simulators [3.4]. Calibrating a set of TCAD simulators is extremely difficult. For example, if the threshold voltage does not match the experimental value, one needs to check the assumptions in both the process and the device simulators, as well as the electrical measurements and the test structure layout. Thus, for a proper TCAD calibration, the requirements are as follows:

- Process/Device: Complete flow and process/device physics must be known
- Fabrication process: Example, Implant temperature affects TED; RTA temperature
- Electrical Test: Example, Electrical vs. optical oxide thicknesses (QM effects)
- Simulation: Example, Model limitations, grid dependence

The problem arises from the fact that process technology is constantly changing, as is the level of physics necessary to simulate it. A careful

calibration of one technology generation does not guarantee that it can predict the next generation. Also the technology developer does not always have time to wait for the TCAD community to develop a model. Despite all the sophistication of today's TCAD, technology developers continue to do much of their development experimentally. Several reasons why process simulation is far from ideal:

- o Some physics is poorly characterized even for standard processes i.e., segregation, defect generation etc.
- o Models for many processes are still in a development stage: silicidation, dislocation loops, cluster formation, details of implant channeling etc.
- o Characterization of processes in non-silicon materials is lagging far behind
- o Many processes (e.g., deposition, etching) depend on equipment

Sources of errors in process simulation are: Insufficient physical models: amorphization or recrystallization effects, cascades in implant, dislocation loops and cluster effects and stress generation. For physically based deposition and etching – almost all rate parameters are equipment dependent and need prior calibration. Most of the parameters of physically based models cannot be measured directly and practically impossible to derive from first principles, e.g., local electronic stopping for MC ion implantation, diffusivities, generation and recombination rates for point defects for advanced diffusion models, oxidation rate decrease in presence of stresses, and segregation coefficients.

Using SILVACO tools, process simulation calibration are generally performed at two levels:

- The first (local) level of calibration allows to tune one or several parameters of a specific model for a specific process step
- The tools and features used for the first level are DeckBuild, Optimizer, Extract, and Auto interface
- The second (global) level of calibration allows to calibrate many parameters of several key models for the whole process
- The second level of calibration uses VWF Automation and Production Tools
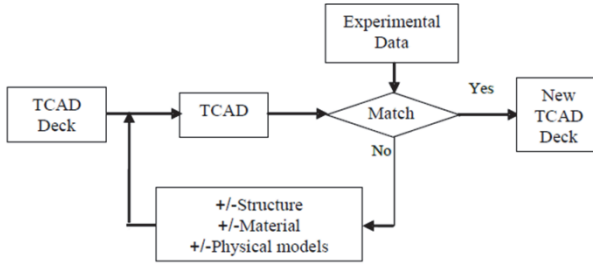
Figure 3.5. Method of calibration flow. After H. Ramakrishnan, Strained silicon technology for low-power high-speed circuit applications, PhD Thesis, Newcastle University, 2008.

TCAD calibration refers to the process of selecting appropriate models and adjusting the model parameters so that the model can predict the measured device data. One of the challenges in process simulation is to reproduce an accurate representation of the semiconductor device. Guidelines for how to calibrate a process flow and device simulator to a particular technology is described using calibration of process and device simulations. A typical calibration flow diagram is shown in Figure 3.5. Although models for a given technology are specified in the process simulator to ensure accurate results, identification of the calibration targets is an important issue. For example, in a bipolar transistor the most important is the Gummel plot. Thus common approach will be to calibrate the parameters that have the greatest effect on Gummel plot. The experimental data obtained using device electrical characterization is compared with an example TCAD deck.

TCAD engineers often need to calibrate some specific parameters to accurately predict the real process. This may be due to the physical models have some simplifications or assumptions that prevent an accurate reflection of the real world. It can also be because of variations in the test results from fab to fab, wafer to wafer and even from die to die. For example, for calibration of a bipolar process/device the strategy could be to tune first (a) the base and collector currents, (b) the base current, (c) the collector current, (d) the base current profile for medium injection, and (e) the base current profile for low injection in all regions and then move to more subtle phenomenon that affect other parts of the base or collector currents. Matching the collector current for all regions

in the Gummel plot is less problematic than matching the base current at the extremes. The process of matching experimental data with example TCAD deck should be continued until a near accurate match is obtained. In the matching process, different mobility models along with different doping profiles and dimensions may be analyzed. Typical calibration procedure for MOSFETs is:

- Adjust the saturation velocity to achieve the measured maximum current
- Adjust total charge to get the threshold voltage seen in measurements
- Set surface trap density/mobility to get sheet resistance seen in measured data

TCAD Calibration Example: MOSFET

The study of scaling is critically important in predicting performance for future technology generations. The scaling work starts with careful calibrations in respect of real devices including device physical dimensions, doping profiles and ultimately electrical characteristics. In the following, a calibration example based on reference [3.5] aims to closely replicate the features in TCAD process simulation and attempts to match consistently the measured electrical characteristics using device simulation. Sentaurus TCAD simulation tools are used. The devices have 35nm physical gate length, NO gas annealed gate oxide, retrograde and super-halo doping profile, and low thermal process S/D extension shallow junctions.

The flowchart of the calibration process is summarized in Figure 3.6. Firstly the simulation matches the process details of real structure as accurately as possible. It includes the structural parameters, process flow steps such as, ion implantation with specific energy and dose and spike RTA. Closely matching retrograde channel doping, the calibration process uses the margin of well doping and halo doping for process simulation flexibility depending on the device simulation error. The simulation steps are carried out for p- and n-MOSFET calibrations. The channel retrograde doping is constructed using multiple implantations and halo doping following the process detail.
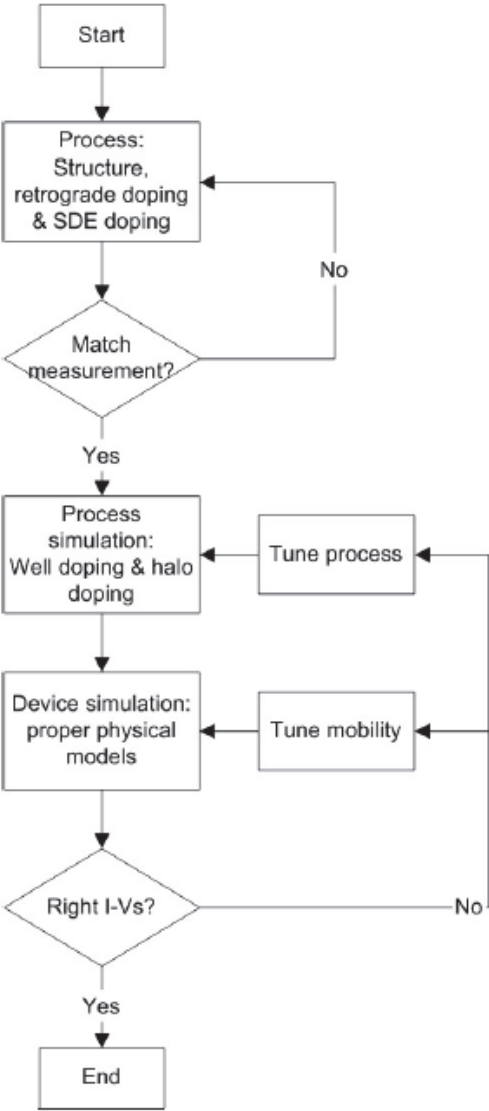
Figure 3.6. Simplified flowchart of systematic simulation calibration methodology. After X. Wang, Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs, PhD Thesis, University of Glasgow, 2010.
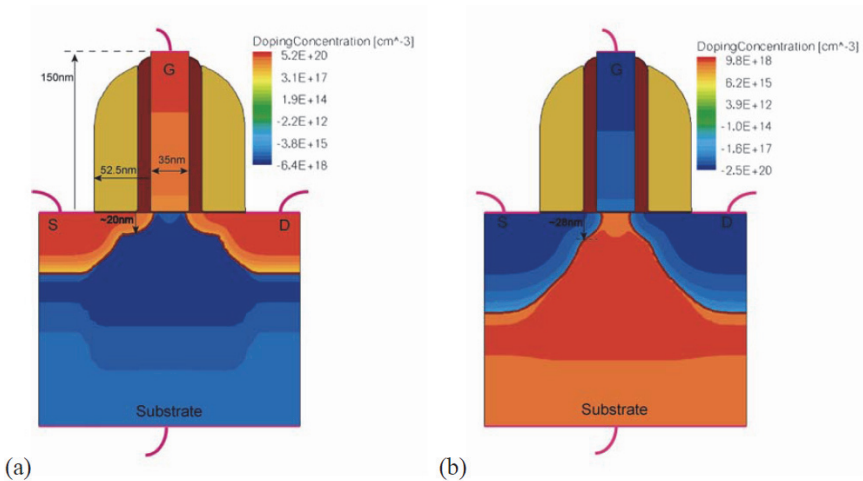
Figure 3.7. Simulation structures of 35nm gate length n-MOSFET (a) and p-MOSFET (b) based on Toshiba experimental data. After X. Wang, Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs, PhD Thesis, University of Glasgow, 2010.

After the first stage, the calibration process focuses on tuning mobility parameters. First the device electrostatics are adjusted using low and high drain voltage biased $I_d - V_g$ characteristics in the sub-threshold regime. Thus, low field mobility is tuned to match the low drain voltage $I_d - V_g$ characteristics. Secondly, high field mobility is tuned to the high drain voltage $I_d - V_d$ characteristics. In simulation the low field mobility tuning is primarily based on phonon scattering and impurity scattering mobility models. The high field mobility tuning adjusts the saturation velocity and the critical field in the field dependent mobility models. The calibration starts with the design of an appropriate. The calibrated structure is illustrated in Figure 3.7.

At the electrical calibration stage the mobility models implemented in the Sentaurus TCAD device simulation tool have been tuned according to experimental data in conjunction with the published mobility models. The calibration starts by adjusting the electrostatics with respect to the low and high voltage $I_d - V_g$ characteristics by tuning the mobility models to match the current magnitude. Figures 3.8 and 3.9 illustrate the calibrated I–V characteristics of the n- and p-MOSFETs, respectively.
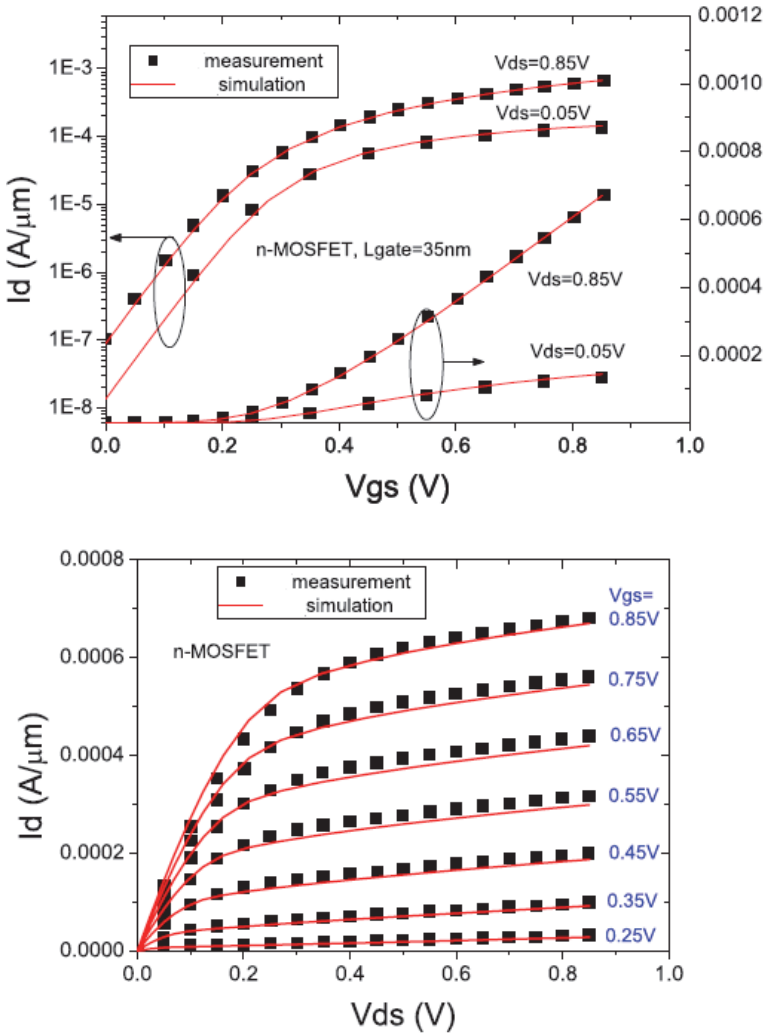
Figure 3.8. $I_d - V_g$ and $I_d - V_d$ characteristics calibrations of Toshiba 35nm gate length n-channel MOSFETs with supply voltage 0.85V. After X. Wang, Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs, PhD Thesis, University of Glasgow, 2010.
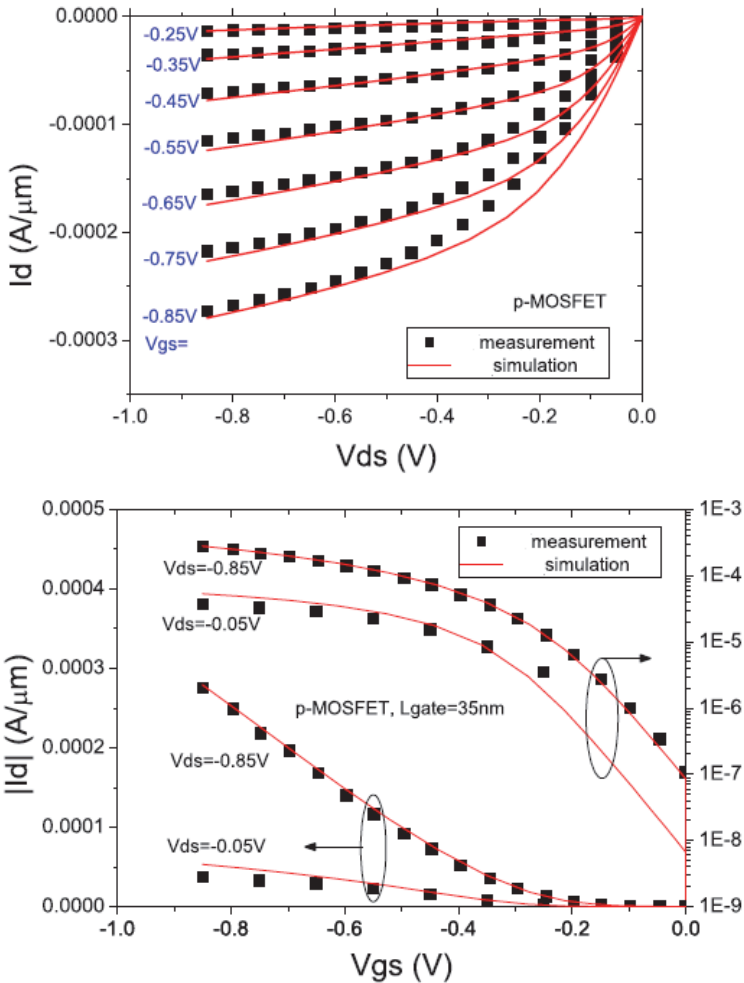
Figure 3.9. $I_d - V_g$ and $I_d - V_d$ characteristics calibrations of Toshiba 35nm gate length p-channel MOSFETs with supply voltage 0.85V. After X. Wang, Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs, PhD Thesis, University of Glasgow, 2010.

Troubleshooting

During simulation mostly one experiences convergence problems. To overcome this one should look into the followings steps:

- Understand the physics of the device. Convergence problems are frequently caused by improper meshing of the structure. A good mesh needs to capture the physics well. For example, depletion regions and impact ionization generation must be captured well to prevent convergence issues. Check for voltage steps (increment); generally small voltages steps are necessary for obtaining initial solutions.

- Start the debugging process by switching off all physical models. Then, switch them on again one-by-one to find out which model causes problems.

- Check if boundary conditions (BCs) are contradictory. For example, an electrode touching a semiconductor and an insulator at a corner will apply an Ohmic BC (default) to nodes touching the semiconductor, and a gate-like BC to nodes touching the insulator. These different BCs will lead to abrupt variations in potential around the corner that can lead to convergence issues, especially if the Avalanche model is switched on. The simplest solution to this problem is to use a work function that leads to a flat band around the corner.

- Lastly, refer to the User Manual for debugging convergence and other problems to find out the reason for the problem encountered and where it is in the device.

Summary

Technology CAD can be used at various stages in the research through manufacturing cycle. A brief review of critical modeling issues for simulation tools was presented. This chapter introduced the different levels to model the electron transport into semiconductors: the quantum-mechanical transport model, the semi-classical Boltzmann transport equation, the energy-based hydrodynamic transport model, and the simplest drift-diffusion transport model. Calibration is a highly iterative, difficult and continuous process. We described how to calibrate a process flow and use a device simulator to a particular technology. Guidelines

are presented for calibration of process and device simulations for a typical bipolar and MOS processes. With extreme scaling down of MOSFETs in high volume manufacturing, it is imperative to develop a systematic TCAD-based methodology to design, characterize, and optimize manufacturability to increase yield.

CMOS Technology

In this chapter, we present the process simulation using ATHENA for a standard LDD MOS process and characterization of n-MOSFETs using ATLAS device simulation. The first stage of simulation is the process simulation where we create the structure of a MOSFET. An overview of the process and device simulation steps is presented below.

4.1 MOS Process Simulation

Main process/device simulation steps

- Step 1: Initialization of the substrate
- Step 2: Gate oxide deposition
- Step 3: Threshold voltage adjustment implant and gate poly
- Step 4: LDD implant and nitride spacer
- Step 5: Source/drain implant
- Step 6: Mirror structure and export to the device simulator
- Step 7: Contact definitions for device simulation
- Step 8: Device simulation

In the following section, a CMOS technology process simulation is carried out following the example provided by SILVACO. The MOSFET is a symmetrical device: the region from the center of the gate to the source on the left is identical to the region from the center of the gate to the drain on the right. Therefore, we need only to fabricate the source region and "mirror" the results to the other side when we are done. In this example, only NMOS is simulated and for PMOS one can follow similar procedure. The simulation example is for a long channel device,

93

meaning that short channel and quantum effects are not considered. Qualitative description of a generic NMOS process used in simulation is described below.

- Substrate Boron doped (100) Si, moderately high resistivity, typically 20 Ohm-cm

The selection of the substrate depends on different applications. Major considerations include wafer type (bulk or SOI), substrate type, resistivity and orientation. SOI wafers are considerably more expensive than bulk wafers. For modern CMOS technologies, a p-type substrate is usually chosen with twin well implant. Although NMOS can be built without introducing p-well, in practice, it is always desirable for process control to build NMOS within a p-well.

- Thermal Oxidation ~ 100Å pad oxide
- LPCVD Si_3N_4 ~ 0.1μm Lithography
- Pattern Field Oxide Regions
- RIE removal of Nitride and pad oxide

In practice, active devices are integrated in a common silicon substrate. In order to electrically isolate these devices, fairly thick SiO_2 is usually grown or deposited between the devices. Oxide is a perfect insulator to provide the necessary isolation. Since LOCOS (Localized Oxidation of Silicon) is relatively old for CMOS technology. The first step of a STI process is to grow pad oxide. A nitride (Si_3N_4) layer is then deposited. This layer is usually highly stressed and produces a large compressive stress in the underlying silicon substrate; this can lead to defect generation. The main purpose of the pad oxide under the nitride layer is to help relieve the stress.

- Channel Stop Implant: 3×10^{12} B/cm^2 60keV
- Thermal Oxidation to grow 0.45μm oxide
- Wet Etch Nitride and pad oxide
- Ion Implant for Threshold Voltage control 8×10^{11} B/cm^2 35keV

Source/Drain implant are high dose implant, they are usually implanted through a thin layer of oxide called screen oxide. The purpose of screening oxide is to prevent channeling and minimize the incorporation

of trace impurities. Implant channeling is caused by the implanted ions have a velocity vector line up with the substrate crystal structure. Ions can go deeper into the silicon substrate without encountering silicon atom, which is undesirable when shallow implantation is desired. A thin amorphous oxide layer can help randomize the directions of implanted ions and minimize this channeling effect.

- Thermal Oxidation To grow 250Å gate oxide
- LPCVD Poly-Si ~ 0.35μm
- Dope Poly-Si to n^+ with Phosphorus Diffusion source

It is well known that gate oxide needs to be of high quality and as such a chemical cleaning is generally used. This removes any remaining oxide and leaves behind a clean silicon surface for dry oxide growth. Gate poly is deposited after the oxide is grown using LPCVD. Generally, the gate poly contact should not be placed on top of the active region as the stress may make the underlying gate oxide and channel region less reliable. The poly doping is done either through in-situ doping or blanket implant after poly deposition.

- Lithography Poly-Si Gate pattern
- RIE Poly-Si gate
- Source/Drain Implantation ~ 10^{16} As/cm^2 80keV
- Thermal Oxidation ~ 0.1μm oxide on poly-Si and source/drain
- LPCVD SiO$_2$ ~ 0.35μm

With the ever decreasing channel length of CMOS technology, if voltage is not properly scaled down as feature size scales, the electric field in the channel will increase dramatically. The LDD is used to grade the doping in the drain region and to produce an $N^+/N/P$ profile between drain and the channel in NMOS and vice versa for PMOS. The LDD also acts like a shallow junction that connects the deeper source and drain. This shallow junction works well to cope with the drain induced barrier lowering effect, usually found in short channel devices. In practice, another implant step called a halo implant is frequently used for advanced CMOS process. This helps alleviate the short channel effect found in sub-micron CMOS devices.

- Lithography Contact Window pattern

- RIE removal of CVD oxide and thermal oxide
- Sputter Deposit Al metal ~ 0.7μm
- Lithography Al interconnect pattern
- RIE etch of Al metallization
- Sintering at ~ 400°C in H_2/N_2 ambient

By now, we have finished the Front End of the Line (FEOL) process steps. The next step is to be devoted to Back End of the Line (BEOL). First, Interlayer Dielectric (ILD) needs to be deposited and contacts will be placed. Finally metals are deposited and patterned on top of the contacts. In real manufacturing, multiple metal layers are involved and copper (Cu) is commonly used.

The first step in simulating a semiconductor structure is defining a mesh. When using Athena for simulation of the fabrication process, the user needs to specify an initial mesh density. After the execution of mesh statements, the mesh setup is shown in Figure 4.1.
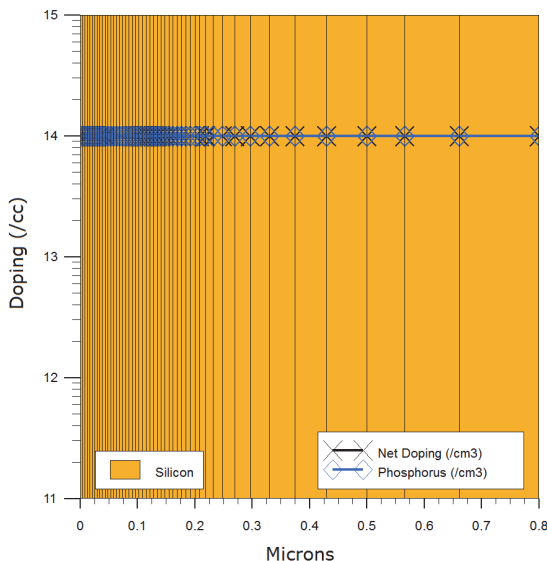
Figure 4.1. NMOS process simulation – mesh generation.

Screening oxide growth and etching are shown in Figure 4.2. This layer is necessary to protect the wafer for the following implantation step.
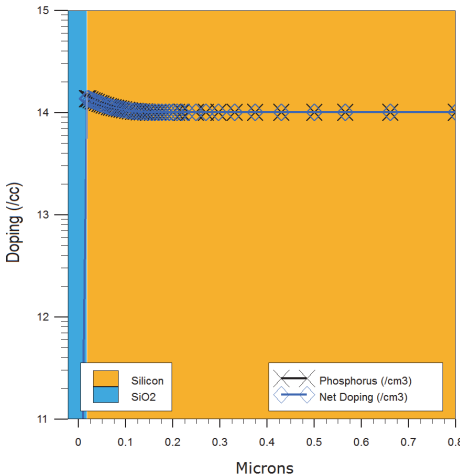
Figure 4.2. Screening oxide growth and etching.
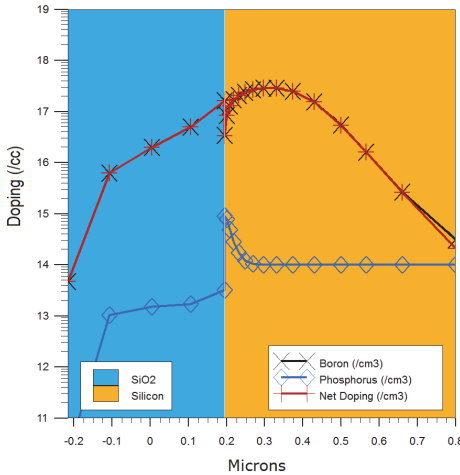
Figure 4.3. P-well formation and implantation.

The next step in the N-channel MOSFET development is implantation of boron to create a P-well in the wafer. The P well is formed using a boron implantation. This implant step is modeled using the default Dual Pearson model.
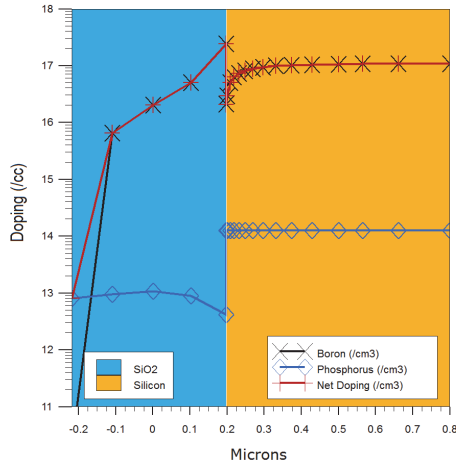
Figure 4.4. P-well formation: doping profile after well drive-in.
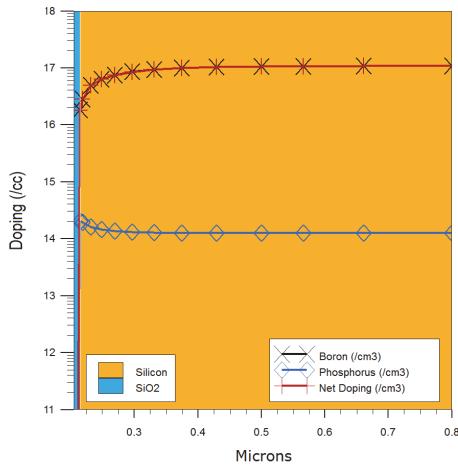
Figure 4.5. After gate oxide growth.

Figure 4.3 shows the nature of the ion implantation step and the peaks at the average penetration depth with a specific doping concentration. The next step is the thermally growth SiO_2 on the wafer. After the simulation, the new structure is plotted in Figure 4.4.

The next step is to perform a sacrificial cleaning which requires oxidation and then the removal of the oxide. This process ensures that the surface layer of Si is free from damage from previous process steps. The 10–12nm gate oxide (see Figure 4.5) is thermally grown in dry oxygen. Next, to accurately adjust the threshold voltage of the N-channel MOSFET device, a low energy boron implantation is performed. After the threshold voltage adjust implant is plotted in Figure 4.6.

The next step in the MOSFET fabrication is the deposition of the polysilicon layer for the gate of the N-channel MOSFET (Figure 4.7). The MOSFET is a symmetrical device: the region from the center of the gate to the source on the left is identical to the region from the center of the gate to the drain on the right. For this reason, in this step, only left half of the full device is simulated. Figure 4.8 shows a two dimensional structure after polysilicon gate formation.
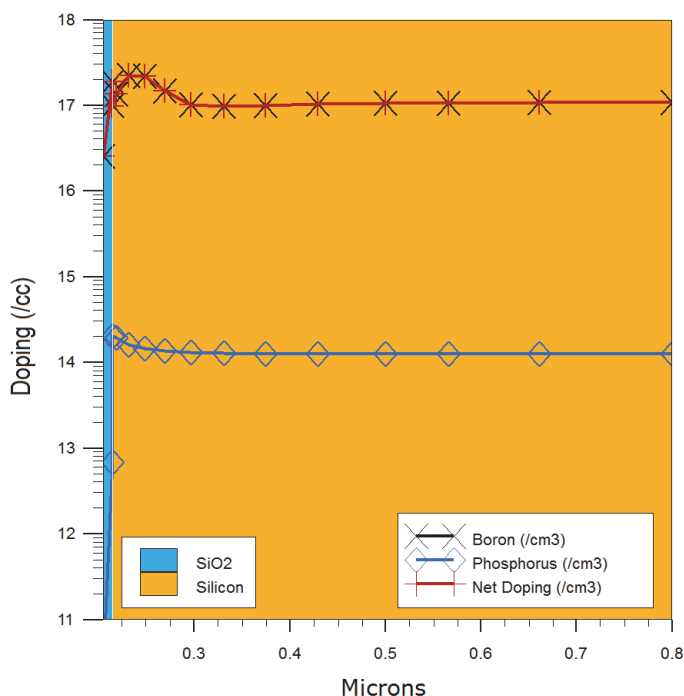
Figure 4.6. Threshold voltage adjustment implant.
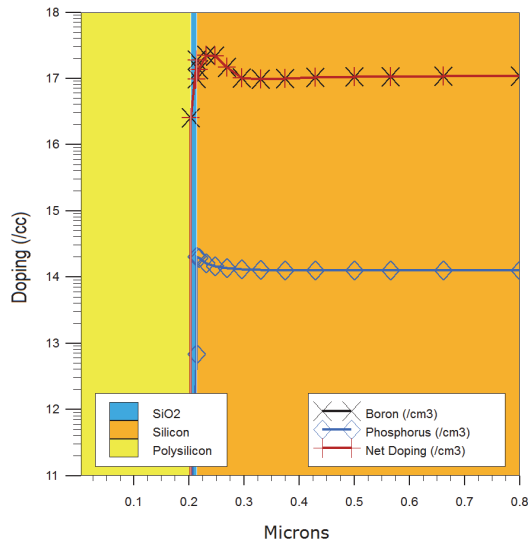
Figure 4.7. Polysilicon deposition.
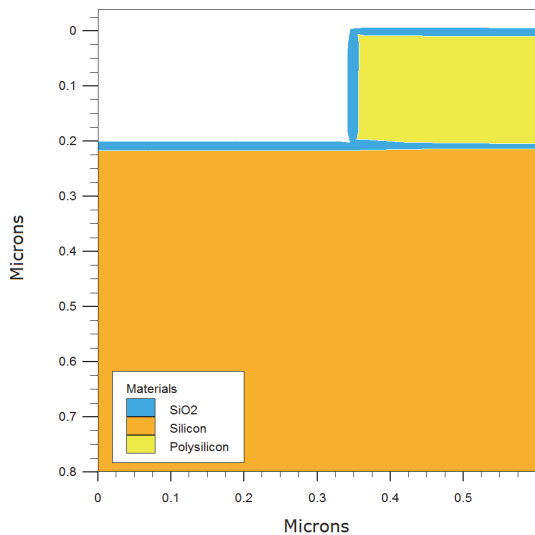
Figure 4.8. Polysilicon etching and gate formation.
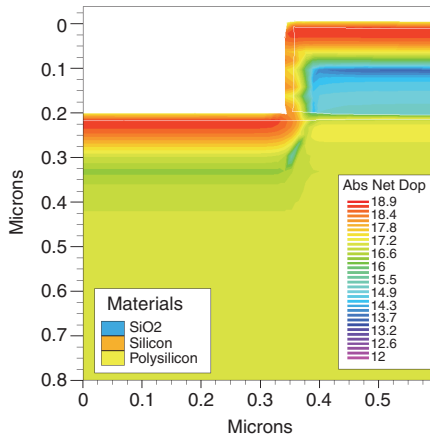
Figure 4.9. N-channel MOSFET structure after LDD implantation.
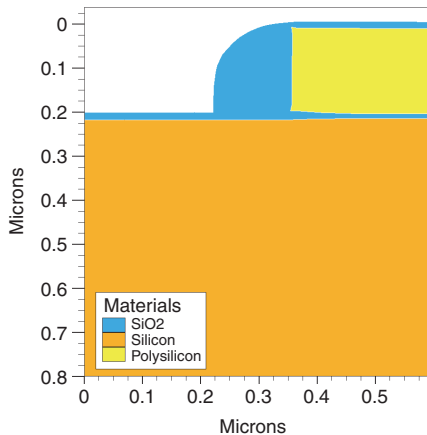
Figure 4.10. LPCVD oxide deposition for spacer formation.

The next step in the MOSFET fabrication is the lightly doped drain/source implantation. The LDD junction can be formed by using low current implantation. It is a shallow junction with very low dopant concentration, extended just underneath the gate. Phosphor implant was used to form the n-type shallow junctions. The LDD regions can now be seen in the resulting structure plot in Figure 4.9. A thick oxide is deposited conformal by low pressure chemical vapor deposition (LPCVD) and anisotropically etched so that oxide spacers at the polysilicon edges are created. The oxide spacer is represented in Figure 4.10.

The next step is source/drain formation. High current and high energy ion implantation forms the heavily doped source/drain junctions, which are keeping apart from the gate by the sidewall spacers. The MOSFET structure after heavy drain/source doping Arsenic implantation is shown in Figure 4.11. The thermal anneal step is required after the ion implantation to electrically activate implanted ions. The thermal anneal moves ions into lattice structure sites of the silicon.
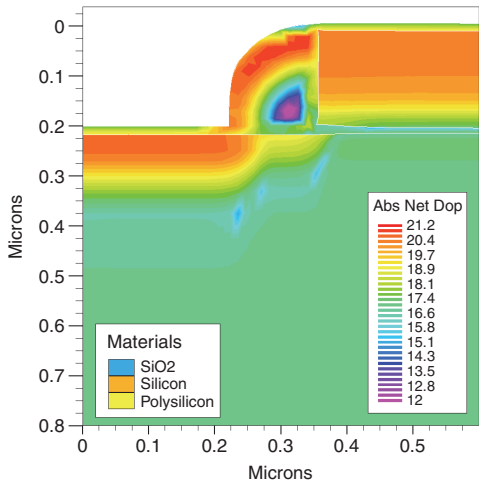
Figure 4.11. N-channel MOSFET structure after heavy drain/source doping Arsenic implantation.
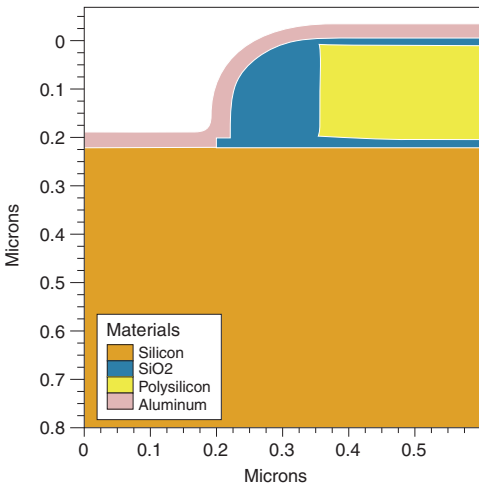
Figure 4.12. Metal deposition.
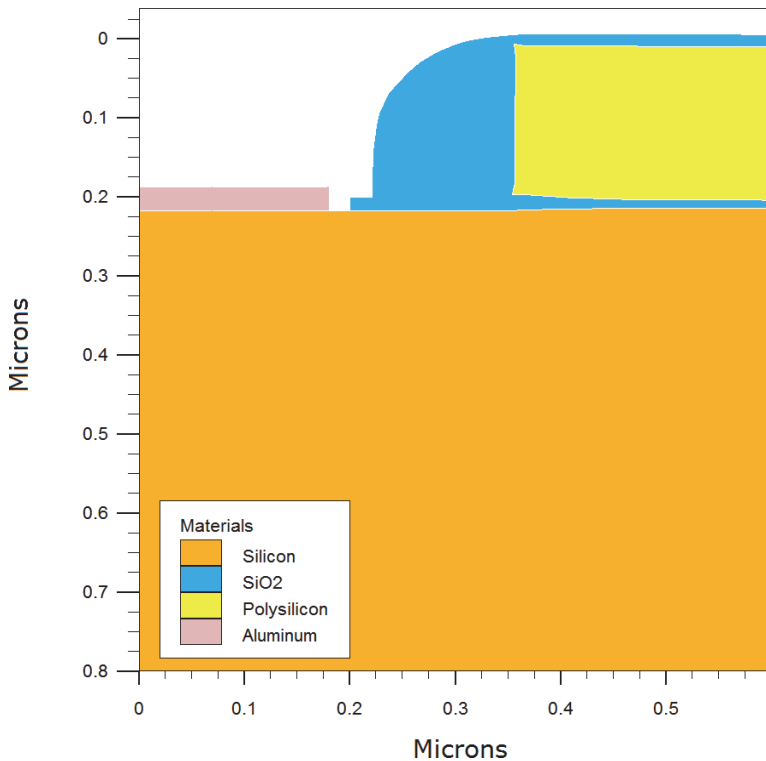
Figure 4.13. Contact opening.

The next step is contact and local interconnect formation which requires depositing and etching the oxide layer above the drain/source region. Aluminum is commonly used for this purpose with its very low resistivity and its adhesion compatibility with SiO_2. Al is deposited on the surface, and then it is etched away, except the one above source/drain region, to form source/drain region electrodes as shown in Figure 4.12. The MOSFET structure after the contact opening is shown in Figure 4.13.

So far, only left half of the full device is simulated. Mirroring of this half device into a full structure is performed using mirror right statement. Finally, the N-channel MOSFET is contacted and the electrical behavior of the device can be analyzed. The full device structure along with net doping, grid and electrodes can now be seen in Figures 4.14a, b and c, respectively.
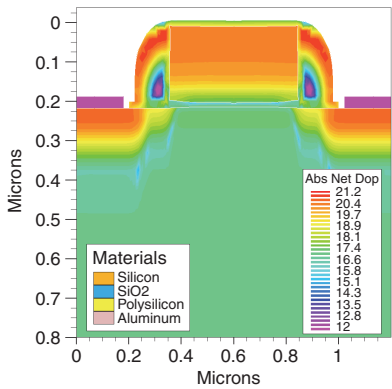
Figure 4.14a. Full n-MOSFET device structure.
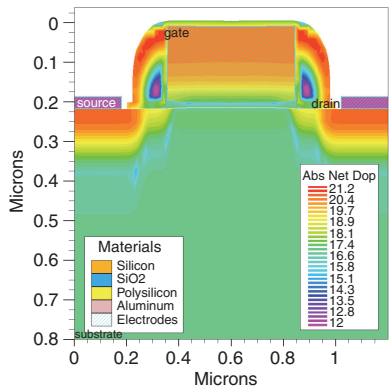
Figure 4.14b. Net doping in the n-MOSFET.
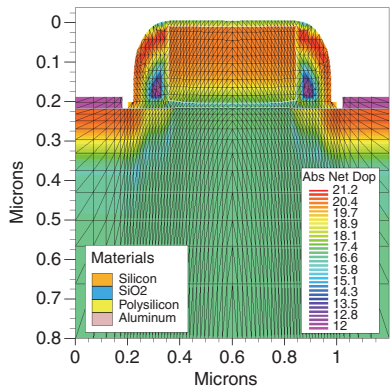
Figure 4.14c. Electrode specifications and grid used for the structure.

4.2   MOS Device Simulation

The complete structure can now be simulated in ATLAS to provide specific characteristics such as the current-voltage characteristics. ATLAS device simulates $I_d - V_d$ to examine family of characteristics. This DC simulation uses the exact set of properties, models, and mathematical solvers. The gate is first ramped to the particular voltage and the drain is swept from 0 to 3V; substrate and source are both grounded. ATLAS simulates $I_d - V_g$ characteristics to examine drain current as a function of gate voltage for low drain bias. Knowing the threshold voltage is important in studying charge injection and clock feed through. The threshold voltage is one of the most important parameters in MOSFET device. Since the threshold voltage determines the requirements for turning the MOS transistor on or off, it is very important to be able to adjust threshold voltage in designing the device.

Standard device models are selected to simulate the DC characteristics of the transistor. The property of the n^+ polysilicon gate is specified. The Philips unified mobility model, velocity saturation within high field regions, and degradation due to surface roughness scattering are all activated to model mobility in silicon. Carrier recombination in silicon is specified using Shockley-Read-Hall recombination with doping dependent lifetime. Charge at the Si-SiO$_2$ interface is defined to have a concentration of 3.0×10^{10} cm^{-3}. Since the simulations pertain to n-MOS transistors, only electron continuity and Poisson's equations are solved.

Many characteristics such as MOSFET $I_d - V_d$ or $I_d - V_g$ simulations require a family of curves to be produced. It is done by obtaining solutions at each of the stepped bias points first, and then solving over the swept bias variable at each stepped point. For example, in MOSFET $I_d - V_d$ characteristics, solutions for each V_g value are obtained with $V_d = 0.0$ V. The outputs from these solutions are saved. The $I_d - V_d$ characteristics obtained from the MOSFETs give a good indication of the performance of the device. Figure 4.15 shows the $I_d - V_d$ characteristics that are obtained from n-channel MOSFET. Gate voltages of 1.1 V, 2.2 V, and 3.3V are applied to the transistor, and the drain current is graphed as a function of the applied drain voltage. Figure 4.16 shows the resulting $I_d - V_g$ characteristics simulated in ATLAS. Breakdown voltage

extraction for the MOSFET may be done by choosing the Selberherr impact ionization model in the simulation. Figure 4.17 shows the high ionization rate at the drain end for a drain voltage of 0.5V.

Device Simulation Using ATLAS   MOSFET OUTPUT CHARACTERISTICS

Extracted n-MOSFET Device parameters from Simulation

nxj=0.174287 um from top of Silicon layer X.val=0.1 gateox=100.163 A X.val=0.05
n++ sheet rho=29.0937 ohm/square X.val=0.05   long channel Vt =0.610 V X.val=0.49
ldd sheet rho=2176.82 ohm/square X.val=0.3   max drain current = 0.000437757 A
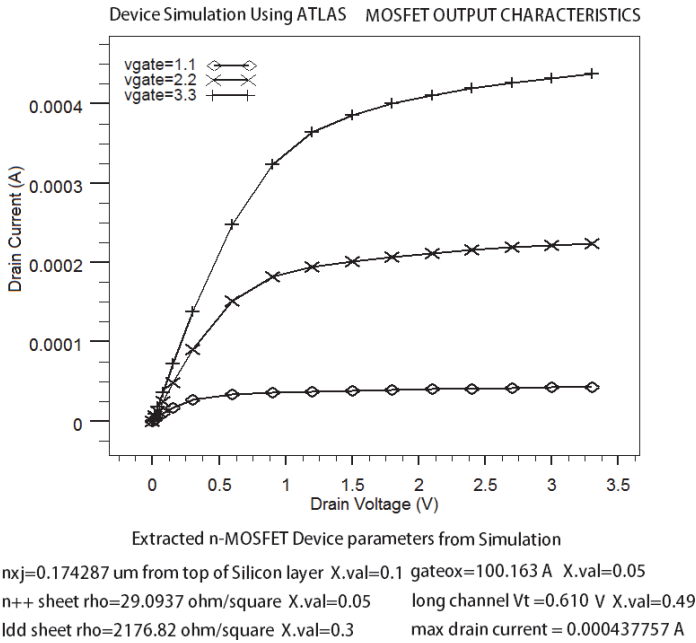
Figure 4.15. Electrical characterization (output characteristics) using device simulation tool ATLAS. Extracted device parameters are also shown.
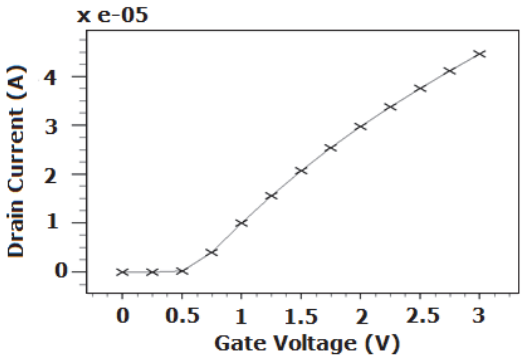
Figure 4.16. $I_d - V_g$ characteristics simulated in ATLAS.
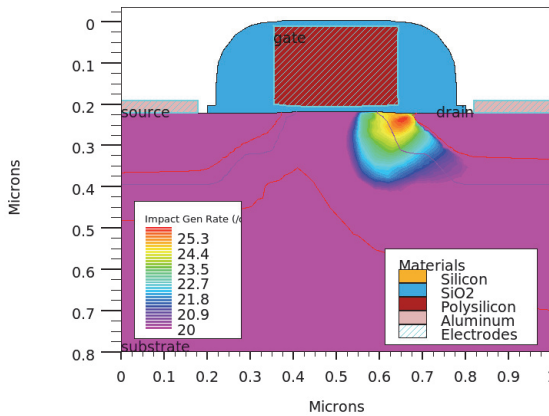
Figure 4.17. Impact generation rate at the drain end for a drain voltage of 0.5V.

4.3 Simulation Studies of Short Channel Effects

In modern MOSFETs, short channel behaviors such as drain-induced barrier lowering, channel length modulation, velocity saturation, etc. impact the transistor on-state characteristics. Device scaling causes unintended effects of device performance. These effects can be categorized according to the different sources:

- Electric field profile changes as two dimensional
- Drain-induced barrier lowering
- Mobility reduction by gate-induced surface field
- Electric field strength becomes very high in the channel
- Velocity saturation
- Impact ionization near drain
- Gate oxide charging
- Parasitic bipolar effect
- Physical separation between the source and the drain decreases.
- Punch through
- Channel length modulation

With aggressive scaling of gate length and gate dielectric thickness, switching operation becomes a big challenge due to difficulty of off state leakage current control. In highly scaled MOSFETs short channel effects occur such as coupling, strong fringing fields, charge sharing effect (CSE) and drain induced barrier lowering. The CSE is due to the increased influence of the depletion region at the source and drain junctions when one scales down the MOSFETs. This leads to a reduction of the depletion charge controlled by the gate and thereby a decrease of the threshold voltage resulting in substantial leakage current increase. The DIBL is due to the electrostatic influence of the drain potential on the source/channel barrier height at higher drain voltage. This phenomenon has been shown to jeopardize deep sub-micrometer device operation by shifting the threshold voltage and increasing the leakage current.

Hot Electron Reliability

Hot carrier is an important short channel phenomenon resulting from the high electric field in the device and causes degradation in device characteristics. Decreasing the channel length in the device without reducing the supply voltage increases the average field in the device. Such high fields are large enough to cause problems in semiconductor devices. Such problems are often called "hot electron" problems because most of them are due to the high energies that electrons (or holes) can reach in high electric fields. When a PMOS transistor is stressed at high voltages a threshold voltage shift is observed. Hot carrier reliability feature simulated for a PMOS is shown in Figure 4.18 with the stressing conditions shown.

The device degradation modeling is done using SILVACO ATLAS. First the structure is loaded and the correct work function and interface fixed charge are set. On the models statement the standard mobility and recombination models are specified. In addition the parameters "hei devdeg.e" are set. "hei" turns on the hot electron injection model for gate current. The "devdeg.e" parameter sets that the gate current is used to calculate device degradation. The "degradation" parameter is used to specify the interface state density and the electron trapping cross section, "sigmae". The density of interface states as a function of position is also defined. The stressing conditions are set at $V_d = -6.0V$ and $V_g = -1.5V$.
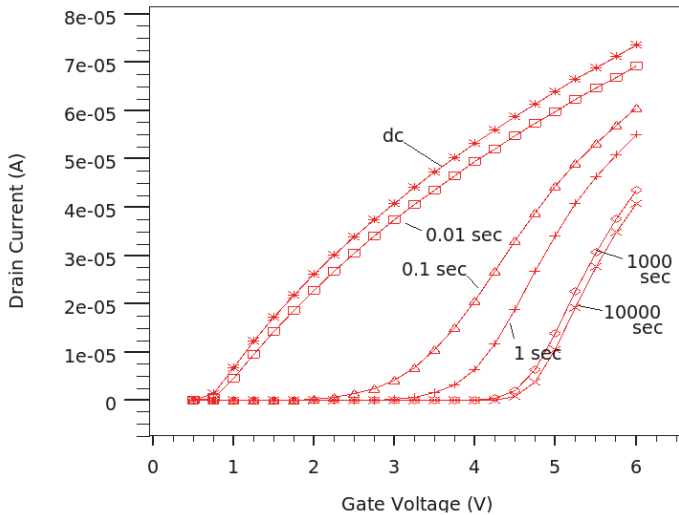
Figure 4.18. Hot carrier reliability simulation for a p-MOSFET.

The device is biased to these voltages in DC mode. Then ATLAS is switched to a transient solution. The transient simulation is set to run for 1000s. At various intervals, data is saved to solution files. The $I_d - V_g$ characteristics from different stress time runs are overlaid in TonyPlot to show the threshold voltage shift caused by the hot carriers trapped at the interface as shown in Figure 4.18.

Poly Depletion

It is usually assumed that the poly-Si gate in a MOSFET is highly doped such that depletion in the gate does not occur. However, in modern CMOS technology a number of processing techniques currently in use can reduce poly doping through various effects such as, dopant evaporation, insufficient implant dose, silicide suck-out and compensation from P^+ implant. Reduced poly doping can result in depletion occurring in the poly Si gate under bias. Poly silicon doping reduction results in work function increase and subsequent voltage drop which causes a loss of current drive in the final device. The effect of reduced poly doping in a MOSFET with a thin gate is to reduce drain current drive and increase the threshold voltage. Poly depletion results in unwanted variation of threshold voltage since some voltage drop across the gate oxide is now dropped across the depletion region as well.

The use of dual n^+–p^+ polysilicon gates is a key advance in modern CMOS technology, since it allows the source and drain regions to be self-aligned to the gate and sets a symmetric and the most suitable threshold voltages for both p- and n-channel bulk MOSFETs. However, when the active carrier concentration in the polysilicon is not high enough to pin the Fermi level at the poly-Si/SiO_2 interface, the band bending in the poly-Si becomes voltage-dependent. As the device is biased such that the Si substrate is inverted and a channel is formed, the poly-Si gate becomes depleted of free carriers and a significant voltage fraction is dropped across the gate; this effect is known as polysilicon depletion. Gate capacitance degradation due to the polysilicon depletion typically accounts for 0.4–0.5nm of the equivalent-oxide thickness of the total gate capacitance at inversion. This is a substantial amount, considering that an equivalent gate oxide of less than 1.5nm (at inversion) is required for the coming technology nodes.

In the following, using a MOS capacitor simulation, we demonstrate the poly depletion effects for a typical sub-micron MOSFET using ATHENA and ATLAS process and device simulators. The example run
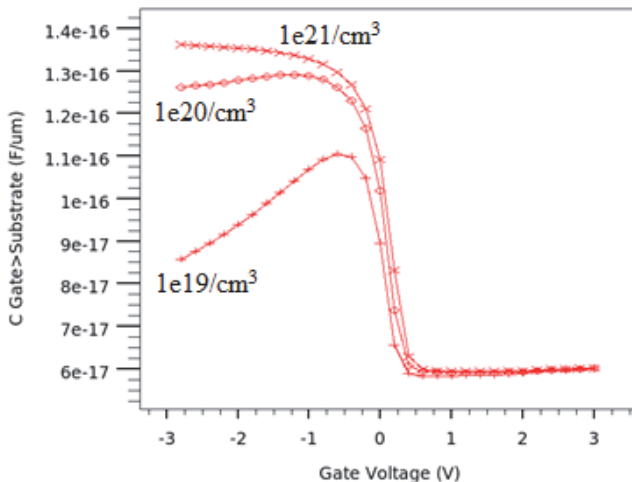
Figure 4.19. Poly depletion effect.

shows Capacitance-Voltage characteristics of a P^+ poly gate doped to a concentration of $1 \times 10^{19}/cm^3$. Then the program is run to compare the same C–V plot for the poly gate doped to $1 \times^{20}/cm^3$ and. By overlaying the three cases the poly depletion effect is shown in Figure 4.19. The undesirable effect shows up in the C–V characteristics as anomalies in the shape of the plot. It is observed that even for a poly gate doped to $1 \times^{20}/cm^3$, depletion occurs in the poly Si gate showing anomalous C–V characteristics which may result in serious consequences if the maximum value from the C–V characteristic is used to measure the thickness of the gate oxide.

Mobility Model Comparison

An examination of the various mobility models available in SILVACO ATLAS is presented below. A comparison is made between three different surface mobility models used in MOSFETs. The three models compared in this example are CVT, Shirahata and modified Watt. The mesh generation and structure definition part of this input file follows the standard syntax. A variable "surf_a" is used to set the mesh spacing at the silicon surface. This mesh spacing is important in resolving the high electric field and carrier concentration gradients at the surface. Overlay of $I_d - V_g$ characteristics and mobility in TonyPlot is shown in Figure 4.20. The extracted mobility in the channel of the MOSFET for the three models is shown in Figure 4.21.
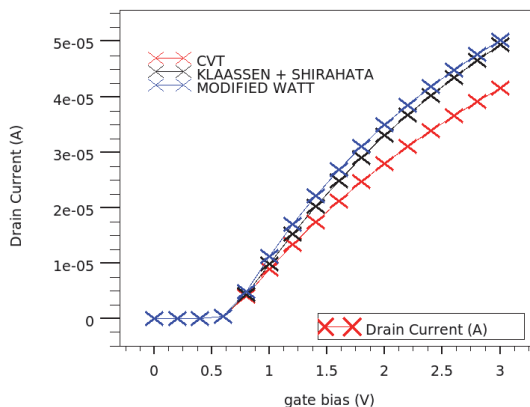
Figure 4.20. Drain current of a MOSFET. Comparison for CVT, SHIRAHATA and WATT mobility models.
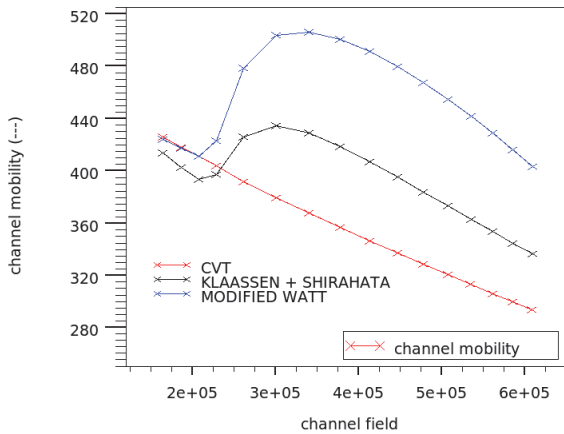
Figure 4.21. Mobility in the channel: comparison of CVT, SHIRAHATA and WATT mobility models.

Comparison of EB and NEB Models

Deep submicron devices should be simulated using the Energy Balance Model due to velocity overshoot and nonlocal impact ionization effects, which could substantially influence device characteristics. This example demonstrates a comparison of $I_d - V_d$ characteristics obtained between the Energy Balance and Nonisothermal Energy Balance Models. Comparison of transport models is made for a Double gate MOSFET including quantum effects. For high current levels (high gate voltages) the thermal self-heating effects can also play an important role by decreasing mobility and impact ionization rate. The first ATLAS run uses Energy Balance Model and the second run uses both the energy balance and heat flow equations. The final $I_d - V_d$ characteristics are compared and overlaid using TonyPlot as shown in Figure 4.22. The impact ionization rate and lattice temperature in the device are shown in Figures 4.23 and 4.24, respectively.
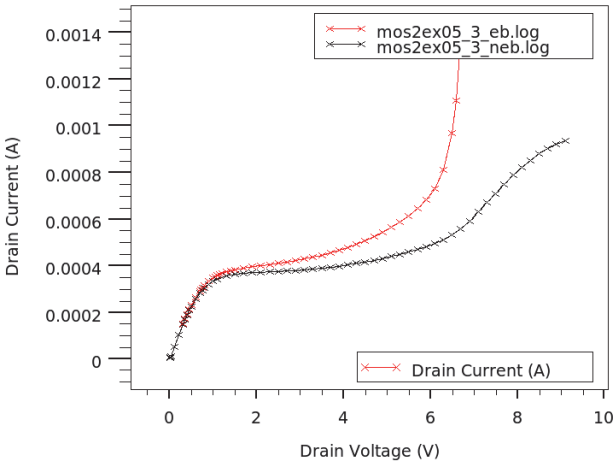
Figure 4.22. Comparison of energy balance and non-isothermal energy balance models.
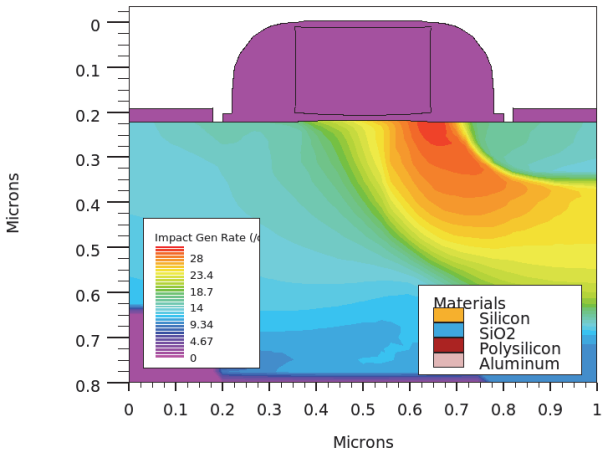
Figure 4.23. Impact ionization rate due to nonlocal impact ionization effects.
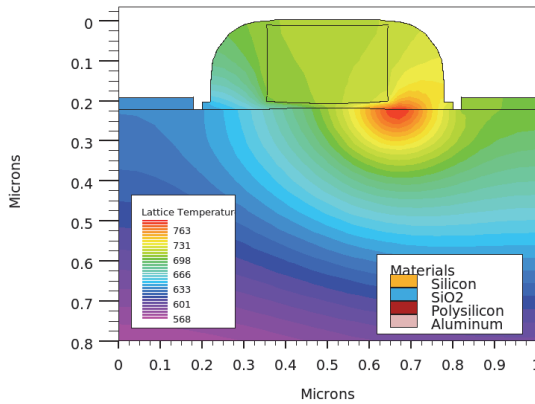
Figure 4.24. Lattice temperature distribution in the device.

4.4   MOSFETs with High-k Gate Dielectrics

Scaling of the gate oxide thickness below 1nm is necessary to control the short channel effects beyond the 65-nm-technology node, but the gate leakage current through such a thin oxide layer becomes intolerably high. Viable solutions to reducing the gate tunneling current and gate capacitance degradation due to polysilicon depletion have been explored through introduction of unconventional materials: high-dielectric constant (high-k) gate dielectrics and metal gate electrodes. In order to understand the behavior of effective mobility, it is important to break down the mechanisms which can affect the inversion layer mobility. Among the scattering mechanisms, coulomb scattering, phonon scattering, and surface roughness are the most prevalent in CMOS devices. Phonon scattering of the inversion carriers originates from the acoustic and optical mode vibrations of the lattice. As the temperature increases, lattice vibrations also increase and thus the phonon scattering rate also increases. At temperatures below room temperature, phonon scattering can be ignored. However, at room temperature, phonon scattering plays an important role on the inversion layer mobility.

Recently, it has been predicted that the remote phonon scattering originating from the gate dielectric cannot be neglected in MOSFET devices with high-k materials. Since the bonds in most high-k dielectrics are metal-oxygen bonds, they are highly polarizable and result in a large static permittivity. Reduced surface mobility observed in channels under hafnium oxide degrade the performance of small transistors. The mobility reduction is considered to be due to Coulombic charge scattering and remote phonon scattering, both of which can be screened by a depth of interfacial oxide. As such it is not just the structure of the transistor's insulating gate that needs to be investigated and optimized but the complete insulator-and-gate-electrode stack. In addition, the high-k gate dielectric material must be compatible with current CMOS fabri-cation process flow and other materials used in the CMOS integrated circuits. The following simulations demonstrate the effects of remote phonon scattering and remote Coulomb scattering on the effective mobility in high-k gate dielectrics.

In the first simulation example we use of the high-k mobility model for determining the mobility reduction due to remote phonon scattering. First we use HfO_2 as gate dielectric and run. Then we use SiO_2 of equivalent thickness and compare the mobility data. We probe the normal electric field and channel charge for calculating the effective channel mobility. Figure 4.25 compares effective mobility vs. perpendicular field for HfO_2 and SiO_2 gate dielectrics. Mobility degradation is clearly observed.
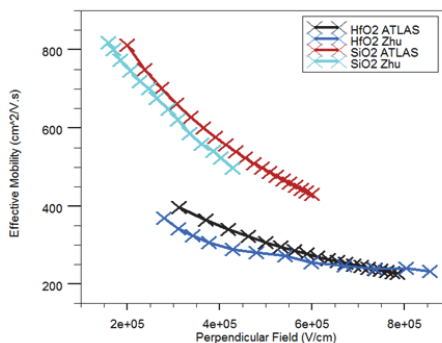
Figure 4.25. Effective mobility vs. perpendicular field for HfO_2 and SiO_2 gate dielectrics due to remote phonon scattering. Data used for comparison purposes are after W. Zhu and T. Ma, Temperature Dependence of Channel Mobility in HfO_2-Gated NMOSFETs, *IEEE Elec. Dev. Lett.*, Vol. 25, pp. 89-91, 2004.
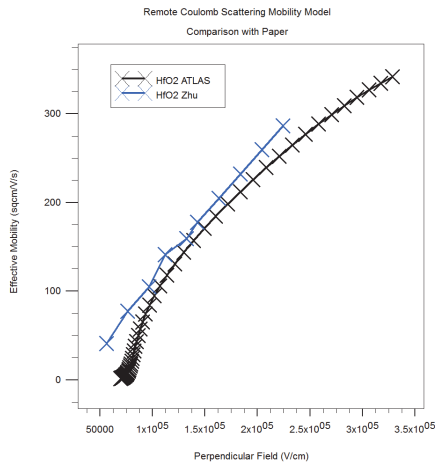
Figure 4.26. Effective mobility vs. perpendicular field for HfO$_2$ gate dielectric due to remote Coulomb scattering. Data used for comparison purposes are after H. Tanimoto *et al*., Modeling of Electron Mobility Degradation for HfSiON MISFETs, *Proc. of SISPAD*, pp. 47-50, 2006.

In the next example, we use of the high-k mobility model for determining the mobility reduction due to remote Coulomb scattering. We probe the normal electric field and channel charge for calculating the effective channel mobility. Figure 4.26 compares simulated effective mobility as a function of perpendicular field.

4.5   Single Event Effects

Radiation-resistant electronics are integral to the aerospace electronics. The operation of the electronic circuits in space presents a host of challenges for device, circuit and system designers. The circuits are subjected to harmful radiations which may disrupt the performance of the circuit temporarily or may even destroy the circuit permanently. The subject of radiation effects in semiconductor devices is complex because of the numerous possible combinations of different types of radiation, radiation damage, and semiconductor devices. It is important to establish a link between basic mechanisms and Electronic Design Automation (EDA) tools, providing an enhanced design flow for radiation-resistant electronics. Radiation resistant electronics are integral to the aerospace and nuclear reactor technologies. The study of radiation effects in semiconductor electronics, and the development of radiation-resistant

integrated circuits, has an active scientific and engineering community that produces a wealth of data and conceptual understanding. The subject of radiation effects in semiconductor devices is complex because of the numerous possible combinations of different types of radiation, radiation damage, and semiconductor devices. Some devices are more sensitive to ionization effects, while other device types are dominated by displacement damage.

When energetic particles hit the sensitive nodes of the electronic circuit, it leaves a trail of electron/hole pairs. These charges are collected by the sensitive node of the circuit, which gives rise to Single Event Effect (SEE). This strike may cause a temporary disturbance in the circuit operation or may even cause a permanent damage in the circuit. The single event effects can be classified into two categories. They may be either destructive SEE or non-destructive SEE. Single event current transients (SET) and single-event upset (SEU) in logic or memory circuits are examples of non-destructive SEE and Single event latchup (SEL) is an example for destructive SEE.

Energy absorbed by electronic ionization in insulating layers, predominantly SiO_2, liberates charge carriers, which diffuse or drift to other locations where they are trapped, leading to changes in the electric fields. Ionization produces three types of effects: total dose, single event, and transient. Displacement damage depends on the Non-Ionizing Energy Loss (NIEL) process. Energy and momentum transfer to lattice atoms depends on the mass and energy of the incident particles. A measure for displacement damage is displaced atoms per volume, and must be specified for a specific particle type and energy. Ionization effects depend primarily on the absorbed energy, and are frequently assumed to be independent of the type of radiation. At typical incident energies, ionization is the dominant absorption mechanism.

An incident particle or photon capable of imparting energy of about 20 eV to a silicon atom can dislodge it from its lattice site. Displacement damage creates defect clusters. Displacement damage manifests itself in two ways; the formation of mid-gap states, and/or a change in doping characteristics. The formation of mid-gap states facilitates the transition of electrons from the valence to the conduction band. In depletion regions, this leads to an increase in the generation current of

reverse-biased pn diodes. In forward biased junctions, or non-depleted regions, mid-gap states facilitate charge loss by recombination. States close to the band edges facilitate trapping, where charge is captured and released after a certain time. A clear understanding of basic mechanisms, and linking this understanding to a design flow methodology, will give designers radiation-knowledgeable device models for the exploration of the design space, with the restrictions given by the radiation environment.

In the following, using the example provided by SILVACO, we simulate the effects of different angles of incidence of an SEU in a MOSFET structure. The MOSFET structure is constructed using DevEdit3D (see Figure 4.27). The structure is then passed to ATLAS for electrical testing. The input file consists of the following parts:

- construction of the device in DevEdit3D
- SEU simulation in ATLAS: particle with normal incidence
- SEU simulation in ATLAS: particle with oblique incidence

The first stage of the input file constructs the MOSFET geometry, material regions, doping profiles, electrodes, and subsequently generates the mesh in 3D. This is done in DevEdit3D by drawing the device regions in interactive mode, and specifying 3D doping distributions. The mesh was generated automatically by specifying basic mesh parameters and constraints with subsequent refinement based on the doping distribution. The simulation is first performed to obtain the condition of the structure prior to the particle hit. This condition is with the drain voltage of 5V, all other electrodes grounded. This condition of the structure is saved in a solution file for the use in the second atlas run. The parameters of the charge track are specified in the "singleeventupset" statement. The general description of the SEU physical model employed in ATLAS3D mode as well as the description of all of the parameters for specifying position and time dependent generation is given in the Appendix to the current example description.
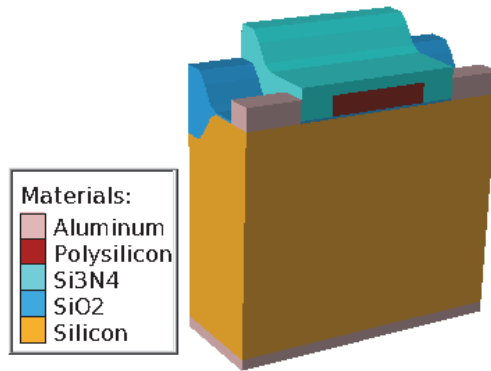
Figure 4.27. DevEdit3D generated MOSFET structure.
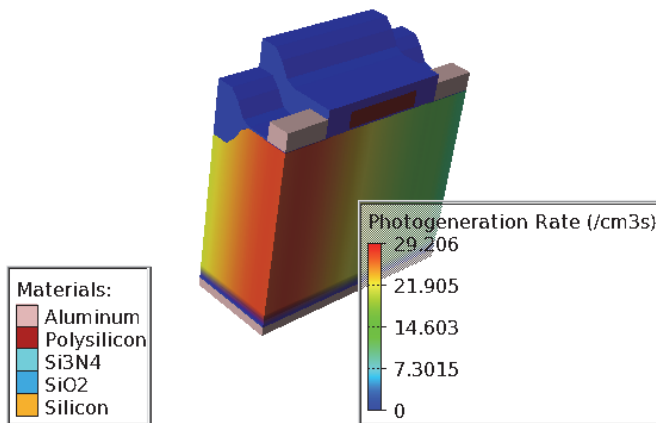
Figure 4.28. Photo generation rate in the device.

The first SEU simulation is performed for a normal particle incidence through the drain region (along Y direction) with a maximum carrier density generated of $1.0e18$ cm^{-3}. The characteristic radius of the track was defined to be $0.05\mu m$, the characteristic time of the Gaussian time dependency – 2 ps, and the time instant of the peak of generation – 4 ps. Figure 4.28 shows the photo generation rate in the device.
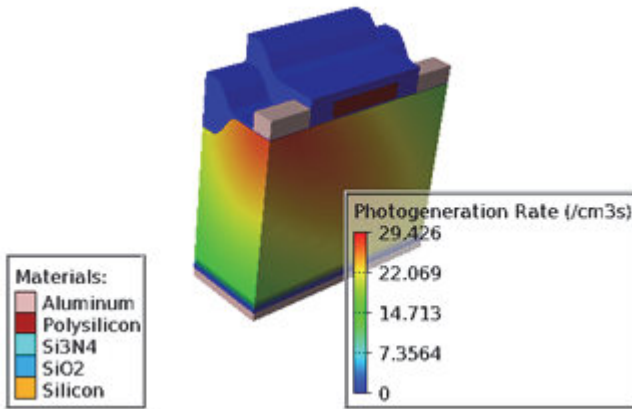
Figure 4.29. Photo generation rate in the device.

The transient simulation is performed to monitor the effects of SEU generation. The initial time steps in the solve statements are defined by the user (e.g., 0.05ps in the first solve statement), while the subsequent time steps are selected automatically. The simulation is performed up to 0.1μs of physical time. To analyze the effect of carrier generation on internal physical distributions an output structure file is saved for the instant of the peak of generation (t = 4 ps). Figure 4.29 shows the photo generation rate in the device.

The simulation then proceeds to the second ATLAS input file for simulating oblique particle incidence. This input file contains the same CONTACT and MODELS definitions as the previous one. The SEU definition of the particle hitting the device along its longer diagonal through the drain region (far left top corner) down to the near right bottom corner of the structure is specified in the SINGLE statement. All other parameters of the SEU are the same as in the previous ATLAS simulation. The condition of the structure of $V_d = 5V$ is loaded and the transient response of the SEU is simulated for the second SEU conditions. Figures 4.30 and 4.31 show the potential distribution and the recombination rate at the drain end in the device.
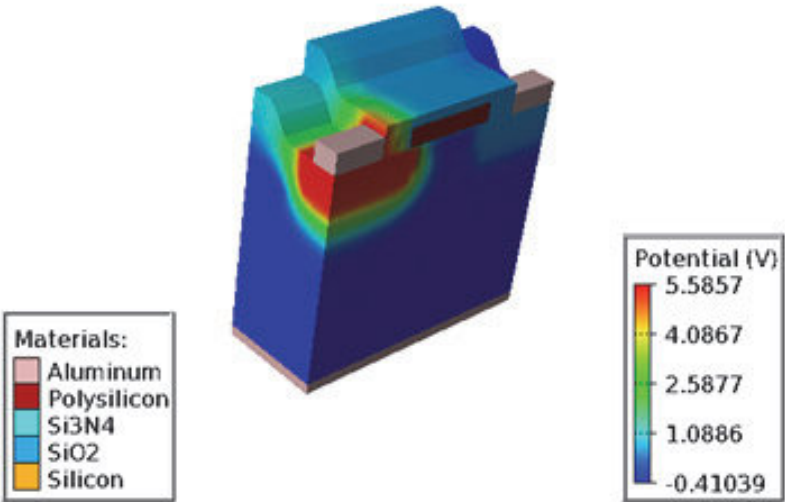
Figure 4.30. Potential distribution in the device.
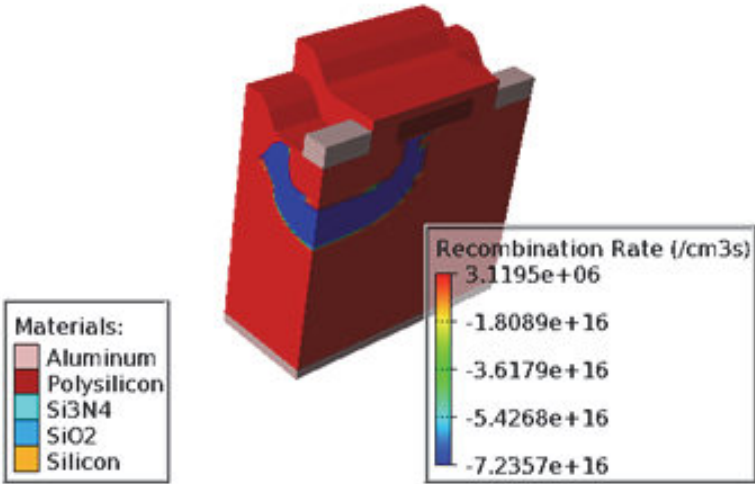
Figure 4.31. Recombination rate in the device.
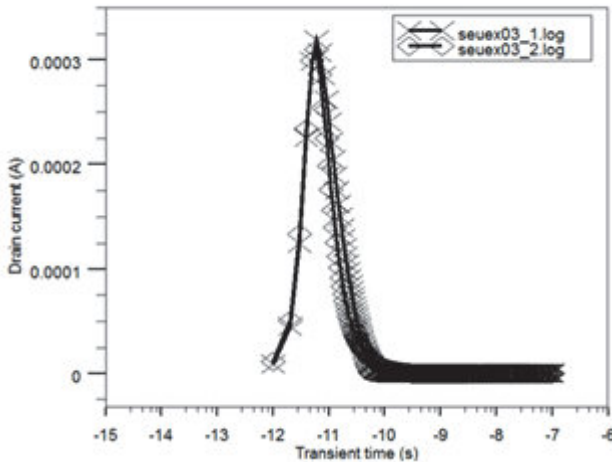
Figure 4.32. Drain transient current characteristics corresponding to the different SEU conditions.

Both drain transient current characteristics corresponding to the different SEU conditions are displayed in Figure 4.32. The peak of the current associated with the drift collection is achieved at approximately 6–7 ps. Note that the peak of generation is taking place at 4 ps. The drift collection continues for about ~100 ps when the charge in the depletion region is practically collected. The peak of the current for the diagonal incidence is calculated to be slightly higher than for the normal incidence. The decrease in current is however, slower for the normal incidence making the total extracted charge greater.

Summary

As semiconductor devices evolve, it is important to understand the fabrication processes and issues that arise with each new generation of transistor technology. In this chapter, we used the SILVACO process simulation tool ATHENA for CMOS process development and ATLAS for device simulation. Commonly observed short channel effects in small MOSFETs are studied using device simulations.

Chapter 5

Stress-Engineered CMOS

At the 65nm process node and beyond, it is evident that stress and strain-based techniques for mobility improvement will dominate traditional geometric scaling to maintain Moore's law trajectories for device performance. Uniaxial stress is now widely used in advanced CMOS technologies to boost transistor performance. Due to the layout dependence, the stress profile in a MOSFET channel is non-uniform. Under the small deformation approximation, one can assume the strain is linearly dependent on stress. The complexity of technology boosters dramatically increase the number of possible design options, making TCAD mandatory for a fast and cheap evaluation of the different device configurations. Presently TCAD is an established method for assessing the impact of process-induced stress variations on transistor performance. In TCAD simulations, the Finite Element Method (FEM) is used to solve stress relations numerically at each mesh point. There is also a strong interaction between the process induced stress and orientation of the substrate. In this chapter, we discuss on the linear elasticity theory to set an overall framework for mechanical stress that is of interest to us.

To understand the impact of stress on transistor performance, we consider two types of transistors in a typical CMOS. Applied mechanical stress affects the band structure of the semiconductor material which in turn affects the mobility and transistor performance. Depending upon the sign and direction of applied stress, stress may be beneficial or harmful, i.e., improving or degrading transistor mobilities. Positive valued stress is known as tensile stress which creates a "stretching" effect, while negative valued stress is known as compressive stress which creates a "squeezing" effect.
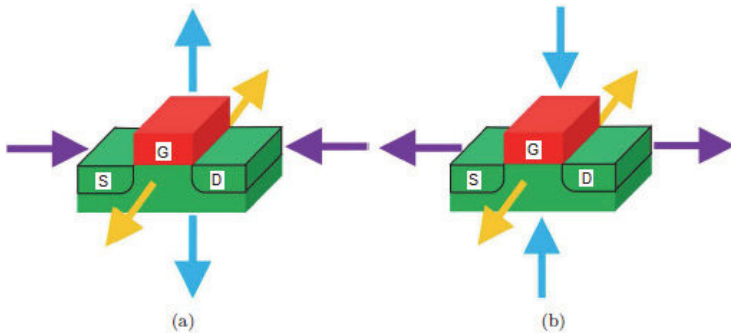
Figure 5.1. Beneficial stress orientations for (a) PMOS and (b) NMOS transistors. The colors corresponding to longitudinal, transverse, and vertical directions are purple, orange, and blue. Arrows pointing inward (outward) indicate compressive (tensile) stress. After S. K. Marella, Performance variations due to layout-dependent stress in VLSI circuits, PhD Thesis, University of Minnesota, 2015.

Figure 5.1 shows the preferred stress directions for NMOS and PMOS transistors. The direction along the transistor channel where current conduction takes place is defined as longitudinal direction, and within the plane of the channel, the orthogonal direction is known as the transverse direction. The vertical direction corresponds to the perpendicular to the wafer surface. The mobility of PMOS transistors is improved under compressive stress along longitudinal direction and tensile stress along the transverse and vertical directions. For NMOS transistors, mobility improves when tensile stress acts along either longitudinal or transverse direction, and compressive stress acts along the vertical direction. The opposite orientation of stress leads to mobility degradation. The actual magnitudes of mobility improvements and degradations can be explained through piezoresistive property of silicon and is due to the combination of stress from different directions.

At the beginning of the 21st century new methods for stress engineering have appeared. In this technique, one uses standard production processes to generate mechanical stress by "processes that induce stress" or "Process-Induced Stress" starting from 65nm CMOS technology node. There are several ways to modify the crystalline periodicity and get a modification of conduction and valence bands of silicon. One such technique is known as virtual substrate. It consists of growing a crystalline Si on a silicon-germanium substrate, thereby changing the size of the crystal lattice which creates strain.

5.1 Substrate Orientation

We shall examine the prospects for engineering crystallographic orientations for microelectronic applications as stress induction is highly dependent on the choice of crystal orientations. We shall focus on two directions: ⟨110⟩ and ⟨100⟩ to the channel with a (001) substrate for both n- and p-MOSFETs. On silicon wafers, the crystal orientation of a substrate is defined using the notch of the substrate. Commonly all the devices should have same direction of flow equivalent current as shown in Figures 5.2a, b, and c. Earlier technology nodes commonly used substrate (001) ⟨110⟩ for a number of reasons. First, the (001) plane is most favorable for the electron mobility and such a substrate is cleaved in the ⟨110⟩ direction. Changing the orientation of the substrate is a major change from the perspective of the application of mechanical stress. Two types of substrate orientations ⟨110⟩ and ⟨100⟩ are considered. Although in relation to deposition, there is no difference but oxidation rate depends heavily on oxidized plane. Other process differences may also take place for example, consumption of silicon will not be same, and boron dose during formation of trench isolation will not be same. This may also affect the location of dopants within the structure which may not be in the same because of channeling effects. Dopant diffusion can be modified by the injection of point defects during oxidation and silicidation.

In terms of mechanics, the two main parameters of elasticity: Young's modulus and Poisson's ratio are highly dependent on the orientation. They are anisotropic in the crystalline silicon which will inevitably play a role when substrate orientation is changed. Influence of orientation on the electrical characteristics of p-MOS devices with the two substrates of different orientations ⟨110⟩ and ⟨100⟩ is shown in Figure 5.3. It is noticed that in leakage current, a 12% increase occurs due to change in mobility. In principle, three kinds of stresses can be applied to the transistor:

- uniaxial stress along one crystallographic direction,
- biaxial stress along two perpendicular crystallographic directions, and

- hydrostatic stress applied uniformly in all three directions.
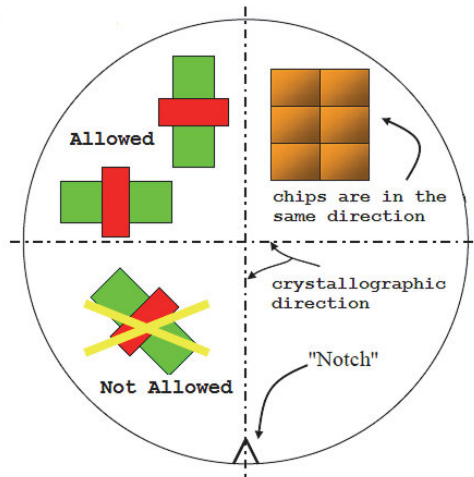
Figure 5.2a. Representative substrate and its notch which serves as a reference to positioning on the chip (right) and the direction of current in the devices (left).
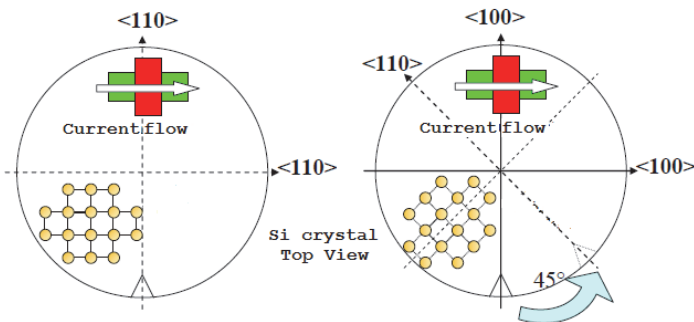
Figure 5.2b. Schematic representation of types of substrate (001) used in 65nm technology node. Orientation ⟨110⟩ (left) and Orientation ⟨100⟩ (right).
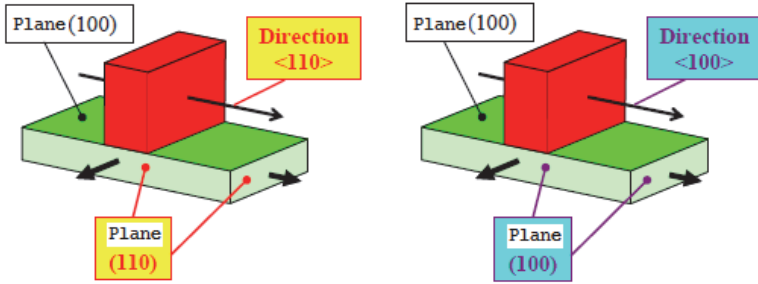
Figure 5.2c. 3-Dimensional representation of transistors with a substrate surface (100). In the case of left with a $\langle 110 \rangle$ orientation and plane (110) and in the case of right with a $\langle 100 \rangle$ orientation and the sidewalls (100).
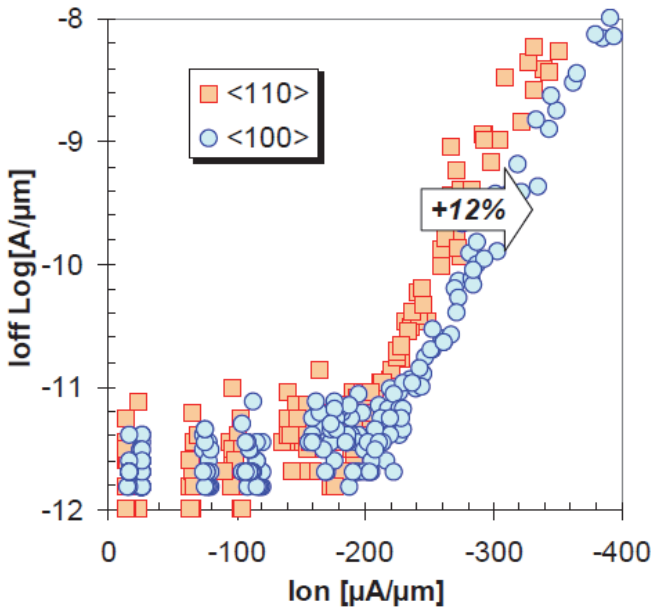
Figure 5.3. I_{ON}–I_{OFF} characteristics of p-MOS devices (W = 1μm) undergone the same manufacturing process but have two different current orientation $\langle 110 \rangle$ and $\langle 100 \rangle$.

5.2   Process Induced Strain (CESL)

We shall study the Contact Etch Stop Layer (CESL) and its consequences on the transistors. This process of inducing stress has become, by far, the most widely used in the microelectronics industry due to its ease of use. The understanding of its functioning will allow us to focus on the sensitivity of the devices according to their dimensions. The Contact Etch Stop Layer for inducing strain is one of the possible solutions amongst all the feasible stress engineering boosters (strained-spacer, silicide, SiGe, and strained-gate). The CESL consists in a nitride layer used to stop the etching of the metallic contact. By modifying the deposition recipe, the intrinsic strain of CESL can be compressive or tensile (see Figure 5.4). The usual thickness of CESL varies from 20nm to 100nm depending on the deposition time, and its thickness uniformity depends strongly on the process tool. CESL stress transmission to channel is a complex process; it is the outcome of several CESL sections acting separately (direct effect) or in association (indirect effect). The impact of CESL on the channel state of stress depends on various parameters, like gate length and height, spacer shape, presence of contact holes, CESL thickness and stress. So far, simple mechanical simulations have been performed to study the effects of CESL on device performance. Stress modeling technique is used aiming at investigating the stress state.
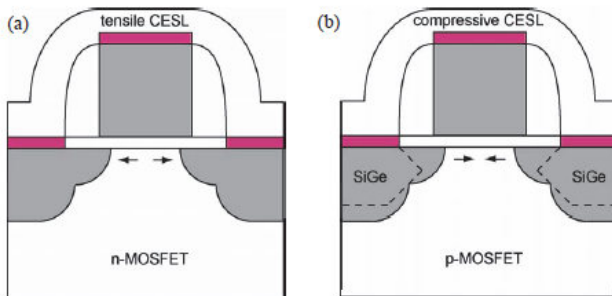
Figure 5.4. Schematic demonstration of strained silicon schemes for (a) n-MOSFETs and (b) p-MOSFETs.

The application of uniaxial strain along the channel has an impact on the electronic structure of the Si channel in a similar way, and therefore enhances electron mobility through the same mechanism as biaxial strain. Theoretical calculation shows that uniaxial strain offers more advantages over biaxial strain such as less band gap narrowing. In the case of holes, compressive stress splits the degenerated valence sub bands, increasing the hole population in the sub band with smaller transport effective mass.

Scaling of threshold voltage requires higher channel dopant density, which in effect decreases carrier mobility. In addition to this, carrier mobility is degraded with increasing effective field, which is a result of gate length scaling. As a result, process induced mechanical stress was adapted by CMOS technologies in sub-90nm nodes to improve carrier mobilities and maintain high drive current.

Embedded lattice mismatched stressors (LMS) in junction regions (SiGe for PMOS, SiC for NMOS), contact etch stop liners and stress memorization techniques (SMT) are intentional stress sources introduced to CMOS processes for this purpose. In addition to these intentional stress sources, there are also unintentional stress sources, such as shallow trench isolation in active silicon, and through-silicon via (TSV) in emerging 3D stacked processes.

In strained silicon technology, the magnitude of stress in the device channel, hence the exact amount of mobility improvement for a device depends on the process parameters of the technology, circuit layout parameters of the device and its neighboring features in the layout. Therefore, two devices with same gate length and width, which are assumed to have same drive strength in a regular Spice simulation, may exhibit considerable within die variations due to nonuniform distribution of stress through the layout. Introduction of strained silicon further increased the sources of process variations and impacted all aspects of circuit design.

Mechanical stress based variations are layout dependent hence they are not random but systematic type of variations. Therefore, efficient characterization and modeling of stress is needed to enable accurate analysis of VLSI circuits, to improve suboptimal circuit performance and

Direction	PMOS	NMOS
Longitudinal (X)	Compressive	Tensile
Lateral (Y)	Tensile	Tensile
Out-of Plane (Z)	Tensile	Compressive
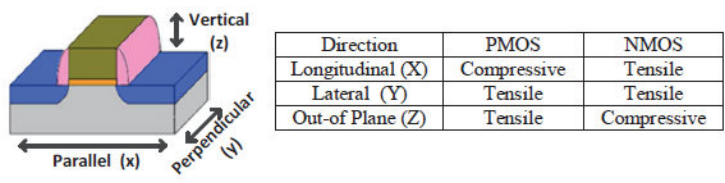
Figure 5.5. Desired stress types for CMOS transistors.
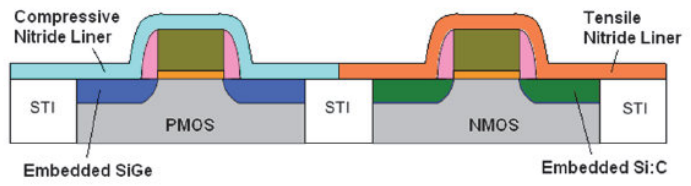
Figure 5.6. Mechanical stress sources in CMOS transistors.

to achieve higher gate density by mitigating uncertain variations and reducing design pessimism. Figure 5.5 summarizes the desired stress types for hole and electron mobility enhancement. Holes and electrons enjoy different type of stress on the channel direction and this introduces extra process steps to the CMOS fabrication.

Mechanical stress sources, used in state of the art CMOS technologies are summarized below and visualized in Figure 5.6. The local strain techniques have the following advantages:

- strain can be independently tailored to optimize performance enhancement for both n- and p-channel MOSFETs,
- the threshold voltage shift is smaller in uniaxially stressed MOSFETs,
- the stress memorization technique (SMT), and
- local stress techniques are cheaper and more compatible with standard CMOS technology.

Process induced stress has become one of the key performance boosters of modern MOSFET devices. All semiconductor manufacturers employ techniques such as the use of tensile nitride capping layers or com-pressive source-drain regions which lead to highly non-homogeneous

stress field within devices. From the TCAD device simulation point of view, handling non constant stress fields is a complex problem. Stress engineering poses significant new challenges for device simulation since important intrinsic material parameters such as mobility, band gap, or quantization phenomena are changed under strain. State-of-the-art full-band Monte Carlo simulators are used for stress simulation for constant strain fields throughout the device which is satisfactory for strained silicon films such as SiGe but does not help in case of process induced strains.

Performance enhancement contributors in each technology node, demonstrate that strain is more and more acting while the devices shrink. There is an increasing need for the use of strain while decreasing transistor dimensions to reach performance specifications. At a first approach, the piezoresistive theory allows one to explain the strain components contributing to effective mobility change for both PMOS and NMOS. From strain to mobility change is distinguished by device type (NMOS or PMOS), silicon orientation and channel direction. Thus to optimize performance, strain engineering need to be applied with the relevant components, keeping in mind that effects can be either cumulative or canceled.

According to the piezoresistive theory, the effective modification of the overall mobility depends on the combination of the whole stress components. The aim of the modeling is to evaluate separately stress components. This allows one to develop comprehensive design and process rules and to forecast transistor performance.

Stress Distribution

The direct effects of each CESL zones are responsible of the Si-channel stress state. Figure 5.7 shows the three direct effects corresponding to the three CESL zones. Figure 5.7 represents the impact of a single CESL zone on the resulting average stress in X, Y or Z direction. The top-CESL leads to compressive stresses in the channel for large gates and its impact decreases with the reduction of the gate width. Thus, no effect is found for narrow gate (L = 65nm). The lateral-CESL leads to com-pressive stresses in the channel also but its impact increases drastically when the gate width decreases. The bottom-CESL adds tensile stresses in
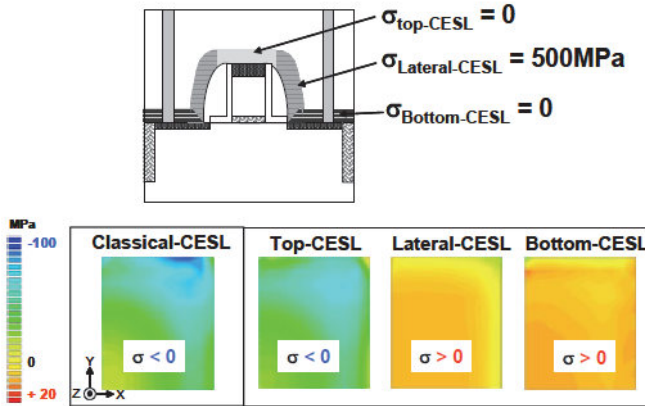
Figure 5.7. Stress distribution schematic representation of the 3 zones (top-CESL, lateral-CESL, bottom-CESL) making up of the whole CESL. After V. Fiori, How do Mechanics and Thermo mechanics affect microelectronic products: Some residual stress and strain effects, investigations and industrial management, PhD Thesis, L'institut national des sciences appliquées de Lyon, 2010.

the channel and its impact increases also drastically when the gate width decreases. Figure 5.8 shows the three indirect effects corresponding to the three CESL zones. For example, the indirect effect of the bottom-CESL applies a vertical force on the Si-channel by pulling on the top-CESL via the lateral-CESL.

By comparing the stress state induced by each of these three CESL zone to the classical CESL, we can conclude that:

- the classical CESL stress in X direction is mainly due to the direct effect of the bottom-CESL,
- the classical CESL stress in Y direction is small due to the contrary effect of each direct effect,
- the classical CESL stress in Z direction cannot be explained by any of the 3 direct effects.

Many possibilities exist for obtaining an engineered CESL. One solution for having a thickness engineered CESL consists in patterning said initial CESL by etching partially and locally the initial CESL. Indeed, the top zone or the bottom zone of the CESL can be removed thanks to an

anisotropic etch to obtain a 3D-patterned CESL. One solution for having an intrinsic stress engineered CESL consists in locally relaxing the stress of the initial CESL for example by germanium implantation. The stress in the channel can be controlled and so optimized by adjusting either the thickness or/and the intrinsic stress of the different CESL zones. Thus, the suitable combination of these three zones permits in particular to lead to the maximum electrical performance for NMOS and PMOS but also according to a wide range of layouts.

Optimization of CESL

The different impact of these three zones can be exploited to reach the best stress configuration that enables us to have the best performance for both NMOS and PMOS devices. In fact, the use of the classical CESL has 2 major drawbacks: In one hand, 2 different intrinsic stresses for the nitride layer must be used to optimize NMOS and PMOS at once: A compressive CESL is needed to optimize PMOS and a tensile CESL to optimize NMOS. So, the performances of NMOS and PMOS transistors cannot be enhanced using the same CESL. On the other hands, the stress map in the channel is affected by the layout. So, it is very hard to ensure the same performance to several transistors. Thus, a simple use of the CESL does not enable to fully manage the stress configuration in the Si channel.

One of the most important design parameters will be the minimum distance between nitride layers on the two gates. As the distance decreases, it becomes critical for the transmission of stress to the transistor. Figure 5.8 shows the geometry patterns of gates with the distance, in case of 65nm and 45nm technology nodes. The simultaneous reduction in the length and height of the gate nitride layers and width of spacers is not possible, resulting in poor stress transfer leading to decrease in expected device performance enhancement. It will be increasingly difficult to maintain a constant thickness for CESL as well.
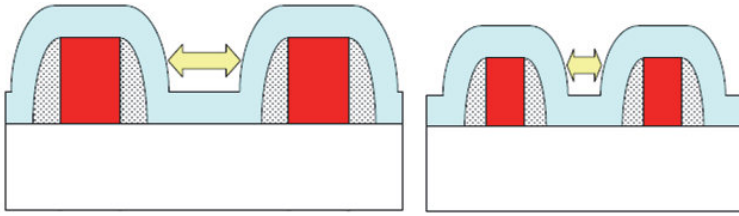
Figure 5.8. Schematic representation of the problem of reduced dimensions for the CESL, left for a 65nm and right for 45nm technology node.
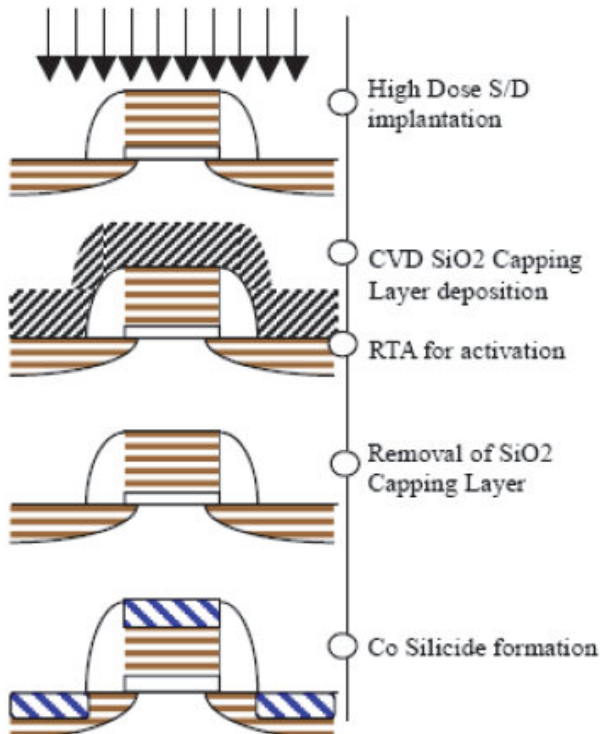
Figure 5.9. Typical process steps used in stress memorization technique.

Below 45nm technology node, the scaling of gate length will also be an issue. For a given technology node, enhancement in device performance gets saturated due to mobility degradation in short channel devices. The future of CESL therefore seems questionable in future technologies.

Stress Memorization Technique (SMT) has been proposed in 2002 to increase the performance of n-MOSFETs. Mechanism of stress transfer and sensitivity of SMT to transistor geometry is examined below. SMT technique is to create mechanical stresses in the gate polysilicon and keep them until the end of the process to enhance the transistor performance. Typical manufacturing process for SMT is shown in Figure 5.9. The performance enhancement in n- and P-MOSFETs is shown in Figure 5.10. For same I_{OFF}, while a 6% increase in performance of n-type transistors is observed with the SMT, a serious degradation occurs in p-MOSFETs.

To model the mechanisms of stress transfer for SMT in the transistor zone, one needs to consider many different phenomena occurring during annealing: materials, change of stress level, grain growth, and recrystallization of the amorphous regions. It has been proposed that the application of a mechanical stress in tension in the nitride layer (or oxide), which creates a compressive stress in the poly gate, as with the CESL layer. Stress is transmitted from the nitride layer to the channel, assuming a purely elastic behavior of different layers. Plastic deformation of the gate that allows one to create the stress.
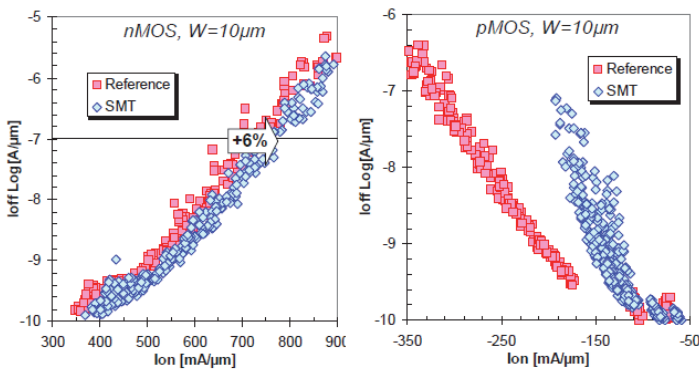
Figure 5.10. I_{ON}–I_{OFF} characteristics of transistors for same process for reference and the SMT (left n-MOS) and right (p-MOS).

5.3   Stress Evolution Modeling

A multidisciplinary effort is required for working on novel CMOS structures in the fields of (a) growth and fabrication of novel substrate materials, (b) device fabrication, (c) characterization, and (d) modeling. It will thus offer a complete perspective from fundamental characteristics of the new materials to device processing and performance. Most commercial process and stress simulators consider deposition not as a physical process govern by a time dependent model but rather a mathematical abstraction in which a new layer just appears on the top of the structure.

The one step approach cannot accurately predict stresses built in the deposited layer and in adjacent areas. In order to accurately predict stresses in thick layers the process should be considered as a series of deposition and relaxation steps to emulate mechanical quasi-equilibrium during this physical deposition process. Using "stress evolution" or "stress history" model implemented in VictoryStress more accurate simulation can be achieved. We use the stress evolution model available in Silvaco tools for simulation of stresses generated during thick layer deposition process and compare the results with standard one-step model. The test structure used in this example is a quarter of a 40-nm MOS transistor.

The simulation is performed in five following steps:

- Formation of a test structure with one layer nitride stressor in Athena
- Calculation of stress in the test structure using the one step model of VictoryStress
- Formation of a test structure with multi-layer stressor in ATHENA
- Calculation of stresses using stress evolution model of VictoryStress and "looping" capability of DeckBuild
- 2D visualization and average stress extraction for both models

The stress evolution simulation uses the same structure with 20 layers at each step while effect of new sub-layer deposition is emulated by
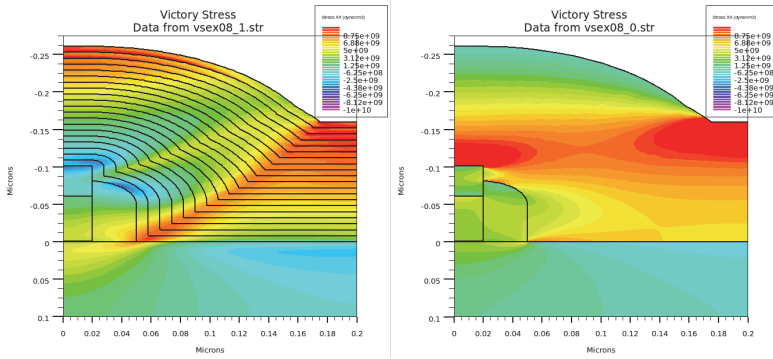
Figure 5.11. Stress distribution simulation using one-step and stress history models.

changing the status of this new sub-layer from "non-active" with properties inherited from "air" to "active" material region with properties inherited from nitride and the intrinsic stress of 1 GPa. Then VictoryStress recalculates stresses in the modified structure after each deposition step.

The simulation results are shown in Figure 5.11 which shows the Sxx stress for two models. In flat areas of deposited layer both models produced similar and essentially uniform stress. In the single layer case a non-uniform stress regions appear only near corners of spacer/nitride. However, in case of multi-layer stress history simulation a highly non-uniform stress field extends diagonally from the spacer/substrate corner to the top of the film. In other words, stresses are significantly higher where the layers have a sharp bent. As a result, the integrated stresses under the gate appear to be 1.4–2 times higher when the stress history method is used. In One-step model simulation the following parameters were used:

- material nitride intrin.sig = 1.0×10^{10}
- material silicon young.m = 1.67×10^{12}
- material silicon poiss.r = 0.28
- material poly young.m = 1.67×10^{12}
- material poly poiss.r = 0.28
- material oxide young.m = 1.0×10^{12}
- material oxide poiss.r = 0.25
- material aluminum young.m = 0.715×10^{12}

- material aluminum poiss.r = 0.35

Next simulation is essentially is an extension of the 2D modeling in 3D. More accurate simulation could be performed by using "stress evolution" or "stress history" model implemented in VictoryStress. The stress evolution simulation uses the same structure with 20 layers at each step while effect of new sub-layer deposition is emulated by changing the status of this new sub-layer from "non-active" with properties inherited from "air" to "active" material region with properties inherited from nitride and the intrinsic stress of 1 GPa. Then VictoryStress recalculates stresses in the modified structure after each deposition step.

The simulation is performed in five following steps:

- Formation of a test structure with one layer nitride stressor in VictoryProcess. The 3D structure created is shown in Figure 5.12.
- Calculation of stresses in the test structure using the one-step model of VictoryStress
- Formation of a test structure with multi-layer stressor in VictoryStress. The 3D structure created is shown in Figure 5.13.
- Calculation of stresses using stress evolution model of VictoryStress and "looping" capability of DeckBuild
- Figure 5.14 shows the 3D and 2D visualization of simulation results for both the models.
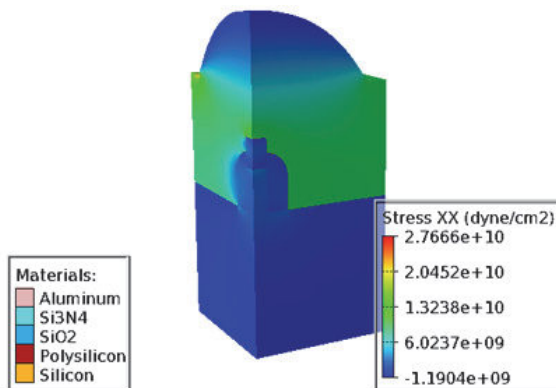
Figure 5.12. Formation of nitride stressor with one-layer model.

In Multi-step model simulation the following parameters were used:

- material silicon young.m = 1.67×10^{12}
- material silicon poiss.r = 0.28
- material poly young.m = 1.67×10^{12}
- material poly poiss.r = 0.28
- material oxide young.m = 1.0×10^{12}
- material oxide poiss.r = 0.25
- material aluminum young.m = 0.715×10^{12}
- material aluminum poiss.r = 0.35

The simulation results are shown in two 3D plots which compare Sxx stresses for two models (Figure 5.12). The Sxx stresses are also compared in two cut planes: Y–Z (vertical plane which roughly corresponds to 2D simulation) and X–Y (horizontal plane just 1nm under silicon surface). It is observed that the stress evolution model predicts higher stress level under the gate. This fact is even better illustrated in the second X–Y plane. The upper-left corner of the cut plane corresponds to the gate area (Figure 5.14).
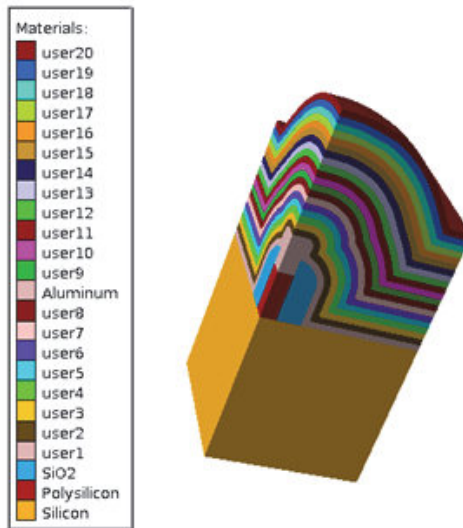
Figure 5.13. Formation of multi-layer stressor.
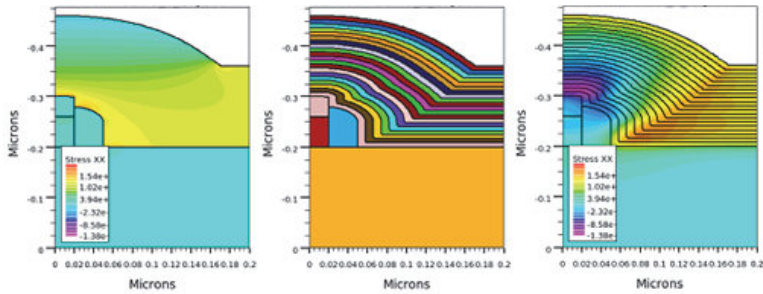
Figure 5.14. 2D and 3D visualization of simulation results (screenshots).

Summary

Currently, strained-Si is the performance booster for extending Moore's law. Local strain approach towards strained-Si engineering perhaps would be the simplest among the possible approaches. 3D TCAD process simulation of stress evolution provides valuable insights for technology development and stress management. Studies on stress engineering, performance and reliability trade-off are carried out for design and technology explorations. VictoryStress was used to study the stress evolution during thick layer deposition process and compare the results with standard one-step model for a 40-nm MOS transistor.

Chapter 6

Heterojunction Bipolar Transistors

The demand for communication systems worldwide has created an enormous market for semiconductors devices for various applications. While scaled CMOS is consolidating its stronghold in the analog and RF domains, the wide gamut of microwave frequencies is being competed for by the various types of III–V heterojunction based semiconductor devices. As silicon devices approach the limits of miniaturization and performance, the driving forces have been to make new devices and the key to the success has been the application of band gap engineering in silicon-based (group-IV) semiconductors. In this chapter, both silicon and III–V compound semiconductor devices are designed and optimized through calibrated 2D and 3D TCAD simulations. The design of an HBT with a collector-up topology, is carried out and optimized for maximizing the high frequency performance. The material system used here is $Ga_xIn_{1-x}P/GaAs$ (with $x = 0.51$ indicating lattice matched composition), which has relatively superior material properties and etching characteristics than the conventional $Al_xGa_{1-x}As/GaAs$ material system.

For the last few decades terahertz technology has raised a strong enthusiasm in the field internal security, especially for use in airport gates. Millimeter wave (30 to 300 GHz) applications have historically been reserved for markets niche as defense, security, aerospace and advanced telecommunications that could justify the high production costs required by semiconductor manufacturing III–V. By the nature of the materials, the frequency performance III–V transistors are unquestionably superior to those of silicon. But recent technological progress on silicon-germanium bipolar transistors make currently very good candidate for the design of circuits dedicated to terahertz applications.

In the field of automobile, radar sensors operating at 77 GHz for a long range and 122 GHz for short-range, enable the accurate assessment speed or distance, thus reducing the risk of collision. Furthermore, the integration on the same chip of different functions (digital, analog and radio frequency) played a decisive role in the development of microelectronics. This functional diversification, called "More than Moore" is one of the three tracks of the current research in the microelectronics sector.

To address these strategic areas, many European research projects are put in place since the 2000s for the development of effective technologies. One of the technologies identified is the Heterojunction Bipolar Transistor based on silicon-germanium whose performance continues to grow in recent European projects: Dotfive, DotSeven and RF2THZ. Based on its performance capabilities, low cost, and capacity for high-integration, silicon-germanium (SiGe) heterojunction bipolar transistor (HBT) BiCMOS technology has established itself as strong technology contender for a host of circuit applications including analog, mixed signal, RF and millimeter-wave. However, as operating frequencies for wireless applications are pushed upward in the spectrum, SiGe HBT technologies face significant challenges at the transistor level as operating voltage limits decrease and performance requirements increase. The BiCMOS technology that combines production, on the same wafer, CMOS transistors for digital/analog functions and bipolar transistors for RF functions makes possible the realization of systems on a single chip with high performance and energy efficiency. Figure 6.1 shows the advancements made in SiGe HBTs during 2002-2013 in terms of maximum frequency of oscillation, f_{max} as a function cutoff frequency, f_T. As III–V technologies have frequencies substantially higher to those of silicon technologies, we have included various III–V technologies in Figure 6.2.
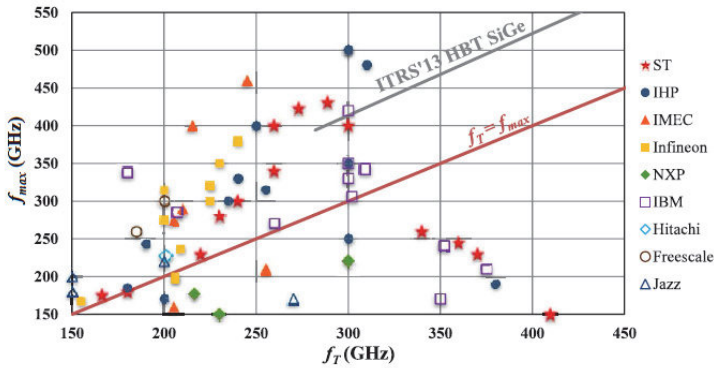
Figure 6.1. State-of-the-art performance of the maximum frequency of oscillation, f_{max} as a function cutoff frequency, f_T of SiGe heterojunction bipolar transistors from major industrial manufacturers and research institutes during 2002-2013. After M. Deng, Contribution a la caracterisation et la modélisation jusqu'à 325 GHz de transistors HBT des technologies BiCMOS, PhD Thesis, Universite de Lille, 2014.
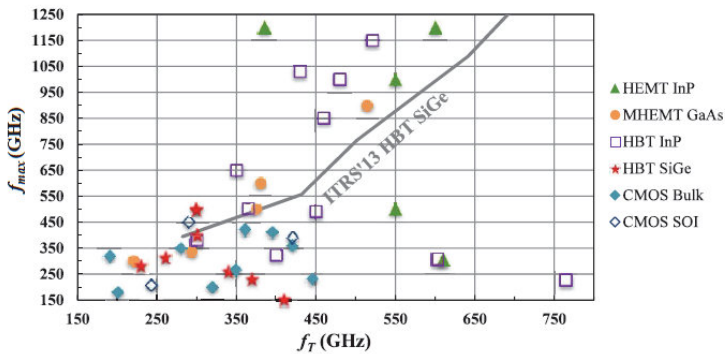
Figure 6.2. State-of-the-art performance of the maximum frequency of oscillation, f_{max} as a function cutoff frequency, f_T for silicon and III–V transistors during 2004–2012. After M. Deng, Contribution a la caracterisation et la modélisation jusqu'à 325 GHz de transistors HBT des technologies BiCMOS, PhD Thesis, Universite de Lille, 2014.
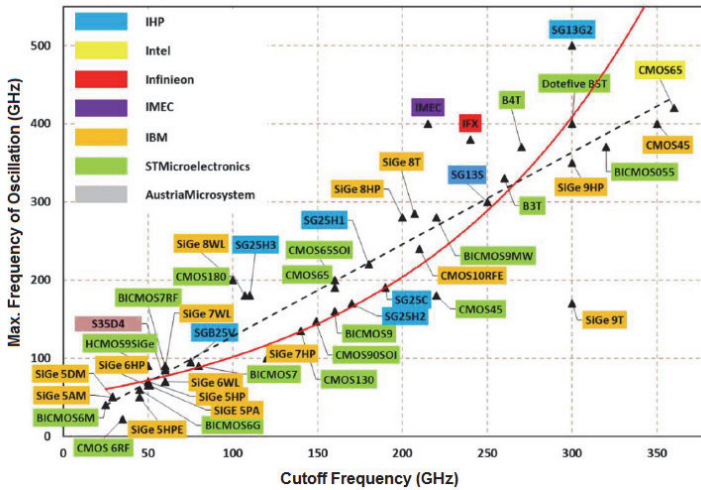
Figure 6.3. Evolution in enhancement in the characteristic frequencies depending on different technology nodes from foundries and research institutes. After A. Serhan, Conception et réalisation de fonctions millimétriques en technologie BiCMOS 55nm, PhD Thesis, Universite Grenoble, 2015.

In Europe, SiGe BiCMOS technologies are the most advanced on the 130nm node and below with CMOS technology. For example, IHP Microelectronics announced a f_T/f_{max} = 300/500 GHz for their SG13G2 technology and STMicroelectronics a f_T/f_{max} = 300/400 GHz for their B5T technology. IBM announced the introduction of 9HP technology SiGe BiCMOS with a couple f_T/f_{max} 300/420 GHz on the node 90nm CMOS technology, thus further increasing the degree of integration. Figure 6.3 shows the evolution in the advancements in the characteristic frequencies depending on different technology nodes from foundries and research institutes. The black triangles represent couples of each technology. The color code is used to identify the foundry. The data includes BiCMOS technologies that date back to 1993 (such as technology S35D4 of AMS), and very recent technologies (such as B55 BiCMOS technology from STMicroelectronics and technology SG13G2 from IHP. The 55nm BiCMOS technology (B55) is a technology developed by STMicroelectronics in RF2THZ project will provide a platform silicon technology for the development of systems Radio Frequency (RF) , millimeter (mm Wave) and Terahertz (THz), such as automotive radar 77 GHz/120 GHz, imaging and detection for biomedical and military applications.

In general, the reduction of the emitter width and neutral base thickness in bipolar transistors give rise to an increase in the maximum frequency of oscillation and cutoff frequency. These frequencies also increase with the doping of the collector region by reducing the resistance of the neutral collector zone. Figure 6.4 shows the evolution of the transition frequency as a function of the width emitter. However, the reduction is accompanied by a reduction of the breakdown voltage of the transistor. Figure 6.5 shows the best performances for each technology wherein the transistor is biased at an optimum current density corresponding at maximum f_T.

A fundamental relationship between the transition frequency and the maximum voltage (for breakdown) which may be applied between the collector and the emitter of the bipolar transistor is known as the Johnson limit. Figure 6.6 shows the avalanche breakdown voltage as a function of the transition frequency of SiGe bipolar transistors in advanced technologies. This study shows that increasing the transition frequency is associated with the reduction of the size of the transistor which imposes a challenge for the breakdown voltage. One can notice that for low transition frequencies, technologies have worked well below the Johnson limit. This is no longer valid for millimeter wave technology which requires more than 250 GHz transition frequency. This decrease in breakdown voltage is accompanied by a reduction of the supply voltage which must not exceed breakdown voltage to ensure the reliability of the transistor.

Regarding the MOS transistors of the maximum drain voltage relative to the voltage power decreases V_d with the gate length L_g to meet the criteria of reliability of scaled transistors. Figure 6.7 shows the nominal supply voltage V_d dependence of the gate width and minimum gate length for CMOS RF technologies. Indeed, the breakdown phenomenon of transistors of the gate oxide can occur when high voltages are applied to the gate and the drain. In addition, the strong electric field between the drain and the source may result in premature aging of transistors by injection of hot carrier in the gate.
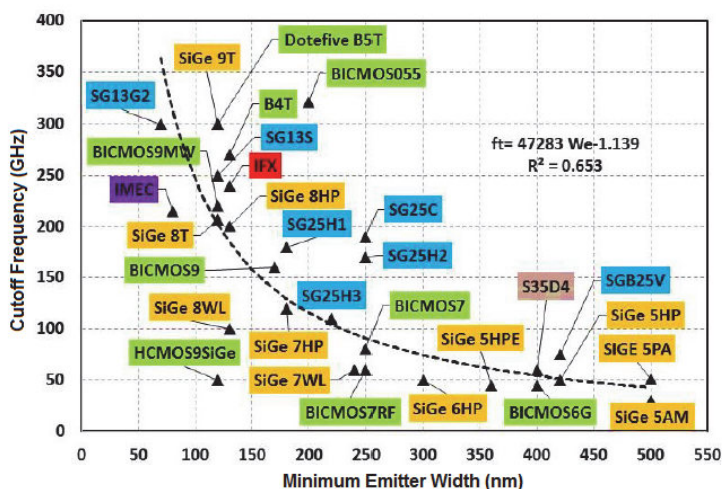
Figure 6.4. Evolution of the transition frequency as a function of the emitter width. After A. Serhan, Conception et réalisation de fonctions millimétriques en technologie BiCMOS 55nm, PhD Thesis, Universite Grenoble, 2015.
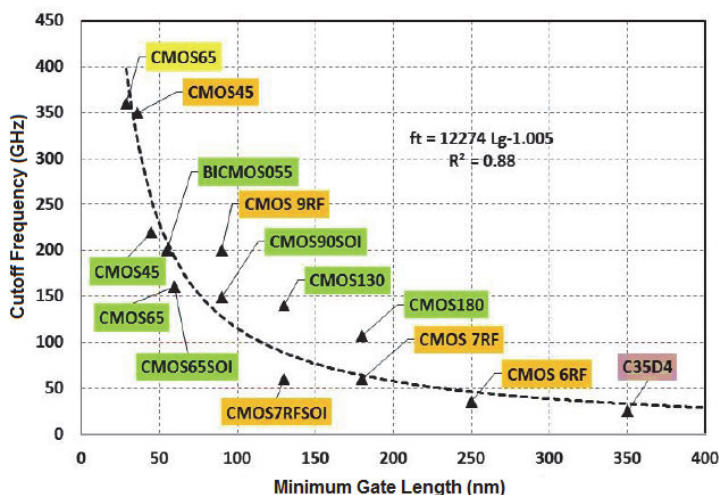
Figure 6.5. Evolution of the transition frequency as a function of the width and minimum gate length. After A. Serhan, Conception et réalisation de fonctions millimétriques en technologie BiCMOS 55nm, PhD Thesis, Universite Grenoble, 2015.
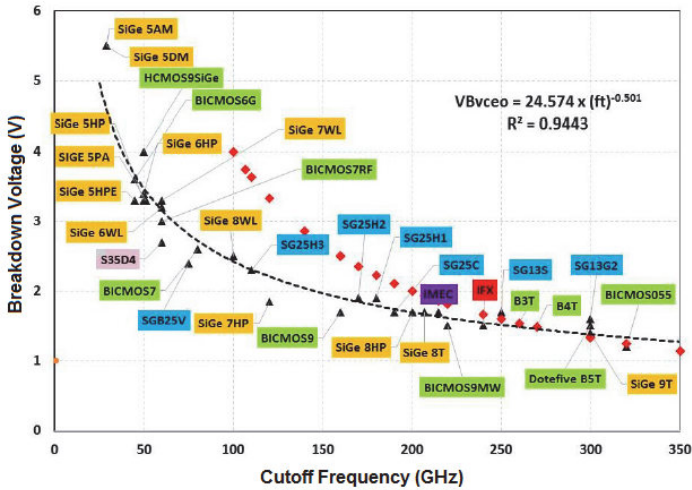
Figure 6.6. Breakdown voltage as a function of the transition frequency of SiGe bipolar transistors in advanced technologies. After A. Serhan, Conception et réalisation de fonctions millimétriques en technologie BiCMOS 55nm, PhD Thesis, Universite Grenoble, 2015.
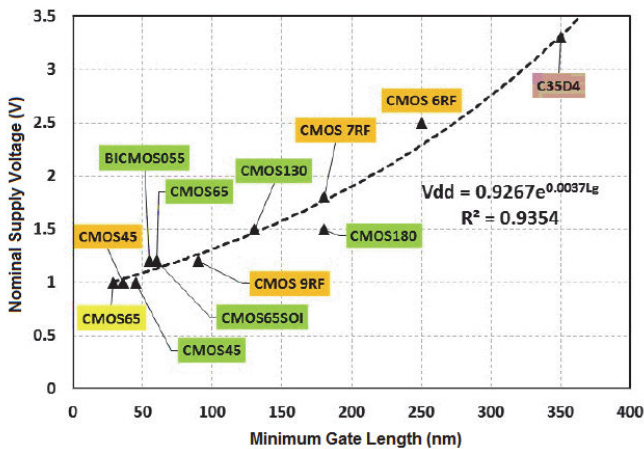
Figure 6.7. Evolution of the nominal supply voltage V_{dd} depending on the gate width and minimum gate length for CMOS RF technologies. After A. Serhan, Conception et réalisation de fonctions millimétriques en technologie BiCMOS 55nm, PhD Thesis, Universite Grenoble, 2015.

The work by Kroemer on the heterojunction opened a new door for achieving higher speeds than anticipated with silicon bipolar transistors. Although the performance advantages of HBTs over BJTs were well understood, no fabrication technologies were available to produce high quality heterojunction until the 1970s, where epitaxial growth of SiGe heterostructures demonstrated by Erich Kasper using molecular beam epitaxy (MBE). With the advent of heteroepitaxy, the concept of strained layers has been extended to include other elemental semiconductors. These developments set the stage for IBM's development of SiGe HBTs in 1987 using MBE.

The successful demonstrations of SiGe HBT technology, in both high-performance digital and analogue circuit applications, are the results of over 15 years of steady research progress from initial material preparations in 1984, through device demonstrations from 1987–1992 to large scale circuit fabrication in 1994 and commercial products in 1998. Since then, the interest in SiGe HBTs from both academia and industry rose progressively through the years causing the figures of merit to increase and reach current state-of-the-art levels. Today's state-of-the-art SiGe pre-production technologies have shown cutoff frequency close to 500 GHz and maximum frequency of oscillation close to 600 GHz as reported recently.

The surging interest in Silicon Germanium (SiGe) devices for use in communications applications has spurred demand for SiGe semi-conductor. As bandwidth needs are ever increasing and wireless is requiring higher frequency parts, the faster is always the better. Recently, the level of activity in SiGe has grown dramatically in parallel with the explosion of the wired and wireless communications IC market. For high power (current) applications, the SiGe HBT can easily achieve in excess of ~2 mA/mm^2 of emitter area, with almost perfect ideality and flat gain over 7 orders of magnitude. The SiGe HBT offers various advantages at lower frequencies in today's wireless applications (e.g., 1.8 or 2.4 GHz). The SiGe HBT technology is used for a variety of high-frequency applications. Furthermore, the higher speed may be traded for low power and a lower noise floor. The key advantages of SiGe HBT technology can easily be extended to achieve lower power dissipation with better noise and reliability than in existing solutions. In addition, SiGe-enhanced BiCMOS is available for highly integrated "system-on-a-chip"

solutions. In this chapter we shall discuss briefly the basic SiGe-HBT device physics and provide an overview of the most common SiGe HBT designs. Detail on SiGe-HBT design is available in reference [1.1]. In the following, the design of SiGeC-HBTs in SiGeC materials platform will be covered.

6.1 SiGeC-HBTs

The first transistor invented in 1947 by Bardeen, Brattain, and Shockley at Bell Laboratories was a germanium-only device. Silicon came into use for solid state devices a few years later, when crystal growing and processing issues with Si could be addressed. Silicon and germanium are completely miscible over the full range of compositions and hence can be combined to form $Si_{1-x}Ge_x$ alloys with the germanium content x ranging from 0 to 1 (0–100%). The equilibrium lattice parameters for Si and Ge are 5.431Å and 5.646Å, respectively. This corresponds to a lattice mismatch to Si of ~4.1%. When an alloy of Si and Ge is formed the equilibrium lattice parameter can roughly be linearly interpolated between Si and Ge, though there is a slight negative deviation from Vegard's law. In 1989, Meyerson and his team at IBM developed the low temperature epitaxy technique to deposit a germanium layer on silicon using ultrahigh vacuum chemical vapor deposition. They fabricated a graded heterojunction an interface between two regions of silicon and Silicon-Germanium alloy layer on top of silicon. The resulting heterojunction was the first step towards SiGe heterojunction bipolar transistors.

When $Si_{1-x}Ge_x$ layer is grown on a silicon substrate, the lattice mismatch at the interface between the $Si_{1-x}Ge_x$ and the silicon has to be accommodated. This can either be done by compression of the $Si_{1-x}Ge_x$ layer so that it fits to the silicon lattice or by the creation of misfit dislocations at the interface. These two possibilities are illustrated schematically in Figure 6.8a. In the former case, the $Si_{1-x}Ge_x$ layer adopts the silicon lattice spacing in the plane of the growth and hence the normally cubic $Si_{1-x}Ge_x$ crystal is distorted (Figure 6.8b). When $Si_{1-x}Ge_x$ growth occurs in this way, the $Si_{1-x}Ge_x$ layer is under compressive strain and the layer is described as pseudomorphic. In the latter case, the $Si_{1-x}Ge_x$ layer is unstrained, or relaxed, and the lattice mismatch at the

interface is accommodated by the formation of misfit dislocations (Figure 6.8c).
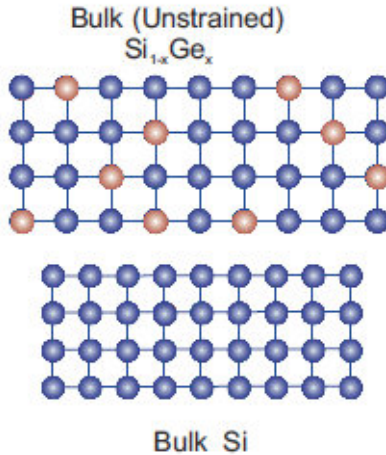
Figure 6.8a. Bulk silicon and unstrained $Si_{1-x}Ge_x$.
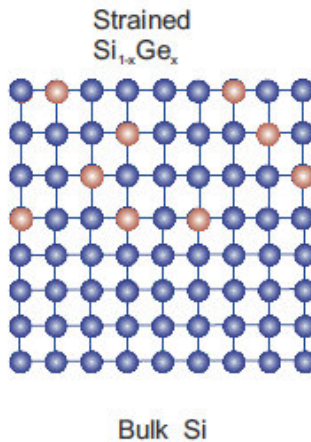
Figure 6.8b. Strained $Si_{1-x}Ge_x$: When the SiGe layer is thin, SiGe is commensurately deposited on the silicon substrate and under compressive strain, with the in-plane lattice constant the same for both materials.
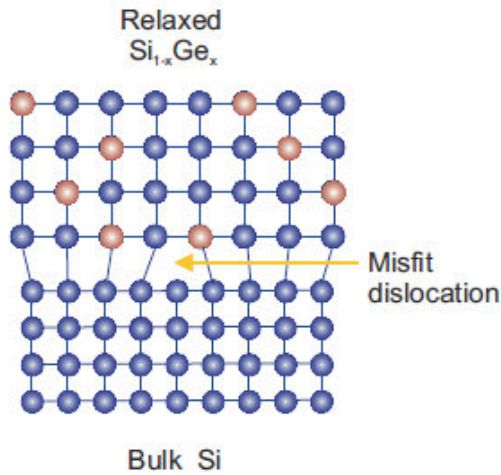
Figure 6.8c. Relaxed Si$_{1-x}$Ge$_x$: When the SiGe layer is thick, SiGe is relaxed by misfit dislocations at the SiGe/Si interface.
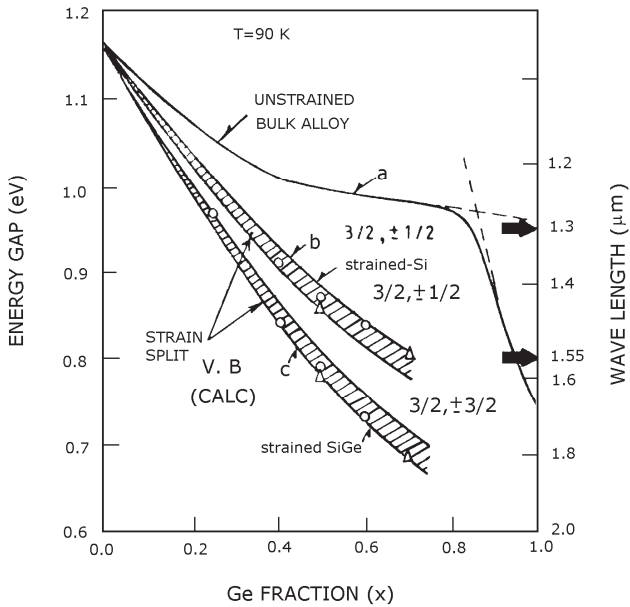
Figure 6.9. Bandgap as a function of Ge content for strained and unstrained Si$_{1-x}$Ge$_x$.

As might be expected, there is a maximum thickness of the $Si_{1-x}Ge_x$ layer that can be grown before relaxation of the strain occurs through the formation of misfit dislocations. This thickness is known as the critical thickness of the $Si_{1-x}Ge_x$ layer, and depends strongly on the germanium content. $Si_{1-x}Ge_x$ alloys have a smaller band gap than silicon partly because of the larger lattice constant and partly because of the strain. Figure 6.9 shows the energy band gap difference between SiGe alloys and Si as a function of the Ge content in the alloy. It can be seen that strain makes a large contribution to the band gap reduction in SiGe.

The essential difference between the SiGe HBT and the Si BJT is best illustrated by considering a schematic energy band diagram (Figure 6.10). For simplicity we consider an ideal, graded base SiGe HBT with constant doping in the emitter, base, and collector regions. The Ge content is linearly graded from 0% near the metallurgical emitter-base (EB) junction to some maximum value of Ge content near the metallurgical collector-base (CB) junction, and then rapidly ramped back down to 0% Ge. The resultant overlaid energy band diagrams for both the SiGe HBT and a Si BJT with the same doping profile, biased identically in forward-active mode, are shown in Figure 6.10.

The incorporation of SiGe layers into Si-based systems provides the possibility to enhance the performance of Si based electronic devices by improving strain-originated carrier transport characteristics and introducing quasi-electric fields from the band gap grading. The insertion of Ge in the epitaxial base of a silicon bipolar transistor produces higher cutoff frequencies, simultaneously reducing noise and power dissipation with the added advantages of enhanced gain and Early voltage. The effect of the bandgap discontinuity in the emitter-base junction was described by Kroemer who showed that if the bandgap in the emitter was wider than that of the base, the minority carriers moving from emitter to the base will see a lower barrier than the minority carriers injected from the base to the emitter, inducing an exponential augmentation in the current gain. Revolutionary nature of SiGe technology has been demonstrated when compared to traditional Si bipolar performance.
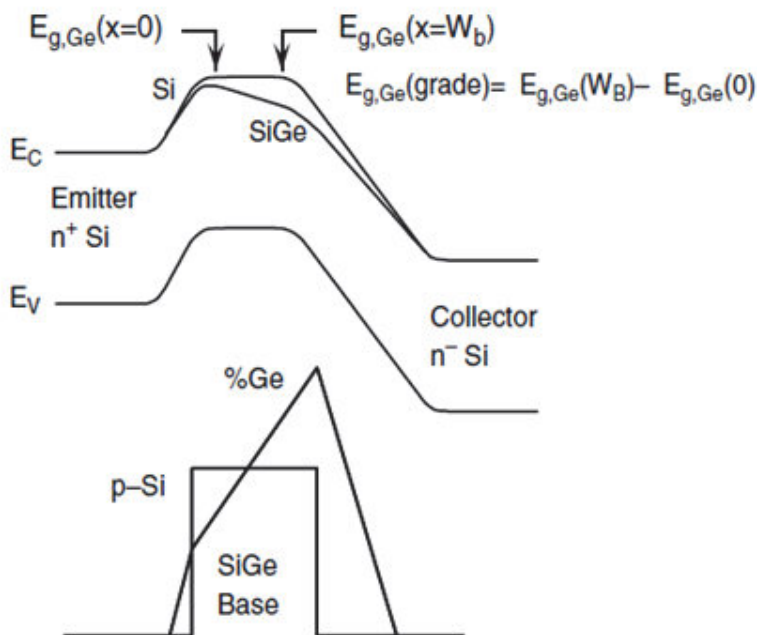
Figure 6.10. Energy band diagram for a graded base SiGe HBT and a Si BJT.

The successful demonstration of SiGe-HBT technology, in both high performance digital and analog (including terahertz frequency) circuit applications are the results of over 25 years of steady research progress from 1984, through device demonstrations during 1987–1992, to circuit fabrication in 1994, and commercial products in 1998 [1.1]. Meanwhile in an effort to improve single chip functionality, despite increased process complexity, BiCMOS processes have been developed to combine the advantages of complementary metal oxide semiconductor (CMOS) and bipolar devices. The key feature of SiGe is the compatibility of SiGe HBT technology with conventional silicon bipolar or BiCMOS processing. In the laboratory cutoff frequencies of up to 650 GHz have been achieved. Using a typical bipolar process as a starting point, the performance enhancement is routinely achieved in either bipolar or BiCMOS processes via advanced device design.

The principle of operation of an HBT is identical to that of the BJT, except that the bandgap of the emitter region exceeds that of the base region by ΔE_g, typically of the order of 0.1–0.2eV. The resultant exponential increase in current gain ($\propto e^{\Delta E_g/kT}$) permits scaling down of the base region to smaller thicknesses (increasing f_T) and higher doping levels to keep the base resistance low (increasing f_{max}). SiGe-HBTs are particularly exciting because of their ability to take immediate advantage of highly developed silicon processing techniques.

Appropriate scaling of the emitter, base and collector regions and optimization of their associated doping profiles are crucial for the design of a SiGe-HBT, for a particular technology generation. A SiGe-HBT offers additional design flexibility in that the bandgap of the base may be tailored by grading the Ge concentration. In spite of increased process complexity, incorporation of Ge into base offers more degree of freedom that relaxes a series of trade-offs in device design. Several advantages of SiGe-HBTs over conventional Si BJTs include: reduction in base transit time, increase in collector current density (hence more current gain), low intrinsic base resistance, and increase in Early voltage.
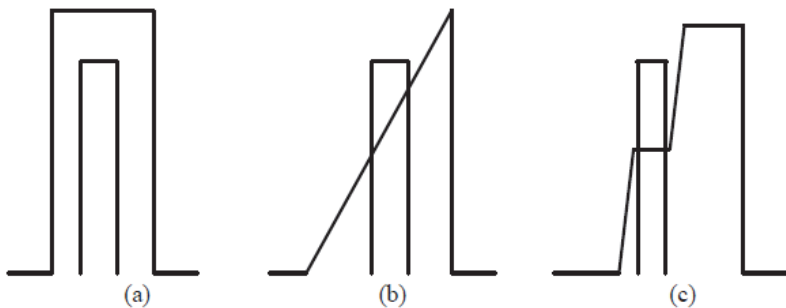
Figure 6.11. Schematic of (a) Box-in-Box, (b) Graded, and (c) Modified Graded base layer designs.

The layer structures for the base region depend on the base design desired. Commonly, there are two SiGe base designs employed, the so-called "Box-in-Box" and "Graded" structures. The graded structure has a couple of different variants. Figure 6.11 shows schematics of three different base design schemes along with the band diagram. The Box-in-Box structure shown in Figure 6.11a is made up of a thin layer of constant boron doping located within a layer of $Si_{1-x}Ge_x$ with constant germanium composition. The thickness of the intrinsic layers of SiGe on either side of the base region is critical to ensure that the base remains inside the germanium region. The Graded structure shown in Figure 6.11b is made up of a graded SiGe layer from a region of high germanium content to a region of no germanium content as the layer is deposited. The base is located within this graded region. One variation shown in Figure 6.11c is to have a graded region from high germanium content to a flat region of medium content and then another graded region to no germanium content. The base region is located in the flat region with medium germanium content. The purpose of this design is to allow any shifting of the e-b junction that might happen, to occur within a region of constant germanium, thus minimizing the change in gain. These graded designs give a band structure that forms a drift region that increases the electron velocity in the base region and will give better high frequency performance due to the reduced transit time. For the base design we chose the "Box-in-Box" structure and Figure 6.12 shows the depth profile.
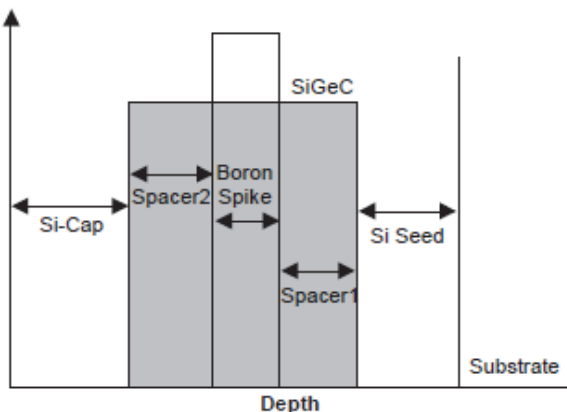
Figure 6.12. Schematic of typical HBT base layer.

While the strong growth of the semiconductor industry over the past decades has been primarily driven by the continual downscaling of CMOS devices, the introduction and optimization of the SiGe HBT has been a major driver of the wireless industry. For the wireless industry, the development of SiGeC HBTs becomes even more important. The most common HBT designs will now be discussed with the focus being the process integration issues and the advantages and disadvantages of each design. There are two main HBT designs that have demonstrated excellent high frequency performance and they differ in the technique used for integrating the base deposition. A brief overview of these designs is presented with only the major processing steps highlighted.

The first HBT design is the so-called selectively epitaxial growth (SEG) base scheme which has demonstrated a balanced f_T and f_{max} values of 155 and 167 GHz, respectively. The final device structure is shown in Figure 6.13a. The most complicated part of this integration sequence is the formation of the selectively grown base region. This involves a careful optimization of the opening formation and the epitaxy growth conditions. The SEG extrinsic base scheme reduces the base resistance by both
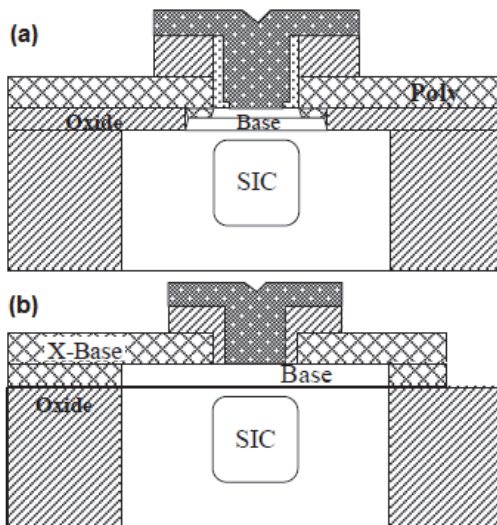
Figure 6.13. Schematic cross section showing a basic (a) SEG base HBT and (b) NSEG base HBT. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.

minimizing the extrinsic base resistance and by avoiding the extrinsic base implantation, which allows the spacer between the emitter and extrinsic base region to be reduced. The second main HBT design is the non-selectively (NSEG) deposited base integration scheme. This design has shown even higher performance than the SEG base design with f_T and f_{max} values as high as 350 GHz and 338 GHz, respectively for differently optimized devices. This HBT design has several different sub-schemes that can give either a self-aligned or non-self-aligned structure, which is a trade-off between complexity and performance. Figure 6.13b shows the final device cross section from NSEG process.

Physical device simulation is of great importance in evaluating and developing an understanding of how a device functions. The system of differential equations comes from solving the Boltzmann transport equation and there are two main combinations of equations used. The first is the drift diffusion approximation and it is made up of three equations, the continuity equations for electrons and holes along with the Poisson equation for the potential. As device dimensions continue to shrink a more detailed approximation is required. This is known as the hydrodynamic approximation and it includes the same three equations of the drift diffusion approximation along with the local carrier temperatures for the holes and electrons. This yields five equations and simulations that employ the hydrodynamic approximation will take more time to obtain a solution. For studying the self-heating effects, we consider non-isothermal device simulations and for these simulations the lattice heat flow equation is also solved. The non-isothermal device simulations will have six total equations and take significantly longer to converge and obtain a solution.

All of the simulations in this chapter are two-dimensional in nature so as to obtain as accurate of a solution as possible in a reasonable amount of time. As will be shown for the non-isothermal simulations, the 2D simulations are not completely sufficient to describe the self-heating effects, but performing full 3D device simulations with the hydrodynamic approximation and lattice heat flow is not realistic considering the large number of mesh points required in the base region and the computing resources available. ISE-TCAD DESSIS device simulator was employed for all simulations. The Philips Unified Mobility model was utilized for all simulations. This model describes the

temperature dependence of the mobility and also accounts for screening of ionized impurities by charge carriers, electron-hole scattering, and the clustering of impurities. The hydrodynamic model was employed for the high field saturation modeling and the Slotboom bandgap narrowing expression was also used in simulation.
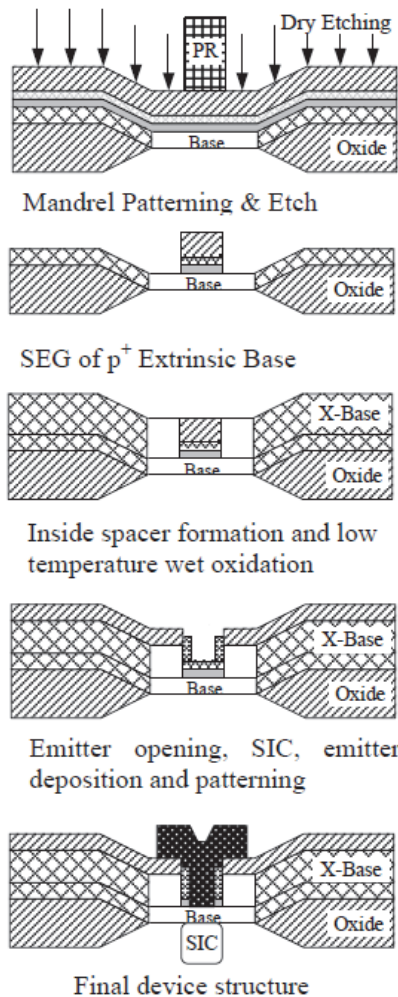
Figure 6.14. Schematic cross sections of the SiGeC-HBT showing the raised extrinsic base design. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.

In the following, we describe a low complexity process flow for the development of SiGeC HBT in the laboratory and perform process and device simulation. Figure 6.14 shows a series of schematic cross sections showing the main process steps followed for the fabrication. Device simulations are performed for the optimization of the selectively implanted collector (SIC). The selectively implanted collector is used to increase the f_T and current drive of a bipolar transistor without significantly increasing the collector-base capacitance. The SEG base design and SIC integration is shown in Figure 6.15. In this design the SIC can either be implanted before or after the selective base deposition, but it is usually done after the base poly and emitter opening have been defined. The other main HBT design is the NSEG base HBT and the integration sequence for this design is shown in Figure 6.16. The effect of the emitter, spacer and SIC widths on the DC and AC performance of the HBT are examined.
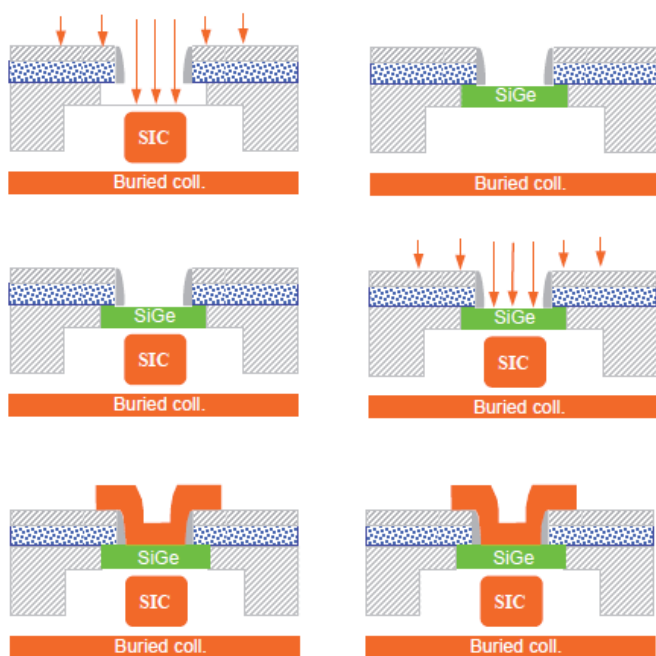
Figure 6.15. SIC integration possibilities in a SEG base HBT design. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.
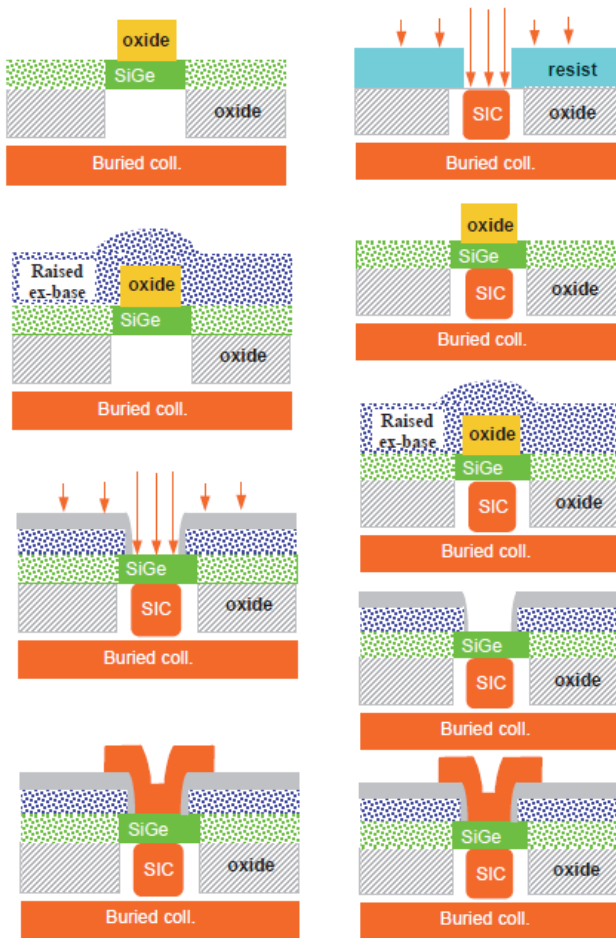
Figure 6.16. SIC integration possibilities in a NSEG base HBT design. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.

As lateral device dimensions continue to shrink the lateral straggle from the implantation will also become more critical to take into consideration when optimizing the SIC. Monte Carlo simulations were performed to further quantify this issue. For identical implantation conditions, energy of 410 keV and a dose of 1×10^{14} cm^{-2}, the emitter opening size was varied from 0.2μm to 0.05μm to see the effect of the lateral straggle. Figure 6.17 shows this comparison. It can be seen that on a scale of

1×10^{17} to 1×10^{19} atoms/cm^3 the SIC is twice as large as the opening for the 0.2μm emitter window and is six times as large for the 0.05μm emitter opening. This demonstrates that as the emitter widths continue to shrink to obtain higher performance devices the SIC process conditions become more critical. A 2D device simulation was constructed within ISE-TCAD with approximately 2000 mesh points modeling the active device region. Three main variables were identified that had the potential to affect the device performance of the HBT. Figure 6.18 shows the full device with mesh points and a zoomed in region identifying the three variables that were varied in the simulations. They are the SIC width, the spacer width, and the emitter width give the best high frequency device performance. It is found that the emitter width had a strong effect on current gain, f_T and f_{max} with all three values increasing as the emitter width was shrunk. By increasing the SIC width the f_T and current gain are found to increase while f_{max} decreased which makes sense when considering the electron current flow and Collector-base capacitance, respectively. The fact that the SIC width can be used to tune the balance between f_T and f_{max} can be seen in Figure 6.19.
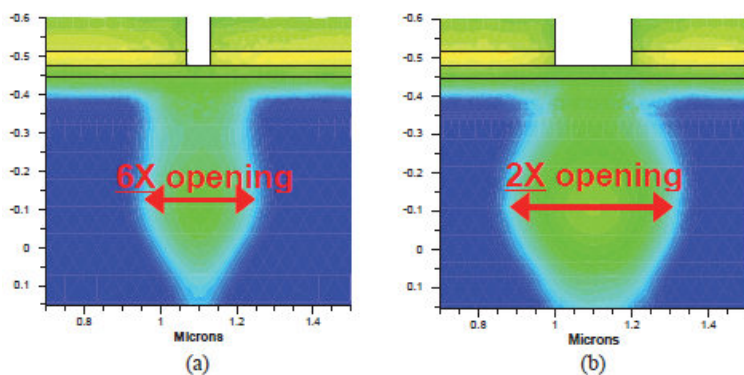
Figure 6.17. Monte Carlo simulations comparing lateral straggle for different emitter widths, (a) 0.05μm and (b) 0.2μm. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.
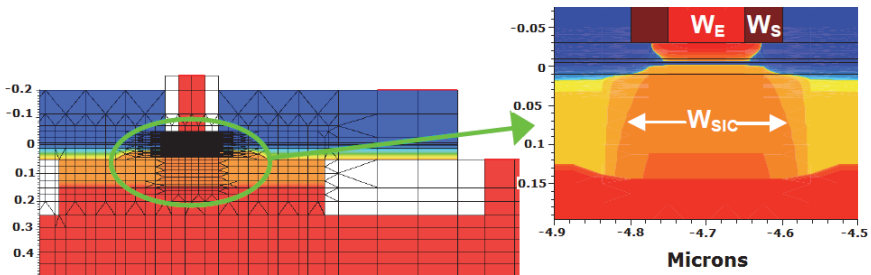
Figure 6.18. Definition of the lateral device dimensions analyzed. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.
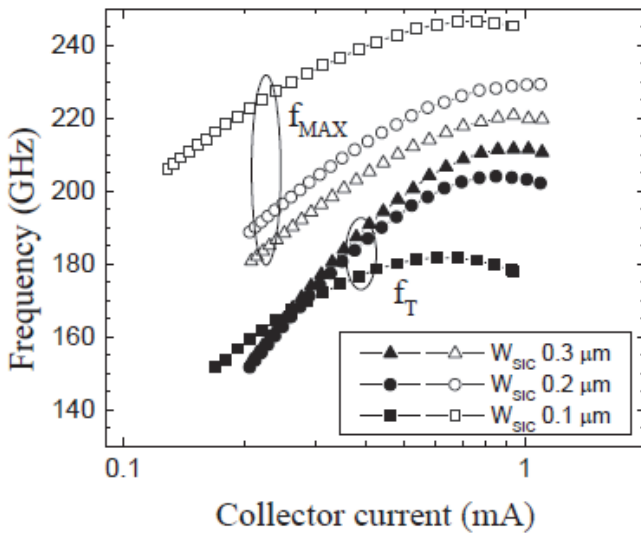
Figure 6.19. f_T and f_{max} vs. I_C for different SIC widths. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.

6.2 Self-heating Effects

In the past self-heating effects have mostly been of concern for high power devices, but as high speed bipolar transistors have continued to be

scaled to obtain higher speeds their current density has also increased, leading to the need for an accurate determination of their junction temperature. Also the decreasing breakdown voltage in aggressively scaled SiGe HBTs has important implications for high frequency large-signal performance in terms of the 1-dB compression point, saturated power, and the power added efficiency. Understanding safe operating limits is important for circuit design, especially the high power circuits. It is also important to be able to perform detailed device simulations that take into consideration the effect of self-heating in order to reduce the development time for future transistor architectures. As SiGe HBT BiCMOS technologies scale for higher frequency performance, circuit designers striving to maintain adequate output power and gain performance must pay careful attention to the complex and increasingly rigorous upper voltage limits for reliable device biasing and dynamic operation – the so-called "safe-operating area".

The fabrication of bipolar transistors on silicon on insulator offers several benefits such as the lowering of the collector-to-substrate capacitance that is important to realizing high speed circuit performance. Using SOI also leads to a higher resistance to substrate noise, lower device parasitic, and improved radiation hardness. In addition, as CMOS continues to scale and SOI is used to a larger extent, the ability to implement the bipolar part of a BiCMOS process on SOI becomes critical. But, while there are several advantages of fabricating bipolar transistors on SOI, there is also one key drawback. This drawback is the well-known effect called self-heating where the heat generated at the base/collector junction must be conducted away from the active device area or else the device will be operating at elevated temperatures that can have an adverse effect on its performance as well as reliability.

Self-heating effects for SiGeC HBTs was studied using the ISE-DESSIS TCAD simulator through 2D and 3D simulations. Figure 6.20 shows the 2D temperature profiles for the two cases with identical isolation structures of STI+DTI+SOI and at identical bias conditions. Electro-thermal simulations were performed with two different boundary conditions at the surface of the vias, an adiabatic boundary and a thermal resistance boundary. As expected the simulation with and adiabatic boundary condition at via surface shows a significantly higher temperature distribution throughout the entire structure. The maximum

temperature at the emitter-base junction is also observed to be approximately 8°C higher for the adiabatic boundary simulation. It is also seen that the more aggressive the isolation structure the more impact the heat transfer through the metallization has. The conclusion is that it is important to include heat transfer through the metallization for device simulations. It is shown that self-heating in SiGe HBTs is very high compared to identical Si BJTs and thus significant electrical performance degradation takes place in SiGe HBTs.
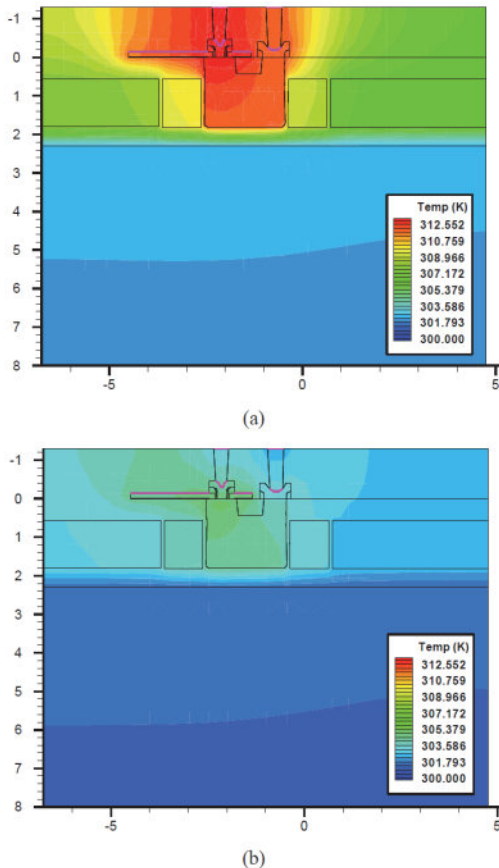
Figure 6.20. 2D cross-sections showing the thermal contours for two different boundary conditions at via surface. (a) Adiabatic and (b) Thermal resistance. After E. Haralson, Device design and process integration for SiGeC and Si/SOI bipolar transistors, PhD Thesis, Royal Institute of Technology, 2004.

6.3 InGaP/GaAs HBT Technology

While scaled CMOS has consolidated its stronghold in the analog and RF, III–V heterojunction based semiconductor devices are dominating in the microwave applications. The InGaP/GaAs III–V material system is emerging as a viable alternative to the conventional AlGaAs/GaAs system owing to the superior electronic and physical properties of InGaP over AlGaAs. Due to absence of Al, the material quality of InGaP is highly improved over that of AlGaAs. The band offsets in InGaP/GaAs HBT technology are more conducive to provide greater emitter injection efficiency. InGaP/GaAs HBTs also exhibit better reliability compared to their AlGaAs/GaAs counterparts. Among these devices, the quest for faster and more powerful transistors has led researchers to investigate innovative topologies. The InGaP/GaAs III–V material system has emerged as a viable alternative to the conventional AlGaAs/GaAs system owing to the superior electronic and physical properties of InGaP over AlGaAs. The following are some advantages of InGaP/GaAs over AlGaAs/GaAs: (a) due to absence of Al, the material quality of InGaP is highly improved over that of AlGaAs. It does not have deep traps and DX centers, which have been a problem in AlGaAs due to its higher reactivity with oxygen, (b) the band offsets in InGaP/GaAs HBT technology are more conducive to providing greater emitter injection efficiency, (c) InGaP does not allow Carbon to form acceptor states like it does in AlGaAs, and (d) probably the most important advantage of InGaP over AlGaAs lies in the etch selectivity of InGaP and GaAs layers with respect to each other.

The Collector-up configuration has some significant benefits that make it attractive for high frequency applications. Since the days of initial design of the first transistor, efficient charge collection was a prime concern for designers that made it imperative to have a larger collector than emitter area. However, this resulted in a significant compromise in high frequency performance of the device owing to the large junction capacitance associated with the larger collector area. In the following, the design of an HBT with a collector-up topology using $Ga_xIn_{1-x}P$/GaAs (with $x = 0.51$) material system will be presented. The material parameters along with the mobility were used for an emitter-up HBT and incorporated into the two dimensional, physically-based, numerical simulator ATLAS from SILVACO Inc.

In a conventional transistor, efficient charge collection is the prime concern for designers that make it imperative to have a larger collector than emitter area. However, these results in a significant compromise in high frequency performance of the device owing to the large junction capacitance associated with the larger collector area. An ingenious solution to this quandary was proposed by Kroemer [1.4] in 1982 by suggesting a structure with a smaller collector area and inverted in configuration, such that the device's collector rests on top of the emitter. In order to retain the efficient charge collection aspect of the emitter-up structures, he proposed the inactivation of that part of emitter-base junction that is not immediately opposite to a part of collector-base junction. Collector-up topologies have significantly improved the high frequency performance of HBTs in material systems as diverse as AlGaAs/InGaAs/GaAs and Ge/GaAs.

Table 6.1. Material parameters of $Ga_{0.51}In_{0.49}P$ used in GaInP/GaAs HBT device simulations. *Data source*: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and Properties". http://www.ioffe.ru/.

Parameter name	$Ga_{0.51}In_{0.49}P$ values
Crystal structure	Zinc Blende (disordered)
Dielectric constant	11.8
Lattice constant (Å)	5.653
Effective electron mass (me)	
G-valley	$0.088m_0$
L-valley	$0.6963m_0$
X-valley	$0.7263m_0$
Effective heavy hole mass (m_h)	$0.7m_0$
Effective light hole mass (m_l)	$0.12m_0$
Energy gap (eV)	1.849
Effective conduction band density of states (cm^{-3})	6.5×10^{17}
Effective valence band density of states (cm^{-3})	1.45×10^{19}
Conduction band offset Ec (meV)	137
Valence band offset Ev (meV)	310
Breakdown field (V/cm)	5×10^5

Table 6.2. Material Properties for $Ga_{0.51}In_{0.49}P$ and GaAs. *Data source*: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and Properties". http://www.ioffe.ru/.

Material properties	InGaP	GaAs
Energy gap Eg (eV)	1.849	1.424
Electron affinity (eV)	4.1	4.07
Effective density of states in CB N_c cm^{-3}	6.5×10^{17}	9.2×10^{17}
Effective density of states in VB N_v cm^{-3}	1.45×10^{19}	1.286×10^{19}
Steady-state recombination lifetime of electrons t_{n0} ns	0.09	5
Steady-state recombination lifetime of holes t_{p0} ns	0.09	3
Saturation velocity of electrons V_{satn} 10^7cm/s	1	2.0
Saturation velocity of holes V_{satp} 10^7cm/s	1	1.8

For device modeling, the simulation is initially done with zero applied bias and subsequently the voltages to be applied are incremented. The sweep of the DC bias yields the Gummel plot of I_B and I_C vs. V_{BE}. For AC analysis, after the desired DC bias point is reached, the AC bias is applied and the AC performance parameters are extracted.
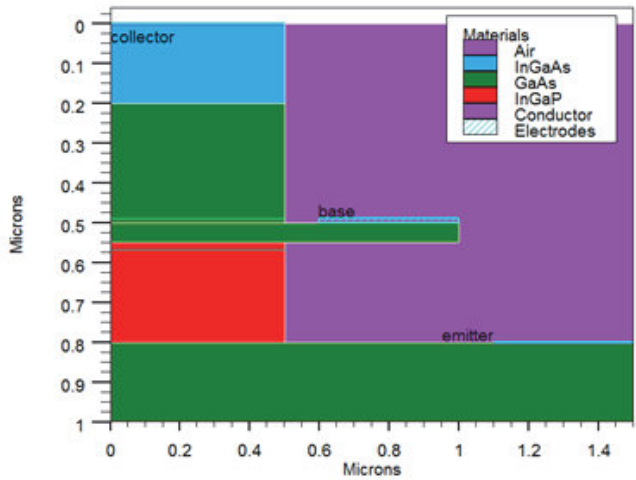
Figure 6.21. The structure of collector-up InGaP/GaAs HBT with partially etched extrinsic emitter.
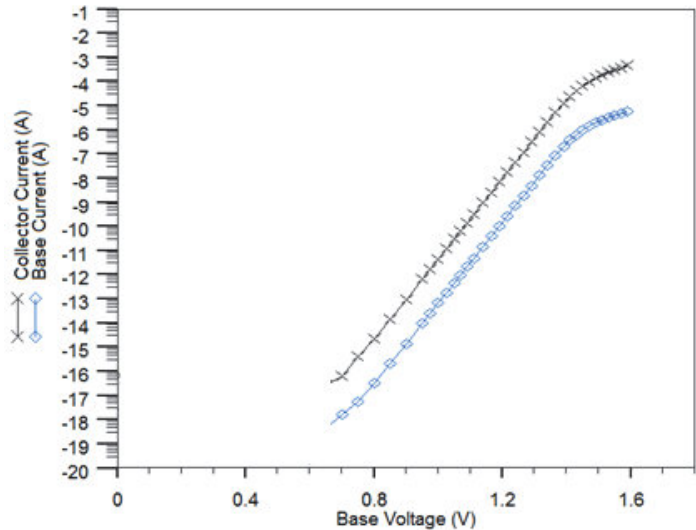
Figure 6.22. Gummel-Poon characteristics for the extrinsic emitter completely etched collector-up InGaP/GaAs HBT.

The effect of undercutting the extrinsic base is that electron injection across the emitter-base junction is limited to the central region beneath

the collector. For the collector-up InGaP/GaAs HBT simulation, a conventional 5 layered structure consisting of a sub-collector, collector, base, emitter and sub-emitter was chosen for simulations. The device's epitaxial structure was then optimized for enhancing both f_T and f_{max} by optimizing each layer's doping and thickness. Figure 6.21 shows the structure of collector-up InGaP/GaAs HBT with partially etched extrinsic emitter. For mobility modeling, Caughey-Thomas mobility models for electrons in InGaP and for electrons and holes in GaAs were used. The device performance parameters such as cutoff frequency f_T, maximum frequency of oscillation f_{max}, small signal current gain and maximum unilateral power gain (U_{max}) are computed as a function of collector current density. The epitaxial structure of the simulated InGaP/GaAs collector-up HBT is shown in Table 6.3.

Table 6.3. Structure of GaInP/GaAs collector-up HBT used in simulation. Data source: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and Properties". http://www.ioffe.ru/.

Material	Layer	Material	Doping (cm^{-3})	Thickness (nm)
InGaAs	Sub-collector	$n^+ In_{0.5}Ga_{0.5}As$	4×10^{19}	200
GaAs	Collector	$n^- GaAs$	9×10^{16}	400
GaAs	Base	$p^- GaAs$	3×10^{19}	30
InGaP	Emitter	$n^- In_{0.49}Ga_{0.51}P$	5×10^{17}	250
GaAs	Sub-emitter	$n^+ GaAs$	2×10^{18}	200

A Gummel-Poon characteristic for the extrinsic emitter completely etched collector-up InGaP/GaAs HBT is shown in Figure 6.22. AC Characteristics for the extrinsic emitter etched collector-up InGaP/GaAs HBT is shown in Figure 6.23. The small signal AC gain is found to be around 30 dB when the extrinsic emitter is completely etched. The device shows a f_T of 139 GHz at 0.8mA and an f_{max} of 233 GHz at 2.5mA.
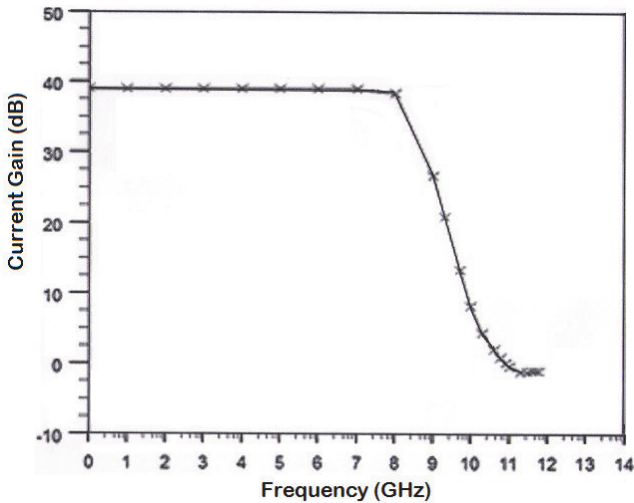
Figure 6.23. AC characteristics of extrinsic emitter etched collector-up InGaP/GaAs HBT.

The primary advantage of using a collector-up HBT is the significant reduction in collector junction area, which greatly minimizes the collector junction capacitance. Compared to conventional emitter up HBTs, the switching times of the device are reduced by a factor of nearly one-third. DC and high frequency characteristics of a typical InGaP/GaAs collector up HBTs were discussed.

6.4 RadHard SiGe-HBTs

Silicon Germanium (SiGe) Heterojunction Bipolar Transistor (HBT) technology has become an important semiconductor technology for both wired and wireless application because of its superior analog and RF performance, together with its CMOS integration capability. SiGe HBT technology has generated considerable interest in the space community due to its robustness to total ionizing dose radiation. The radiation effect of SiGe HBT technology is of interest to spacecraft designers to insert the latest technology into the space-borne system. Clearly, a space qualified IC technology must demonstrate sufficient radiation immunity to support high-speed circuit applications as well possess total dose tolerance. SiGe HBTs have been found robust to displacement damages

and total ionizing dose radiation but recent experiments show that SiGe HBTs are prone to single event effects (SEE). Thus it is important to study the single event effects on SiGe HBTs. To understand SEE in SiGe HBTs, one must use calibrated three dimensional (3D) device simulation to assess the charge collection characteristics of SiGe HBTs. In general, charge collection is found maximum for a strike at the emitter center and reduced for strikes outside deep trench isolation. The positional dependency of charge collection is studied, particularly regarding the drift and diffusion components. The SiGe HBT has a planar self-aligned structure. The SiGe HBT has a poly emitter contact, silicided extrinsic base and deep and shallow trench isolations. The SiGe base is grown using ultra-high vacuum chemical vapor deposition (UHVCVD). The NPN layers of the intrinsic transistor and the p-type substrate form an n-PNP multilayer structure, making the charge collection more complicated. Figure 6.24 shows a schematic device cross section of a first generation (f_T = 50 GHz) SiGe HBT.
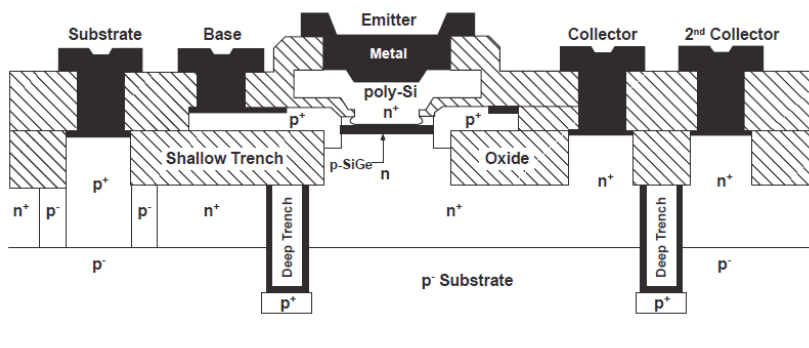
Figure 6.24. Schematic device cross section of a first-generation SiGe HBT. After M. Varadharajaperumal, 3D simulation of SEU in SiGe HBTs and radiation hardening by design, PhD Thesis, Auburn University, 2010.
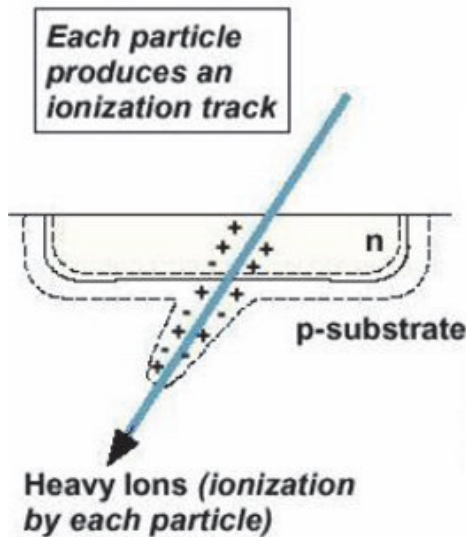
Figure 6.25. Illustration of SEU in a pn junction.

When highly energetic particles hit the sensitive nodes of the electronic circuit, it leaves a trail of electron/hole pairs. When an ionizing particle passes through the target material, electrons and holes are released along the path of the ionizing particle. The width of this track depends on the energy of the carriers, which depends on the energy of the incident particle. These charges are collected by the sensitive node of the circuit, which gives rise to Single Event Effect (SEE). Figure 6.25 illustrates the single event upset in a pn junction. There are two mechanisms of charge generation. In the first type the primary ion generates enough electron/hole pairs to cause an upset which is shown in the figure. In the second type the primary ion has no enough energy to generate the required amount of charge to cause an upset. But the primary ion reacts with the target material to produce secondary ions which can generate enough charge to cause an upset. Figure 6.25 shows the formation of carriers along the ion path. The electrostatic potential is disturbed in the junction and this disturbed field extends deep in the substrate. The disturbed field collects charge deposited deep in the substrate

The particle strike may cause a temporary disturbance in the circuit operation or may even cause a permanent damage in the circuit. The

single event effects can be classified into two categories. They may be either destructive SEE or non-destructive SEE. Single event current transients (SET) and single-event upset (SEU) in logic or memory circuits are examples of non-destructive SEE and single-event latchup (SEL) is an example for destructive SEE. 3D simulation is the only viable way for studying charge collection in SiGe HBTs. In the following, we present 3D simulation of Single Event Upset in IBM 8HP SiGe heterojunction bipolar transistor. The device simulation starts with the construction of the device. Simulations are performed on the IBM 8HP SiGe HBT which is constructed using the Synopsys TCAD tool MESH. The 3D device is constructed based on the layout and the vertical divisions of the device.

Figure 6.26 gives the schematic cross section and the layout of the regular 8HP device. The device is divided into layers and each layer is
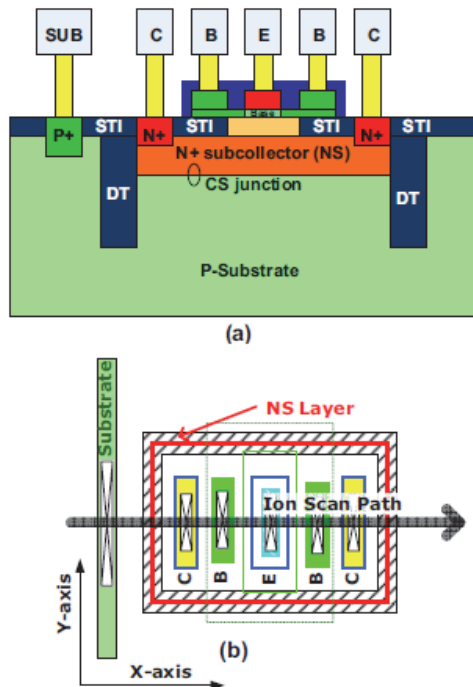
Figure 6.26. (a) Schematic cross-section and (b) schematic of the layout of a SiGe HBT (not to scale). After M. Varadharajaperumal, 3D simulation of SEU in SiGe HBTs and radiation hardening by design, PhD Thesis, Auburn University, 2010.

divided into vertical divisions showing the materials in each layer. The horizontal view of the materials in a layer can be seen with a layout of the device. The device structure can be divided into five layers as seen in the figure. The bottom layer consists of silicon substrate only. The fourth layer consists of the deep trench oxide and silicon and it can be called as the deep trench (DT) layer. The middle layer consists of the shallow trench and silicon and is called as the shallow trench (ST) layer. The SiGe layer is on the top of the ST layer. Above the SiGe layer lies the polysilicon emitter. The boundaries of different regions in 2D can be specified as polygons or multiple rectangles. Figure 6.27 shows the 3D view of the regular 8HP HBT device. A 2D cut is made along the center of the emitter to get the detailed view of the intrinsic portion of the device. Figure 6.28 shows the resulting 2D cross section. The device has a p-type substrate. There is the N^+ buried layer, which is the N^+ sub-collector (NS) of the regular HBT.
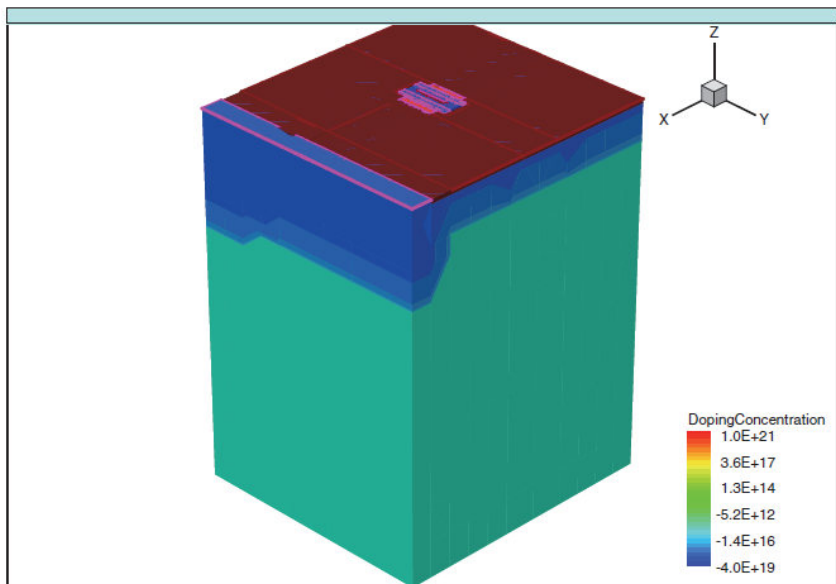
Figure 6.27. 3D view of the 8HP regular HBT. Color indicates doping profile. After M. Varadharajaperumal, 3D simulation of SEU in SiGe HBTs and radiation hardening by design, PhD Thesis, Auburn University, 2010.
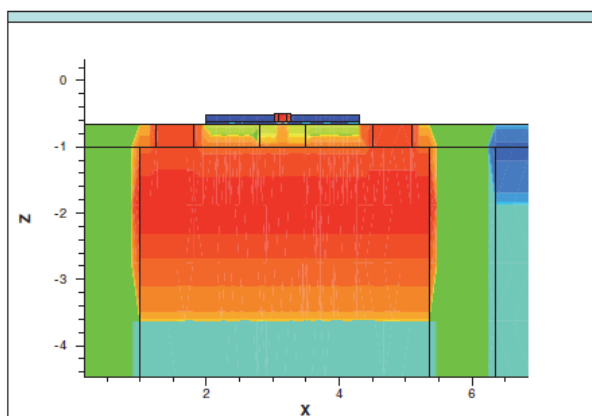
Figure 6.28. 2D view of the 8HP regular HBT (cut made at the center of the emitter). Color indicates doping profile. After M. Varadharajaperumal, 3D simulation of SEU in SiGe HBTs and radiation hardening by design, PhD Thesis, Auburn University, 2010.

When a heavy ion passes through a semiconductor material, it keeps losing energy along its path. Depending on the energy of the heavy ion and depending on the target material, the heavy ion comes to a rest in the device after losing its energy (shallow strike) or the heavy ion traverses the entire device (deep strike). The total path traveled by the particle in the target material before it comes to a rest is called the range of the particle. The Linear Energy Transfer (LET) is the measure of the energy lost by the ion, which can be defined as the energy lost by the particle per unit path length as it passes through the target material. The LET defined above has a unit of MeV cm^2/mg. The LET of the ion can be related to the charge deposition per unit length because it takes a certain amount of energy to release carriers for a particular material. In silicon one electron-hole pair is released for every 3.6 eV energy released by the incident ion. The density of silicon is 2328 mg/cm3. Hence an LET of 97 MeV cm^2/mg corresponds to a charge deposition of 1 pC/μm. LET should be specified in pC/μm for SEU simulations. In DESSIS, a constant LET or a variable LET can be specified. A normal incident deep ion strike through the emitter with a uniform LET of 0.1 pC/μm (10 MeV-cm^2/mg) is simulated. The target material used in simulation is silicon.

When a heavy ion passes through the device it creates a trail of electron-hole pairs. These charges are collected by the electric field present in the

device by drift and by diffusion. The point of strike is specified along with the direction of the ion track. In DESSIS [2.6], the generation rate due to heavy ion is computed by the following relation

$$G(l, w, t) = \text{LET}(l) \times R(w) \times T(t)$$

where l is the distance from the incident point and w distance from the track. A charge track containing a high density of electrons and holes is created when a heavy ion passes through the semiconductor material. This charge track is simulated in the simulation tool. The point of strike and the direction of strike is specified in the simulation. The charge track was generated for a period of 10 picoseconds using a Gaussian waveform. The 1/e characteristic time scale is 2 picoseconds and the 1/e characteristic radius is 0.1μm. The peak of the Gaussian occurs at 2 picoseconds.

The physical models selected in a simulation influence the accuracy of the result to a larger extent. So one needs to be very careful in selecting the physical models. The most important models selected in this simulation are Phillips unified mobility model, Auger recombination and bulk Shockley-Read-Hall (SRH) recombination and surface SRH recombination. The Phillips unified mobility model is used because it is more accurate for bipolar devices. Auger recombination is turned on when there is a heavy concentration of carriers. The velocity saturation parameter is also turned on.

The heavy ion generates a heavy charge density along its path in the target. The deposited charges are collected through drift and diffusion by the terminals of the device. The CS junction plays a major role in charge collection. If the ion track passes through the junction electric field, the very high charge density generated by the ion collapses the junction field. Figure 6.29 shows the current and charge collection characteristic for a 0.5×10 μm^2 emitter 5HP HBT for an ion strike at the emitter center. The CS junction is reverse biased at −5.2 V. The charge characteristic is obtained by integrating the current characteristic. The charge collection starts by 0.05 ns and the collector and substrate collects most of the charges. The collector collects electrons and the substrate collects holes.
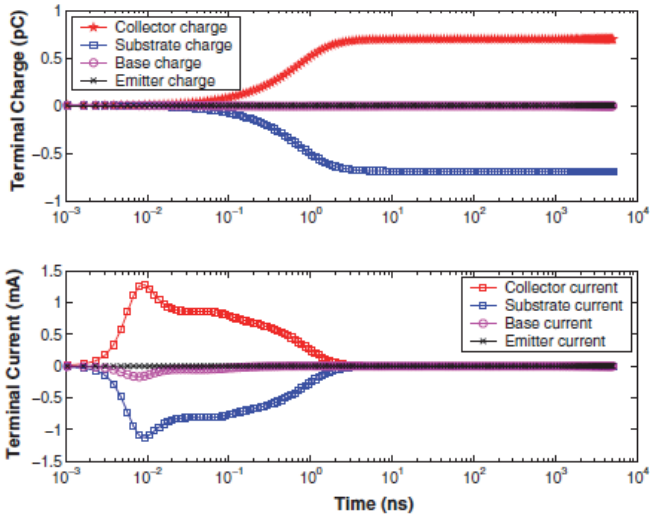
Figure 6.29. Terminal current/charge for an ion strike at emitter center for the 0.5×10 μm^2 5HP HBT. After M. Varadharajaperumal, 3D simulation of SEU in SiGe HBT and radiation hardening by design, PhD Thesis, Auburn University, 2010.

Summary

Through a comparison of various available BiCMOS technologies from semiconductor manufacturers, it is shown that SiGe bipolar technology is a very good candidate for the sub-terahertz circuit applications (frequency >100 GHz). In this chapter, both silicon and III–V compound semiconductor devices are designed and optimized through calibrated 2D and 3D TCAD simulations. The design of an HBT with a collector-up topology is carried out and optimized for maximizing the high frequency performance. The material system used was $Ga_xIn_{1-x}P/GaAs$ (with $x = 0.51$ indicating lattice matched composition), which has relatively superior material properties and etching characteristics than the conventional $Al_xGa_{1-x}As/GaAs$ material system. Self-heating effects for SiGeC HBTs was studied using 2D and 3D simulations. Single event upset was simulated in 5HP and 8HP SiGe HBTs. SiGeC HBT technologies as well as SiGe have enabled mm-wave and THz applications are foreseeable. In fact, now-a-days SiGe technologies have passed the 600 GHz barrier in terms of maximum frequency of oscillation.

Chapter 7

Stress-Engineered HBTs

In advanced semiconductor devices especially in MOSFETs, strain engineering technology is being used as an additional degree of freedom to enhance the carriers transport properties due to band structure changes and mobility enhancement [7.1]. As discussed in Chapter 6, the European joint research project dotfive partners IMEC and IHP Microelectronics have achieved SiGeC HBTs with maximum frequency of oscillation f_{max} of 400 GHz through completely different architectures. While the IMEC an IHP Microelectronics approach for 400 GHz f_{max} has been through fully self-aligned (FSA) SiGeC architecture, ST Microelectronics approached through a conventional double-polysilicon FSA selective epitaxial growth (SEG) Si/SiGeC HBT. The speed improvement compared to previous SiGe HBT technologies originates from the reduced specific collector-base capacitance and base resistance and scaling of the device dimensions.

With the continuing reduction of silicon integrated circuits, new engineering solutions and innovative techniques are required to improve bipolar transistors performance, and to overcome the physical limitations of the device scaling. Strained-silicon technology has become a strong competitor in search for alternatives to transistor scaling and new materials for improved devices and circuits performances. The continuous demand for bipolar devices having higher cutoff frequency, it becomes imperative to develop new bipolar architectures suitable for high frequency and power applications. Mobility of charge carriers in a bipolar device can be increased by creating mechanical tensile strain in the device in the direction of electrons flow to improve electron's mobility, and by creating mechanical compressive strain in the device in the direction of holes flow to improve hole's mobility.

SiGe HBTs is realized by the pseudomorphic growth of SiGe on a Si substrate, which allows engineering of the base region to improve performance. In this way the base has a smaller energy band gap than the emitter, which increases the gain. The energy band gap of SiGe reduces with increasing Ge composition, but the maximum Ge composition is limited by the amount of strain that can be accommodated within a given base layer thickness. Various techniques and efforts have been proposed to improve the performance of bipolar devices: a) introduction of a grading germanium into the base, b) introduction of carbon to improve 1D doping profile, and c) reduction of emitter width.

The main focus of this chapter is to examine the impact of strain engineering technology on NPN-SiGe-HBT device's electrical properties and frequency response, through exploring new device concepts and new device architectures that are based on strain engineering technology using TCAD tools. The benefit of using strain engineering for bipolar transistor technology will be highlighted. In the following, the impact of strain engineering technology applied on NPN-Si-BJT/NPN-SiGe-HBT devices on the electrical properties and frequency response will be discussed.

7.1   Strained-Si SiGe HBTs

Recent experimental and theoretical works point out uniaxially and biaxially stressed Si and SiGe substrates as a possible technique for the new generation HBT devices. Uniaxially and biaxially stressed Si and SiGe substrates can be used for the fabrication of new generation HBT devices. Experimentally, wafer bending and virtual substrate techniques are used to introduce stress. Wafer bending induces uniaxial stress while virtual substrates create biaxial stress over the device. Four-point bending apparatus has been used to apply a uniaxial stress in the range of -200 MPa to $+200$ MPa on SiGe HBT devices fabricated in 0.18μm self-aligned SiGe BiCMOS technology with an emitter area of $0.2 \times 10.16 \mu m^2$. The results show that the performances of SiGe HBTs are changed with the stress level. The changes in the collector current, base current, current gain, and the breakdown voltage were found to be linearly dependent upon the mechanical uniaxial stress level for the entire range of

−200 MPa to +200 MPa. The strain-polarity dependence of the collector current, base current, and current gain was positive under uniaxial compressive stress, whereas that of the breakdown voltage was negative.

The starting point was the growth of the sub collector, collector, and base and emitter layer epitaxially as shown in Figure 7.1a. This was followed by etching of the material surrounding the emitter, down to the collector to form mesa isolation, so the emitter could be isolated from the base and collector contacts, as shown in Figure 7.1b. An N-type collector link was implanted using P at a dose of 5×10^{15} cm^{-2} and energy of 20 keV and an extrinsic base region was subsequently implanted using BF$_2$ at a dose of 1×10^{15} cm^{-2} and an energy of 35 keV as illustrated in Figure 7.1c. The next step was low thermal oxide (LTO) deposition (400nm) to isolate the structure and define the collector and emitter window, as shown in Figure 7.1d. Heavily doped n$^+$ polysilicon (P, 5×10^{19} cm^{-3}) was deposited for both contacts, Figure 7.1e, followed by RTA step, which was 900°C for 10s. Etching the polysilicon was necessary to isolate the emitter poly from the collector poly, see Figure 7.1f. LTO deposition (400nm) was performed to define the emitter, base and collector window contacts, as illustrated in Figure 7.1g. The process was concluded with the deposition of the Al base, emitter and collector contacts with a TiW barrier layer and a forming gas anneal. The final transistor structure is shown schematically in Figure 7.1h.

In order to explore the impact of parameter design space on the current gain, devices were fabricated with different values of E$_W$ emitter window width, L$_E$ emitter window length, D1 distance to the collector and D2 distance to the base. Figure 7.2 shows a schematic diagram of a bipolar device and the parameter design space. The schematic of the device structure used in simulation is shown in Figure 7.3.
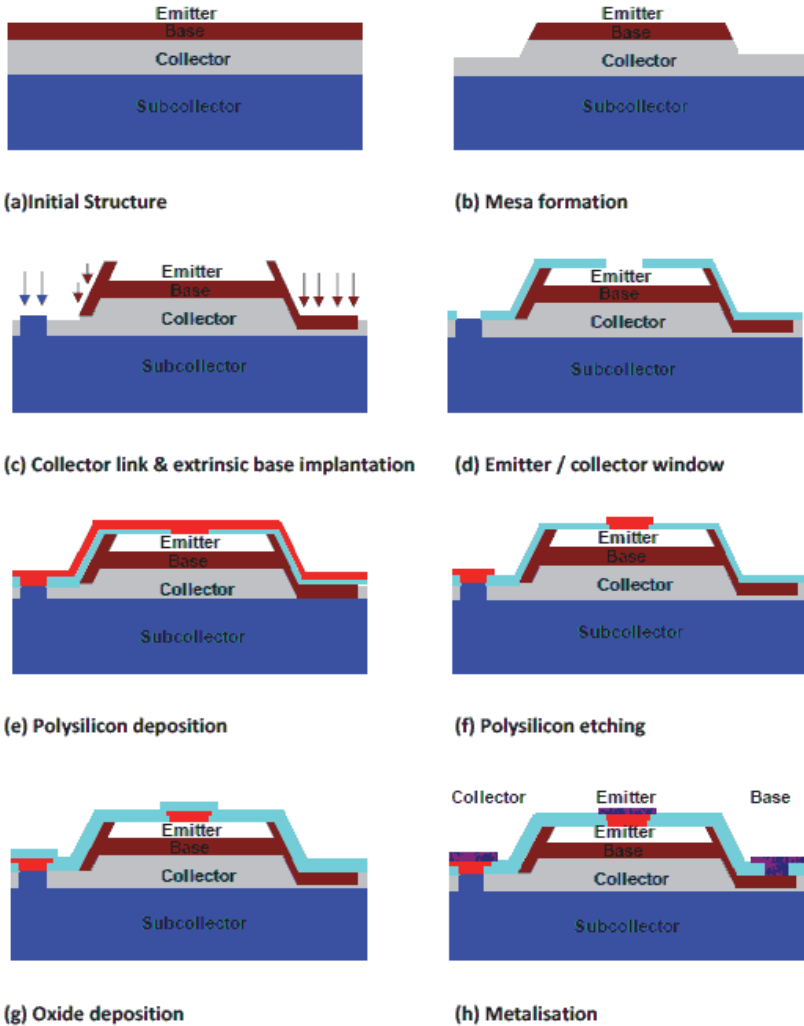
Figure 7.1. Simplified process flow, Si BJTs, SiGe HBTs and strained-Si HBTs were fabricated using the same process flow. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

Figure 7.2. Schematic diagram illustrating different parameters design space. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.
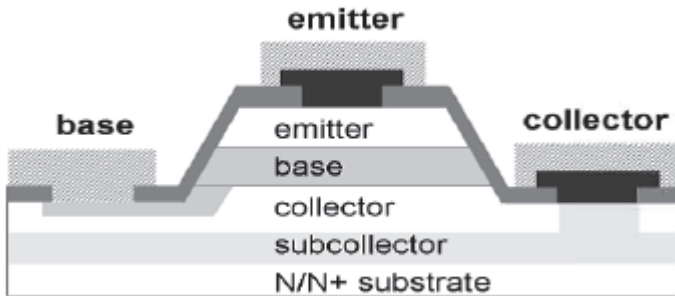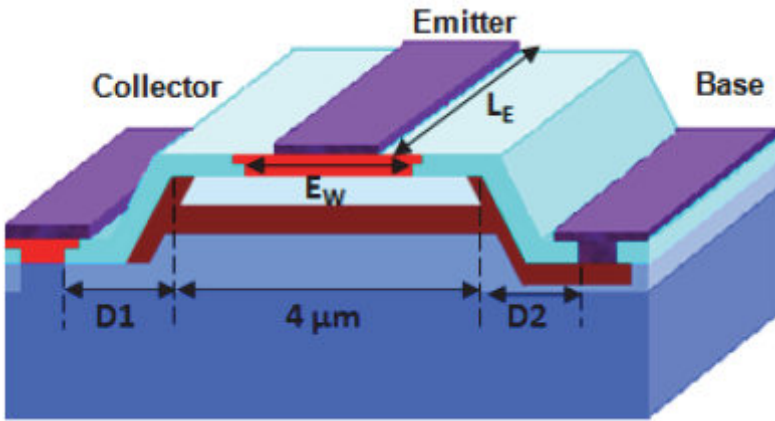
Figure 7.3. Schematic of the final strained Si HBT structure. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

Figure 7.4 shows a comparison of the collector current for strained-Si HBTs, SiGe HBTs and Si BJTs. The SiGe HBTs exhibits a higher collector current compared with Si BJTs, as expected. However, the collector current in the strained-Si HBTs is increased considerably compared to both devices. This is mainly due to the band gap of the base layer which decreases from 1.17 eV for the Si BJTs to 1.04 eV for the SiGe HBTs and 0.94 eV for the strained-Si HBTs. The conduction band
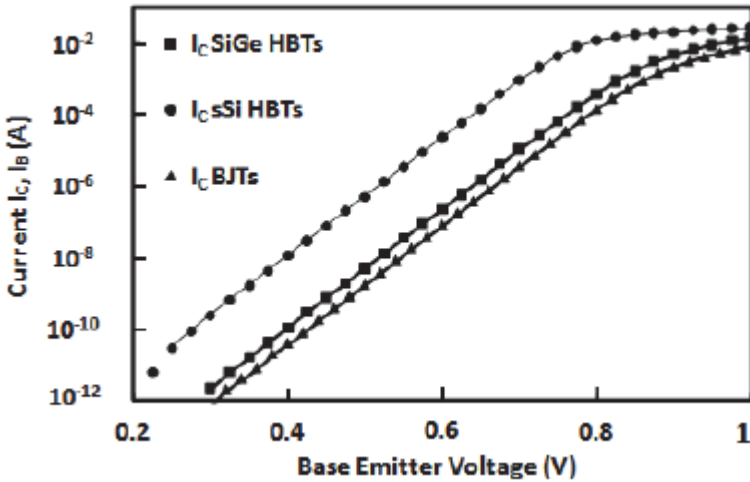
Figure 7.4. Collector current for strained-Si HBTs, SiGe HBTs and Si BJTs @ $V_{BC} = 0$ V. $W_E = 1\mu m$ and $L_E = 10\mu m$. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

in both the SiGe HBTs and the strained-Si HBTs exhibits a discontinuity at the base-collector junction; this discontinuity causes a reduction in I_C.

Figure 7.5 shows that the base current for the SiGe HBTs and the Si BJTs is same. However, the strained-Si HBTs exhibits a slightly higher base current. The conduction band discontinuity is slightly smaller for the strained-Si HBTs compared to SiGe HBTs. The strained-Si HBT is formed from a relaxed SiGe (collector), a compressed SiGe (base) and a tensile strained Si, which serves as the emitter. The strained Si has a smaller band gap compared with relaxed Si which increases the intrinsic carrier concentration. This results in the strained-Si HBTs having a higher base current compared with the other devices. The emitters in both the Si BJTs and the SiGe HBTs are formed from relaxed Si.
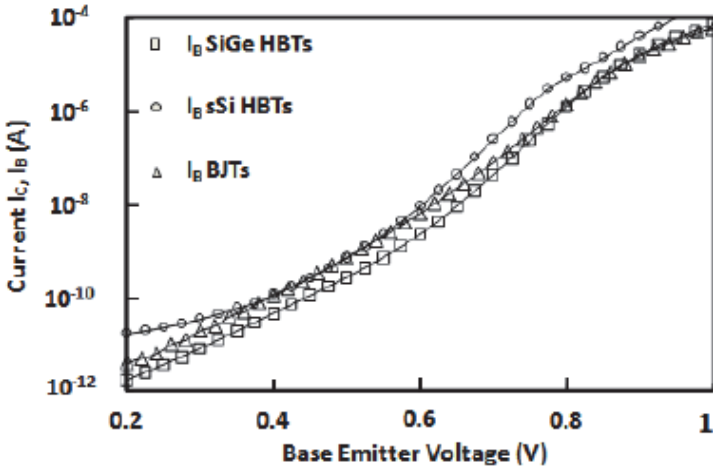
Figure 7.5. Base current for strained-Si HBTs, SiGe HBTs and Si BJTs @ V_{BC} = 0 V. W_E = 1μm and L_E = 10μm. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

Figure 7.6. Current gain β vs. BE voltage V_{BE}. β is increased by more than one order of magnitude in the strained-Si HBT, as compared with the SiGe HBT and the Si BJT. W_E = 1μm and L_E = 10μm. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

The current gain is calculated from the ratio of the collector current and the base current. Figure 7.6 shows a plot of current gain β as a function of base-emitter voltage V_{BE} at $V_{CB} = 0$. It is plotted on a logarithmic scale and shows that the maximum value of β is 3700, 334 and 135 for the strained Si HBTs, SiGe HBTs and Si BJTs respectively. This represents an improvement in the gain of the strained-Si HBTs by $11\times$ compared with the conventional SiGe HBTs and an improvement of $27\times$ compared with the conventional Si BJTs. Figure 7.6 also shows that the maximum current gain is achieved at lower VBE for the strained-Si HBTs. In other words the amplitude of the input signal needed for the strained-Si HBTs in order to operate (at maximum β) is low compared to the other devices. This high gain can be traded off against lower base resistance. The gain improvement is largely attributed to the difference in energy band gap between the emitter and the base, but the conduction band offset between the base and the collector is also important for the collector current level.

A comparison of the common-emitter characteristics for strained-Si HBTs, SiGe HBTs and Si BJTs at $I_B = 3\mu A$ is shown in Figure 7.7. The collector current in the strained-Si HBTs is $7\times$ larger than in the SiGe HBTs and $22\times$ larger than in the BJTs at a collector-emitter voltage of $V_{CE} = 1$ V. Clearly, Si BJTs and SiGe HBTs do not exhibit self-heating. If the self-heating in the strained-Si HBTs were to be reduced, then the enhancements in IC would be even larger. The strained-Si HBTs exhibit a high collector current which results in high power dissipation. This power is translated into heat, mainly in the base collector depletion region, where the current and the electrical field are high. The thermal conductivity of the SiGe layer is known to be small compared to Si. The high power dissipation and the low thermal conductivity of SiGe in the SRB result in an excessive increase of the temperature. The high temperature causes a strong lattice vibration which decreases the electron/hole mobility leading to a reduction in the collector current. The extracted Early voltage for the Si BJTs and SiGe HBTs were 6.0V and 1.6V, respectively. The reduced VA is due to the base modulation.

The BV_{CEO}, measured at base open were found to be 2.5, 2.7 and 4.5V for strained-Si HBTs, SiGe HBTs and Si BJTs, respectively. In this case the avalanche breakdown of the base-collector junction is further influenced by transistor action in the common-emitter mode of operation, since the holes generated by impact ionization are pulled back into the

base region which results in an additional component of base current. This additional base current causes an even larger flow of electrons through the base and into the collector due to the current gain of the device. This current is larger for strained-Si HBTs, since it exhibits the highest current gain, compared to Si BJTs and SiGe HBTs. This larger flow of electrons in the base collector junction causes an even larger generation of electron-hole pairs which causes the strained-Si HBTs to have a small BV_{CEO}. The figure of merit $\beta \cdot BV_{CEO}$ describes the ability to offer simultaneously high β and high BV_{CEO} for given device. The strained-Si HBTs exhibited a much higher $\beta \cdot BV_{CEO}$ (9250V) than both the SiGe HBTs (900V) and the Si BJTs (600V). The $\beta \cdot BV_{CEO}$ for strained-Si HBTs is 15× that of Si BJTs, which confirms that strained-Si HBTs is a good platform for high performance HBTs.
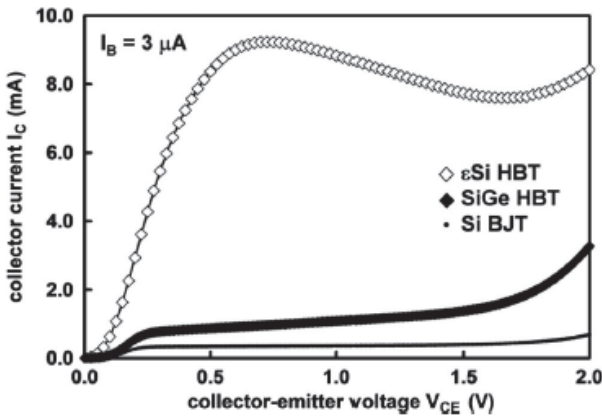
Figure 7.7. Comparison of collector current I_C vs. collector–emitter voltage V_{CE} characteristics for strained-Si HBTs, SiGe HBTs, and Si BJTs at $I_B = 3\mu A$. Self-heating is observed for strained-Si HBTs, but no SHEs are observed on Si BJTs and SiGe HBTs. $W_E = 1\mu m$ and $L_E = 10\mu m$. After M. Fjer, Strained Si heterojunction bipolar transistors, PhD Thesis, Newcastle University, 2011.

7.2 Strained-Si PNP SiGe HBTs

For high-speed analog and mixed-signal circuit applications, the use of a complementary pair of NPN and PNP devices is advantageous. However, PNP Si BJTs have traditionally suffered from low current gain and low cutoff frequencies, in comparison with equivalent NPN devices. Relaxed SiGe virtual substrate technology has attracted much interest, utilizing SiGe to create active heterojunction channel CMOS devices and all-SiGe NPN HBTs. As for future virtual substrate PNP HBTs, they can be designed and optimized in an identical manner to conventional SiGe PNP HBTs, retaining, for example, the same strained-$Si_{1-x}Ge_x$ base grading deep into $Si_{1-y}Ge_y$ relaxed alloy substrate (collector) for a smooth transition. The TCAD simulation methodology used consist of the following steps:

- To construct a PNP SiGe HBT structure with SiGe base that consists of strained-Si base layer on relaxed SiGe virtual substrate.
- To study the effects of doping dependence and base width variation on current gain and cutoff frequency

In the following, we will discuss the design of a strained-Si PNP SiGe HBT on virtual substrates. Strain relaxed buffers (SRB) act as "Virtual Substrates" (VS) on which one can grow layers of a semiconductor with different lattice constant to the original substrate. In the case, a virtual substrate HBT uses relaxed SiGe layer on which silicon epitaxial layer is deposited to have biaxial strain. This enhances the mobility of carriers across the base region. The production of defect free virtual substrate is a critical issue. The possible defects in a virtual substrate are either threading dislocation or surface roughness. Figure 7.8 shows the schematic of the device cross-section of the PNP SiGe HBT on virtual substrate simulated in SILVACO ATLAS framework. It is generated from the process simulation of the PNP HBT using the process simulator ATHENA from Silvaco. The $p^-/p+$ doped SiGe relaxed alloy substrate (virtual substrate) serves as the collector. The n^+ base and p-type emitter regions are formed assuming a two-step pseudomorphic growth of tensilely strained Si layer on a virtual substrate with a strained-Si cap as the emitter. A base peak doping concentration of about 1×10^{19} cm^{-3} and a

total base width of 25nm were obtained after process simulation. Subsequently, instead of the conventional polysilicon deposition step used in Si bipolar technology, a thin 100nm p^+ SiGe (relaxed) layer is selectively grown on top after the emitter contact opening. The proposed novel PNP HBT design on virtual substrate is inherently free from the detrimental C-B valence band barrier effects. The peak doping concentration in the emitter SiGe layer is $1 \times 10^{20} cm^{-3}$. The N^+ base contact doping is done by a self-aligned phosphorous implant.
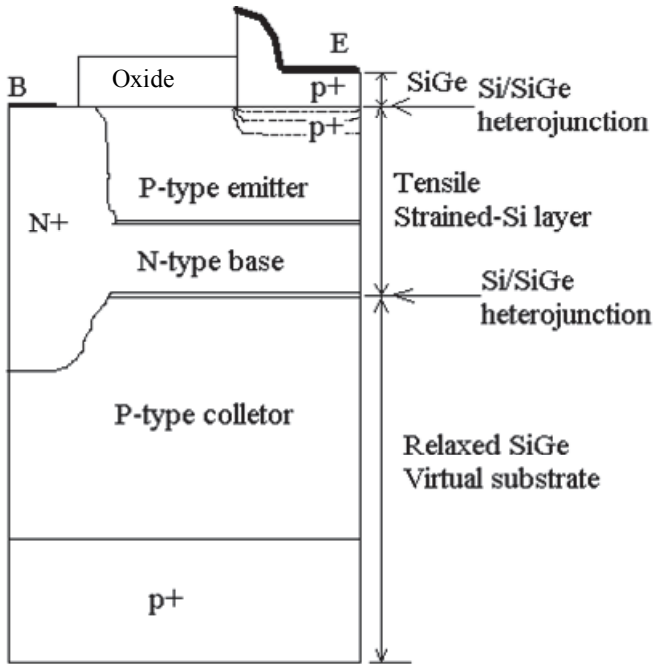
Figure 7.8. 2D cross-section of the schematic of the simulated virtual substrate PNP HBT device with strained-Si/SiGe (relaxed) hetero-interface at the base–collector junction and emitter bulk contact. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.

The base-collector (BC) junction is represented by a strained-Si/SiGe (relaxed) heterojunction with favorable band offsets. The smooth energy band transition occurs because the strained-Si layer on the SiGe relaxed buffer produces a type II band alignment with both conduction band offset ΔE_c and valence band offset ΔE_v having the same sign (negative

with reference to SiGe). Since these offsets follow the direction of the built-in B/C junction potential difference, the potential spike appearing in valence band of the SiGe-base HBT is not present in the virtual substrate HBT. In the case of the SiGe/Si hetero-interface, the type I band alignments yield a small negative ΔE_c offset, but a large ΔE_v opposing the built-in B/C potential difference and producing C/B valence band barrier which impedes the flow of the carriers across the base/ collector junction. It may be that a second strained-Si/SiGe (relaxed) heterojunction in the virtual substrate PNP HBT is situated at the emitter contact with large ΔE_v. If the PNP HBT structure has a strained-Si/SiGe hetero-interface at the E/B junction as that of a conventional SiGe PNP HBT, the band offset ΔE_v would create a huge blocking valence band barrier for holes injected from the emitter to the base, which will totally sacrifice collector current. Instead, the second heterojunction is moved to the emitter contact. Because of this, two band offsets ΔE_c and ΔE_v appear at the emitter contact of the virtual substrate PNP HBT. The influence of ΔE_c is small, although ΔE_v presents a significant potential barrier for the minority holes to overcome, which they can only overcome by thermionic field emission.

The DC and AC characteristics of the virtual substrate PNP HBT are obtained utilizing the 2D device simulator ATLAS from Silvaco. For the simulation of virtual substrate HBT, we have started from the process simulator ATHENA where the entire structure was created. This structure was then passed on to ATLAS for device simulation. For carrier transport parameters, we have used Klaassen's built-in silicon mobility model. The increase in charge carriers low-field mobility in the strained-Si base layer, were modeled using the adopted model from reference [7.2]. The other models used are band-gap narrowing model, the auger and SRH recombination model, the concentration dependent and low-field mobility model. Figure 7.9 shows the simulated PNP HBT device structure obtained from ATHENA. The carrier mobility model used is incorporated in ATLAS via the C-interpreter module.
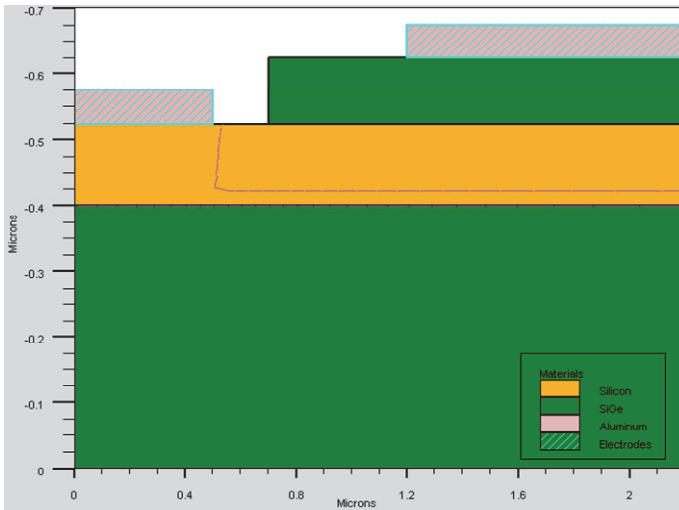
Figure 7.9. Simulated PNP strained-Si on virtual substrate HBT structure. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.
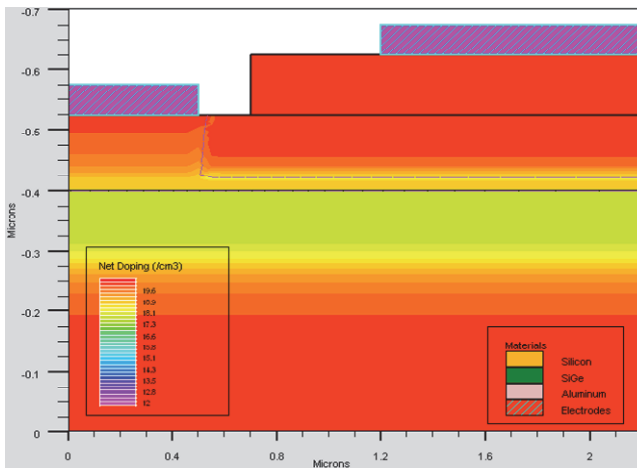
Figure 7.10a. Doping contours of simulated PNP strained-Si on virtual substrate HBT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.
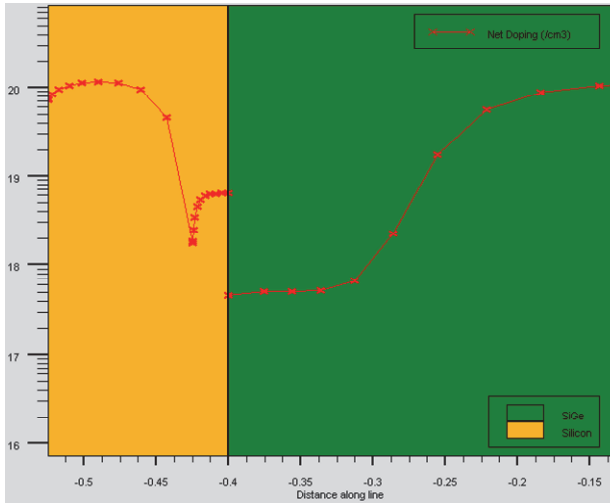
Figure 7.10b. Doping profile of simulated PNP strained-Si on virtual substrate HBT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.

Figures 7.10a and b show the doping profile of the simulated PNP virtual substrate HBT structure. The Peak base doping is about 5×10^{18} cm^{-3}, whereas, the emitter doping is kept at a peak of 1×10^{20} cm^{-3}. The base width is kept to be 0.025μm. Both the emitter and the base are made of strained-Si. This is done to prevent the detrimental valence band barrier to prevent carrier flow in to base region. The emitter contact is moved towards the SiGe/Si hetero-interface, where minority carrier recombination velocity is much less compared to the surface recombination velocity encountered in the hetero-interface. This improves the current gain. The current gain of the HBT for 15% Ge fraction is compared with the identical Si-BJT as shown in Figure 7.11. Figure 7.12 shows the comparison between the strained-Si HBT with 15% Ge and the identical Si-BJT, both showing enhancement in current gain.

Figure 7.13 shows the Gummel plot comparison between the 15% Ge strained-Si HBT with that of the identical Si-BJT. The plot reveals that even with only 15% Ge, the design is superior compared to the identical Si-BJT, as both current gain and collector current increases, while base current decreases for the strained-Si on virtual substrate HBT. In

Figure 7.14, the current gain variation is shown as a function of Ge mole fraction. With Ge mole fraction increase, the current gain increases. For the designed PNP virtual substrate HBT, this gain variation is found to be small. The peak gain obtained for 50% Ge is about 115, which agrees well with the available data for strained-Si HBT current gain [7.3]. The current gain is much higher compared to the current gain obtained for the identical Si-BJT (which was found to be 22.5). The variation of cutoff frequency with Ge fraction has been found to be small with change in Ge content. With increasing Ge content, the maximum oscillation frequency is found to increase. The peak cutoff frequency obtained for 50% Ge is 40.4 GHz, which is much higher compared to the f_T value obtained for identical Si-BJT (21.3 GHz).
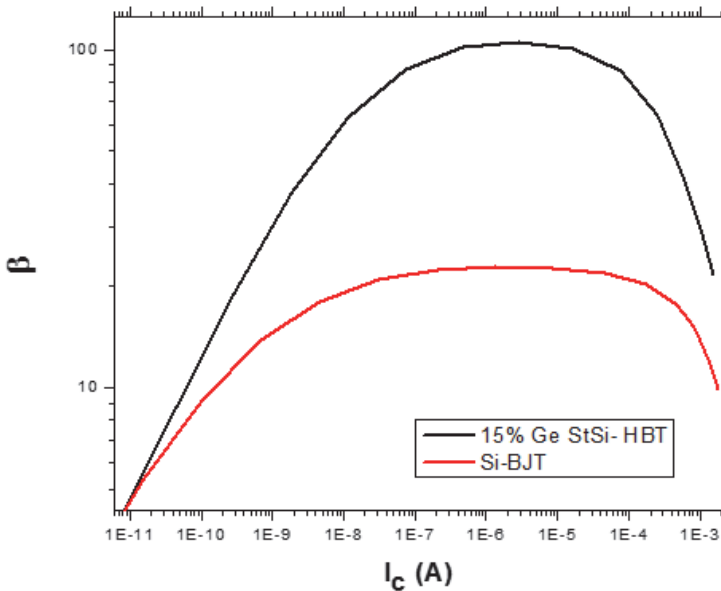
Figure 7.11. Current gain comparison between the simulated PNP strained-Si on virtual substrate HBT with that of the identical Si-BJT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.
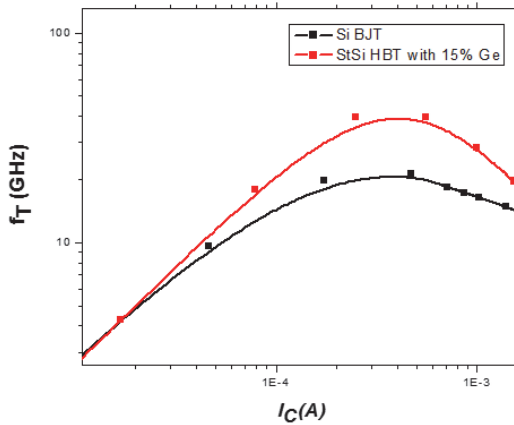
Figure 7.12. Cutoff frequency comparison between the simulated PNP strained-Si on virtual substrate HBT with that of the identical Si-BJT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.
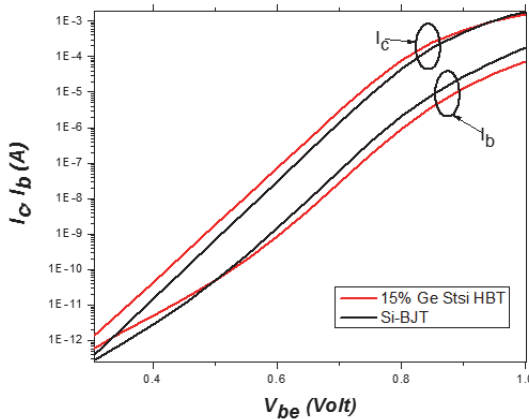
Figure 7.13. Gummel plot comparison between the simulated PNP strained-Si on virtual substrate HBT with that of the identical Si-BJT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.
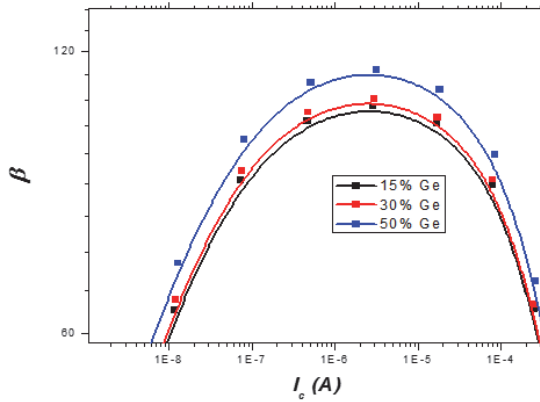
Figure 7.14. Current gain variation with increasing Ge fractions (15%, 30%, 50%) for the simulated PNP strained-Si on virtual substrate HBT. After C. Mukherjee, Strain engineering in heterojunction bipolar transistors, MTech Thesis, IIT Kharagpur, 2010.

To generate the strain inside the device two approaches have been used. The first approach was through introducing strain engineering technology principle at the device's base region using SiGe extrinsic stress layer. The second approach was through introducing strain engineering technology principle at the device's collector region using strain layers.

7.3 Si NPN BJT with Extrinsic Stress Layer

The impact of introducing a SiGe stress layer formed over the extrinsic base layer, and adjacent to the intrinsic base of NPN-Si-BJT device on the electrical properties and frequency response has been investigated using TCAD modeling. Following the reported in reference [7.4], process simulations are performed using Sentaurus TCAD software tools to build the device structure, and to calculate the associated mechanical stress due to the existence of the extrinsic stress layer using anisotropic elasticity model. The major process flow steps are described. The process simulation starts by the fabrication of shallow and deep trench isolation,

then the trenches are filled with silicon oxide to the same level as the surface of the doped silicon substrate. The main process steps are:

- Shallow and deep trenches isolation
- Deposition of etch-stop material and a thin layer of polysilicon
- Formation of base region
- Deposition of oxide and nitride layers
- Formation of emitter mandrel and recesses
- Deposition of SiGe extrinsic stress layer
- Emitter opening and formation of sidewall spacers
- Formation of T-shaped emitter

The final device structure obtained from process simulation is shown in Figure 7.15. As usual only one half of the device is used in device simulations as shown in Figure 7.16. The simulated results on the effects of interposing the extrinsic stress layer in the device is shown in Figure 7.17; the isocontour lines represent the stress values generated in the x-direction (Sxx) and y-direction (Syy) of the device due to the existence of the extrinsic SiGe stress layer.
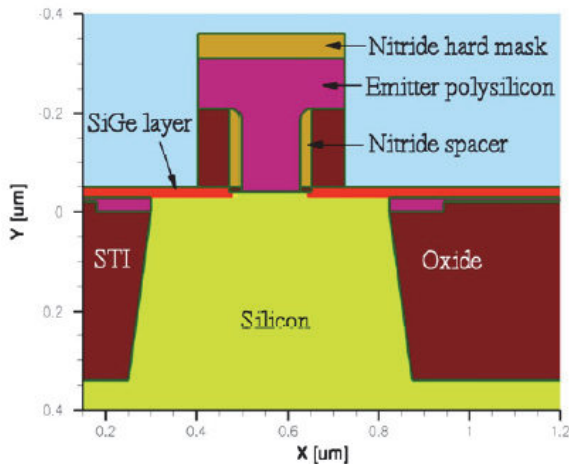
Figure 7.15. The final device schematic after the process simulation. After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.
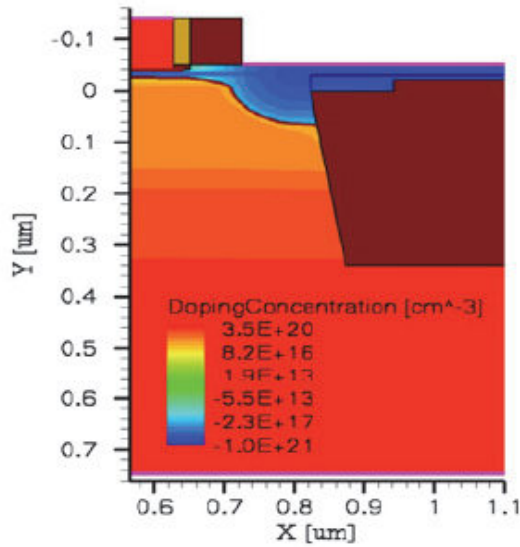
Figure 7.16. Cross-section of the one half of the device used for device simulation. After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.
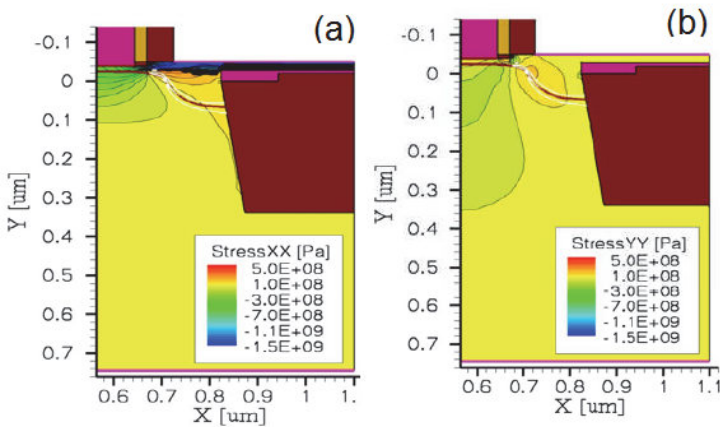
Figure 7.17. Cross section of the one half the device region of interest, the isocontour lines represent the stress Sxx (a) induced in the x-direction and Syy (b) induced in the y-direction of the structure. After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.

Sentaurus TCAD software tools have been used to perform the two-dimensional device simulations using hydrodynamic transport model (HD), where the carrier temperature equation for the dominant carriers is solved together with the electrostatic Poisson's equation and the continuity equations. Other standard silicon models, such as Philips unified mobility model, high field saturation mobility model, Shockley-Read-Hall recombination model, Auger recombination model, piezoresistive model for calculating mobility enhancement due to the applied stress, bandgap narrowing model and default parameter files, are all included in the simulation file. The doping profiles at the emitter, the base and the collector region of the device were supplied in a data file. Figure 7.18 shows the forward Gummel plots obtained by simulating both NPN-Si-BJT incorporating SiGe extrinsic stress layer at the base region, and a standard conventional NPN-Si-BJT device. Simulation results show that introducing the extrinsic stress layer in the device will increase the collector current by almost three times, resulting in an enhancement of the maximum current gain in comparison with the conventional BJT. The transit frequency f_T and the maximum oscillation frequency f_{max} are plotted vs. the collector current I_C for both structures as shown in Figure 7.19 and Figure 7.20.
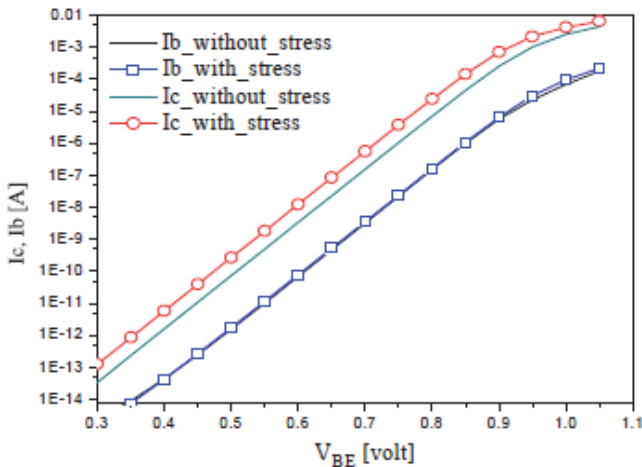
Figure 7.18. Comparison of forward Gummel plots for both conventional BJT, and BJT with stress layer (Ge = 25%, W_E = 130nm). After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.
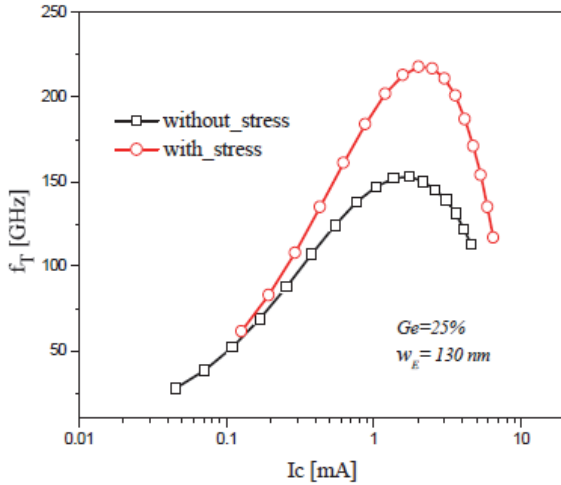
Figure 7.19. Cutoff frequency as a function of collector current for both devices (Ge = 25%, W_E = 130nm). After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.
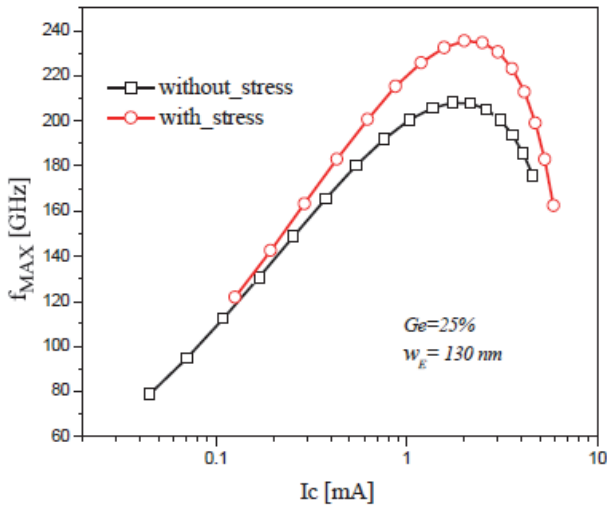
Figure 7.20. Maximum oscillation frequency as a function of collector current for both devices (Ge = 25%, W_E = 130nm). After M. Al-Sadi, TCAD based SiGe HBT advanced architecture exploration, PhD Thesis, Universite Bordeaux I, 2011.

The device simulation results show that bipolar device with extrinsic stressor layer exhibit better high frequency characteristics in comparison with an equivalent standard conventional device. An approximately 42% of improvement in f_T, and 13% of improvement in f_{max} in NPN-Si-BJT with extrinsic stress layer have been achieved. These improvements are mainly due to the enhanced vertical electron mobility, which can be fully accounted to the impact of interposing the extrinsic stress layer in the device.

The impact of changing the Ge content at the extrinsic stress layer on the stress values generated inside the device (i.e., Sxx, Syy, and Szz), maximum dc current gain, f_{max} and f_T has been studied and the results obtained are shown in Figure 7.21. As shown in the Figure 7.21a, increasing the Ge content at the extrinsic stress layer will increase the stress values generated inside the device, which in turn improves the high frequency characteristics of the device and enhances the current gain. These improvements are related to the increase of the lattice constant difference between the silicon substrate and the SiGe stress layer, which will increase the stress values induced at the base, resulting in a decrease of the conduction band energy and hence the total bandgap energy. This decrease in the bandgap energy will improve the electrons injection efficiency from emitter to collector, consequently enhancing the maximum current gain (Figure 7.21b) and high frequency characteristics of the device (Figure 7.21c).

Unfortunately, increasing the Ge content at the stress layer will also increase the misfit dislocations between the silicon substrate and the stress layer, which may cause a degradation of the device performance. Therefore, the Ge content at the stress layer must be controlled and chosen carefully to avoid such problems. Moreover, the impact of changing the device's emitter width on the device performance has been studied. The result illustrates that increasing the emitter width will decrease the stress values induced at the base region of the device as shown in Figure 7.21d, causing a limited enhancement in device performance. Application of strain on the base region of the NPN-SiGe-HBT device have shown that is less efficient in comparison with the BJT as the SiGe base is already stressed due to the existence of Ge at the base.
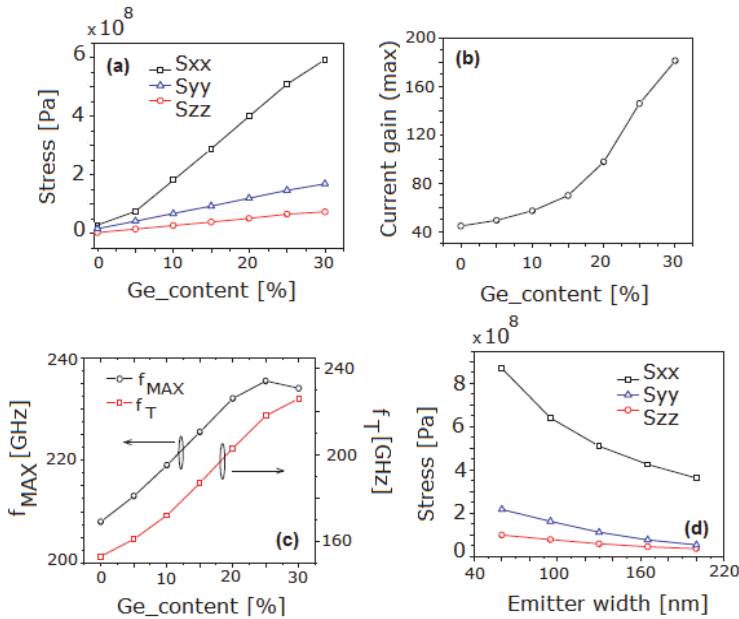
Figure 7.21. (a) Variation of the stress values generated inside the device with Ge content at the stress layer (W_E = 130nm), (b) variation of the maximum current gain with Ge content at the stress layer (W_E = 130nm), (c) variation of f_T and f_{max} with Ge content at the stress layer (W_E = 130nm), and (d) variation of the stress values generated inside the device with the device emitter width. After M. Al-Sadi, TCAD Based SiGe HBT Advanced Architecture Exploration, PhD Thesis, Universite Bordeaux I, 2011.

Summary

New device concepts and novel device architecture for bipolar device performance enhancement via strain engineering technology have been explored using TCAD modeling. The impact of strain engineering in NPN-Si-BJT and NPN-SiGe-HBT devices on the electrical properties and frequency response has been studied in detail. This chapter has focused on the material and electrical characterization of strained-Si HBTs and compared the results with the co-processed SiGe HBTs and Si BJTs. Using a strain-relaxed buffer to grow strained-Si in the device has led to significant improvement in the performance of the bipolar transistor. A strained-relaxed buffer has been used to improve SiGe heterojunction bipolar transistor, through the increase of the amount of Ge incorporated in the base. Simulation results showed that the strained

silicon BJT/HBT devices exhibit better DC performance and high frequency characteristics in comparison with equivalent standard conventional BJT/HBT devices. An approximately 42% improvement in f_T and 13% improvement in f_{max} have been achieved for BJT device employing strain at the base region.

An enhancement of the collector current by nearly three times in strained silicon BJT device has been attained. It has been shown that utilizing a strain layer at the device's collector region has resulted in a ~9–14% improvement in f_T and 7–12% improvement in f_{max} in comparison with a conventional NPN-SiGe-HBT.

Chapter 8

FinFETs

A significant challenge associated with the downscaling of MOSFETs is the poor electrostatic control of a single gate contact over the channel of ultra-short devices called short channel effects. For digital applications the most relevant scaling issues include prevention of short channel effects, controlling the leakage currents through the gate dielectric in the on- and off-state of the transistor, improving the inverse sub threshold slope for faster and more efficient switching of the devices and reducing the power voltage supply. Double-gate FinFET transistors are recognized as one of the most promising successors of traditional planar bulk devices in the sub-22nm regime due to the significantly reduced leakage current, excellent short channel behavior, and compatible fabrication process with existing SOI or bulk technology. FinFETs have significantly better current drive capability than bulk devices but has a higher gate capacitance.

In the post-planar transistor era, chip power continues to be a major challenge and the need for low-voltage transistors is ever more important. Development of nanoscale MOSFETs has given rise to alternative device concepts, like ultra-thin body MOSFETs, FinFETs, or nanowire MOSFETs in order to complement the ongoing scaling. Currently, there is a great interest in FinFET technology for lower cost and good compatibility with planar CMOS [8.1, 8.2]. FinFET was originally developed on SOI, but recently, there is strong interest in forming FinFET on bulk Si for lower cost and better compatibility with planar CMOS. Intel's 22nm CMOS announced in 2011 (Ivy Bridge) is the first commercially available bulk FinFET technology which opened a new era of 3D CMOS for the low power applications. In contrast to conventional planar devices, FinFETs feature multiple gates surrounding the channel which consequently lead to a higher current drive per unit

gate area. Bulk FinFET has the advantages of cheaper wafer cost, ease of combination with conventional bulk CMOS. FinFET has around 2× gate capacitance for the same device size leading to increased power, higher on current. Optimal device sizing is very critical to harness the power for power reduction.

Figure 8.1 shows a comparison between conventional transistor architecture and the FinFET. In FinFET technology, the transistors are raised from the substrate into structures known as fins. The gate wraps around the transistor channels from three directions, thus providing better electrostatic control over the channel. However, the three dimensional nature of the transistor architecture results in unwanted variations in channel stress in different directions. Multi-gate device architectures enable improved electrostatic control of the channel region. This implies better control of short channel effects and a better scalability with less extreme channel and gate oxide thicknesses.

Multi-gate architectures has been introduced by Intel for the 22nm technology node can help suppress SCE, even at short gate lengths; deliver near-ideal sub threshold slopes; and reduce drain-induced barrier lowering. Possible substrate types of FinFET are shown in Figure 8.2.
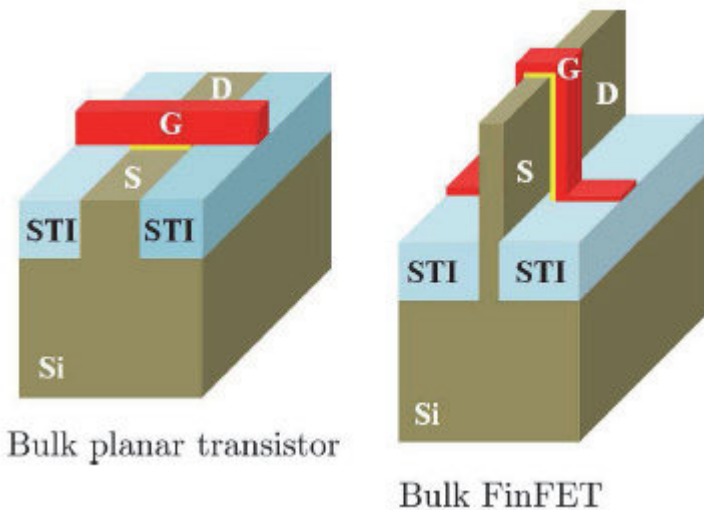
Figure 8.1. Comparison of bulk planar transistor and bulk FinFET. The gate oxide is shown in yellow regions.
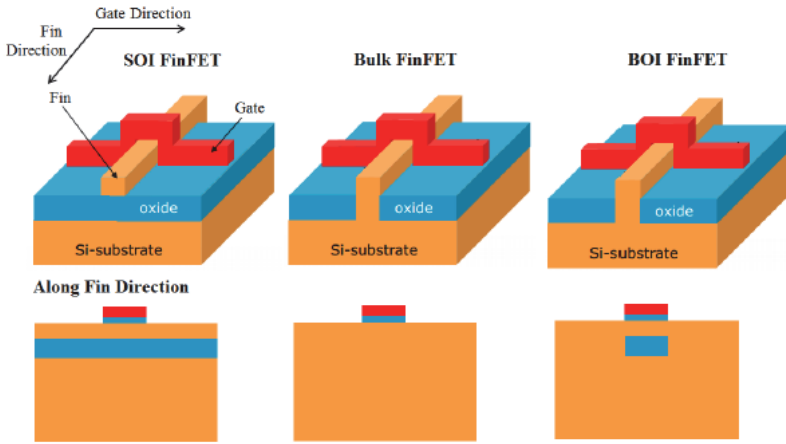
Figure 8.2. Possible substrate types of FinFET. After P. Feng, Design, modeling and analysis of non-classical field effect transistors, PhD Thesis, Syracuse University, 2009.

These favorable features make SOI FinFET promising towards microprocessor performance improvement and energy conservation at 14nm node. The 3D fin can be fabricated on silicon-on-insulator or by standard bulk CMOS technology. The bulk FinFET structure has a lower wafer cost and wafer defect density compared to the SOI technology. It also eliminates the floating body effect and improves heat dissipation. In comparison, SOI FinFET demonstrates an easier fin patterning, a higher I_{ON}/I_{OFF} ratio and a better immunity to variability issues while scaling down the dimension with a 10nm gate length.

The FinFETs, which are vertical double-gate MOSFETs, were firstly developed by researchers at the University of California, Berkeley. The conductive channel is a thin wall with the "fin" shape wrapped by gate material, enabling good electrical control of the carriers from more than one direction (as shown in Figure 8.3a). The device is considered as vertical because two gates locate vertically on the two sides of the channel. The channel length is decided by the thickness of the gates. Using this design, the source-drain leakage current of the transistor is dramatically reduced, and the short channel effect is alleviated. Based on the FinFETs, multi-gate technologies have been developed by leading semiconductor companies. The structure of the tri-gate transistors built

by Intel is shown in Figure 8.3b, where multiple fins are connected together. The tri-gate transistors have been applied in Intel's products since the commercialization of the 22nm technology and are considered to be the promising architecture for future scaling. The non-planar feature of FinFET encourages the exploration of the multi-gate structure. The triple-gate structure actives the top gate control by shrinking the corresponding gate dielectrics. Triple-gate and double-gate transistor are recognized as feasible structures for fabrication compared to the other four types of MOSFET.

Low leakage transistors are essential for mobile and other power constrained platforms. In this chapter we shall examine how fin shape impacts transistor leakage in bulk tri-gate n-FinFETs with thin fins. In particular, the interplay between fin body doping and fin shape is examined and exploited to design minimal leakage FinFETs. By combining triangular fins with existing gate-source/drain underlap multi-threshold techniques, design of ultra-low-power n-FinFETs with high performing I_{ON}/I_{OFF}, threshold voltage, and sub threshold swing is discussed.
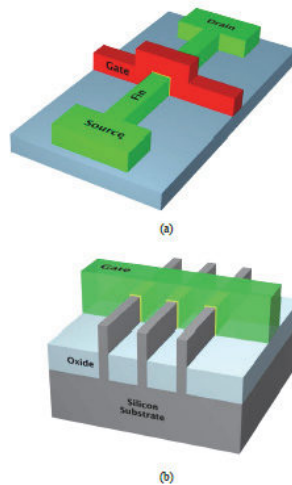
Figure 8.3. (a) The structure of FinFETs. (b) Tri-gate technology by Intel where multiple fins are connected together. After L. Han, Investigation of gate dielectric materials and dielectric/silicon interfaces for metal oxide semiconductor devices, PhD Thesis, University of Kentucky, 2015.

In this chapter we shall address briefly the manufacturing issues of advanced CMOS process and present process solutions to the new and novel set of problems faced in 28nm planar and 22/14nm FinFET technology. Recent development efforts on FinFET technology are discussed. FinFETs utilize a fin-shaped body perpendicular to the wafer surface to carry the current. The fin is sandwiched between the front gate and back gate, and is made very thin to geometrically suppress the short channel effect. In the ideal case the bodies of both FDSOI and FinFET are undoped, which eliminates random dopant fluctuation as a component of threshold variation. Fin uniformity is critical for SCE and variability control. FinFETs on SOI advantages are:

 Electrostatic benefits

- Low voltage, high threshold operation
- Low doping benefits like FDSOI
- High mobility
- Low RDF
- Some challenges similar to FDSOI
- Epitaxial source/drain and access resistance

FinFET technology parameters are:

- Fin Width – determines short channel effect
  - Fin Height – limited by etch technology
  - Trade-off: layout efficiency vs. design flexibility
- Fin Pitch – determines layout area
  - Limits S/D implant tilt angle
  - Trade-off: performance vs. layout efficiency

FinFETs show the following advantages:

- Better subthreshold swing
- Better short channel control
- Negligible body effect
- Excellent SCE control
- Scalability
- Double-gates are self-aligned

- Insensitivity to channel doping

Limitations:

- Gate material
- Contact/Series resistance
- Area efficiency (fin pitch)

8.1 Strain Effects on FinFETs

Hole transport in the inversion layer of silicon p-MOSFETs under arbitrary stress and device surface orientation is discussed in this section. Piezoresistance coefficients are calculated and measured at stress up to 300 MPa via wafer-bending experiments for stresses of technological importance: uniaxial compressive and biaxial tensile stress on (001)- and (110)-surface oriented devices. The results show that biaxial tensile stress degrades the hole mobility at low stress due to the quantum confinement offsetting the strain effect. Uniaxial stress on (001)/⟨110⟩, (110)/⟨110⟩, and (110)/⟨111⟩ devices improves the hole mobility monotonically. Unstressed (110)-oriented devices have superior mobility over (001)-oriented devices due to the strong quantum confinement causing smaller conductivity effective mass of the holes. When the stress is present, the confinement of (110)-oriented devices undermines the stress effect, hence the enhancement factor for (110)-oriented devices is less than (001)-oriented devices. Hole mobility enhancement saturates as the stress increases. At high stress, the maximum hole mobility for (001)/⟨110⟩, (110)/⟨110⟩, and (110)/⟨111⟩ devices is comparable.

Piezoresistance coefficients are widely used as an effective approach characterizing the resistance change at low stress. The piezoresistance coefficients are obtained through the linear regression of the measured resistance vs. stress. The actual strain in the devices is measured through the resistance change of a strain gauge mounted on the sample, and via the laser-detected curvature change of bent wafer. Both measured and calculated results show that under uniaxial longitudinal stress, (110)/⟨111⟩ devices have the largest piezoresistance coefficient,

followed by the (001)/⟨110⟩ devices. The piezoresistance coefficient of (110)/⟨110⟩ devices is the lowest. Under uniaxial transverse stress, the piezoresistance coefficients are smaller than longitudinal stress. The biaxial tensile strain increases the channel resistance and hence degrades the hole mobility at low stress.

Surface and channel orientation dependence of electron and hole mobility has been investigated experimentally. It has been reported that for p-type devices with ⟨110⟩ channel, the mobility is the highest in (110)-oriented and lowest in (001)-oriented p-MOSFETs. The calculated hole mobility vs. the effective electric field of unstressed (001)/⟨110⟩ and (110)/⟨110⟩ Si p-MOSFETs are compared with experimental mobility characteristics in Figure 8.4.
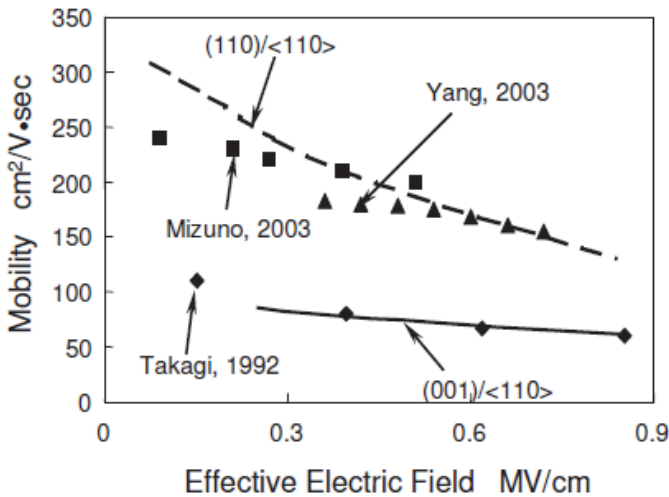
Figure 8.4. Hole mobility vs. inversion charge density for relaxed silicon. Both measurements and simulation show larger mobility on (110) device. After G. Sun, Strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.
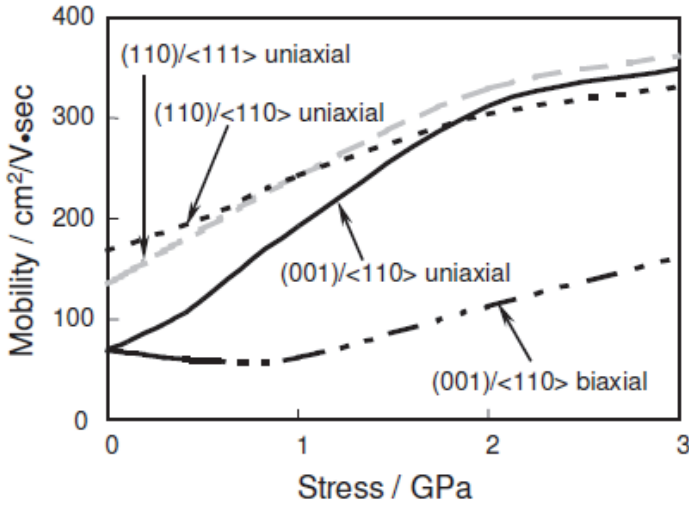
Figure 8.5. Hole mobility vs. stress with inversion charge density $1\times10^{13}\text{cm}^{-2}$. The enhancement factor is the highest for $(001)/\langle110\rangle$ devices and lowest for $(110)/\langle110\rangle$ devices. At high stress (3 GPa), three uniaxial stress cases have similar hole mobility. After G. Sun, Strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.
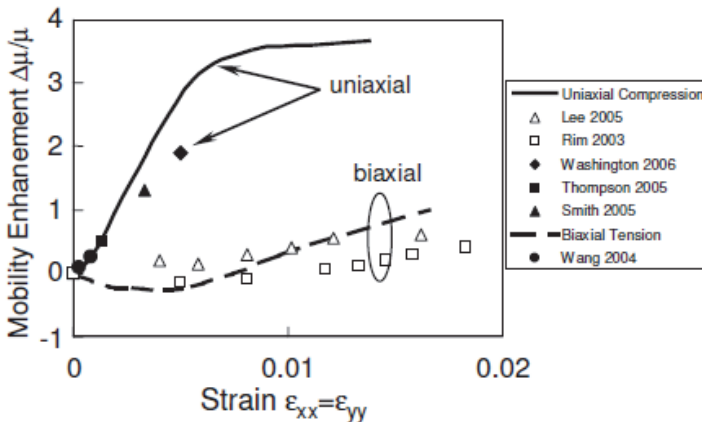
Figure 8.6. Calculated strain induced hole mobility enhancement factor vs. experimental data for (001)-oriented p-MOSFET. After G. Sun, Strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.

Figure 8.5 shows the hole mobility vs. (up to 3 GPa) stress at inversion charge density $p_{inv} = 1 \times 10^{13} cm^{-2}$ and channel doping density $N_D = 1 \times 10^{17}$ cm^{-3}) for $(001)/\langle 110 \rangle$, $(110)/\langle 110 \rangle$, and $(110)/\langle 111 \rangle$ p-MOSFETs. Uniaxial compressive stress improves the hole mobility monotonically as the stress increases. The hole mobility enhancement saturates at large stress (3 GPa). Under uniaxial longitudinal compressive stress, the maximum hole mobility enhancement factor is 350% for $(001)/\langle 110 \rangle$ p-MOSFETs, 150% for $(110)/\langle 111 \rangle$ p-MOSFETs, and 100% for $(110)/\langle 110 \rangle$ p-MOSFETs. At 3 GPa uniaxial stress, (001) and (110) p-MOSFETs have comparable hole mobility. Under biaxial tensile stress, the maximum hole mobility enhancement factor is about 100%. Calculated strain-induced hole mobility enhancement factor of (001)-oriented p-MOSFETs is shown in Figure 8.6 comparing with experimental data. Good agreement is found between the calculated and measured data.
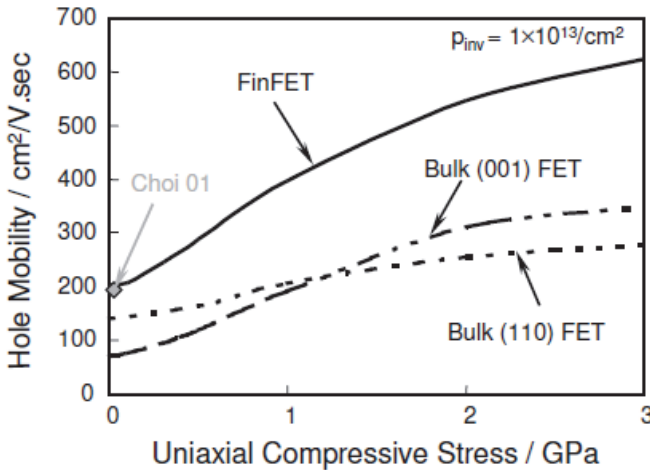
Figure 8.7. Hole mobility of FinFETs under uniaxial stress compared with bulk (110)-oriented devices at charge density $p = 1 \times 10^{13} cm^{-2}$. After G. Sun, Strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.
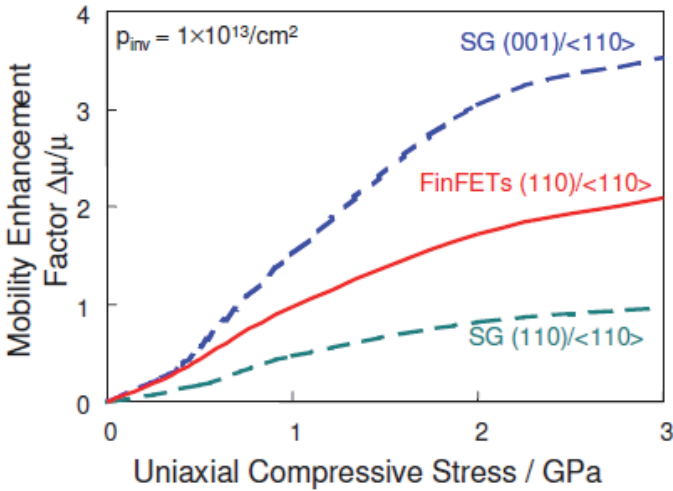
Figure 8.8. Hole mobility enhancement factor of FinFETs under uniaxial compressive stress at charge density $p = 1 \times 10^{13} \text{cm}^{-2}$. After G. Sun, Strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.

The total hole mobility of the FinFET with respect to the stress is shown in Figure 8.7, comparing with the single-gate (110)- and (001)-oriented p-type devices at the inversion charge density of $1 \times 10^{13} \text{cm}^{-2}$. When 3 GPa uniaxial compressive stress is applied to a FinFET, about 300% enhancement of the mobility is expected, compared to only 200% enhancement for a bulk (110)-oriented transistor as shown in Figure 8.8. Even though the (001)-oriented p-MOSFET shows greater relative enhancement (over 400%), the absolute mobility is still lower than that of the FinFET due to its low mobility with no stress.

To understand the hole mobility difference between FinFETs and traditional single gate (110)/⟨110⟩, the hole mobility gain contribution from effective mass change and phonon scattering rate change is shown in Figure 8.9. It shows that phonon scattering rate change is the main factor to improve the hole mobility for both FinFETs and bulk p-MOSFETs. Both the effective mass and phonon scattering rate for FinFETs change are larger than single gate (110)/⟨110⟩ devices, which leads to higher mobility enhancement. Smaller surface roughness
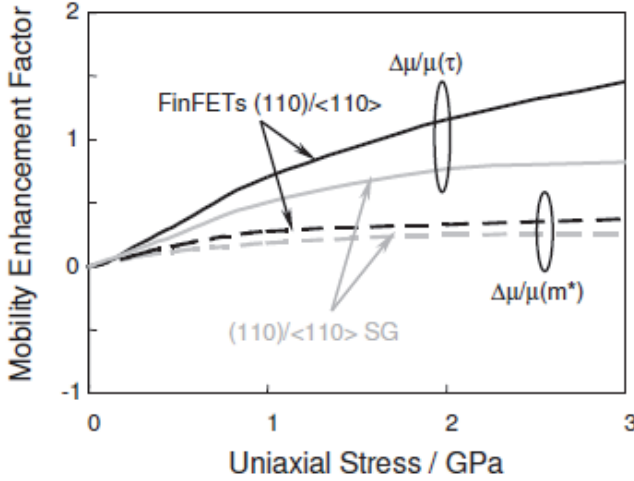
Figure 8.9. Hole mobility gain contribution from effective mass and phonon scattering suppression under uniaxial compression for (110)/⟨110⟩ FinFETs compared with SG (110)/⟨110⟩ p-MOSFETs at charge density $p = 1 \times 10^{13} \text{cm}^{-2}$. After G. Sun, strain effects on hole mobility of silicon and germanium P-type metal-oxide-semiconductor field-effect-transistors, PhD Thesis, University of Florida, 2007.

scattering rate due to small electric field in FinFETs also contributes to the higher mobility enhancement.

8.2 Stress-Engineered FinFETs

Since the 90nm CMOS technology node, the strained nitride capping layer is used as a stress-engineering booster enabling transistors improvement. The effectiveness of the nitride layers constrained with the source/drain SiGe will be discussed for MOSFETs. In the following, we present a simulation study explaining how the strained nitride capping layer transmits its intrinsic stress to the Si-channel for improving the device performance. FinFETs are an attractive replacement for MOSFETs due to their superior electrostatics compared to their planar counterpart. Strain-engineered FinFETs provide higher drive current along with immunity to short-channel effects. Uniaxial tensile strain combined is found to give the best electron mobility enhancement in

n-FinFETs. Besides extending the fin-based design, augmenting the FinFET for improved performance allows one to evolve the process infrastructure for a few more technology nodes. In the following, we use SILVACO tools to create a 3D FinFET using VictoryCell, calculate the internal stress using VictoryStress, and analyze the electrical characteristics using VictoryDevice.

First, VictoryCell is used to generate the FinFET structure from the specifications provided in a layout file. The generated FinFET structure is of prismatic type. Stress analysis is performed by the VictoryStress tool in which material command defines different elastic properties of the materials used. For this particular example FinFET orientation is $(100)/\langle100\rangle$, i.e., (100) silicon wafer and $\langle100\rangle$ direction of the fin orientation will used. VictoryCell was used to build up a FinFET structure with a 50x50nm fin with 1µm in length. The fin was deposited on a SiO_2 substrate layer and a 2nm gate isolation layer separated it from the 50nm polysilicon gate crossing it at right angles. A 100nm Si_3N_4 capping layer was deposited on top of the structure. Figure 8.10 shows the 3D FinFET. Stress analysis is performed using VictoryStress by one-step model simulation (see Section 5.3) and once the stress analysis is complete the three-dimensional structure is saved and is plotted using Tonyplot3D as shown in Figure 8.11. Figure 8.12 shows the stress Sxx in the nitride layer. Figure 8.13 shows the stress developed in the fin itself.

Effects of stress/strain on the electrical properties of a 50nm bulk silicon FinFET is studied next. In the simulation example, a 50nm bulk silicon FinFET is considered. The structure created by VictoryCell is then imported in VictoryStress which was used to perform a stress analysis over the whole FinFET device structure. The Si_3N_4 capping layer was set to have uniform hydrostatic tensile stress of 1 GPa using the command material nitride intrin. sig $= 1\times10^{10}$. The evaluation of mobility enhancement factors along the fin is based on a full 3D piezoresistive model as discussed in Section 3.3.
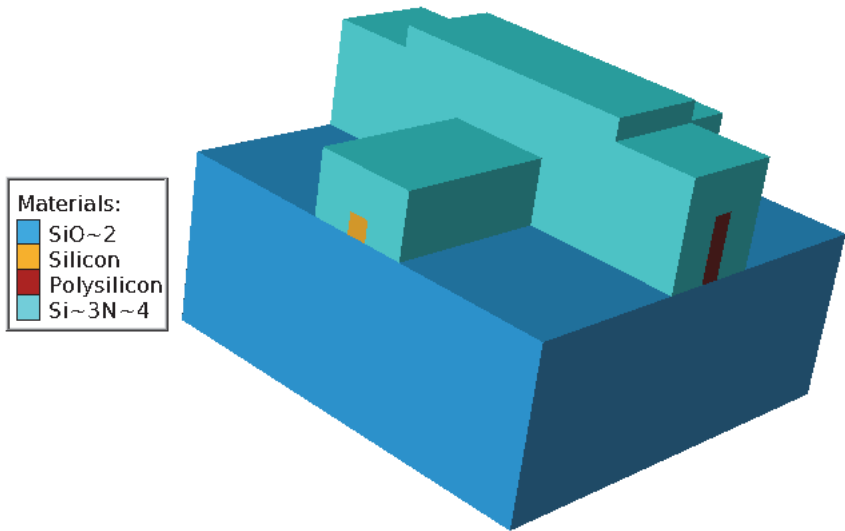
Figure 8.10. VictoryCell generated 3D FinFET structure.
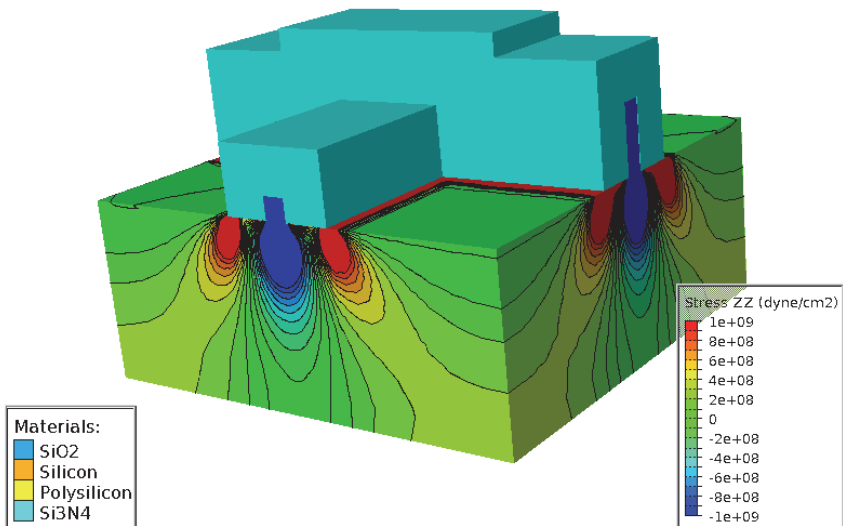
Figure 8.11. VictoryStress generated stress transfer in the FinFET.
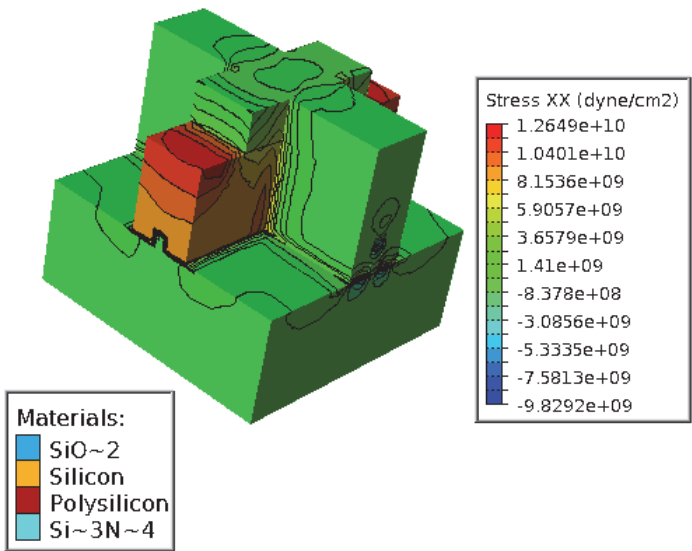
Figure 8.12. Stress distribution in the total structure.
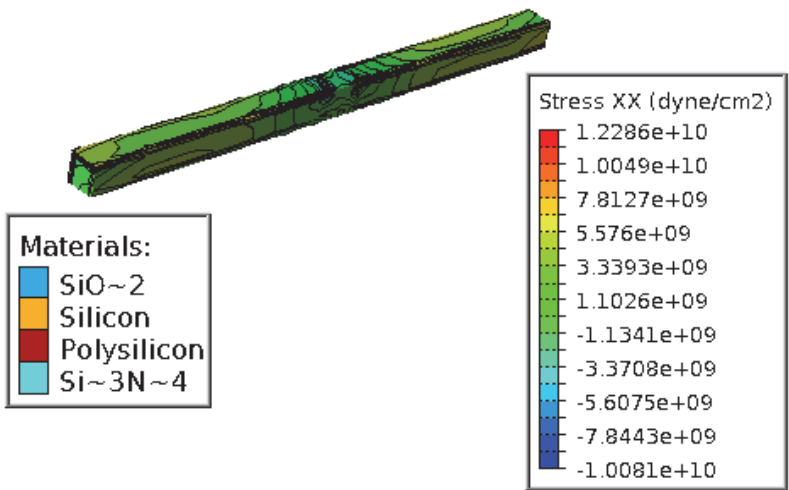
Figure 8.13. Stress distribution in the fin.

The structure, along with the stress/strain and mobility enhancement data, was then imported in VictoryDevice for device simulation and

analysis. Source and drain electrodes were defined at either end of the Silicon fin, the Polysilicon region was defined to be the gate, and we put a substrate contact at the bottom of the device. The work function of the gate used in simulation was 4.17 eV. In simulation, besides the standard silicon mobility models such as, CVT, it is possible to use the strain dependent mobility enhancement models such as "nhance" and "phance" were defined in the mobility statement for this simulation. The mobility enhancement models apply the second order mobility enhancement tensor (calculated by VictoryStress) directly to the low field mobility. This results in directionally dependent (anisotropic) electron and hole mobilities.
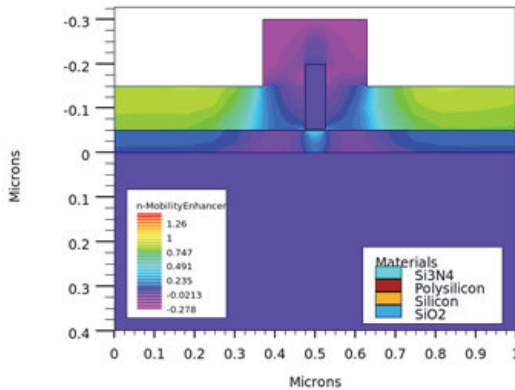
Figure 8.14. 2D cut plane electron enhancement factor in FinFET.
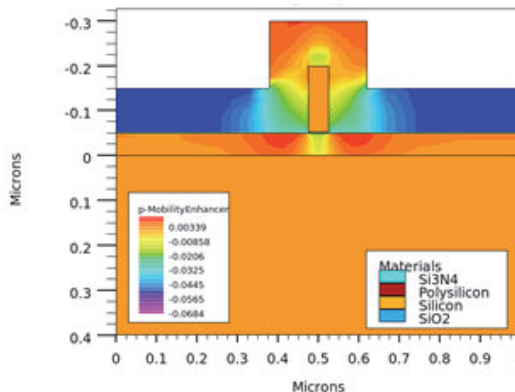
Figure 8.15. 2D cut plane hole enhancement factor in FinFET.
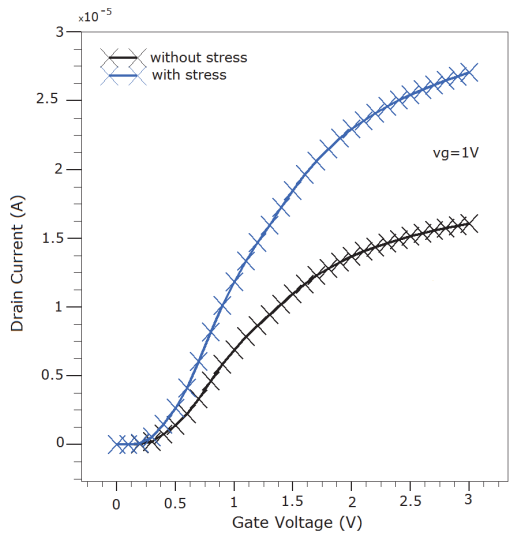
Figure 8.16. Electrical characterization (I_d–V_g) using device simulation tool VictoryDevice.
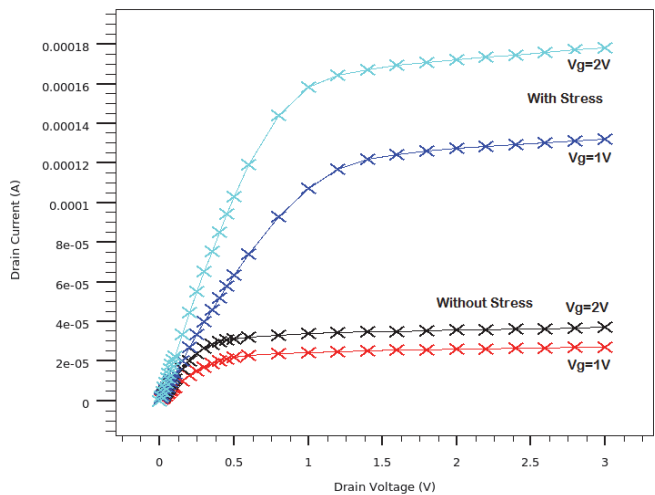
Figure 8.17. Electrical characterization (I_d–V_d) using device simulation tool VictoryDevice.

Figures 8.14 and 8.15 show the 2D cut plane electron and hole enhancement factors. Device simulation results are shown in

Figures 8.16 (I_d–V_g plot) and 8.17 (I_d–V_d plot) in which the effects of stress/strain on drive current is found to increase as expected.

8.3 Stress Simulation in n-FinFET

The 3D stress distributions in the structures were simulated with the following values of intrinsic stress: 1GPa (tensile) for liners and S/D regions in n-type device and −1GPa (compressive) for both types of stressors in p-type devices. This value of intrinsic stress roughly corresponds to compressive $Si_{0.90}Ge_{0.10}$ and tensile $Si_{0.99}C_{0.01}$ (Young's modulus E = 187 GPa and Poisson ratio ν = 0.28 were used). The elastic properties of the liners correspond to those of Si_3N_4. The elastic properties of silicon fins correspond to those of isotropic silicon. Though VictoryStress can take into account anisotropic elastic properties of silicon and other materials, this anisotropy has a very small effect in silicon because its tensor of elasticity varies insignificantly for different crystallographic orientations. However, the mobility enhancements depend on silicon crystal orientation because the tensor of piezoresistivity varies significantly for different crystallographic orientation. It has been shown that piezoresistance mobility model can describe the stress impact on transistor performance with good accuracy. The piezoresistance model provides accurate stress-dependent mobility values (within about 20%) at stress levels below 1 GPa. The following relations were used for the calculation of n- and p-mobility enhancement factors for $(100)/\langle100\rangle$ and $(110)/\langle100\rangle$ crystallographic orientations in silicon:

$$\mu_n^{100} = 1.0 - (-1.022*\sigma_{xx} + 0.534*\sigma_{yy} + 0.534*\sigma_{zz})$$
$$\mu_p^{100} = 1.0 - (0.066*\sigma_{xx} - 0.011*\sigma_{yy} - 0.011*\sigma_{zz})$$
$$\mu_n^{110} = 1.0 - (-0.311*\sigma_{xx} - 0.175*\sigma_{yy} + 0.534*\sigma_{zz})$$
$$\mu_p^{110} = 1.0 - (0.718*\sigma_{xx} - 0.663*\sigma_{yy} - 0.011*\sigma_{zz})$$

In these relations the stresses are in units of GPa and coefficients of piezoresistivity are in 1/GPa. The most common method of introducing desirable stresses in the transistor channel is the deposition of high tensile or high compressive films of nitride type materials. A tensile film (when intrinsic stress value is positive) improves drive current in

n-MOSFETs due to enhancement of electron mobility whereas a compressive film (when intrinsic stress value is negative) enhance the hole mobility for p-MOSFETs. The following simulation example shows stress simulation in a single n-FinFET. We start with process simulation which includes following steps:

- Definition of geometrical domain and region limits for subsequent consistent mesh formation
- VictoryCell process steps
- Cartesian mesh formation and saving the structure for stress calculation
- Stress calculation
- Visualization and analysis of simulation results, extraction of average stresses and mobility enhancement factors

Common method of stress engineering is by the introduction of so-called Source/Drain stressors (SiGe or SiC). The structure in this example allows to have such stressor and is called "nstressor". The thickness of the stress liner over the gate is very important as it determines the amount and the stress distribution inside the fin. The FinFET contains the vias which are placed over the metal contacts (aluminum). For simplicity the length of each via (for source and for drain) was chosen the same as the length of stressors (0.3μm). Two important considerations to form an optimal mesh for Manhattan-type structures are: a) reasonable number of mesh points (spacing) should be generated across areas of interest such as, gate and fin and b) it is preferable that in each direction a mesh line coincides with material region boundary.

For simulation the following n-FinFET parameter were used:

gate_length = 0.06
gate_height = 0.05
stress_liner_height = 0.05
fin_width = 0.03
fin_height = 0.06
fin_length = 0.8+$gate_length
oxide_box_height = 0.6
oxide_top_height = 0.15
spacer_thick = 0.005

met_height = 0.005
stress_value = 1.0e10

The following process steps were used for the creation of FinFET structure using VictoryCell. The 3D view of n-FinFET device on BOX substrate and active area of device with Sxx stress distribution is shown in Figure 8.18.

- Formation of the fin
- Source side local stressor (i.e., compositional SiGe)
- Drain side local stressor
- Oxide spacer deposition
- Gate formation
- Deposition of metal contact thin layer
- Deposition of stress liner (n-liner)
- Deposition of the top oxide layer
- Etching of top oxide and stress liner for the via formation
- Deposition of metal (aluminum) for formation of via

By default the absolute value of the initial intrinsic stress is 1GPa; for the n-FET it is positive; for the p-FET it is negative. The following material parameters are used for user-defined materials in the structure:
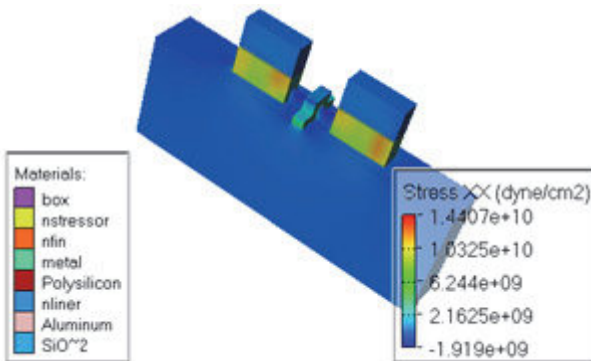
Figure 8.18. The 3D view of n-FinFET device. Only BOX substrate and active area of device with Sxx stress distribution is shown.

material Silicon nitride ("nliner") young.m = 3.89×10^{12}
material Silicon nitride ("nliner") poiss.r = 0.3

material oxide ("box") young.m = 6.6×10^{12}
material oxide ("box") poiss.r = 0.2

material silicon ("nfin") young.m = 1.67×10^{12}
material silicon ("nfin") poiss.r = 0.28

material silicon ("nstressor") young.m = 1.67×10^{12}
material silicon ("nstressor") poiss.r = 0.28
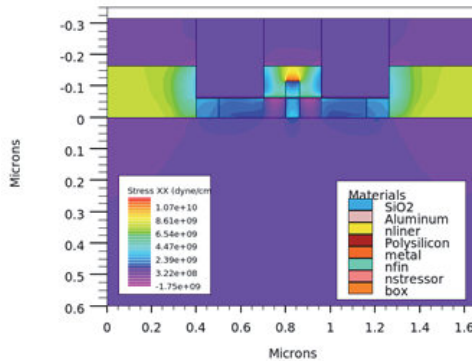
Figure 8.19. 2D side wall Sxx stress distribution through center of n-FinFET.
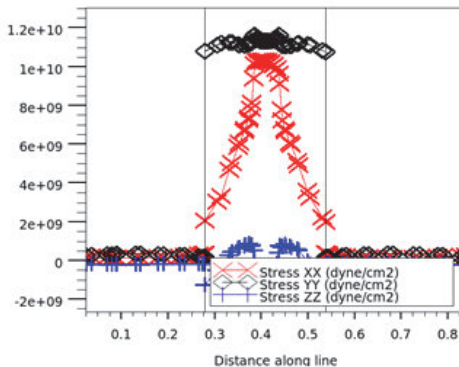
Figure 8.20. 1D top side stress distribution under the gate.

For visualization and analysis of simulation results, extraction of average stresses and mobility enhancement factors 2D cut plane is desirable. Since the simulation results are in 3D, the only convenient way of analyzing them is to extract figures of merit from inside using cut planes within the 3D structure. These 2D planes are generated using "–cut" option of Tonyplot3D. In this example, the two planes selected are parallel to the fin length–one along sidewall of the fin and another along the top of the fin. The exact position of the cut could be set manually using Cut plane setting in Cut plane view menu of Tonyplot3D. The generated cut plane structures with 2D Sxx distribution and 1D stress distribution are shown using TonyPlot. A 2D side wall Sxx stress distribution through center of n-FinFET is shown in Figure 8.19. 1D top side stress distribution under the gate used for calculation of average stress is shown in Figure 8.20. The 2D top side Sxx stress distribution through center of n-FinFET is shown in Figure 8.21. 1D conformal stress distribution under the gate used for calculation of average stress is shown Figure 8.22.
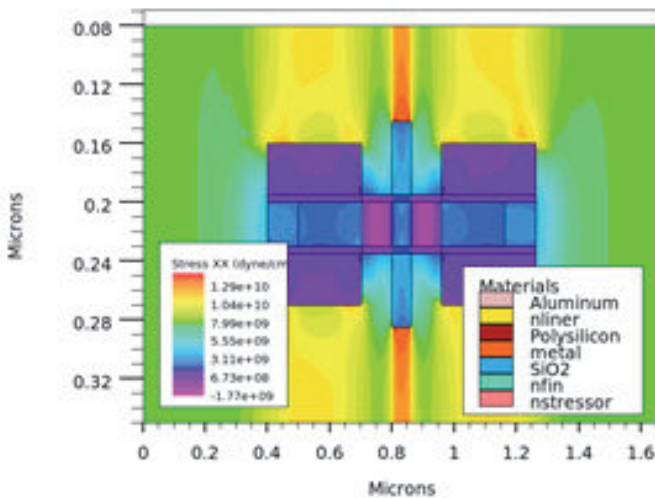
Figure 8.21. 2D top side Sxx stress distribution through center of n-FinFET.
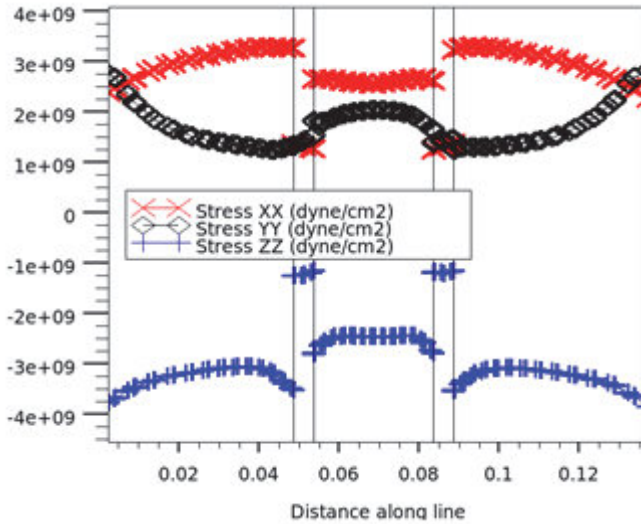
Figure 8.22. 1D stress distribution under the gate used for calculation of average stress.

Average stresses calculations were performed using "factor" (100 in this case) which is required for obtaining integrated stresses in GPa. The Sxx, Syy and Szz-components in the side walls (Figure 8.22) and top side of the fin under the gate were computed by integration of corresponding stress distributions along the cut 5nm below the fin-oxide boundary under the gate. Mobility enhancement calculations were performed using the piezoresistivity coefficients for $(100)/\langle100\rangle$ n-Fin orientation (in cm^2/GPa) values:

nx1=−1.022
ny1=0.534
nz1=0.534

The piezoresistivity coefficients for $(110)/\langle100\rangle$ n-Fin orientation (in cm^2/GPa) values used are:

nx2=−0.311
ny2=−0.175
nz2=0.534

Mobility enhancement factors were calculated at the side wall and top of the Fin under the gate. Mobility Enhancements for n-FinFET with 2 fin orientations were found to be weakly dependent on fin orientation, as the enhancement factors are different because the piezoresistivity coefficients are different. For the simulation of a p-FinFET device having the same structure as n-FinFET, the stresses can be calculated using the same procedure but with different piezoresistivity coefficients.

The piezoresistivity coefficients for $(100)/\langle 100 \rangle$ p-Fin orientation (in cm^2/GPa) are:

px1=0.066
py1=−0.011
pz1=−0.011

The piezoresistivity coefficients for $(110)/\langle 100 \rangle$ p-Fin orientation (in cm^2/GPa) are:

px2=0.718
py2=−0.663
pz2=−0.011

The mobility enhancement factors for 2 substrate orientations are:

n_factor_side_100=1.25044
n_factor_top_100=1.25501
n_factor_side_110=1.22897
n_factor_top_110=1.23924

It is clear that due to stress generated by the stressors mobility enhancement takes place in individual n-FinFETs. However, when different devices (for example p- and n-FETs) are placed in close proximity (for example in a cell) with both n- and p-liners, noticeably different stresses and mobility enhancements can be obtained from simulation. This happens because of the local nature of S/D stressors and the interaction effect for different S/D stressor configurations or for denser cell layouts. 3D process and stress simulation is useful for the detailed performance analysis and optimization of various stress engineering schemes by varying geometry of stressors, substrate

orientation, and material composition of each device as well as by changing location and density of the individual devices inside the cell layout.

8.4 FinFET Design Optimization

3D TCAD FinFET simulations are time consuming and must be used judiciously. FinFET optimization is difficult because parameterized compact SPICE models needed to explore the transistor leakage dependency on fin shape are not available in the literature. Following reference [8.3], three-dimensional simulation of 20nm FinFETs with round fin corners and tapered fin shape is taken up to minimize transistor leakage and study the impact of fin shape on leakage. The transistor base is kept fixed and the top fin width is varied, generating various trapezoidal fin shapes. The fin body doping is optimized first. Device structure used in simulation is based on reported experimental 22nm bulk n-FinFET as shown in Figure 8.23. The key geometry parameters used in simulation correspond to Intel's bulk FinFET as given in Table 8.1. Illustrations of the input parameters that define the geometry and the doping are shown in Figure 8.23. The corner radius of the rounded fin is set to $\frac{1}{2} W_{top}$ to minimize corner effects.

Table 8.1. n-FinFET model geometry parameters. After B. Gaynor, Simulation of FinFET electrical performance dependence on Fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.

Parameter	Description	Value
L	Gate length	34nm
H	Fin height	35nm
W_{bottom}	Active fin bottom width	15nm
W_{top}	Active fin top width	15nm

Sentaurus Device provides advanced stress models to take into account the sub-band modulation by the stress. The TCAD simulations using Sentaurus Device (SDevice) included the following models. Physical models used for stress simulation include:

• Stress-dependent deformation of band structure

- Strained effective mass and density-of-states
- Sub-band stress-dependent electron/hole mobility
- Stress-independent carrier saturation velocity

The drift-diffusion models used included adjusted carrier velocity saturation values suggested in:

- Philips unified mobility
- High-field mobility saturation
- Inversion and accumulation layer mobility with auto-orientation

The FinFET structure offers better gate control, but it increases the gate-induced drain leakage (GIDL) in the off-state. So, the Schenk
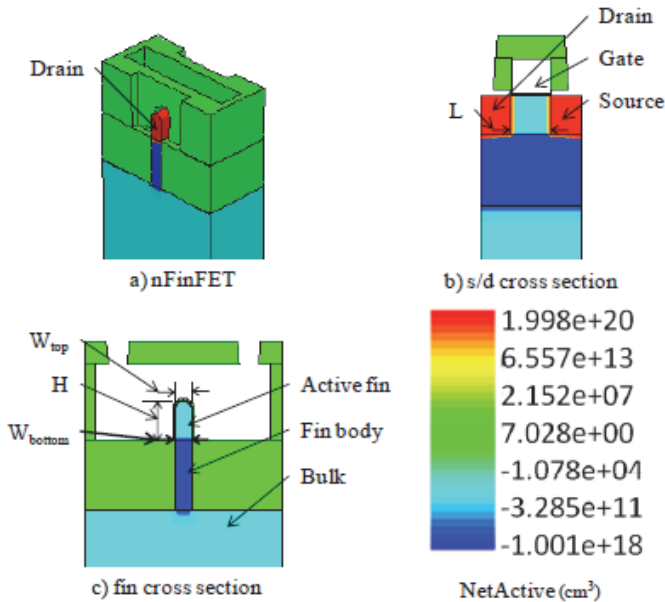
Figure 8.23. Doping concentration of n-FinFET structure: a) isomorphic view, b) source/drain cross section cut at middle of fin, and c) fin cross section cut at middle of channel. After B. Gaynor, Simulation of FinFET electrical performance dependence on fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.

band-to-band tunneling model is selected to simulate the leakage mechanism. The inversion layer mobility has surface orientation dependency. The inversion and accumulation layer mobility model used has auto-orientation accounted for the sidewall-angle dependent surface orientation of the fin, with model parameters calibrated to experimental data. When Si channel thickness becomes thinner than approximately 10nm, low-field mobility is modulated due to the quantum-mechanical confinement effect.

Considering that the width of the top part of the fin goes down to 5nm, the thin-layer mobility model is used. The thin-layer mobility models used were calibrated with parameters to capture quantum-mechanical confinement effects. The top and side surfaces of FinFETs have different crystal orientations. Therefore, the auto-orientation framework is important to consider orientation-dependent quantum correction. The density gradient quantum correction model with auto-orientation captures orientation-dependent quantum corrections calibrated to the solution of the Poisson–Schrödinger equations by the Sentaurus Band Structure.

In simulation the p-type fin body doping concentration is swept from 1×10^{16}–1×10^{19} per cc. The active fin is undoped with a p-type concentration of 1×10^{15} per cc. The results are shown in Figure 8.24. The optimal I_{OFF} is achieved with a fin body doping concentration of 1×10^{18} per cc. The optimal I_{ON}/I_{OFF} ratio also corresponds to a fin body doping concentration of 1×10^{18} per cc, indicating no disproportionate saturation current degradation due to the fin body doping. V_t changes 32mV over the simulated range, with a value of 407mV at the optimum design point. V_t is extracted using the maximum transconductance method with $V_d = 0.05$ V. The sub threshold swing (SS) is 70.5 mV/dec at the optimal design point. The SS is extracted at a gate voltage of 0.01 V.

The interaction of fin body doping and fin shape is examined first. I_{OFF} for the n-FinFET with maximum W_{top}=15nm (rectangular) and minimum W_{top}=1nm (triangular) is plotted as a function of fin body doping as shown in Figure 8.25. The triangular n-FinFET exhibits a 70% reduction in I_{OFF} over the rectangular design. The optimal fin body doping profile for the triangular n-FinFET is 5×10^{17} per cc, less than the optimum doping for the rectangular n-FinFET.
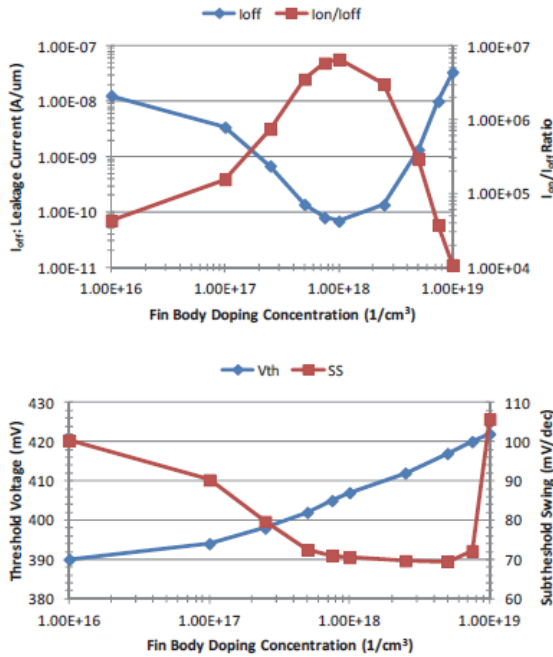
Figure 8.24. I_{OFF} and I_{ON}/I_{OFF} ratio of rectangular n-FinFET as a function of fin body doping (top). V_t and SS of rectangular n-FinFET as a function of fin body doping (bottom). Active fin is undoped with concentration=1×10^{15} per cc. After B. Gaynor, Simulation of FinFET electrical performance dependence on fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.
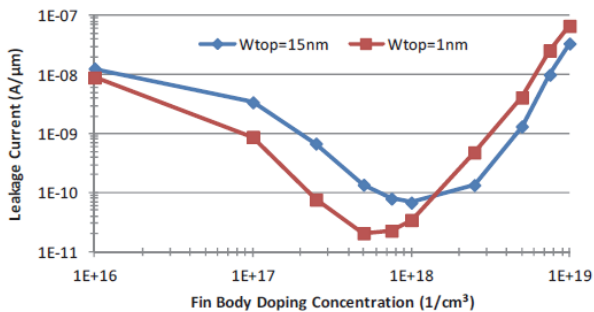
Figure 8.25. I_{OFF} of rectangular and triangular n-FinFET as a function of fin body doping. Active fin is undoped with concentration=1×10^{15} per cc. After B. Gaynor, Simulation of FinFET electrical performance dependence on fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.

The optimal doping concentration for the triangular n-FinFET is reduced due to the shape-dependence of the two dominant and competing leakage mechanisms: SCE and BTBT. As seen in Figure 8.26(a–b), the leakage through and below the active fin increases with the lower doping concentration; however, Figure 8.26(c–d) demonstrates that the BTBT for the triangular n-FinFET is significantly reduced. The optimal design point for the triangular n-FinFET shows a net leakage performance improvement due to the trade-off of non-optimal SCE-induced fin leakage for significantly reduced BTBT.
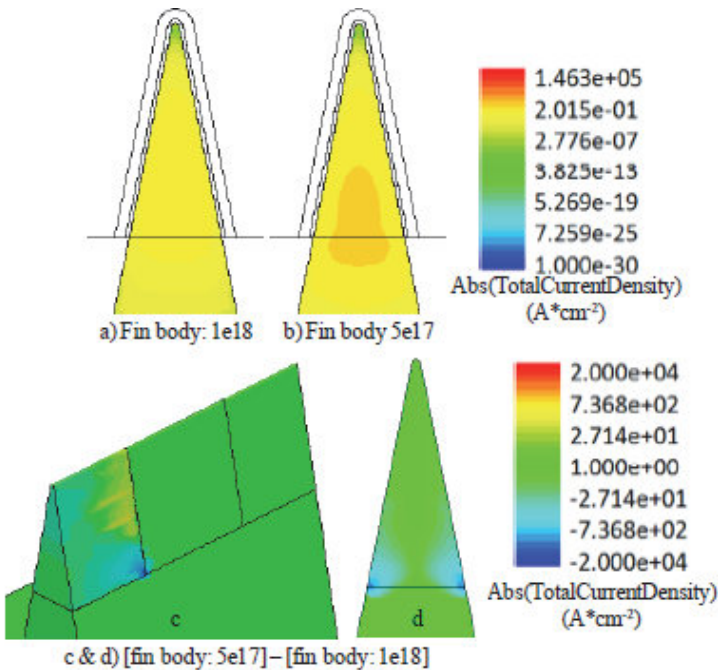
Figure 8.26. I_{OFF} current density distribution of triangular n-FinFET with a) fin body doping concentration=1×10^{18}/cm^3, fin cross section cut at middle of fin, b) fin body doping concentration=5×10^{17}/cm^3, fin cross section cut at middle of fin, c) current density distribution difference between the two n-FinFETs, and d) cross section of c, cut at the drain/channel interface. While a slight increase in leakage current through the fin is observed due to increased doping, the reduction in BTBT at the base of the active fin yields an overall leakage improvement for the triangular n-FinFET with fin body doping of 5×10^{17}cm^{-3}. After B. Gaynor, Simulation of FinFET electrical performance dependence on Fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.

To examine the impact of fin shape on I_{OFF}, W_{top} is swept over the range of 1–15nm. It is seen that I_{OFF} decreases as W_{top} decreases except for the n-FinFET with 1×10^{18} per cc fin body doping when W_{top} is less than 5nm. It may be noted that the leakage current is more sensitive to W_{top} with the lower doping profile. To obtain the optimal leakage performance for various fin shapes, the doping concentration and fin shape must be considered simultaneously to optimize performance.

Figure 8.27 shows the current density distribution of n-FinFET with rectangular, trapezoidal and triangular fin cross sections provides information by which fin shape reduces I_{OFF}. As W_{top} decreases, the leakage current is forced towards the center of the fin volume. Thus the current density is greatly reduced with the thinner fin. It is observed that the maximum current density (Figure 8.27) for the triangular n-FinFET is not in the top, rather it is pushed into the fin volume due to quantum confinement. Figure 8.27 provides insight into the means by which fin shape reduces I_{OFF}. As W_{top} decreases, the leakage current is forced into the center of the fin volume. The current density is greatly reduced with the thinner fin. The maximum current density (Figure 8.28) for the triangular n-FinFET is not in the top, narrowest portion of the fin. Instead, the maximum current density distribution is pushed into the fin

a) W_{top}=15nm   b) W_{top}=7nm   c) W_{top}=1nm
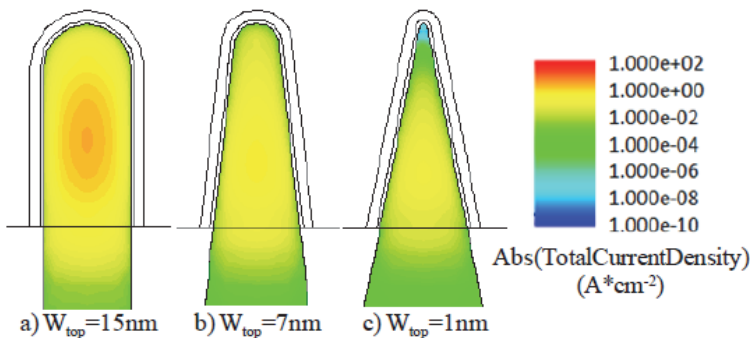
Figure 8.27. I_{OFF} current density distribution of n-FinFET with a) rectangular fin cross section (W_{top} = 15nm), b) trapezoidal fin cross section (W_{top}=7nm), and c) triangular fin cross section (W_{top} = 1nm), cut at middle of channel. Rounded top corners with corner radius=W_{top}/2, fin body doping concentration = 1×10^{18}cm^{-3}. After B. Gaynor, Simulation of FinFET electrical performance dependence on Fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.
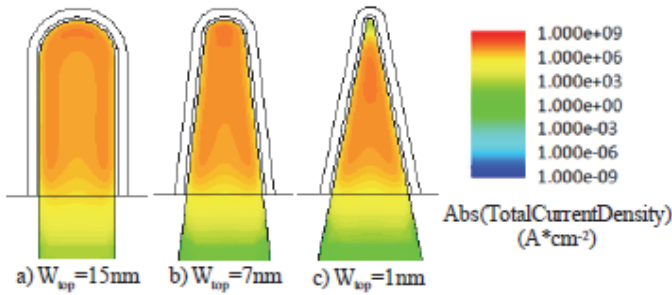
Figure 8.28. I_{ON} current density distribution of n-FinFET with a) rectangular fin cross section (W_{top} = 15nm), b) trapezoidal fin cross section (W_{top} = 7nm), and c) triangular fin cross section (W_{top} = 1nm), cut at middle of channel. Rounded top corners with corner radius = $W_{top}/2$, fin body doping concentration = 1×10^{18}/cm^3. After B. Gaynor, Simulation of finFET electrical performance dependence on fin shape and TSV and back-gate noise coupling in 3D integrated circuits, PhD Thesis, Tufts University, 2014.

volume due to quantum confinement. Corner effects have been minimized by maximizing the corner radius for each W_{top}.

8.5 Variability in FinFETs

Introduction of the multi-gate FET technology has prolonged Moore's law by solving the problem of threshold variability caused by Random Dopant Fluctuations. Strain is extensively utilized since 90nm node which include nitride cap layer, e-SiGe, and STI. The major problem in the future technology scaling is the variations in process parameters that are interpreted as imperfections in the development process. Less frequently discussed in the scientific world is the variability of devices. It is a local form of process variation, meaning that two transistors may have significantly different threshold voltages, or dopant concentrations for instance. As we approach atomic length scales smallest variations in the device structure might result in a fail of the device. Hence, maintaining the device parameters across the chip and from chip to chip will be very critical for mass production.

In the following, we present current technology scaling challenges for FinFETs. First we explain why variation exists and the major sources of

variation. The devices are more sensitive to the environmental changes of temperature and supply voltage as well as to ageing. The statistical variability sources: random discrete dopants, gate line edge roughness and poly-silicon granularity are simulated after careful calibration of fabrication processes and electrical characteristics for n- and p-MOSFETs with 10nm physical gate length.

FinFETs are susceptible to threshold voltage variability due to Line Edge Roughness (LER) due to requirement of very narrow fin widths for required electrostatic control of channel. In FinFETs, fin-width variation may take place due to fin edge roughness (FER), together with traditional variability sources including random discrete dopant (RDD), gate edge roughness (GER), and possible metal gate granularity (MGG), statistical variability is still of great concern for FinFET process and performance. Process variations are classified into subgroups: random or systematic (within-die and die-to-die). When device dimensions are very small, statistical variations come into play. For example, in 10nm technologies, the expected number of dopants in the channel is in the order of 100 atoms. These variations are totally random, and there is no better solution than to observe them statistically. Another type of process variations can be classified as systematic. These variations are deviations from their nominal projected value, but they are characterized by the correlation between devices. They depend on the space and distinct patterns. An example of systematic variations is silicon surface flatness variations caused by the layout density variation during Chemical Mechanical Polishing.

From the physical sources of process variations, we consider the most dominant two; the Random Dopant Fluctuations and Line Edge Roughness. Random Dopant Fluctuations is caused by the imperfection in the development process resulting from the random number of dopant atoms in the channel. For ultra-small devices it is much harder to control the total number of dopants in the channel. This variation in the number of dopants in the transistor channel results in the variation of the threshold voltage. Figure 8.29 illustrates the effects of the random dopant fluctuations. Each small dot present one dopant atom. It can be observed that this number is very small comparing to the channel dimensions. It is seen that doping fluctuations, fin-edge and gate-edge roughness, and gate work function fluctuation are occurring in concert.
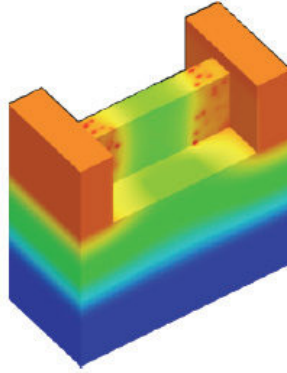
Figure 8.29. FinFET Random Dopant Fluctuations. After Z. Jaksic, Cache memory design in the FinFET era, PhD Thesis, Universitat Politècnica de Catalunya, May 2015.
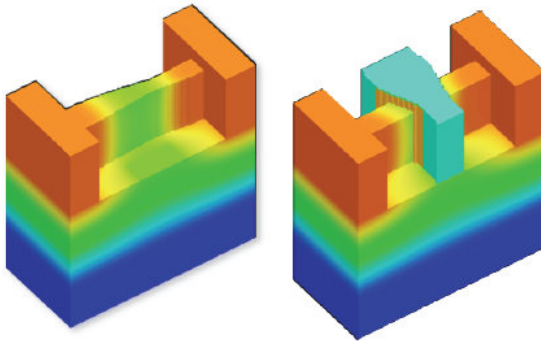
Figure 8.30. FinFET Line Edge Roughness: (left) Fin Edge Roughness (FER) and (right) Gate Edge Roughness (GER). After Z. Jaksic, Cache memory design in the FinFET era, PhD Thesis, Universitat Politècnica de Catalunya, May 2015.

Gate patterning introduces a non-ideal gate edge; this imperfection is referred as Line Edge Roughness (LER) which is expected to be a significant source of variation for 10nm technology node. For FinFET technologies, LER can be considered as Gate Edge Roughness (GER) and Fin Edge Roughness (FER) as shown in Figure 8.30. The features are almost same through the technology nodes, and the variations caused by LER are going to dominate threshold voltage variation for the 10nm SOI FinFETs.

In addition to the process variations described above, any change in environmental conditions, also causes power and performance deviations from their nominal projected values. Common sources of environmental variations are voltage and temperature. For a given power density, the silicon temperature is a function of its thermal conductivity. In a bulk CMOS technology, the generated heat spreads through the silicon substrate as well as the wires. However, in a silicon-on-insulator technology the poor thermal conductivity of the buried oxide causes most of the heat to be carried away primarily along the wires, which causes the temperature to increase at a faster rate. Voltage variations existing in different parts of the chip lead to unbalanced heating of the chip.

Device ageing also affects circuit performance and the most common source of ageing is due to bias temperature instability (BTI). In general, BTI strongly depends on the following factors:

- Voltage – The higher the operating voltage, the higher is the BTI degradation. Therefore, lower operating voltages are desired. However, at low V_{DD} the susceptibility to V_t variations increases.
- Temperature – Reports on BTI show that degradation is higher for higher operating temperatures.

The main characteristic of tri-gate FinFET devices (Figure 8.31) is that the gate wraps the channel from 3 sides. This increases the effective channel width that gives better control of the short channel effects, and a high dopant concentration is not necessary. Because of their 3D structures, tri-gate FinFETs achieve better channel control and higher I_{ON}/I_{OFF} is reached with lightly doped channels. Lightly doped channels reduce the effect of Random Dopant Fluctuations (RDF) which is the greatest source of variability in classical bulk technology. However, some sources of variability caused by Line Edge Roughness (LER) and Metal Grain Granularity (MGG) still remain. Variation in circuit parameters resulting from the effects of Fin Edge Roughness (FER) and Gate Edge Roughness (GER) still pose significant challenge for these devices and its effects are going to worsen as the channel length approaches 10nm. The device is designed to meet the requirements set by ITRS roadmap for high-performance applications. The case study is designed for a 10nm channel length (L); fin height (H_{fin}) and fin thickness (T_{fin}) are 12.5nm and 5nm, respectively. Effective oxide

thickness (EOT) is 0.585nm. Detailed transistor parameters are presented in Table 8.2. Plots have been obtained by TCAD simulations when the variability of RDD, FER, GER, metal gate granularity (MGG) and interface trapped charge (ITC) are incorporated.
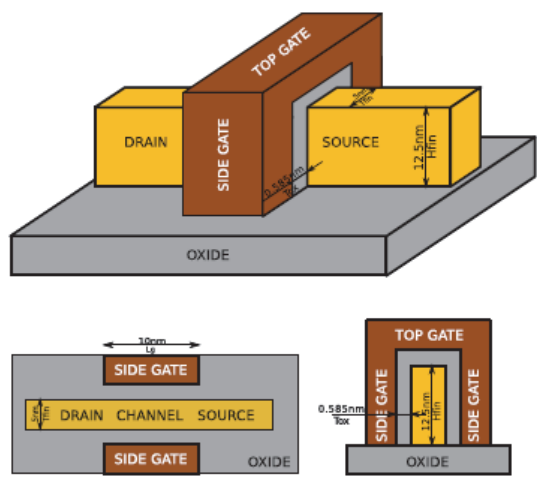
Figure 8.31. 3D structure of SOI Fin-FET considered for simulation. After Z. Jaksic, Cache memory design in the FinFET era, PhD Thesis, Universitat Politècnica de Catalunya, May 2015.

Table 8.2. The FinFET parameters used in simulation.

Technology Node (nm)	11
L (nm)	10
EOT (nm)	0.585
T_{fin} (nm)	5
H_{fin} (nm)	12.5
N_{SD} (cm^{-3})	3×10^{20}
N_{CH} (cm^{-3})	1×10^{15}
V_d (V)	0.8
I_{OFF} (nA/μm)	97
I_{dsat} (μA/μm)	1958
SS (mv/dec)	73.2
DIBL (mV/V)	47

Due to its structure, the simulation of FinFET devices demands certain innovations in the simulation approach especially when process variability is introduced. The detailed simulation procedure, technology assumptions and individual or combined effects of variability source on the device characteristics can be found in reference [3.5]. Numerical simulations based on finite element method can be combined with Monte Carlo simulations to predict the impact of the device variation on the circuit performance. However, these simulations are very time-consuming and applying them to a circuit with a large number of transistors is difficult.

Alternative simulation procedure based on analytical model comprising of a set of complex mathematical equations that describes physical process of the device has been proposed. In general data obtained from the simulations (or physical measurements of real devices) are fitted by set of mathematical equations by carefully setting appropriate parameters. The author has proposed a procedure to obtain variability characteristics of the 10nm FinFETs from TCAD simulations in order to incorporate them in an HSPICE BSIM-CMG card for time-efficient simulation without compromising results.

Simulation study of statistical reliability for PBTI degradation in n-channel fully depleted SOI MOSFETs featuring 22nm, 16nm, and 11nm gate lengths and considering the metal gate-first technology has been reported [8.4]. RDF, LER, and MGG have been included as a statistical variability source in the fresh transistors. Three degradation levels with average interface trapped charge sheet densities of 1×10^{11} cm^{-2}, 5×10^{11} cm^{-2} and 1×10^{12} cm^{-2} are considered in this study. As an example, a typical potential profile corresponding to the trap charge sheet density of 1×10^{12} cm^{-2} in the presence of RDF and LER variability sources is shown in Figure 8.32a and combined RDF, LER and MGG variability sources is shown in Figure 8.32b. The spikes at the sites of trapped charges reduce the local carrier concentrations and thus cause local and global current reduction.
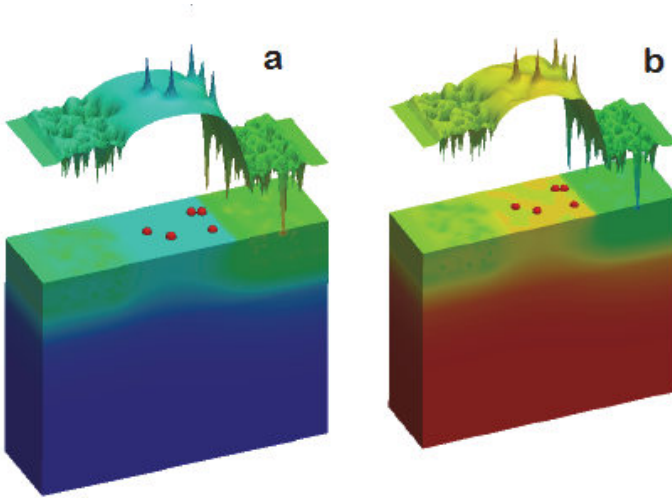
Figure 8.32. The electrostatic potential of 22nm FD-UTB SOI n-MOSFET including RDF and LER with interface-trapped charge density of 1×10^{12} cm^{-2} (a) and combined RDF, LER and MGG at interface-trapped charge density of 1×10^{12} cm^{-2} (b). The trapped charges are shown in red color (W/L=1). After A. S. M. Zain, Scaling and variability in ultra-thin body silicon on insulator (UTB SOI) MOSFETs, PhD Thesis, University of Glasgow, 2013.

The GSS "atomistic" simulator GARAND is used to investigate the statistical variability and reliability. For the FinFET, 1000 instances are analyzed to minimize statistical error (approximately 95% confidence for 3σ). Nominal supply voltage is $V_d = 0.8V$. The first step is to load 1000 TCAD transfer characteristics data (1000 I_d–V_g plots obtained by the TCAD variability simulation of the FinFET devices). Next, we pick up a random value of standard deviation for different parameters in HSPICE model card (H_{fin}, T_{fin}, L and EOT). Then, perform HSPICE DC simulation with 1000 Monte Carlo samples and compare frequency distributions of I_d for every value of gate-source bias. This procedure is repeated iteratively until the difference (or similarity) between the outputs is below a certain threshold. After the iterative process, this procedure achieves an error of less than 5% for 3σ and less than 2% for 5σ. Simulation methodology followed was:

3D process simulation
    LER is introduced at gate patterning
    Highly tensile cap layer is deposited as main stressor in n-MOSFET

Drift-diffusion device simulation
  Doping, vertical and lateral
  Field dependent mobility
  Density gradient quantum corrections

Strained devices:
  + Stress mobility enhancement

Since BSIM-CMG is surface-potential-based model, there is no threshold voltage parameter, as such an alternative for process variability in simulation was considered. Process variation in H_{fin} (fin height) and T_{fin} (fin thickness), EOT (oxide thickness), L (channel length). We assume a normal distribution of these random variables with the mean value equal to the nominal value of the device parameter. A normal distribution of these random variables with the mean value equal to the nominal value of the device parameter was assumed. The procedure to determine the standard deviations of these random variables that would approximate process variability with the highest accuracy is shown in Figure 8.33.
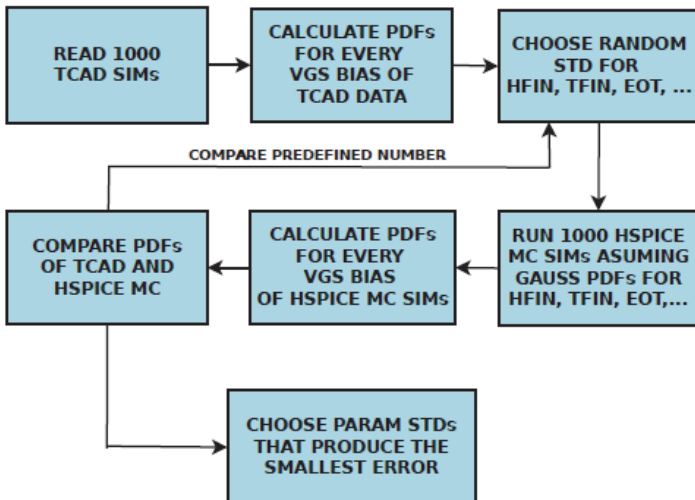
Figure 8.33. FinFET variability calibration procedure. After Z. Jaksic, Cache memory design in the FinFET era, PhD Thesis, Universitat Politècnica de Catalunya, May 2015.

Summary

Strain/stress engineering has been actively pursued since 2002 to enhance transistor performance. Stress engineering techniques with very high stress levels require accurate models for studying the stress effects. To overcome the short channel effects present in conventional 2D transistors, transistor architectures have now evolved from conventional planar structures to three-dimensional structures. In this chapter, a comprehensive simulation based scaling study of FinFETs is presented. It is observed that fin shape significantly impacts leakage in bulk tri-gate n-FinFETs with thin fin widths when the fin body doping is optimized to minimize leakage. Fin shape can be used to implement multi-threshold n-FinFETs. A combination of 3D process simulator VictoryCell and 3D stress simulator VictoryStress has been employed for stress analysis of individual FinFETs and also complex cell structures. It is important to simulate stresses not just in an individual device but in the whole cell.

According to ITRS, in the next 10–15 years, the nominal supply voltage of the semiconductor circuits is going to be 0.7V for high performance devices. Optimization of triangular fin with gate-source/drain underlap may lead to the design of ultralow power FinFETs. We have explored the fin shape dependence of leakage in bulk tri-gate n-FinFETs with thin fins when the fin body doping profile is optimized to minimize leakage. Fin shape is evaluated as a technique to construct multi-threshold n-FinFETs without increasing chip area consumption.

The statistical variability of electrical characteristics, due to intrinsic parameter fluctuation sources, in contemporary and scaled FinFETs is systematically investigated. 3D strained devices with LER have been simulated based on calibrations. Strain variability due to LER has been demonstrated. Strain enhances the variation of current while it indeed improves the drive current. 3D simulation of strain development in tri-gate FinFETs, coupled with the understanding of strained planar MOSFET physics, are used to explain the strain enhanced tri-gate FinFET performance.

C

Advanced Devices

In this chapter, advanced device design techniques are discussed. In order to mitigate the transistor scaling issues of planar Si CMOS transistors various technology boosters are currently in use. Strained silicon is introduced in order to improve the mobility of the device, whereas high-k metal-gate technology is employed to reduce the gate leakage current. With the introduction of fully-depleted silicon on insulator (FDSOI), parasitic and leakage current in the substrate are reduced. Various advanced CMOS techniques are applied to enhance the quality of the transistors in order to achieve the performance targets set by Moore's law. In addition, various new technologies based on entirely new materials (nanowires and nanotubes) are being explored as a replacement for Si-CMOS for sub-5nm node. In the recent years, International Technology Roadmap for Semiconductors has shifted its focus from More-Moore (Moore's law scaling) to More-than-Moore (MtM) [9.1]. Instead of focusing strictly on the performance of CPU, MtM emphasizes on the overall integration and on the efficient implementation of every component. Beyond CMOS, emerging technologies will reach device dimensions below 5nm and will use nanowire transistors, quantum devices, carbon nanotubes, graphene, or molecular electronics.

9.1 More-Moore to More-than-Moore

In formulating his famous law, Gordon E. Moore made an empirical observation on the doubling of the number of transistors per CPU every year. Moore's law provided a clear direction and timeline for innovation in the semiconductor industry for the last five decades. Relentless focus

on Moore's Law, guided by the scaling rules set by Dennard, has provided ever-increasing transistor performance and density (see Figure 1.1). According to Dennard scaling, the oxide thickness (T_{ox}), transistor length (L_g) and transistor width (W) are scaled by a constant factor (1/k) in order to provide a delay improvement of 1/k at constant power density. As transistor scaling entered the nanometric regime (physical gate length below 100nm), the classical transistor scaling could not meet the scaling rules set by Dennard. Short channel effects pose a major challenge to the current transistor architecture. For instance, by reducing the length of transistor, the off-state leakage current is increased due to the drain induced barrier lowering and degraded sub threshold Slope. Furthermore, scaling of gate oxide thickness is needed to improve the electrostatic control of the channel. An oxide thickness of 1nm is required for the 10nm node in order to attain sufficient electrostatic control over the channel. On the other hand, such thin atomic layers of gate oxide comes at a cost of increase in the gate leakage current. Practical considerations on leakage limit the physical gate length to ~15–20nm. Moreover, decreasing the gate pitch decreases the stress enhancement for n- and p-MOSFETs thereby decreasing the mobility of the carriers.

Evolution of Device Structures

Figure 9.1 illustrates the two roadmaps that will be crucial for deployment of future nanotechnologies. On one hand, transistor scaling is continued as this leads to high performance CPU, memory and logic. On the other hand, MtM aims at providing power efficient systems. One of the key enabler for realizing MtM systems is three dimensional integration. Scaling the transistor in the nanometric regime worsens short channel effects, increases device parasitic and increases gate leakage current. In order to mitigate these drawbacks, various transistor structures are being investigated for advanced technology nodes. These transistor structures can be broadly classified based on the method of electrostatic confinement over the channel.

- Planar transistor structure with enhanced electrostatics (e.g., Strained-Si, High-k, UTB).
- Multiple gate transistor structure (e.g., double-gate, tri-gate, FinFET).
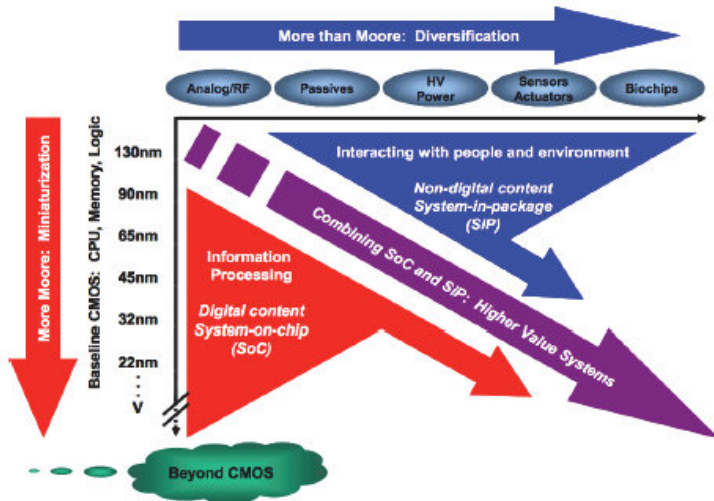
Figure 9.1. More than Moore. "Whereas More Moore may be viewed as the brain of an intelligent compact system, 'More-than-Moore' refers to its capabilities to interact with the outside world and the users." [*Source*: ITRS]

• Gate-All-Around (GAA) transistor structure (e.g., nanowire FET).

Strained Silicon

This technique has been widely adopted by the industry as the performance of the transistor is improved without any further shrinking of the transistor gate length by introducing lattice strain into the Si channel. By inducing mechanical strain in the channel region, the carrier transport properties of the NMOS and PMOS transistors are enhanced. By placing an active silicon region on a substrate layer with a larger lattice constant, strain is induced. Figure 9.2 illustrates a silicon layer on Silicon Germanium lattice. This modifies the band-structure within the active silicon region. This modification leads to a lower scattering probability and thus, to a higher mobility in the channel. By using silicon with 20% Ge portion, the electron mobility can been enhanced by 70% leading to a speed improvement around 30% [1.3].
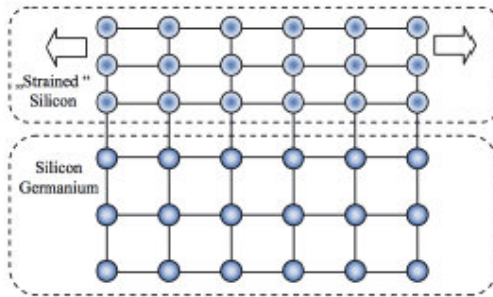
Figure 9.2. Illustration of straining of silicon by means of silicon germanium.

Strained Si is being implemented in nearly all 90nm, 65nm, and 45nm technology nodes. Though the gate length is kept constant with this technique, the transistor density can be increased by scaling down the transistor pitch. However, beyond few technology nodes, this will be limited by parasitic resistance and capacitance between the gate and source/drain contacts.

Multi-gate Transistor Structures

In order to minimize the increasing short channel effects beyond 22nm process, a number of multiple gate FETs have been developed. The trend towards multiple gate transistor started from the idea of a double-gate device based on SOI material. Figure 9.3 shows various multiple gate transistor structures that have been demonstrated over the last two decades. FinFET: The FinFET device shown in the Figure 9.3 is a manufacturable and cost-effective version of a double gate device with which we can realize double gate devices on a bulk-Si substrate. Unlike traditional transistor channel, the channel of a FinFET device is vertical to the plane of the substrate.

Tri-gate: The tri-gate device shown in the Figure 9.3 is an extension of a FinFET device (with a double gate) to a three gate structure. Tri-gate devices have gates around three sides of the device which improves the electrostatic control over the channel thereby reducing the short channel effects. When compared to FinFETs, there is no gate-blocking layer on the top of the gate.
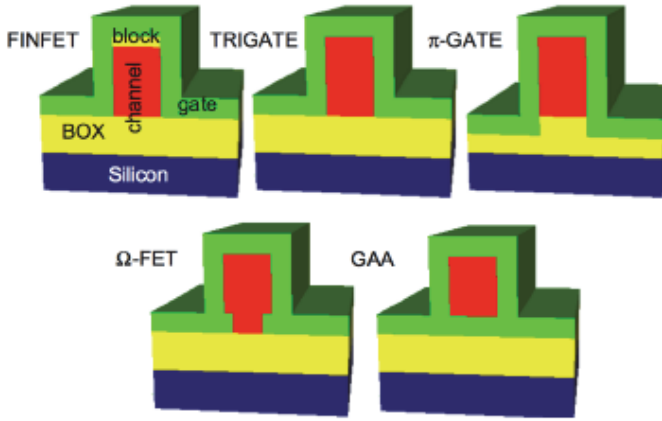
Figure 9.3. Types of multiple gate architectures. After S. K. Bobba, Design methodologies and CAD for emerging nanotechnologies, PhD Thesis, École Polytechnique Federale de Lausanne, 2013.

π-Gate and Ω-Gate: Both π-gate and Ω-gate are extensions of tri-gate structure for further improving the electrostatic control of the channel. In the case of π-gate, the gate is extended below the channel region which creates a virtual back gate thereby reducing the leakage current from the drain. In the case of Ω-FET, in addition to the tri-gate structure, the Ω-gate underlaps also the fourth side of the transistor channel. Similar to π-gate structure, this device has a similar effect in reducing the channel leakage thereby improving the electrostatic control of the channel.

Gate-All-Around (GAA) FET: When compared to the rest of the multi-gate devices, Gate-All-Around devices are comprised of a gate which wraps entirely around the channel, thereby providing full two dimensional confinement over the channel.

One of the most popular multiple-gate FETs is a tri-gate transistor. Intel employed this transistor at 22nm process node. When compared to UTB SOI device, a tri-gate device provides better electrostatic control over the channel as well as near ideal sub-threshold slope. One of the main challenges in making a tri-gate device is manufacturing the fins so that they are uniform.

High-k Metal-Gate

Scaling the thickness of the gate oxide (T_{ox}) improves the electrostatic control of the channel. A T_{ox} of 1nm is required for the 10nm node in order to attain sufficient electrostatic control over the channel. On the other hand, such thin atomic layer of T_{ox} comes at a cost of exponential increase in the gate leakage current. Typically, FETs with 20nm gate length have leakage current densities in the order of 10^{-2}–10^{-1}A/cm^2. Gate leakage current is mitigated by incorporating gate oxides with high dielectric constant (called as high-k), thereby allowing to increase the oxide thickness while ensuring good electrostatic control of the channel. The thicker the gate oxide, the lower is the gate tunneling current. Among the various high-k dielectrics being investigated, HfO$_2$ is widely adopted by the industry for process nodes below 45nm. For instance, the gate leakage current is reduced by a factor of more than 10^4 for a 20nm device.

9.2 Emerging New Materials

As silicon CMOS technology reaches its fundamental scaling limits, alternative materials such as high mobility III–V compound semiconductors have proven to be strong contenders for extending high performance logic. However, most promising demonstrations of III-V HeteroFETs/HEMTs have micron-scale source/drain spacing despite gate lengths on the nanometer scale. III–V semiconductor devices have historically relied on alloyed ohmic contacts which require large spacing to prevent shorting between the source and drain after alloying, where contacts can diffuse up to hundreds of nanometers. This severely limits the scalability of III-V logic technology. Non-alloyed contacts offer a practical route to greatly reduce the III–V device footprint for application in future technology nodes.

The rising III-V compound semiconductor materials, which have superior transport properties, high breakdown electric fields or high thermal and/or high heat capacity and thermal conductivity, are suitable for applications that are driven more by performance and less by cost

such as high power and high frequency application and where silicon technology cannot meet the performance requirements such as high dynamic range or low noise figure. In the past, the utilization of III–V semiconductor confronts the unavailability of the substrate, the lack of the appropriate gate dielectric for good interface condition, the absence of reliable enhancement-mode device, and the difficulty in the heterogeneous integration of III–V compound with silicon CMOS on the same substrate. However, as the aggressive scaling pushing the silicon to its limit, scientists and researchers have been actively investigating the device physics behind the new materials, advancing new techniques for III–V substrate development and fine epitaxial growth quality. Besides, the efforts of the silicon device fabrication progress also benefit the III–V compound process techniques. As the research and development endeavors to the semiconductor compounds over decades, InGaAs and GaN compound materials are recognized as the promising candidates in the semiconductor roadmap as the replacement materials of the silicon. The former, featuring high electron mobility and comparable bandgap to silicon, is encouraging for ultra-high speed, ultra-low power and ultra-high frequency applications.

Thereby InGaAs attracts the semiconductor industry for the use in CMOS beyond the silicon theoretical limit in sub-22nm technology regime. InGaAs shows high breakdown electrical field and high electron mobility, has a great potential in RF transistor, power amplifier and high voltage switching device applications. It is expected that compound semiconductors will continue to advance through a combination of gate length scaling and more importantly epitaxy or bandgap engineering. Furthermore, new approaches in the integration of high performance III-V transistors and high-density digital circuitry are also being explored for high performance chips which are equipped with diversified functionalities that are out of reach for the conventional silicon technologies. Gallium Nitride (GaN) High Electron Mobility Transistors (HEMTs) are on the leading edge of wide bandgap technology and have the performance characteristics to dominate in high power – high bandwidth applications.

GaN falls into the category of wide band gap semiconductors (band gap energy, E_g = 3.44 eV) along with other materials such as diamond and SiC. A wide energy band gap generally translates into an ability to

support high internal electric fields before electronic breakdown occurs. III–nitride materials such as AlGaN and GaN possess polarized wurtzite crystal structures, having dipoles across the crystal in the [0001] direction. In the absence of external fields, this macroscopic polarization includes spontaneous and strain-induced (piezoelectric) contributions. The primary effect of polarization is an interface charge (polarization charge) due to abrupt divergence in the polarization at III–nitride hetero interfaces. In general, device simulators provide model for polarization effects, ranging from placement of explicit interface charges to full-volume polarization fields computed from stress distributions in the device. Different simulation case studies are presented in this chapter. These include optimization for various applications, high-temperature performance analysis, and transconductance collapse investigation of normally-off devices.

Knowledge of the material system is a key requirement for proper device modeling. Established physical models for the lattice and thermal properties of the materials as well as models describing relevant effects are included with respect to HEMT specifics and material properties. Several GaN-based devices are simulated using the physical models and model parameters. Performance predictions for down-scaled devices in high-temperature operation are shown. The best option for modeling polarization in these devices is the strain model that computes spontaneous and piezoelectric components based on the local mole fraction and assumes pseudomorphic growth on top of a buffer layer with a certain lattice constant. All model parameters are mole fraction dependent and, by default, linear interpolation is applied between values for side materials such as GaN, AlN, and InN. As a side effect of using a self-consistent polarization model, all hetero interfaces where significant polarization divergence is observed would also exhibit strong polarization charges. However, these charges at interfaces below the GaN buffer layer are likely to be compensated by charged defects and are typically irrelevant to the electrical operation of devices. Simulations involving wide-bandgap materials are indeed more prone to convergence problems due to the numeric challenges in performing floating-point operations with extremely large and extremely small numbers such as the concentrations of majority and minority carriers in wide-bandgap semiconductors.

9.3 High Electron Mobility Transistors

The first GaN based transistors were realized in the early to mid-1990 and since then have been extensively researched and developed for high-power high-frequency applications as well as for high-voltage power switches. In this section, we discuss III-Nitride materials and material systems on which HEMTs are based. HEMTs based on GaN are commonly used for high-power and high frequency applications, such as mobile communications and Monolithic Microwave Integrated Circuits (MMICs) are commercially available. However, ultra-broadband communication systems require further device optimization. Recent findings for the band structure and other relevant scattering mechanisms are necessary for the simulations. New transport models suitable for III-V materials are being developed and need to be implemented in the device simulator. High electron mobility transistors based on GaN have demonstrated power density which is well beyond the capability of GaAs FETs. Unlike the strain technology in Si devices, stressing methods have not been intentionally used during device fabrication to alter GaN HEMT performance. Nevertheless, the AlGaN layer is under large biaxial tensile stress due to its lattice mismatch from the bottom relaxed-GaN substrate. In addition, since both the AlGaN and GaN are piezoelectric materials, the lattice deforms when an electric field is present. Under typical GaN HEMT operation conditions, the total stress in the AlGaN layer is approximately 3GPa or higher.

To explore the full range of applications of the AlGaN/GaN HEMT, it is also worth investigating its potential as a high voltage power switch. This is because; GaN has a breakdown field which is 3 times of that in Si or GaAs. In the following, we study the influence of the gate to drain distance on the breakdown voltage with and without silicon substrate. The structure consists of 20nm AlGaN with x.comp = 0.2 and 1μm GaN buffer layer. The device structure with silicon substrate is shown in Figure 9.4. Donor and acceptor traps are included in the simulation. Gaussian profile acceptor traps are used with a peak concentration close to the bottom of the buffer layer to mimic the fact that the material quality improves toward the surface of the device. Donor Interface traps at the AlGaN/Nitride interface and acceptor interface traps at the GaN/Silicon interface are also included during the simulation.
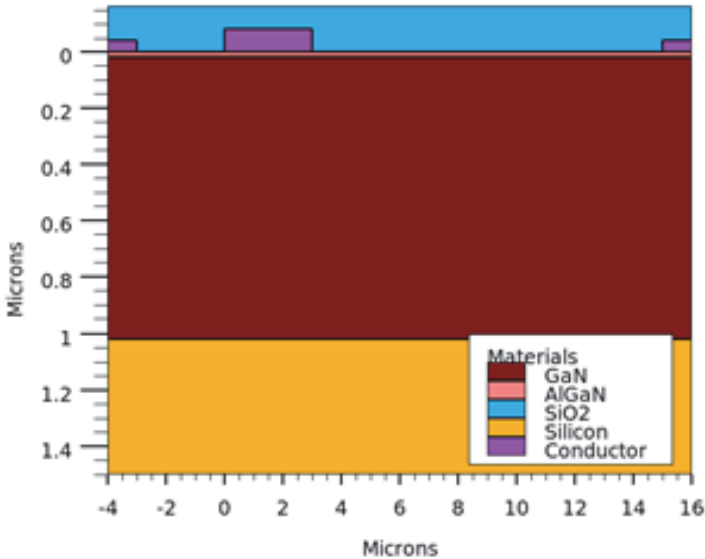
Figure 9.4. Device structure with silicon substrate.

Forward characteristics simulation reveals a threshold voltage of around −2V. Slow transient simulation is used to simulate the breakdown voltage characteristic. The deck is parameterized so that we can make variation of the distance between the gate and drain (LGD). DeckBuild DoE capability is used to automate simulations of BV vs. LGD with and without substrate. As shown in Figure 9.5 an almost linear increase of BV vs. LGD when silicon substrate is not present. An increase of the gate current is observed in Figure 9.6 near breakdown, and the value is of the same order as the drain current. This result indicates that BV is defined by an avalanche phenomenon between the gate and the drain as shown in Figure 9.7.
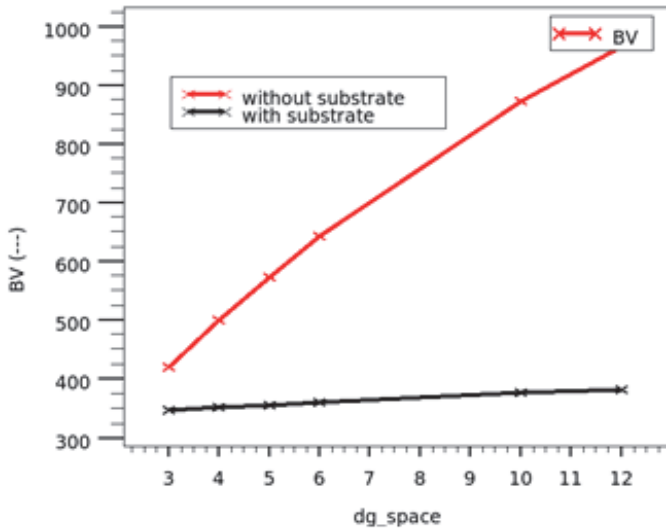
Figure 9.5. Breakdown voltage vs. gate to drain distance with and without silicon substrate.
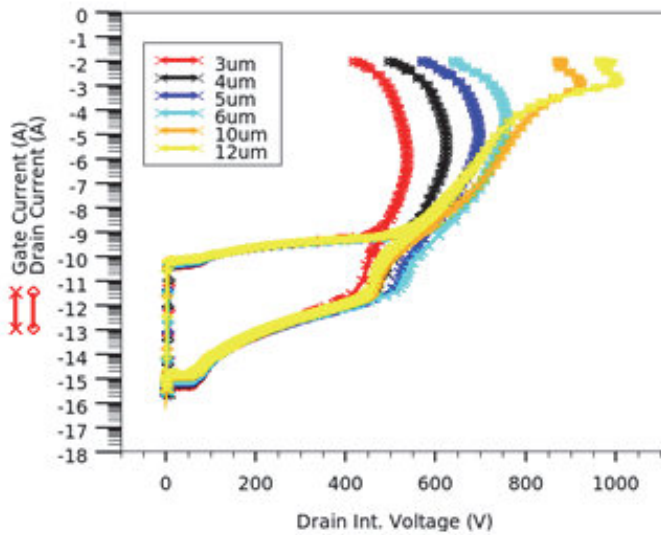
Figure 9.6. Drain and gate current vs. drain voltage.
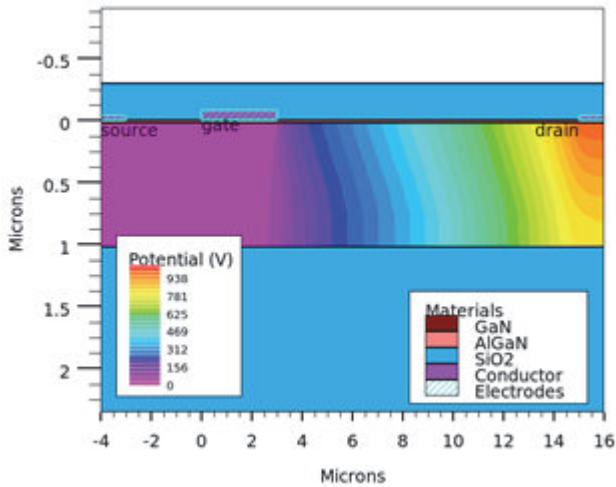
Figure 9.7. Potential distribution as a function of gate to drain distance.

In the following, we examine the effect of field plate on avalanche breakdown an AlGaN/GaN HEMT via simulation. The device under consideration consists of a simple 0.4μm gate length GaN HEMT augmented with a field plate used to increase the breakdown voltage. This simulation loops on increasing field plate lengths to examine the relationship between field plate geometry and breakdown voltage. The simulations steps are:

- Construction of the heterojunction structure using ATLAS
- Material and models parameter specification
- Simulation of $I_d - V_d$ and breakdown characteristics

The device is described using the ATLAS structural syntax. The region statements are used to define the AlGaN and GaN regions. The Al composition fraction (x.comp = 0.295) is used in simulation. The models statements included: field dependent mobility, SRH recombination and Fermi-Dirac statistics. In order to simulate avalanche breakdown, the impact ionization-generation model was turned on along with the impact selb statement in which the Selberherr impact ionization model is activated. In the simulation, the beam statement is used to specify an optical source of carrier pair generation in addition to the thermal

generation provided by SRH. The drain voltage is ramped to 1200V. At that point the drain boundary is changed to a current boundary using the contact current statement. Figure 9.8 shows the breakdown voltage as a function of field plate length. The impact generation rate in the device is shown in Figure 9.9.
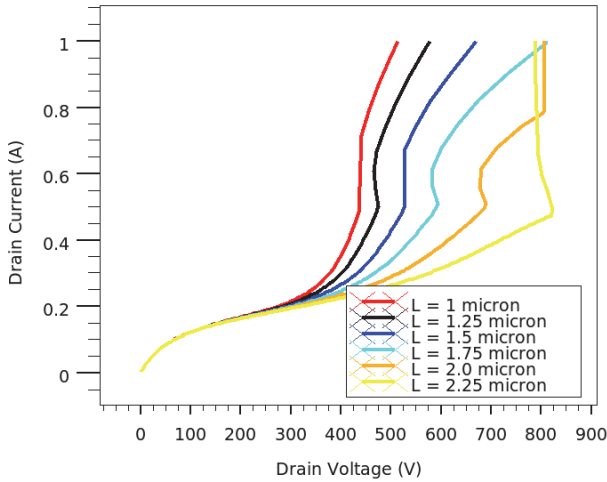
Figure 9.8. Breakdown voltage as a function of field plate length.
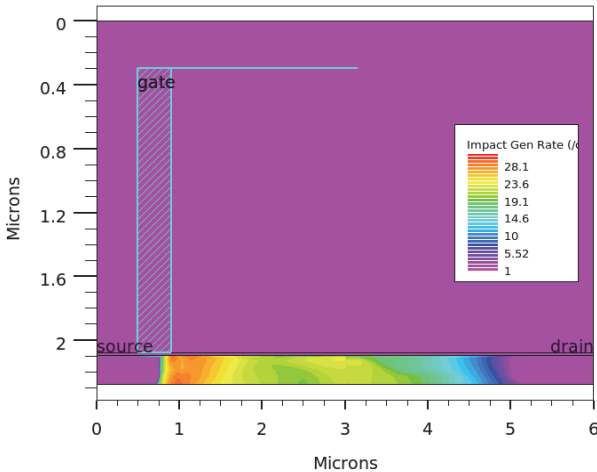
Figure 9.9. Impact generation rate in AlGaN/GaN device.

9.4   Stress Effects in AlGaN/GaN HEMT

The device under consideration is a 3D AlGaN/GaN HEMT simulation taken from Silvaco. The main idea of GaN based power devices is to use epitaxial strain to create 2DEG. The 2D structure is first obtained by ATHENA and then extruded in 3D using DEVEDIT. 3D stress distribution is simulated by VictoryStress. The polarization model that supports dependency on loaded strain tensor has been used in simulation. The model is enabled by the TEN.PIEZO flag of the MODELS statement. The polarization model that supports epitaxial strain due to lattice mismatch is enabled by the CALC.STRAIN flag on the MODELS statement. When both models are set in the simulation both the imported strain and the lattice mismatch calculation are accounted for. The lattice and imported strain dependent components of polarization can be independently scaled using the scale factors TENSO.SCALE and PIEZO.SCALE on the MODELS statement. When enabling the TEN.PIEZO flag spontaneous polarization is automatically included in the calculation. The size of the spontaneous component can be scaled using the PSP.SCALE parameter of the MODELS statement.

After the initial solution is obtained the drain voltage is ramped to 1V, then the $I_d - V_g$ characteristic is extracted from $V_g = -8$ to $V_g = 1.0V$. The device under consideration is a 3D AlGaN/GaN HEMT shown in Figure 9.10. 3D stress distribution is simulated by VictoryStress. This example produces stress distributions via the stress-liner made of nitride as shown in Figure 9.11. Note that the stressor in this example is used to illustrate strain polarization on top of piezo and spontaneous polarization. The effect of strain is seen on $I_d - V_g$ characteristic is shown in Figure 9.12.
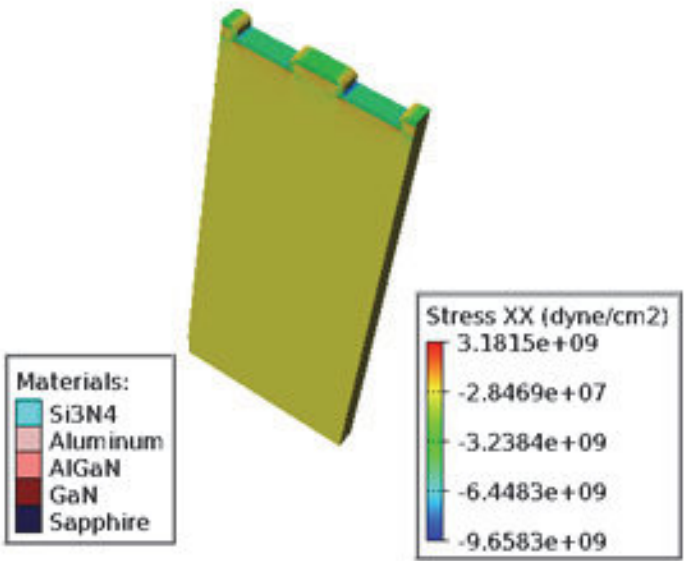
Figure 9.10. 3D AlGaN/GaN HEMT structure.
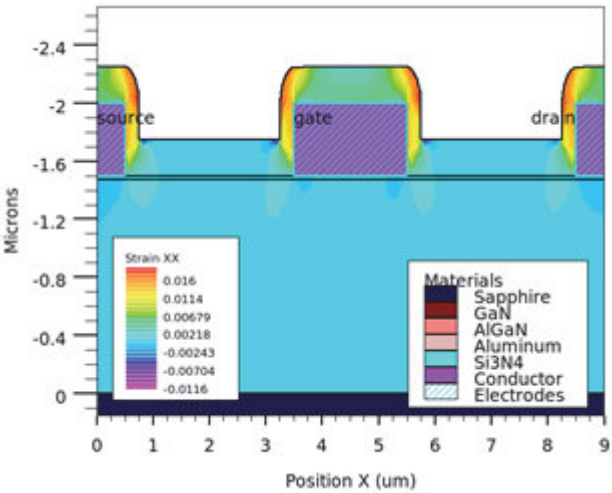
Figure 9.11. Strain xx generated by a stress-liner made of nitride.
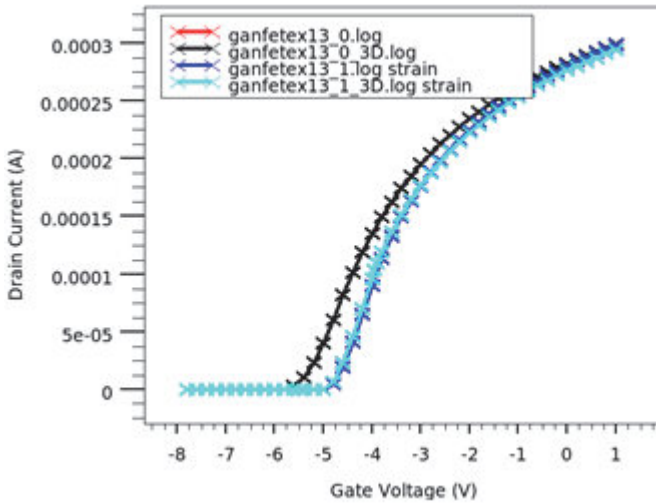
Figure 9.12. Drain current vs. gate voltage for before and post-stress device.

9.5 AlGaN/GaN Hetero-FETs

III-V materials have the potential to replace Si in the future, due to its higher electron mobility as well as wider choice of bandgap. Gallium Nitride (GaN) high electron mobility transistors are promising for high-power applications. Of the wide band gap materials, GaN is particularly appealing due to its ability to form heterojunction to wider band gap semiconductors such as aluminum gallium nitride (AlGaN) or aluminum nitride (AlN) (up to 6.2 eV). In doing so, a 2-dimensional electron gas (2DEG) forms at the interface due to large polarizations in the material which provides a highly dense, majority carrier channel with large electron mobility (usually and most preferably with undoped materials). These properties can be exploited to make devices which are capable of handling or providing high output power that can be operated as power switches or comfortably up to 10 GHz for power amplifier applications respectively. High voltage microwave AlGaN/GaN HFETs operating under high-power conditions suffer from degraded RF performance and linearity due to a nonlinear resistance effect in the gate-source region. During RF operation, the nonlinear resistance is due to the onset a of

space-charge limited current transport mechanism within the device channel. A temperature dependent, impact ionization initiated RF breakdown model in the 2DEG channel of AlGaN/GaN HFETs has been reported. When operating these devices in RF power amplifier circuits, impact ionization in the channel has a significant effect upon gain saturation, power-added efficiency and output power. Popular III-nitride-based devices such as HFETs for RF and power-switching applications are built in epitaxial structures formed by multiple layers with constant mole fractions.

AlGaN/GaN hetero-junction field effect transistors devices have great potential for RF power applications. The material properties of GaN have enabled the manufacture of high power devices capable of withstanding voltage and currents levels and higher frequencies well beyond that of silicon and gallium-arsenide devices. The large band-gap of GaN enables a large breakdown voltage since the material is able to sustain large electric field magnitudes. High power operation is enabled by the capacity of GaN devices to carry a large channel current in conjunction with the large bias voltages. The AlGaN/GaN material system forms a hetero-junction which results in the establishment of a two-dimensional electron gas channel at the material interface with charge densities $\sim 1 \times 10^{13}$ cm^{-2}. The mobility in the channel can range up to $\mu_n \sim 2000$ cm^2/V.s and the electron velocity can reach $v_{sat} \sim 2.5 \times 10^7$ cm/s leading to a high channel current and the ability to operate at higher frequencies. In the following, TCAD based investigations of the internal device behavior are taken up. The design issues for AlGaN/GaN HFETs are involved for the followings:

- time dependent degradation of drain current due to gate-surface leakage
- current collapse due to trapping at the gate edge on the surface and in the AlGaN layer
- non-reversible sudden increase in gate leakage current due to damage to the AlGaN layer at high fields
- breakdown in the conducting channel under high-voltage conditions due to impact ionization

This case study taken from Silvaco demonstrates simulation of a normally-off AlGaN/GaN transistor with conductivity modulation

resulting from hole injection from a p-AlGaN gate to the AlGaN/GaN heterojunction. In simulation, calculation of spontaneous and piezo-electric polarization, strain for the material system, composition dependent models for bandgap, electron affinity, permittivity, density of state masses, recombination, impact ionization, heat capacity, refractive index, low and high field mobilities were considered. GaN specific impact ionization and field and temperature dependent mobility models and phonon-assisted tunneling model were used.

The p-AlGaN gate allows normally-off operation. The ATLAS generated device structure is shown in Figure 9.13. When the gate voltage increases and reaches the built-in pn junction voltage at the gate (around 3.5V) holes inject into the channel and generate the equal number of electrons increasing the 2DEG. Electrons with high mobility will reach the drain under the effect of the electric field whereas the holes will stay since their mobility is much lower than the electron. The current is thus modulated by the number of holes injected. It is also interesting to notice that this device exhibits a double peak in the transconductance characteristic, a second proof of hole injection as shown in Figure 9.14.
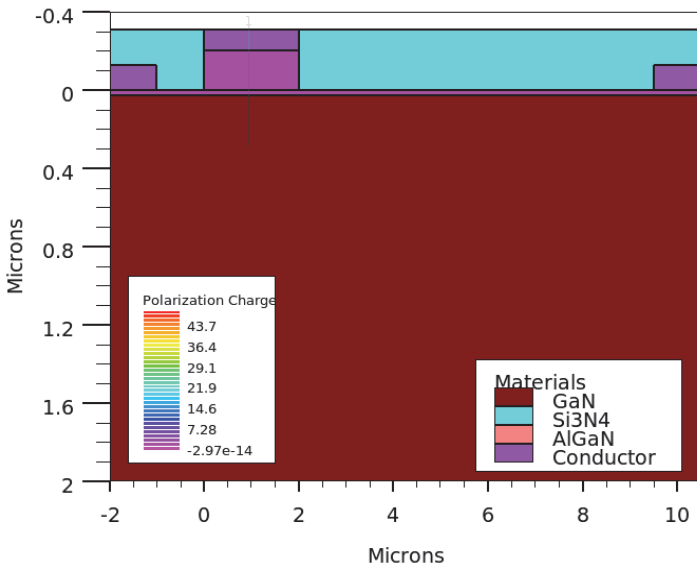
Figure 9.13. AlGaN/GaN HFET structure used ATLAS simulation.

Electron injection from the channel to the gate is limited by the AlGaN/GaN heterojunction. As a consequence no current offset is observed at zero drain since gate current is very low. This device exhibit a threshold voltage around 1V, a maximum drain current at $V_d = 10V$ and $V_g = 5V$ of 200mA/mm and a breakdown voltage of 640V as shown in Figures 9.15, 9.16, 9.17, respectively.
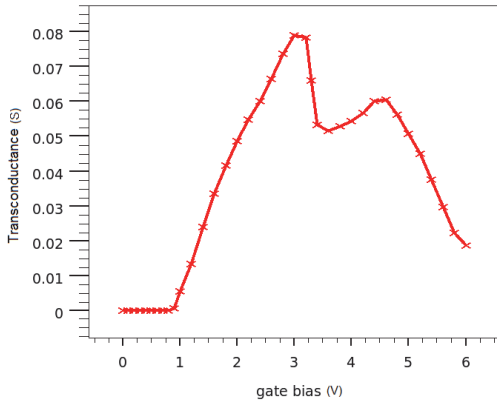
Figure 9.14. Transconductance vs. gate voltage characteristics.
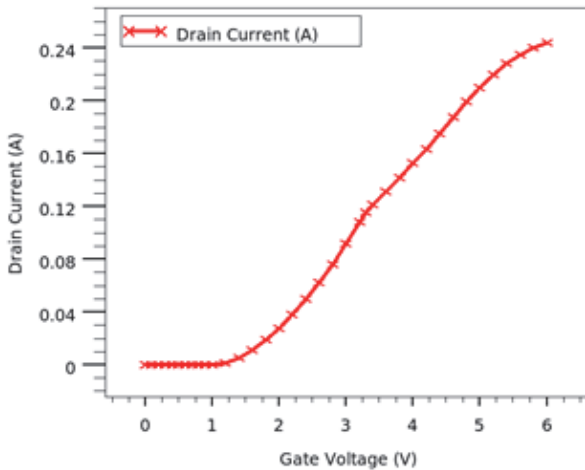
Figure 9.15. Drain current vs. gate voltage characteristics.
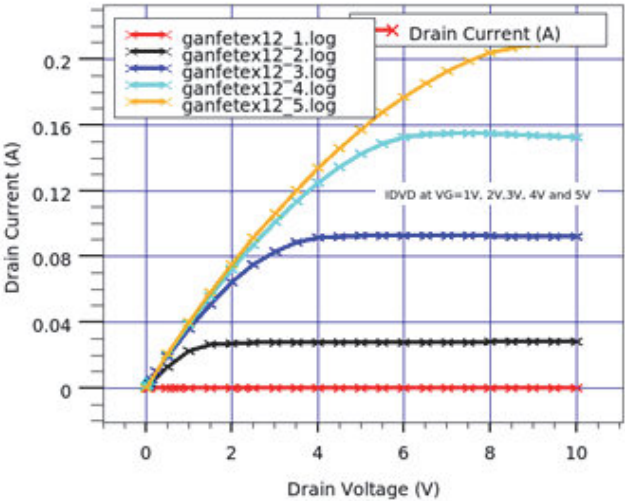
Figure 9.16. Drain current vs. drain voltage characteristics.
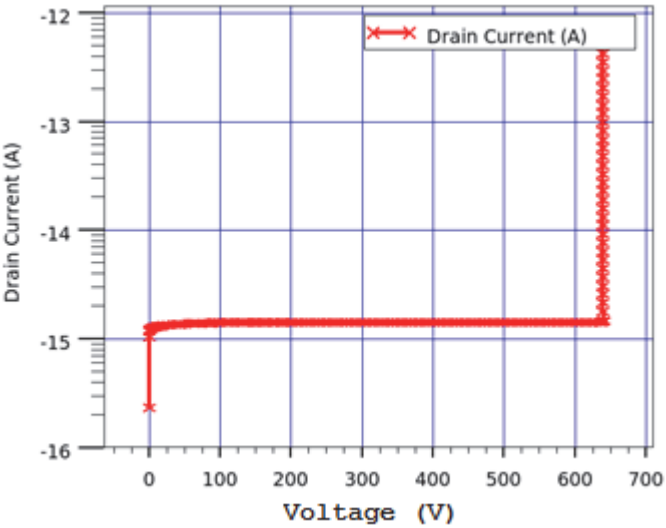
Figure 9.17. Breakdown voltage characteristics.

9.6 Ultrathin Body MOSFETs

Bulk devices require a heavy channel doping to control short-channel effects. With a heavily doped channel, carrier mobility is severely degraded due to impurity scattering and an increased transverse electric field. Above a concentration of 10^{18}cm^{-3}, mobility is expected to be noticeably affected by channel dopants. In UTB and DG devices, short-channel effects are controlled by a thin silicon film, thus allowing for gate-length scaling down to the 10nm regime without the use of channel dopants. Partially depleted (PD) SOI was the first SOI technology introduced for high performance microprocessor applications. The buried oxide (BOX) layer serves to provide for reduced source/drain junction capacitance. PD/SOI MOSFETs offer a good trade-off between power and performance. This approach replaces the bulk silicon of a normal transistor with a thin layer of silicon built on an insulating layer, creating a device that is often called ultrathin body silicon-on-insulator (UTB SOI), also known as a fully depleted SOI.

Fully depleted (FD) SOI MOSFETs achieve superior control of short channel effects and therefore are beneficial for low-power technologies which have stringent leakage specifications. FD/SOI MOSFETs rely on a thin body to suppress leakage, which presents news challenge for transistor fabrication, mainly the tight control of the physical body thickness and doping profiles. Series resistance in the source and drain extension regions is another issue. An elevated source/drain structure is needed to mitigate this parasitic resistance, but this structure results in increased gate-to-drain capacitance (the Miller capacitance) and thus degrades circuit performance.

Figure 9.18a illustrates an UTB SOI transistor. An SOI isolation layer separates the thin active device layer and the main substrate. Due to the full depletion of the body, there is no room for unwanted current paths to form under the channel like in a traditional bulk device. This results in reduced parasitic and leakage currents in the substrate. The root of UTB SOI technology for planar electrostatic confinement dates back to 1980s. UTB SOI devices are quite similar to conventional planar CMOS transistors, thereby making them easier to manufacture. For UTB MOSFETs, short-channel effects are strongly dependent on the body

thickness (T_{Si}), and can be evaluated by the ratio of the gate length (L_g) to the body thickness. SCEs are sufficiently suppressed when the ratio between the two device dimensions (L_g/T_{Si}) is larger than 4 and this criterion has been tested for different experimental devices. This value may be further reduced by scaling the gate-oxide thickness or increasing the channel doping concentration. Nevertheless, metal gate work-function engineering is necessary to adjust and control the threshold voltage of UTB MOSFETs. In UTB devices, the body thickness should be as thin as possible for ultimate scalability; however, the series resistance of the source and drain regions could limit transistor drive currents.

UTB SOI transistors are ideal for low-power applications as there is a possibility for body biasing with the thin buried oxide (BOX). By applying a small voltage to the silicon substrate below the BOX, we can alter the channel properties, reducing the electrical barrier that stops current flowing from the source to the drain. As a result, less voltage needs to be applied to the transistor gates to turn the devices on. When the transistors are not needed, the bias voltage can be removed, which restores the electrical barrier thereby reducing the leakage current of the device. The main challenges of UTB SOI include variation in the thickness of the thin silicon film and also the difficulties in inducing strain in the channel. Figure 9.18b shows an ultrathin body double-gate SOI MOSFET.

a. Ultrathin body SOI MOSFET   b. Double-gate MOSFET
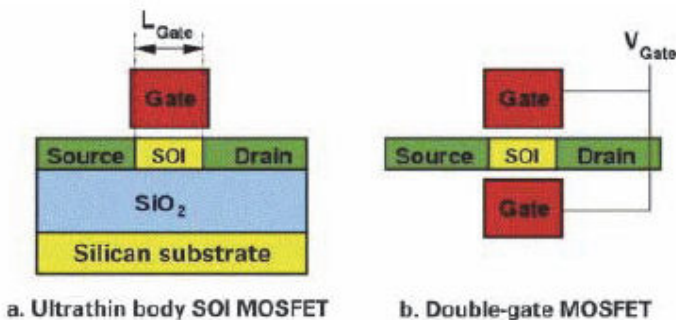
Figure 9.18. (a) Ultrathin body SOI MOSFET (b) Double-gate SOI MOSFET.

Stress simulation in extremely thin SOI CMOS

With significant device performance enhancement, fully depleted SOI with extremely body (ETSOI) is a viable option for future CMOS technology owing to its superior short channel control, inherent low device variability due to undoped channel, and compatibility with mainstream planar CMOS. ETSOI has been studied extensively as a viable option for continued scaling of CMOS technology owing to its superior control of short channel effect with negligible dopant fluctuation. Low GIDL and low threshold voltage variability make high-k/metal-gate ETSOI devices attractive for low power applications. Further performance enhancement is achieved by stress coupling between stress liners and the channel by using faceted RSD. However, one concern for ETSOI performance is that embedded stressors such as e-SiGe cannot be used because the silicon layer is too thin.

In the following example, incorporation of two strain techniques, viz., (a) SiC RSD for n-MOSFET and SiGe RSD for p-MOSFET, and (b) enhanced stress liner effect coupling with faceted RSD have been used to boost performance. Using the new process flow and the stress boosters, drive current enhancements in n-MOSFET and p-MOSFETs have been observed. Typical ETSOI CMOS flow includes:

- Device formation (Implant-free)
- Isolation (Simple)
- Well implant (No)
- Halo (No)
- Random dopant fluctuation (Minimal)
- No Mask levels to form S/D
- HK/MG stack (mid-gap)
- Resist shadowing (No)
- Gate-height scaling (Easy)
- Implant damage (No)
- ETSOI amorphization (No)
- Resist strip (Easy)

The nominal gate length for this simulation is 26nm. The stress simulation has been performed for a half of the device due to its

symmetry. The following geometry parameters can be changed in the simulation in order to investigate the influence of the device geometry on stress distribution, and, therefore, on device performance:

- the gate length;
- the gate height;
- the RSD height;
- the thickness of stress liner;
- the gate oxide thickness;
- the thickness of silicon on insulator (SOI);
- the thickness of bottom oxide;
- proximity of the bottom of RSD to the gate;
- proximity of the top of RSD to the gate (in case of faceted RSD)
- the width of simulation domain.

The following parameters were used to vary the simulation conditions:

- slope of RSD edge;
- RSD serves as a stressor;
- nitride stressor;
- poly as a stressor;
- materials for RSD.

All above parameters may also be used in design of experiments. The 2D structure was obtained from ATHENA simulation and subsequently used in VictoryStress. The isotropic model for stress simulations is used in this simulation since the structure is 2D. The elastic properties of Si, SiGe, and SiC are considered in simulation. Each run of this example produces stress distributions for three types of stressors:

- first is for RSD regions either of SiGe with negative intrinsic stress or of SiC with positive intrinsic stress;
- the second is the stress-liner made of nitride;
- the third one is the stress memorized in the gate.

In case of vertical RSD, the initial structure, RSD-stressor, stress-liner, and poly gate stressor are created and the stress distribution in these structures are shown Figures 9.19, 9.20, 9.21, and 9.22, respectively.
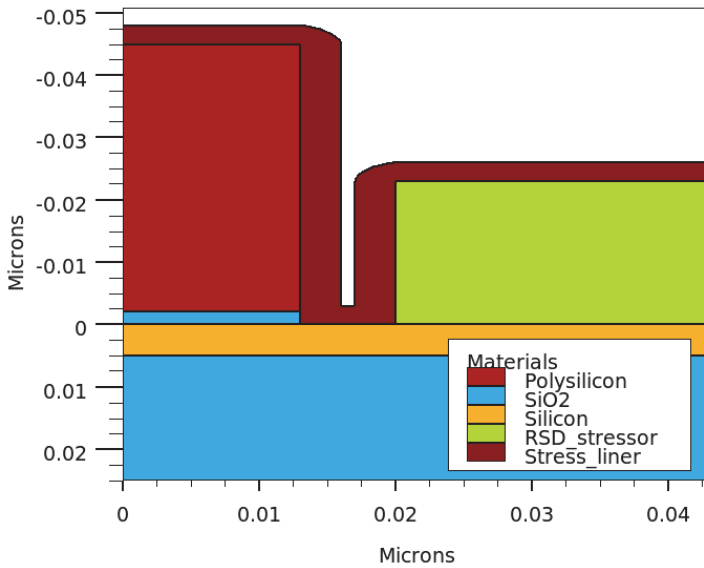
Figure 9.19. Vertical RSD the initial structure.
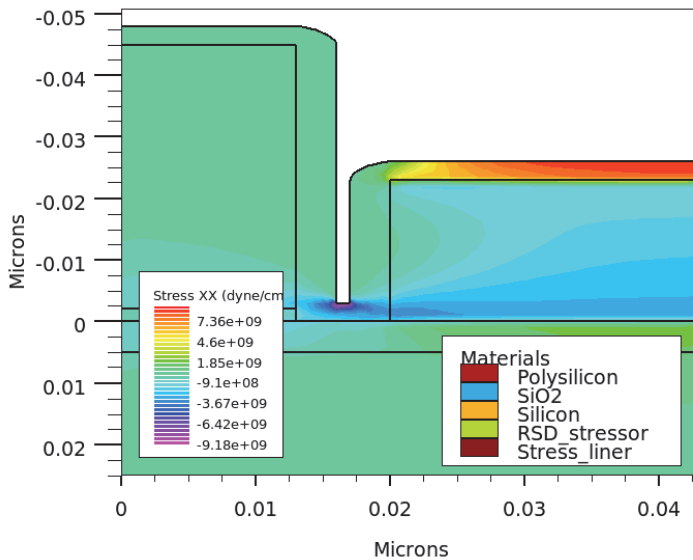
Figure 9.20. Stress distribution in vertical RSD stressor.
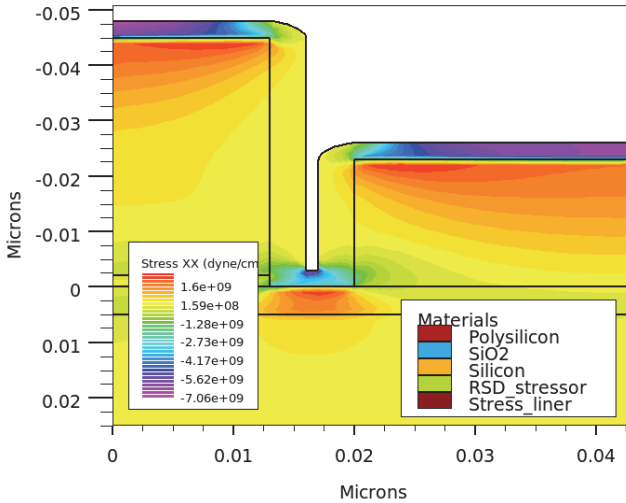
Figure 9.21. Stress distribution in vertical RSD stress-liner.
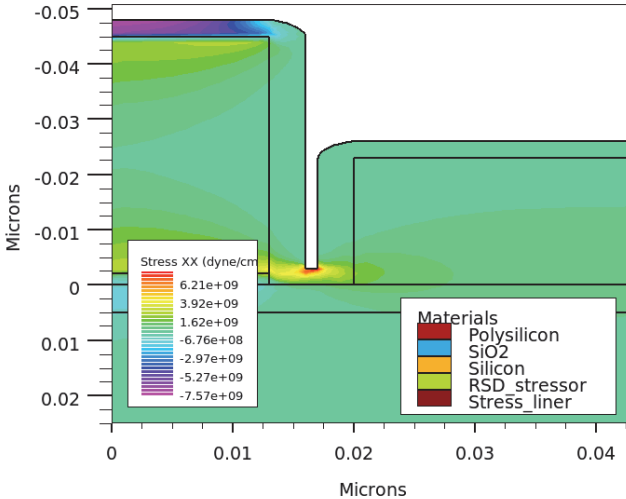
Figure 9.22. Stress distribution in vertical RSD poly gate stressor.

In case of faceted RSD, the initial structure, RSD-stressor, stress-liner, and poly gate stressor are created and the stress distribution in these structures are shown Figures 9.23, 9.24, 9.25, and 9.26, respectively.
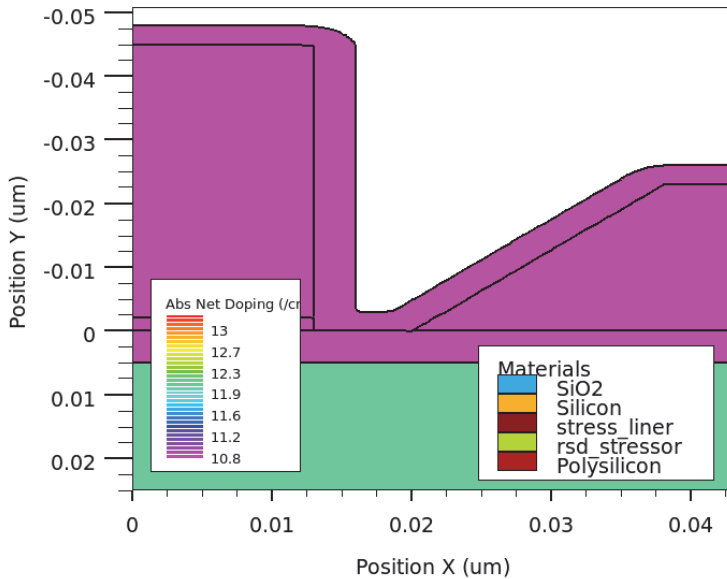
Figure 9.23. Faceted RSD the initial structure.
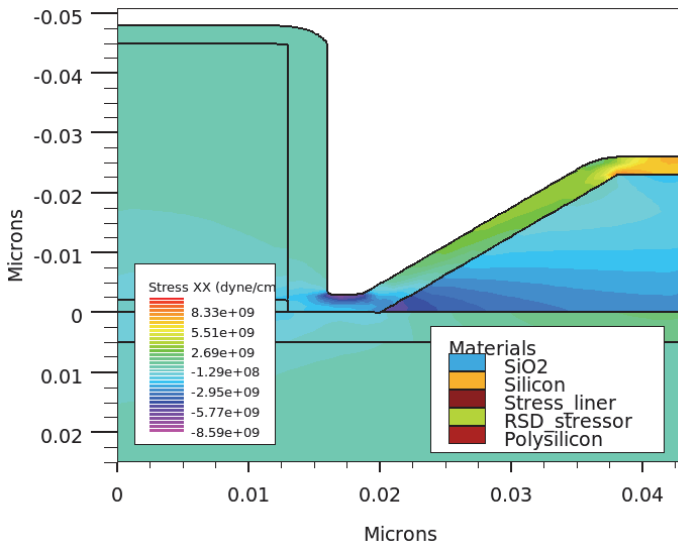
Figure 9.24. Stress distribution in faceted RSD stressor.
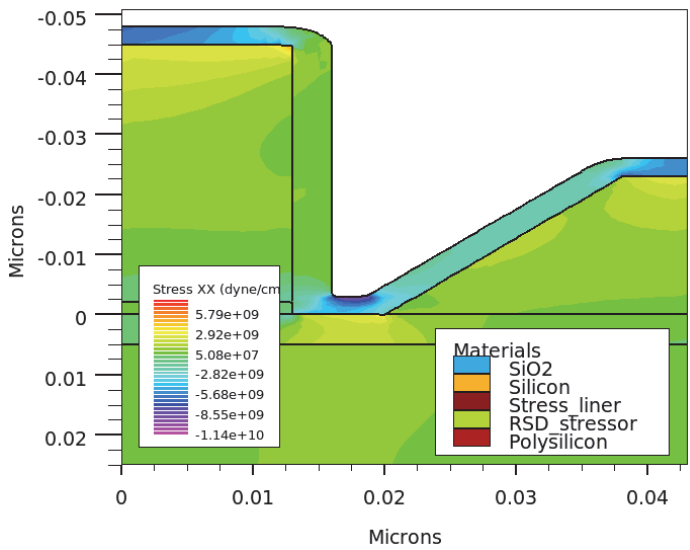
Figure 9.25. Stress distribution in faceted RSD stress-liner.
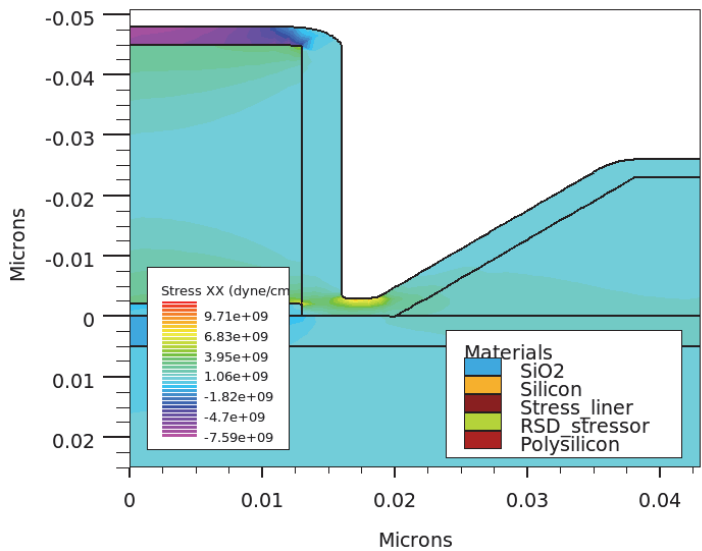
Figure 9.26. Stress distribution in faceted RSD poly gate stressor.

The successful integration of SiGe RSD for P-FET and SiC RSD for N-FET provides considerable channel stress as high as 500 MPa and therefore performance improvement takes place. It may be noted that in ETSOI, the conventional vertical RSD prevents close proximity of stress liners to the channel, thereby hindering efficient stress coupling. The 2D structure is created in ATHENA and subsequently used in VictoryStress (see Figure 9.27). In the simulation, we compare the I–V characteristics obtained using ATLAS, with stress-liner made of nitride with two geometries of RSD (vertical and faceted). Along with the conventional silicon mobility models (CVT and SRH), we used the strain dependent mobility enhancement models nhance and phance in the mobility statement in this simulation. The mobility enhancement models use the second order mobility enhancement tensor (calculated by VictoryStress) directly from the low field mobility. This results in directionally dependent (anisotropic) electron and hole mobilities. For the model to take effect, the enhancement factors was included in the structure file simulated by VictoryStress. This is done by setting "nhance" and "phance" defined in the stress statement of VictoryStress. Electrical simulations show that the effect of stress/strain results in higher drive current increase with faceted RSD. It is seen that faceted RSD structure provides improved stress transfer resulting in 40% increase of hole mobility and 15% enhancement of drive current as shown in Figures 9.28 and 9.29, respectively.
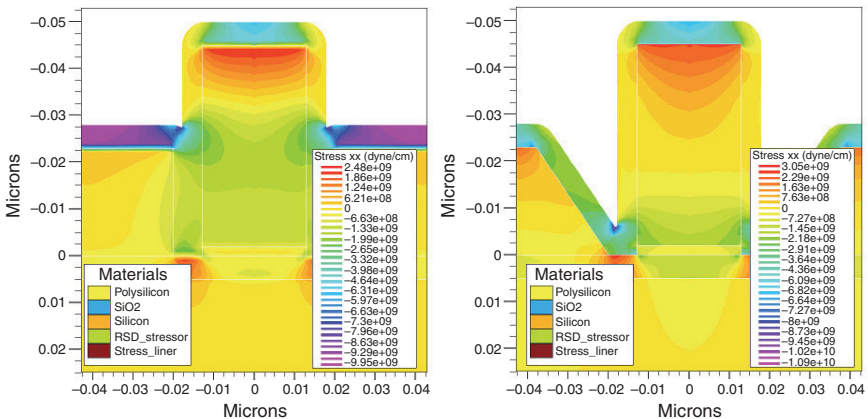
Figure 9.27. ATHENA/VictoryStress generated stress distribution in vertical and faceted RSD structures.
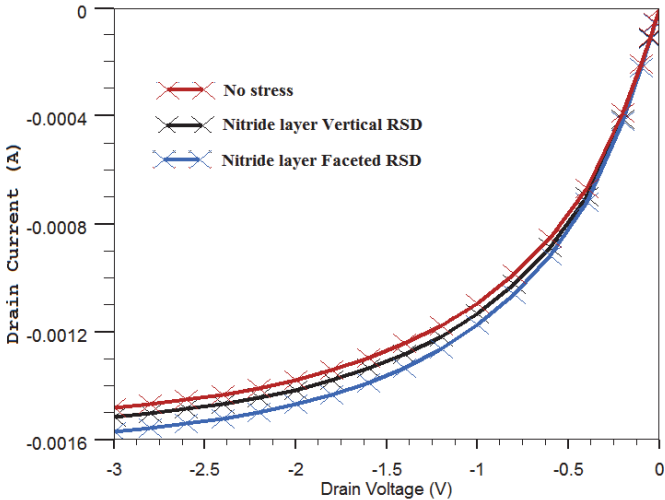
Figure 9.28. Comparison of $I_d - V_d$ characteristics for vertical and faceted RSD ETSOI MOSFETs at $V_g = -3V$.
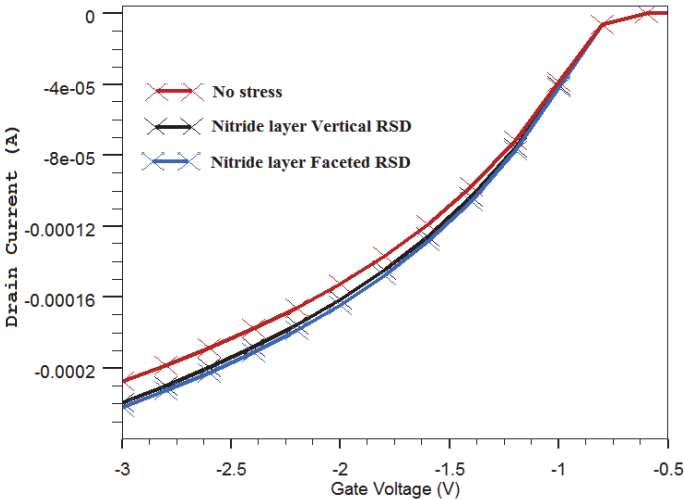
Figure 9.29. Comparison of $I_d - V_g$ characteristics for vertical and faceted RSD ETSOI MOSFETs.

9.7 Tunnel FET

Advanced structures such as the FD-SOI MOSFET and FinFET can certainly help to improve the performance of a transistor and allows for further technology scaling. Nevertheless, CMOS has a fundamental limit on energy per operation. Governed by Boltzmann statistics, the theoretical minimum subthreshold swing of a MOSFET is 60 mV/dec at room temperature. As a result, a new switch with dramatically different carrier injection mechanism needs to be explored. In this Section, we explore the tunnel field effect transistor (TFET) as an alternative switching device to overcome the fundamental limit in subthreshold swing and hence the energy efficiency of CMOS. TFETs rely on carrier injection via band-to-band tunneling (BTBT) and the absence of thermal (kT) dependence allows for the sub threshold swing to be steeper than 60 mV/dec.

Any prospective MOSFET-replacement device should have a SS that is smaller than 60 mV/dec limit in order to provide for lower energy computing. For the same leakage current, this new device with a smaller SS will be able to achieve a higher on-state drive current (Figure 9.30a), resulting in a lower energy per operation for a given operating frequency (Figure 9.30b). There are currently a number of device candidates that have the potential for a very steeply switching operation; these include tunneling FETs, the negative capacitance MOSFET, and nano-electro-mechanical (NEM) relays. The tunnel FET, in particular, has emerged as a strong MOSFET-replacement candidate due to its close similarity to a conventional MOSFET. Figure 9.31 shows a basic comparison between the MOSFET and tunnel FET. Because carrier injection in a TFET relies on quantum mechanical tunneling instead of thermionic emission over a potential barrier, the TFET can theoretically achieve a SS smaller than 60 mV/dec at room temperature. Optimization of TFET performance is still an area of active research.
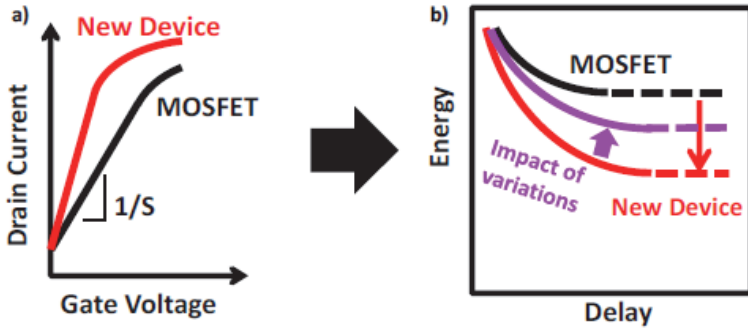
Figure 9.30. a) A new logic device with a smaller subthreshold swing (i.e., turning on/off more steeply) as compared to the MOSFET allows for higher ON-state drive current for the same OFF-state leakage current. b) Energy per operation vs. delay can be lowered for a steep switching device. However, the impact of non-idealities such as process-induced variations will reduce the energy savings. After N. Damrongplasit, Study of variability in advanced transistor technologies, PhD Thesis, University of California, Berkeley, 2014.
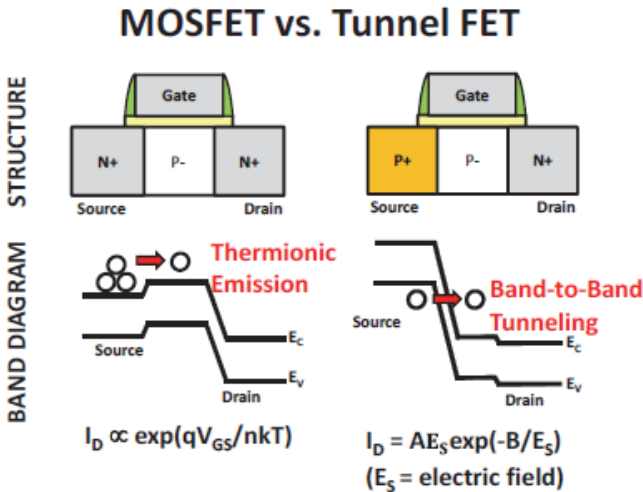
Figure 9.31. Comparison between MOSFET and TFET structures, and their operations. After N. Damrongplasit, Study of variability in advanced transistor technologies, PhD Thesis, University of California, Berkeley, 2014.

TFETs promise low leakage currents with steep sub threshold slopes, not limited by the 60 mV/decade limit of MOSFETs. However, they suffer from much lower ON currents due to low tunneling probabilities. The inclusion of SiGe and delta doping along with surround gate topologies show promise in improving ON currents. Further, the large parasitic capacitance due to the gate oxide and metal directly being on top of the source region needs to be reduced. This can be done by incorporating low-k materials which are thermally stable to allow metal anneal after deposition of this low-k material.

Unlike metal oxide semiconductor field effect transistors, TFETs can in principle operate at sub-threshold slopes of less than 60 mV/dec at room temperature. This results in higher I_{ON}–I_{OFF} ratio at reduced supply voltages thus enabling aggressive supply voltage scaling. Further due to the presence of the source side tunnel barrier the off current, I_{OFF} can be significantly smaller compared to MOSFET and is limited only by the reverse biased diode leakage current. This can significantly reduce the dynamic (proportional to square of the supply voltage) and static power dissipation in TFET (proportional to supply voltage and off current, I_{OFF}). Due to this unique feature of TFET, it is important to understand the tunnel devices to develop energy-efficient devices for future high performance and low power computing systems.

In the following, the switching behavior of TFET is studied through simulations. Analysis of the electrical characteristics of homo-junction TFET gives key insight into its device operation and identifies the critical factors that impact its performance. A TFET design with the planar Germanium-source TFET is considered. The structure extends the benefits of line tunneling further by employing a small bandgap material (i.e., Ge) in the source region of the device to enhance the tunneling current. It will be shown that by employing Germanium only in the source region of the TFET, I_{ON}/I_{OFF} ratio can be greatly enhanced to provide for significant energy savings compared to CMOS. Figure 9.32a and (b) show the cross-sectional schematics of an n-channel MOSFET and TFET, respectively. The only structural difference associated with a TFET is in the asymmetrically doped source and drain regions (p+−i−n+) vs. symmetrically doped source and drain regions (n^+−pn^+) of a MOSFET. The sub threshold swing observed for the state-of-the-art MOSFET technology is approximately 90 to 100 mV/dec. This number is expected to degrade with advancing technology nodes due to problems

(i.e., short channel effect) associated with scaling down the transistors to extremely small dimensions. This will further prevent the reduction of the threshold and supply voltages.

The operation of a TFET also relies on gate-voltage modulation of the channel potential for the injection of carriers (electrons for an n-channel and holes for a p-channel MOSFET) from the source into the channel region. However, in contrast to an injection over the potential barrier (i.e., thermionic emission) in a MOSFET, the carriers are injected into the channel through the potential barrier via a process called band-to-band tunneling (BTBT) (Figure 9.32d). The main advantage associated with BTBT is that the energy band gap cuts off the Boltzmann "tail" of the electrons in the p-type source region (holes for the n-type source region) (Figure 9.32d). BTBT is a quantum mechanical phenomenon in
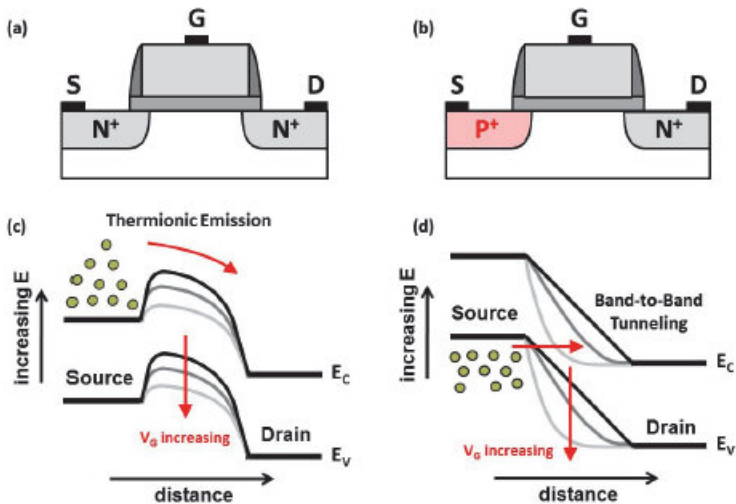
Figure 9.32. Schematic cross-sections for n-channel (a) MOSFET and (b) TFET. (c) The operation of a MOSFET is based on gate-voltage modulation of the channel potential for the injection of carriers over the barrier height from the source into the channel region through a process called thermionic emission. (d) The operation of a TFET is also based on gate-voltage modulation of the channel potential, but the carriers are injected into the channel through the potential barrier via band-to-band tunneling. After S. H. Kim, Germanium-source tunnel field effect transistors for ultra-low power digital logic, PhD Thesis, University of California, Berkeley, 2012.
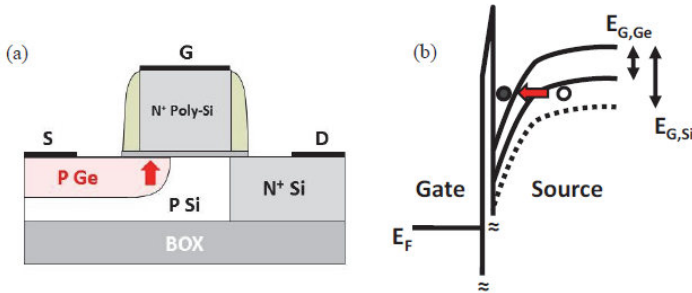
Figure 9.33. (a) Schematic cross section of the planar Ge-source n-channel TFET. (b) The corresponding energy band diagram along the perpendicular direction of the gate-to-source overlap area showing the band bending of Ge vs. Si. After S. H. Kim, Germanium-source tunnel field effect transistors for ultra-low power digital logic, PhD Thesis, University of California, Berkeley, 2012.

which the electrons "tunnel" across the energy gap of a semiconductor. From a physical perspective, BTBT is the plucking of electrons from covalent bonds formed between the semiconductor atoms. Sentaurus Device will be used for design optimization study.

The selective use of Ge only in the source region of the device enhances the on-state drive current of the TFET as compared to a Si device. Ge has approximately half the bandgap as compared to Si (0.66 eV vs. 1.2 eV) and smaller effective mass ($0.06m_o$ vs. $0.2m_o$), which provide for smaller B factor and hence exponentially larger I_{ON}. In other words, for the same gate-overdrive, Ge allows for more energy band overlap and smaller tunneling distance vs. Si (Figure 9.33b). In addition to the I_{ON} benefit arising from the use of Ge-source, I_{OFF} can be effectively suppressed due to the presence of Ge-Si heterojunction (Ge source to Si channel) which improves the overall I_{ON}/I_{OFF} ratio of the Ge-source TFET as compared to an all-Si or all-Ge device. In the off state ($V_g = 0V$ and $V_d = V_{DD}$), the n-channel TFET is reverse biased and the dominant source of I_{OFF} comprise of the reverse-bias p-i-n diode leakage.

In order to gain better understanding of the device and perform further design optimization study, TCAD (SDevice) was used to investigate the impact of various physical parameters of the device. This section will highlight the optimization methodology and propose a nominal planar Ge-source n-channel TFET. The cross-sectional schematic of the $L_g =$

30nm Ge-source n-channel TFET structure used to perform the design optimization study is shown in Figure 9.34. The embedded-Ge source region is heavily doped (10^{19} cm^{-3}) p-type, the Si body region is doped p-type, and the Si drain region is heavily doped (10^{19} cm^{-3}) n-type. The underlying buried oxide layer is 200 nm thick. The physical gate length is 30nm and the gate dielectric equivalent SiO$_2$ thickness (EOT) is 1nm, relevant for state-of-the-art CMOS technology. N$^+$ poly-Si gate (4.0 eV work function), 8nm wide silicon-nitride gate-sidewall spacers, and 5nm gate-to-source overlap (L$_{Ge}$) are used. Fixed charge at the gate-dielectric/Ge interface, ~10^{11} q/cm^2 is assumed.

The most crucial design parameters are the source doping and the gate-to-source overlap area, both of which determine the on-state drive current and the steepness of the turn-on characteristics. In order to maximize the I$_{ON}$/I$_{OFF}$ ratio for the largest energy savings, it is not only important to maximize I$_{ON}$, but also crucial to suppress I$_{OFF}$ as much as possible. Figures 9.35a and b show the simulated transfer characteristics for various body thicknesses (TBODY), for light body doping (NBODY = 10^{15} cm^{-3}) and moderate body doping (10^{18} cm^{-3}), respectively. Generally, the leakage floor (I$_{OFF}$) is reduced with moderate body doping. The I$_{OFF}$ dependence on TBODY changes with NBODY, which suggests that the dominant leakage mechanism also changes with body doping.
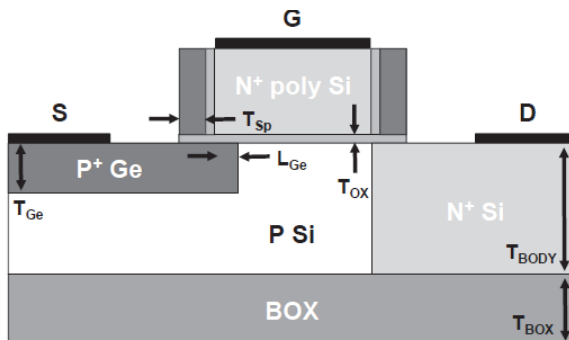
Figure 9.34. Cross sectional schematic of the planar Ge-source n-channel TFET used to perform design optimization study in Sentaurus Device. L$_g$ is 30nm and other physical parameters are varied to investigate their impact. After S. H. Kim, Germanium-source tunnel field effect transistors for ultra-low power digital logic, PhD Thesis, University of California, Berkeley, 2012.
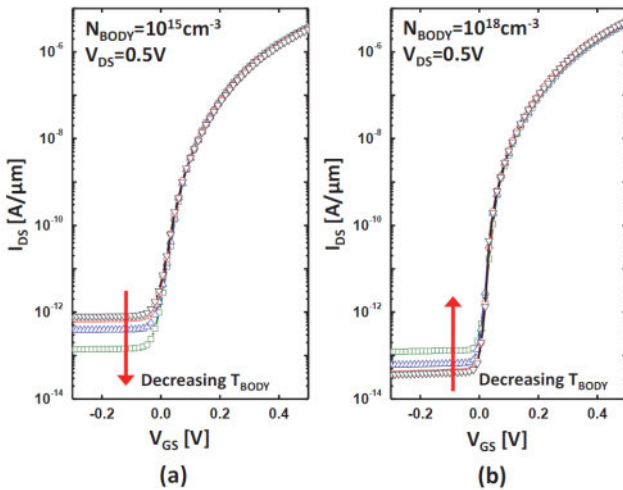
Figure 9.35. Comparison of simulated transfer characteristics for various body thickness values (TBODY = 20, 30, 50, and 100nm). (a) NBODY =10^{15} cm^{-3}. (b) NBODY = 10^{18} cm^{-3}. After S. H. Kim, Germanium-source tunnel field effect transistors for ultra-low power digital logic, PhD Thesis, University of California, Berkeley, 2012.
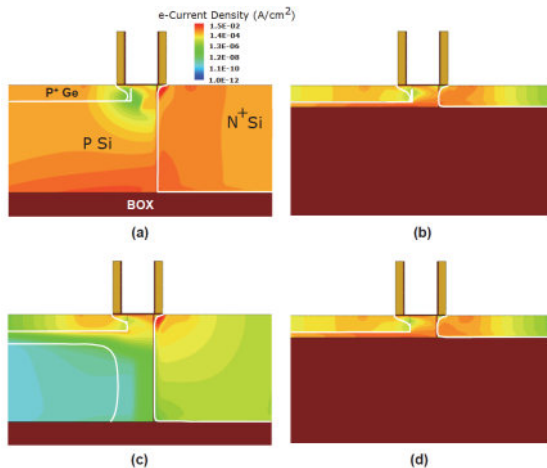
Figure 9.36. Contour plot of electron current density in the OFF-state for (a) NBODY = 10^{15} cm^{-3} and TBODY = 100nm, (b) NBODY = 10^{15} cm^{-3} and TBODY = 20nm, (c) NBODY = 10^{18} cm^{-3} and TBODY = 100nm, and (d) NBODY = 10^{18} cm^{-3} and TBODY = 20nm. The edges of the depletion regions are denoted by white lines. The dominant leakage path depends on the body doping concentration and body thickness. After S. H. Kim, Germanium-source tunnel field effect transistors for ultra-low power digital logic, PhD Thesis, University of California, Berkeley, 2012.

Figure 9.36 shows simulated off-state current contour plots for various combinations of TBODY and NBODY. Since lower body doping corresponds to larger vertical and lateral depletion widths and larger minority carrier concentration, the TFET off-state leakage is dominated by the reverse diode current of body-drain junction for NBODY = 10^{15} cm^{-3} (Figure 9.36a). As TBODY is scaled down (Figure 9.36b), the body-drain junction area is reduced and the leakage floor correspondingly decreases (Figure 9.35a). For NBODY = 10^{18} cm^{-3} and TBODY = 100nm, the body is only partially depleted and its minority carrier concentration is lower, so that the reverse diode current of the body-drain junction is negligible (Figure 9.36c). The TFET off-state leakage is dominated by the surface component. If TBODY is reduced such that it becomes fully depleted, however, the situation becomes similar to that for the lightly doped body: body-drain reverse diode current becomes significant (Figure 9.36d) and the leakage floor correspondingly increases (Figure 9.35b). These results indicate that if moderate body doping (10^{18} cm^{-3}) is used, then it is not necessary to scale down the body thickness for optimal performance; rather, it is preferable to avoid a fully depleted body region because it results in larger sub-surface reverse-diode leakage current.

9.8 Nanowire MOSFETs

Gate-All-Around devices offer the best potential solution to electrostatic confinement challenges. Nanowires are an extreme case of GAA devices. One possibility is to extrapolate the FinFET concept by using a vertically stacked nanowire device that is completely surrounded by a cylindrical gate. The superior performance of these devices comes from a high I_{ON}/I_{OFF}, due to the gate-all-around structure, which improves the electrostatic control of the channel, thereby reducing the leakage current of the device. Typical process integration schematic of nanowire MOSFETs is shown in Figure 9.37.

Nanowire FETs are fabricated starting from 300nm (001) SOI or sSOI wafers with ~10–15nm Si layer and 145nm BOX, and following the process steps shown in Figure 9.37. FD-SOI-based device fabrication results in undoped i.e., high purity channel. For PMOS enhancement, a process to form SGOI is performed. The devices are patterned by using

the top-down approach, consisting of mesa isolation and hybrid DUV/e-beam lithography followed by a resist trimming process. This is done in order to achieve NW structures as narrow as 10nm (top-view width). An Hf-based high-k/metal gate stack (HfSiON/TiN) is then processed. After the gate pattering, first nitrided spacer (SiN) is formed, and then raised S/D were realized by epitaxial Si (or SiGe for PMOS improvement), in order to get low parasitic resistance. Next, S/D implantation with phosphorus for NMOS and with boron for PMOS, respectively, is performed twice. By the first lightly doped drain and second highly doped drain implantation processes, SCE and hot-carrier generation are effectively reduced. Compressive CESL is deposited to enhance the performance in some PMOS devices. After dopants activation and silicidation of S/D regions, the fabrication ended with a standard back-end of line (BEOL) process.

9.9 Nanowire Transistors

Nanowire transistors feature a better gate control because they are small and the gate can be very close to the channel. In the following, based on work reported in reference [9.2], nanowire FinFETs are studied using

Figure 9.37. Process integration schematic of MOSFETs. After M. Koyama, Electrical characterization of interface properties in nano-scaled MOSFET devices based on low-frequency fluctuations, PhD Thesis, Universite Grenoble Alpes, 2015.

simulation. Schrödinger Equation Monte Carlo in Three Dimensions (SEMC-3D), a simulator that handles quantum transport and scattering in 3D nano-scale MOSFETs has been used in simulation. Schrödinger Equation Monte Carlo (SEMC) technique bridges the gap from fully phase-coherent quantum transport to semi classical transport for a physically rigorous treatment of quantum transport and inelastic scattering processes such as optical and acoustic phonon scattering. SEMC has been described as a nonequilibrium Green's function (NEGF) approach which has become the standard for quantum transport modeling.

SEMC solves the Schrödinger equation for modeling a coupled carrier-phonon many-body system. While NEGF formalism provides physical accuracy, the Monte Carlo technique specific to SEMC provides computational efficiency. SEMC-3D provides a new insight into transport phenomena in nanoscale MOSFETs, particularly with regard to interplay among scattering, quantum confinement and transport, and strain. SEMC-3D has been employed for investigating the carrier transport phenomenon in nanoscale 3D devices such as NW MOSFETs. The effects of different kinds of strain on the carrier transport in the Si NW n-MOSFETs are also studied using SEMC-3D.
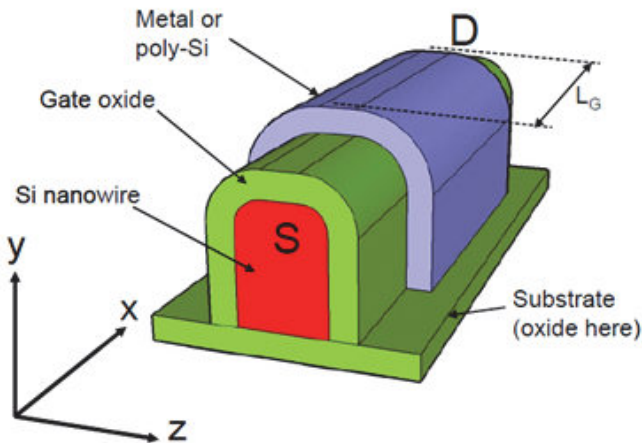
Figure 9.38. Si nanowire n-MOSFET used for SEMC-3D simulation. The coordinate system adopted is also shown. After K.-M. Liu, Schrödinger equation Monte Carlo-3D for simulation of nanoscale MOSFETs, PhD Thesis, University of Texas at Austin, 2008.

Figure 9.38 shows schematically the type of device structure considered, a Si n-channel FinFET or nanowire (NW) MOSFET with a 10nm long channel, 5nm source and drain lengths, and the cross section, as shown in Figure 9.39, consists of a 3nm channel width and 4nm height with rounding at the top and 1nm gate oxide thickness. The source and drain extensions are doped with $10e^{20}$ cm^{-3} and the channel is undoped to avoid the random dopant fluctuation effects. The number of incident carriers from each side S/D varies from hundreds to thousands; 20 sub bands are considered.

Figure 9.40a shows the $I_d - V_g$ characteristics with and without scattering for $V_d = 0.4V$. Figure 9.40b shows the $I_d - V_d$ characteristics with and without scattering for $V_g = 0.4V$. From Figures 9.40a and 9.40b, the continued degradation of drive current and transconductance due to phonon scattering are evident even for such a short channel device. When we consider a higher drain voltage of 0.5V with this device structure, ballistic simulation also exhibited an artificial "anti-screening" of the otherwise small potential well in front of the source-to-channel barrier as shown in Figure 9.40a.
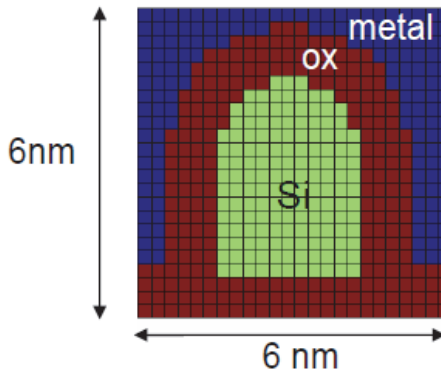
Figure 9.39. The cross section of Si nanowire n-MOSFET used for SEMC-3D simulation. After K.-M. Liu, Schrödinger equation Monte Carlo-3D for simulation of nanoscale MOSFETs, PhD Thesis, University of Texas at Austin, 2008.
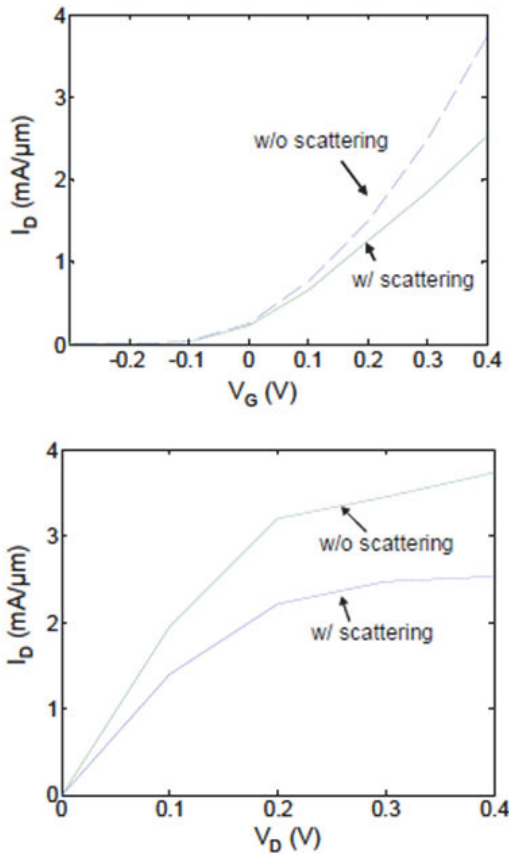
Figure 9.40. (a) The $I_d - V_g$ characteristics in linear scale with and without phonon scattering for $V_d = 0.4V$. (b) The $I_d - V_d$ characteristics with and without phonon scattering for $V_g = 0.4V$. After K.-M. Liu, Schrödinger equation Monte Carlo-3D for simulation of nanoscale MOSFETs, PhD Thesis, University of Texas at Austin, 2008.

Effects of adding surface roughness scattering in simulation are studied next. For a bias condition, $V_g = V_d = 0.4V$, simulations are performed for four conditions regarding scattering for comparison: (1) without scattering, (2) with phonon scattering only, (3) with surface roughness scattering only and (4) with phonon and surface roughness scattering. Figure 9.41 shows the carrier density profiles vs. position along the channel under these conditions as a function of the scattering model. Surface roughness scattering increases the scattering rate, but it does not
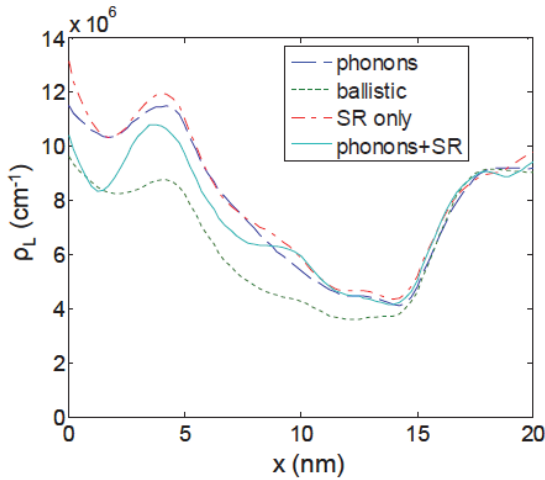
Figure 9.41. The carrier densities vs. position along the channel under 4 conditions regarding scattering: (1) without scattering, (2) with phonon scattering only, (3) with SR scattering only and (4) with phonon and SR scattering. The bias condition is $V_g = V_d = 0.4V$. After K.-M. Liu, Schrödinger Equation Monte Carlo-3D for Simulation of Nanoscale MOSFETs, PhD Thesis, University of Texas at Austin, 2008.

decrease the drain current further when combined with phonon scattering.

Most studies of strain effect on carrier transport either use semi classical approaches or consider the ballistic transport without scattering. SEMC-3D provides a means to investigate the strain effect in Si NW n-MOSFETs based on a quantum transport framework including scattering and 2D quantum confinement. In the following, SEMC-3D has been used to examine the effects of uniaxial and biaxial strain on electron transport in Si NW n-MOSFETs. To simulate tensile strain is applied and the strain strength is 0.5%.

Figure 9.42 shows the $I_d - V_g$ characteristics of these two types of strain in ballistic transport and with phonon scattering. It is observed that in the ballistic case, the effect of strain is not significant because of the pre-existing degeneracy splitting by strong quantum confinement in NW MOSFETs. However, in the case considering phonon scattering, the effect of strain is notable (about 15% increase in drain current at $V_g = V_d$
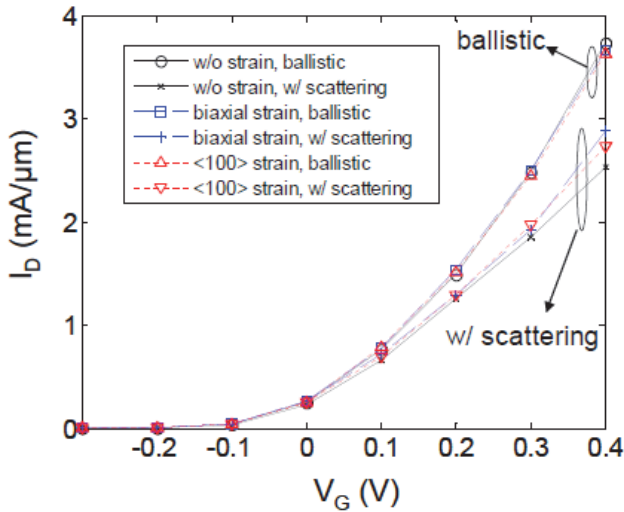
Figure 9.42. The $I_d - V_g$ characteristics of 0.5% tensile biaxial strain and 0.5% tensile uniaxial $\langle 100 \rangle$ strain with and without considering phonon scattering under $V_d = 0.4V$. Both types of strain have $\langle 100 \rangle$ channel orientation. The $I_d - V_g$ characteristics of the same device without strain are also shown as reference. After K.-M. Liu, Schrödinger equation Monte Carlo-3D for simulation of nanoscale MOSFETs, PhD Thesis, University of Texas at Austin, 2008.

$= 0.4V$ for biaxial strain). The effect of biaxial strain is slightly stronger than uniaxial $\langle 100 \rangle$ strain.

9.10 FinFETs with GST Stress Liners

Transistors with pure Ge channels are being investigated for implementation in next generation microelectronic devices. The ultimate advantage of working with Ge is the carrier mobility enhancement, making it attractive for high-speed circuit applications. Low field electron mobility in Ge is more than double that of Si (3900 vs. 1500 cm^2/Vs) and the increase is four-fold for holes (1900 vs. 450 cm^2/Vs). Despite the intrinsic speed advantages of implementing Ge transistor technology, Ge has not established a strong presence as an electronic material for ubiquitous microelectronic application because it does not form a stable oxide, which is critical for gate electrode formation, limiting its utility in traditional MOS manufacturing methods.

However, the research work on strained Ge transistors, especially Ge MuGFETs or NW FETs, is very limited. While most of strained Ge p-MOSFETs were fabricated on the biaxially strained Ge layer grown on $Si_{1-x}Ge_x$/Si substrates, some of the strained Ge transistors were realized by the thermal condensation of $Si_{1-x}Ge_x$ on SOI substrates. Extremely high mobility Schottky gate modulation-doped field-effect transistors (MODFETs) have been reported. Biaxially-strained Ge p-MOSFET with a gate stack comprising Si/SiO_2/low-temperature oxide/poly-Si, showing a peak hole mobility of 1160 cm^2/Vs. In recent years, high performance strained Ge p-MOSFETs with high-k/metal gate (HKMG) were demonstrated by various research groups, which offers possible solutions for the integration of strain on Ge transistors in the ultra-scaled technology nodes.

Biaxially strained Ge p-MOSFETs was reported to have 4 times higher hole mobility than the state-of-the-art strained Si MOSFETs at a hole density NS of 5×10^{12} cm^{-2}. As compared to biaxial strain, uniaxial strain could enhance the hole mobility more effectively. Uniaxially or asymmetrically strained Ge NW FETs exhibit even higher hole mobility than the biaxially strained Ge p-MOSFETs. Stress liners can also induce asymmetric strain in the transistor channel with simple integration process. However, so far there has been no report on the realization of Ge FETs with stress liners technologies.

A new strain engineering concept involving volume contraction of the liner material has been reported [9.3]. Phase change material GeTe is used as a stress liner, exploiting its property of volume contraction when phase-changed from the amorphous state (α-GeTe) to the polycrystalline state (c-GeTe). Following reference [9.3], simulation of strained p-channel FinFETs with GeTe liner stressor are discussed. A finite element method simulation has been performed to study the impact of GeTe-induced strain on the Si valence band structure. GeTe liner stressor results in reduction in hole effective mass and the band dispersion between HH and LH near Γ point. Electrical characterization of Si p-FinFETs with and without GeTe liner stressor was carried out. Significant drain current enhancement has been observed for FinFETs with c-GeTe liner stressor over the control devices.

In the following, we present the integration of Germanium Antimonite Telluride (GST) liner stressor with Ge gate all around nanowire FETs formed on GeOI substrates. Ge NWs with wire width down to 3.5nm were fabricated for the control of short channel effects. Si CMOS compatible process including low-temperature Si passivation, HKMG, and self-aligned metallic Schottky-Barrier nickel germanide (NiGe) source/drain (S/D) was used to fabricate the high performance Ge NW transistors. With the integration of GST liner stressor, significant enhancement of drain current and peak transconductance could be achieved for Ge NW p-FETs. 3D stress simulation for GST-strained Ge is performed to study the mechanism of the hole mobility enhancement. The stress simulation showed the good scalability of GST liner stressor.

9.11   GST for Strain Engineering

Figure 9.43 shows a 3D schematic of a GST stress liner wrapped around a GAA FET formed on GeOI substrate. When amorphous GST (α-GST) is crystallized to polycrystalline GST (c-GST), the volume of the GST layer is reduced (Figures 9.44a and b). Therefore, the Ge NW wrapped around by the GST liner is squeezed. Figures 9.44c and d illustrate the key concept of this structure using cross-section schematics of the transistor in the A-A′ plane [(110) plane, cutting through gate line and perpendicular to fin) and B-B' plane [(Π0) plane cutting through fin and perpendicular to gate line], respectively. The liner is amorphous when first formed over the FinFET. When GST undergoes phase change or crystallization from α-GST to c-GST, the volume contraction causes it to constrict or tighten its grip on the FinFET structure. In Figure 9.44c, the contraction of GST liner results in a downward force on the S/D regions, causing S/D regions to expand laterally. Thus, the channel region under the gate is compressed laterally, i.e., along [110] direction [source (S)-to-drain (D)]. In Figure 9.44d, the contraction of GST liner causes the Ge NW to be compressed in both transverse fin and vertical directions.
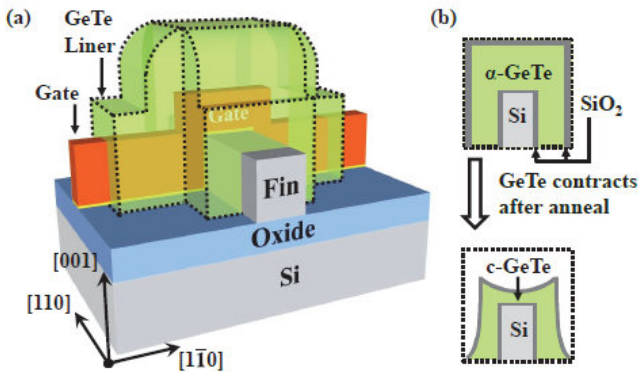
Figure 9.43. Three-dimensional schematic of a Ge GAA NW FET wrapped by a $Ge_2Sb_2Te_5$ (GST) liner stressor. The GST liner stressor comprises a bottom Al_2O_3 isolation layer, a GST liner layer, and a SiO_2 capping layer. The Source (S)-to-Drain (D) direction is along [110] axis, the TaN gate line is along [Π10] axis and the out-of-plane direction is along [001] axis. After C. Ran, Strain engineering for advanced silicon, germanium and germanium-tin transistors, PhD Thesis, National University of Singapore, 2014.
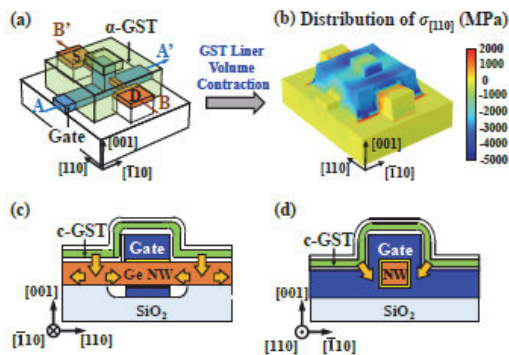
Figure 9.44. (a)–(b) Numerical simulation study of the channel stress and strain induced by the GST liner stressor. As GST changes phase from amorphous to polycrystalline state, its volume is reduced, squeezing the channel region. (c) 2D illustration of the stress transfer mechanism in BB′ Plane. Due to the contraction of GST layer, a downward force would exert on the S and D regions. Thus, the two regions tend to expand laterally, resulting in a contraction in the channel region. (d) 2D cross-section on AA′ plane showing that GST liner would squeeze the Ge nanowire in both transverse fin and vertical directions. After C. Ran, Strain engineering for advanced silicon, germanium and germanium-tin transistors, PhD Thesis, National University of Singapore, 2014.

Model Dimensions	
Gate Length L_G = 150 nm	Gate Height H_G = 100 nm
Fin Width W_{NW} = 10 nm	Fin Height H_{NW} = 30 nm
GST Thickness T_{GST} = 70 nm	NW Length L_{NW} = 500 nm
Buried Oxide Thickness T_{BOX} = 120 nm	

Figure 9.45. A 3D finite element method simulation model was established for the stress simulation based on the framework of thermo elasticity. The dimensions of the model used for simulation are shown in the inset. After C. Ran, Strain engineering for advanced silicon, germanium and germanium-tin transistors, PhD Thesis, National University of Singapore, 2014.

Strain Simulation

In this section, 3D strain simulation was performed to study the effect of the GST liner stressor on the channel strain profiles. Figure 9.45 shows a schematic of the simulated structure. The geometric parameters (Figure 9.45 inset) were determined from the fabricated device structure, i.e., gate length L_g = 150nm, gate height = 100nm, NW width = 10nm, NW height = 30nm, and NW length = 500nm. The NW FET is conformally covered by a layer of amorphous GST (α-GST) at a thickness Tα-GST of 70nm. Upon crystallization, the GST volume contracts by ~6.5%, giving the crystallized GeTe (c-GeTe) a thickness T_c-GST of 65nm.

The boundary conditions of the model were set as follows. The upper surfaces normal to [100] axis was set to be free surfaces. The bottom surface of the substrate was fixed with the zero displacement in the vertical or [001] direction. The GST film was allowed to relax in both S-to-D ([110]) direction and transverse wire ([110]) direction. The lateral dimensions of the buried oxide and Si substrate were set to be large

enough so that the stress near the substrate boundaries is negligible as compared to the stress in the channel region. The Young's modulus for Si, SiO$_2$, TaN, and GST are taken as 165 GPa, 70 GPa, 187 GPa and 59 GPa, respectively. The stiffness constants used to convert the stress tensors to the strain tensors are C$_{11}$ = 155.6 GPa, C$_{12}$ = 21.5 GPa, C$_{13}$ = 47.5 GPa, C$_{33}$ = 129.1 GPa, C$_{44}$ = C$_{55}$ = 66.5 GPa, and C$_{66}$ = 41.0 GPa. The contraction of the GST layer was simulated using the framework of thermo elasticity.

As shown in Figure 9.45, plane CC′ is cut through the NW along [110] or S-to-D direction, and is 1nm away from the NW sidewall which is close to the peak of the inversion charge density. Figure 9.46 shows the strain profiles of ε[110] in the S-to-D direction and ε[110] in the transverse NW direction. ε[110] is ~−0.8% and ε[110] is ~0.5% in the transistor channel region.
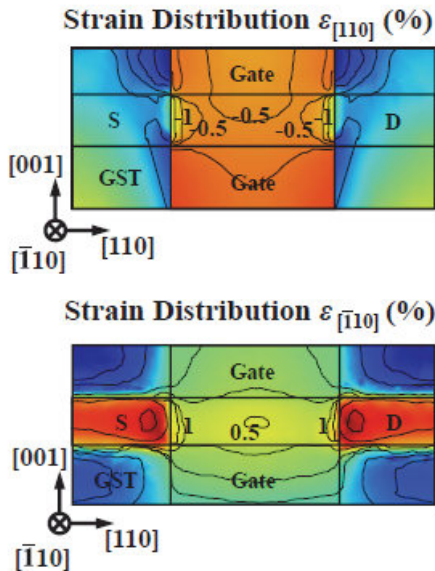
Figure 9.46. Channel strain in both longitudinal [110] and transverse [110] directions cut along A-A′ plane and is 1nm within the NW sidewall. The dimensions of the simulated structure were kept the same as the fabricated device. The Ge channel region is compressive strained in [110] and is lightly tensile strained in [110]. After C. Ran, Strain engineering for advanced silicon, germanium and germanium-tin transistors, PhD Thesis, National University of Singapore, 2014.
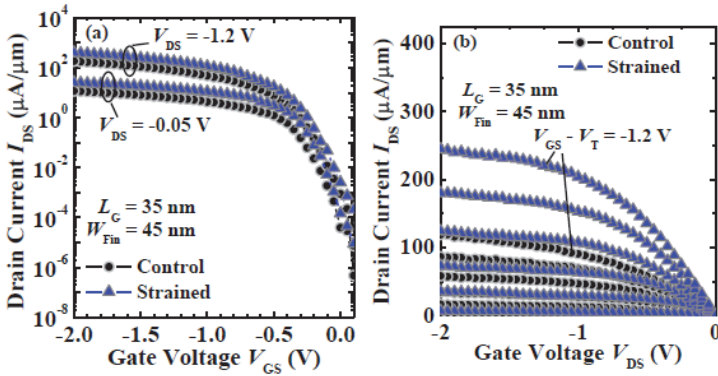
Figure 9.47. (a) $I_d - V_g$ characteristics of strained (with c-GeTe) and control p-FinFET with L_g = 35nm and W_{Fin} = 45nm, measured at V_d of −0.05V and −1.2V, showing comparable SS and DIBL and (b) $I_d - V_d$ characteristics of the same pair of devices as shown in (a), measured at ($V_g - V_t$) of −0.2V to −1.2V in steps of −0.2V. I_d enhancement of ~110% is observed at $V_g - V_t = -1.2V$. After C. Ran, Strain engineering for advanced silicon, germanium and germanium-tin transistors, PhD Thesis, National University of Singapore, 2014.

Figure 9.47a shows the $I_d - V_g$ characteristics of a pair of control and strained FinFETs. For both devices, L_g is 35nm and W_{Fin} is 45nm. The drain current I_d is normalized by the effective device width WEFF which is given by $2H_{Fin} + W_{Fin}$. The fin height H_{Fin} is 35nm. The devices have comparable sub threshold swing of ~68 mV/V and drain induced barrier lowering of ~92 mV/decade. A slight threshold voltage shift could be observed for FinFET with c-GeTe liner stressor due to the strain induced bandgap narrowing and change in the density of states. Figure 9.47a shows the $I_d - V_d$ characteristics of the same pair of devices as shown in Figure 9.47. Drain current enhancement of ~110% could be observed for the strained FinFET over the control at a gate overdrive $V_g - V_t = -1.2V$.

9.12 Future Directions

CMOS utilizing high mobility Ge/III–V channels on the Si platform is expected to be one of the promising devices for high-performance and low power logic LSIs in the future. To realize the III–V/Ge logic CMOS, there are many critical issues including the technologies of the gate

stacks, the channel formation, low resistivity S/D formation, formation of UTB channels with high immunity against SCEs, and CMOS integration. However, the device-process-integration technologies for next generation Ge/III–V n- and p-MOSFETs have not yet been established. As CMOS technology evolves towards the 7nm technology node, alternative channel materials have been widely explored to overcome the limitation of the silicon channel. Germanium is the one of the most promising candidates. Therefore, it is necessary to perform TCAD analysis of SiGe FinFETs to determine the best performance, design, and manufacturability trade-offs. To perform 3D analysis of 7nm SiGe FinFETs one needs to include the following features:

- Germanium concentration effect on the diffusion
- Stress effect on the diffusion
- SiGe stress-relaxed buffer (SRB) layer model
- Device Monte Carlo model with electronic band structure and scattering mechanisms
- Corrected drift-diffusion model with ballistic mobility and adjusted saturation velocity

The steps include; creating the FinFET, modeling stress evolution, germanium and dopant diffusion during the process flow, quantum correction for the device Monte Carlo simulation, Monte Carlo simulation of the FinFETs, and drift-diffusion simulation with ballistic transport models. The tool flow of this starts with Sentaurus Process (SProcess) to calculate the geometry, the doping, and the stress. Then, Sentaurus Visual (SVisual) measures the channel stress. Sentaurus Process (SProcess) generates the structure and the mesh for the Sentaurus Device Monte Carlo simulation. The next Sentaurus Mesh node (x-slice) cuts the 3D FinFET structure at the middle of the channel and generates a 2D FinFET cross section for the 2D Sentaurus Device simulations of the quantum-correction steps.

While silicon based nanodevices will continue to dominate consumer electronics for this decade, it is well-understood that conventional Moore's Law scaling must come to an end sometime by the next decade, due to a combination of on chip power dissipation and speed limitations. It is therefore highly probable that new materials and devices will take

the place of Si-CMOS and related devices, which have dominated the market for the past 40 years.

Two of the most promising contenders for carbon electronics, as a future replacement for Si-CMOS, are Graphene and Carbon nanotube based transistors. Both these technologies are compatible with the current CMOS process flow. Two-dimensional graphene films have generated a huge interest recently as an alternative for channel replacement material in MOSFET structures. Graphene films are well known to behave as high-mobility zero band-gap semiconductors with high carrier mobilities. Researchers have already demonstrated high-speed devices in the range of 300 GHz and are expected to go up to 1 THz. This opens applications in the RF-analog range. From an integration point of view, graphene devices are planar and compatible with CMOS process. However, one of the main drawbacks for VLSI compatibility is the "Zero band gap". Because of the zero bandgap, devices implemented on large-area graphene channels cannot be switched off and therefore are not suitable for logic applications.

Bandgap can be induced in graphene by cutting the material into thin ribbons or applying an electric field to bilayer graphene. When patterned to sufficiently small ribbon widths, the graphene ribbons begin to display a finite band gap resulting from quantum confinement. Opening a band gap requires nano ribbons with sub 5nm width coupled with very well-defined edges. Variation in edge roughness has a huge impact on the mobility of the device. On the other hand, bilayer graphene requires applied voltages of around 100V to create a bandgap of about 0.25 eV which is simply not feasible for IC applications. However, Graphene is highly desirable in other venues like Optoelectronics, NEMS, and Spintronics.

One of the close contenders for Graphene is a mono-layered material called Molybdenite (MoS_2). MoS_2 has a direct bandgap (1.8 eV) from the start and does not need to be made into nano ribbons for semiconductor applications. Single-layer molybdenite is a direct-bandgap semi-conductor, unlike silicon, which has an indirect gap. It is easier to make devices like LEDs, solar cells and photo detectors and any other photonic devices with direct rather than indirect gap semiconductors.

Summary

Advanced MOFET structures have been developed to enable continued transistor scaling beyond the limit of the planar bulk MOSFET, but they also pose new challenges for integrated circuit manufacturing. TCAD modeling of the state-of-the-art devices with III–V and wide bandgap material have been performed in this chapter. Tunnel FETs are a promising alternate to MOSFETs for low power design due to its ability to scale threshold voltage and hence the supply voltage, without increase in I_{OFF} currents. In this chapter, we have presented a planar Germanium-source n-channel TFET which overcomes the fundamental limits in poor swing and low drive current associated with the conventional source-to-channel lateral tunneling device design. It has been shown that employing Ge only in the source region and introducing gate-to-source overlap area are effective methods to improve I_{ON}/I_{OFF} ratio for low supply voltage operation.

Simulation of strained p-channel FinFETs with GeTe liner stressor was also considered. The strain effect of phase-change liner stressor GeTe on Si p-FinFETs by simulation and electrical characterization was discussed. The strain induced by the GeTe liner reduces the hole effective mass and increases the energy separation between the light and heavy hold bands, both of which are beneficial for the hole mobility enhancement. GST liner stressor is a promising technique to boost device performance for Ge p-MOSFETs in future technology nodes.

Chapter 10

Memory Devices

Semiconductor memory is an indispensable component of modern electronic systems. Static Random Access Memory (SRAM) is used as a cache memory in personal computers since it offers the fastest write/read speed among all memories. However, a single SRAM cell consists of 6 transistors, so SRAM chip density is very low, although 4T SRAM cells have been demonstrated. SRAM memory can retain the stored information as long as the power is on, drawing very little current. However, the information will be lost when the power is turned off, so SRAM is not a nonvolatile memory.

A Dynamic Random Access Memory (DRAM) cell consists of one transistor and one capacitor. It is superior to SRAM in many aspects except that the write speed is slower in the DRAM than in the SRAM. However, its cell size is much smaller than that of SRAM and thus it is a low cost commodity memory device. Compared to flash memory, DRAM has much faster program/read speed with very low operating voltage, while flash memory needs 1μs to 1ms programming time and high programming voltage. Unfortunately, DRAM is a volatile memory. The data retention time is about 100ms in DRAM while it is 10 years in flash memory: a DRAM cell needs refreshing frequently to maintain its data, so its power consumption is significant.

Memory chips with low power consumption and low cost are essential for portable electronic devices such as cellular phones. These applications require the memory to have ten years data retention time and as such nonvolatile memory devices are indispensable. There are mainly four types of nonvolatile memory technology: flash memory, Ferro-electric Random Access Memory (FeRAM), Magnetic Random Access

Memory (MRAM) and phase change memory (PCM). Flash memory is most suitable for nonvolatile applications for the following reasons:

- Flash memory can achieve the highest chip density. A flash memory cell consists of only one transistor. A FeRAM memory cell generally consists of one transistor and one capacitor, while a MRAM cell needs a transistor and a magnetic tunnel junction. Phase change memory is expected to be a promising nonvolatile memory as well; however, its memory cell consists of one resistor and a bipolar junction transistor.

- Flash memory possesses the multi-bit per cell storage property. Four distinct threshold voltage states can be achieved in a flash memory cell by controlling the amount of charge stored in its floating gate.

- Flash memory fabrication process is compatible with the current CMOS process and is a suitable solution for embedded memory applications. A flash memory cell is simply a MOSFET cell, except that a poly-silicon floating gate (or Silicon Nitride charge trap layer is sandwiched between a tunnel oxide and an inter-poly oxide to form a charge storage layer.

All other nonvolatile memories require integration of new materials that are not as compatible with a conventional CMOS process. It is easier and more reliable to integrate flash memory than other nonvolatile memories with logic and analog devices. Since flash memory possesses these three key advantages, it has become the mainstream nonvolatile memory device now-a-days. However, flash memory exhibits some disadvantages as well. The device has a slow program/erase speed and requires a high voltage to program/erase its data. Additionally, its endurance also needs to be improved, although 10^5 program/erase cycles is enough for most applications. By applying a nitride layer (as a charge trap or charge storage layer) instead of floating gate, Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) has low pinhole density, which induce leakage and longer retention time. SONOS can also achieve the low-power and low-voltage performance because it can be applied lower voltage to program and erase.

Although the minimum feature size of CMOS devices have shrunk to 15nm with an equivalent gate oxide thickness (EOT) of 0.8nm, semiconductor flash memory scaling is far behind CMOS logic device scaling. For example, the EOT of the gate stack in semiconductor flash memory is still more than 10nm. Moreover, semiconductor flash memory requires operation voltages of more than 10V, while the operation voltage of CMOS logic has been scaled to about 1V or even less. However, the use of high-k dielectrics may solve the oxide scaling issues, it results in mobility and reliability degradation. By applying the high-k material as a charge trap or charge storage layer, i.e., Silicon-Oxide-high-k-Oxide-Silicon (SOHOS) has low pinhole density, which induced leakage and longer retention time.

To alleviate the tunnel-oxide design trade-off for floating-gate memory devices, a single-transistor memory-cell structure with discrete nanocrystal charge-storage sites embedded within the gate dielectric has been proposed. Nanocrystal memory cells feature program/erase characteristics comparable to conventional stacked-gate memories, excellent endurance and long-term non-volatility in spite of the thin bottom oxide (5.5–7nm), further underlying the interest in silicon nano-crystal memories for NVM applications.

The purpose of this chapter is to present simulation and modeling of SONOS and SOHOS memory devices using Sentaurus tools. We shall use technology CAD tools to design novel SOHOS structures which are expected to inherit the good data retention characteristics when one uses high-k dielectrics. High-k films trap electrons in the spatially isolated deep energy level, and thus possess the discrete traps advantages. Critical issues, such as, data retention characteristics and operating voltage of SOHOS structures using HfO_2 as dielectrics in non-volatile flash memory devices are investigated through TCAD simulation. The variation of the electric field and the injected charge as a function of device parameters will be studied. Process simulation is used to fabricate memory devices using high-k dielectric Aluminum Nitride (AlN) as charge trapping layer. Memory devices with Al-based (AlN) charge trapping layer show excellent charge storage characteristics than that of Si_3N_4 layer, demonstrating that high-k (AlN) based memory devices with SiO_2 as a blocking oxide can be useful for future flash memory device applications.

10.1   Non-volatile Flash Memory Devices

The basic operating principle of floating-gate nonvolatile semiconductor memory devices is the storage of electrons on a floating polysilicon gate that is isolated from the rest of the device by high-quality dielectrics (Figure 10.1). Since the floating gate controls at least part of the underlying transistor channel, the charge on this gate will directly influence the current in the channel. In other words, by storing electrons on the floating gate, the current can be switched off, leading to a first logical state (e.g., a logical 0). On the other hand, by removing the electrons from the floating gate, the channel will be able to transport current from drain to source and a second logical state is obtained (e.g., a logical 1). For external current control, a second gate, referred to as the control gate, is used. This implies that the external threshold voltage of the memory cell has to be defined from this control gate. In summary, the external threshold voltage of the memory transistor can be modified to switch between two distinct values, while the readout or sensing operation uses a readout voltage that lies somewhere between these two threshold voltages (Figure 10.2).

The schematic cross section of a floating gate device is shown in Figure 10.1. The upper gate is the control gate, the lower one, completely surrounded by dielectric is the floating gate. Oxide-Nitride-Oxide (ONO) is used to be the interpoly oxide uses because of its high dielectric constant that will increase the coupling ratio of electric field to the floating gate.
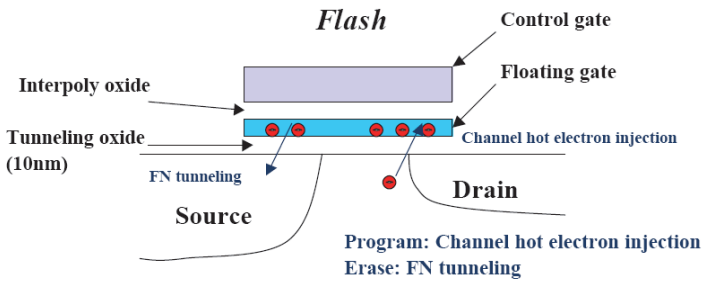
Figure 10.1. The schematic cross section of a floating gate Flash memory device.
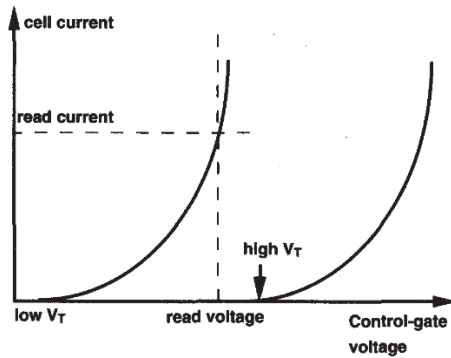
Figure 10.2. Sensing principle for floating-gate devices.

During programming, the stacked gate device transfers electrons to the floating gate through the oxide under the floating gate using drain-side channel hot electrons injection (CHEI). The high voltage on drain and control gate induces the channel hot electrons. A portion of the channel hot electrons with the right direction and sufficient energy to overcome the interface of silicon and gate oxide are injected into gate oxide. Although many channel hot electrons enter the gate oxide, few of them are transferred to the floating gate, causing a low programming efficiency. During erasing, the stacked gate cell transfers charge from the floating gate through the tunnel oxide, i.e., the Fowler-Nordheim tunneling mechanism pulls the electron out of the floating gate.

10.2 Silicon-Oxide-Nitride-Oxide-Silicon

Nowadays, Silicon-Oxide-Nitride-Oxide-Silicon (SONOS) becomes the most popular charge-trapping device because of its complete compatibility with existing advanced CMOS technology. SONOS has the better scaling capability than that of floating gate device. And it can be scaled down beyond 0.18μm. By applying the nitride layer instead of floating gate, SONOS has low pinhole density, which induced leakage and longer retention time. SONOS can also achieve the low-power and low-voltage performance because it can be applied lower voltage to
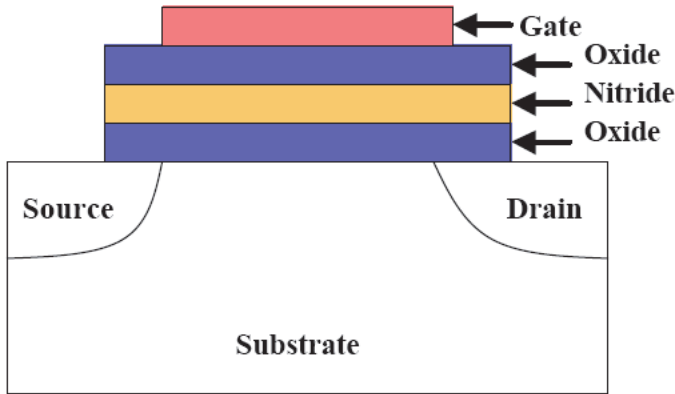
Figure 10.3. Cross section of a SONOS device.

program and erase. Figure 10.3 shows the simple structure of SONOS. During programming, either Fowler–Nordheim (FN) tunneling or channel hot electron injection (CHEI) can be used.

The Fowler-Nordheim current density is given by:

$$J \propto E_{inj}^2 \exp\left(-\frac{E_c}{E_{inj}}\right)$$

where

$$\alpha = \frac{q^3}{8\pi\phi_b}\frac{m}{m^*}$$

and

$$E_c = \frac{4\sqrt{2m^*}\phi_b^{3/2}}{3\eta q}$$

where ϕ_b is energy barrier at the injecting interface (3.2eV for Si–SiO$_2$), E_{inj} is electric field at the injecting interface, q is charge of a single electron (1.6×10^{-19} C), h is Planck's constant, m is free electron mass (9.1×10^{-31} kg), m^* is effective mass of an electron in the band gap of SiO$_2$, and $\eta = h/2\pi$. It should be noted that the tunnel current density is totally controlled by the field at the injecting interface, and not by the characteristics of the bulk oxide. Once the electrons have tunneled through the barrier, they are traveling in the conduction band of the oxide with a rather high saturated drift velocity of about 10^7 cm/s.

At sufficient drain bias, the minority carriers that flow in the channel of a MOS transistor are heated by the large electric fields seen at the drain side of the channel and their energy distribution is shifted higher. This phenomenon gives rise to impact ionization at the drain, by which both minority and majority carriers are generated. A consequence of carrier heating occurs when some of the minority carriers gain enough energy to allow them to surmount the SiO_2 energy barrier. If the oxide field favors injection, these carriers will be injected over the barrier into the gate insulator and give rise to the so-called channel hot- electron injection current. The lucky electron model assumes that an electron is injected into the gate insulator, if it can gain enough energy in the lateral field without undergoing a collision, by which energy could be lost. The lucky electron model is formulated as:

$$I_g = C\left(\frac{\lambda E_m}{\phi_b}\right)^2 \exp\left(\frac{-\phi_b}{\lambda E_m}\right)$$

where ϕ_b is zero oxide field energy barrier at the interface (3.2 eV), E_m is peak lateral electric field at the drain junction, C is pre exponential coefficient, and λ is electron mean free path. The effective electron temperature model assumes that the electrons are heated and become an electron gas with a Maxwellian distribution with an effective temperature that is dependent on the electric field. The gate current can then be calculated as the thermionic emission of heated electrons over the interface energy barrier.

When electrons tunnel through the silicon/oxide potential barrier or gain sufficient energy to overcome that barrier (hot electrons), the properties of the oxide film gradually deteriorate because of structural damage generated in the oxide. When the accumulated damage becomes too large, the oxide abruptly loses its insulator properties, which is defined as oxide breakdown.

The write/erase cycling endurance (or endurance for short) of a floating gate or charge trap device is the major reliability characteristic because it describes the gradual degradation of the cell and, therefore, its lifetime in terms of the number of write/erase cycles that can be applied before failure subjected to repeat write/erase cycling. For a single cell, the intrinsic endurance is usually monitored by measuring the threshold voltage window as a function of the number of applied write/erase cycles (shown in Figure 10.4). As long as the high threshold voltage is higher than the readout voltage, the leakage current in this state will be smaller

than the channel current at which the cell's threshold is defined. On the other hand, the low threshold voltage is usually a measure for the readout current of the cell and, depending on the type of cell considered, the threshold voltage value should not increase above a certain value in order to guarantee a sufficient readout current.

The most basic requirement for a nonvolatile memory is the ability to contain its information in the absence of an external power supply. Therefore, charge retention can be defined as "the ability to retain valid data over a prolonged period of time under storage conditions." This is usually characterized by monitoring the threshold voltage window vs. storage time at high temperature. Practice, temperature-accelerated tests are used, which means that the charge loss enhanced by using elevated testing temperatures (typically 250°C to 300°C). The charge loss through the bottom oxide and the lateral migration of trapped charges in the silicon nitride layer are mainly considered for the data retention loss in the localized storage SONOS. The program-state threshold voltage drops as the retention time increase and the erase-state threshold voltage increase as the retention time increases. Threshold voltage of program-state drops due to the trapped charge escape by Frenkel–Poole emission and oxide trap assisted tunneling. According to the tunneling front model, the threshold voltage shift can be expressed as:
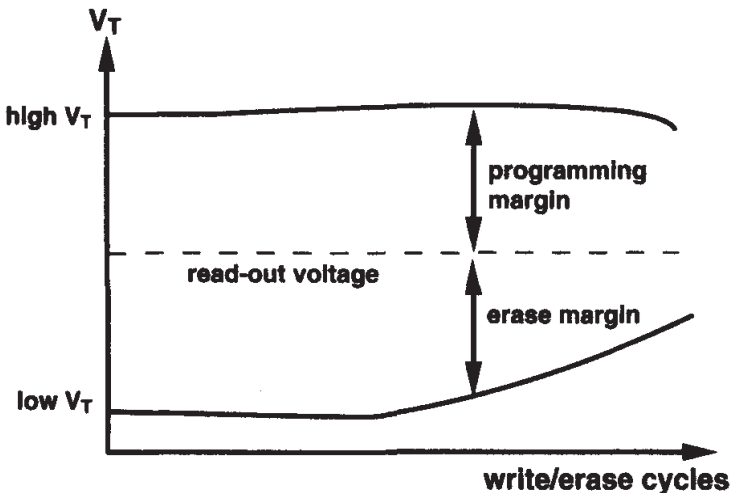
Figure 10.4. Endurance characteristics of a floating-gate device.

$$\Delta V_t(t) = 2.3 \frac{1}{C_{ONO}} \frac{q\hbar N_{OX}}{2\sqrt{2m\phi_{OX}}} \log(t)$$

where ϕ_{OX} is the energy level of the positive trapped charge, N_{OX} is the charge density.

10.3 Silicon-Oxide-High-k-Oxide-Silicon

Since in the programmed state of SONOS memory devices electrons are stored in silicon nitride layer, the data can be lost if electrons leak away from the injection region. As a result, the programmed state threshold voltage decreases. In the scaling process, if the gate oxide becomes too thin, excessive tunneling current may occur. To assure the same required capacitance of the MOS gate stack in devices, a dielectric with a higher dielectric constant than that of SiO_2 (3.9) is necessary. In such case, a thicker gate layer might be used that can reduce the leakage current flowing through the structure. Thus the high-k memory structures have drawn increasing attention in application for non-volatile flash memory devices due to superior charge trapping properties compared to the conventional poly-silicon floating gate. The high-k material sandwiched between the two silicon dioxide (SiO_2) layers to form the Silicon-Oxide-High-k-Oxide-Silicon (SOHOS) structure (shown in Figure 10.5), stores charges in spatial deep level traps, making it less vulnerable to a single defect in the tunnel oxide. This significantly helps to minimize the discharge of the memory cell.
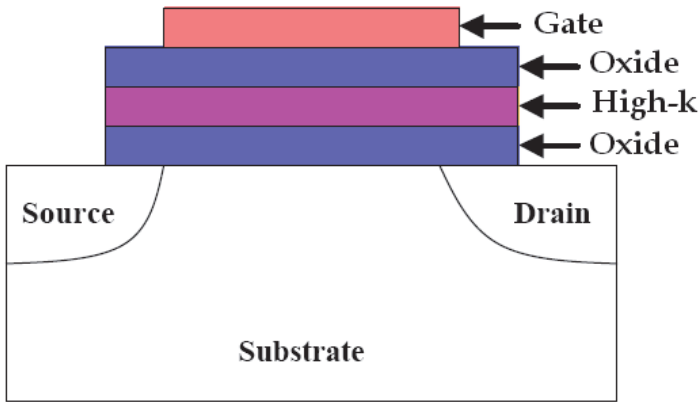
Figure 10.5. The cross section of SOHOS device.

During the write (program) operation of the SOHOS n-channel transistor device, electrons tunneled through the tunnel oxide and are stored in the deep level traps. During the erase operation under negative gate bias, electrons trapped in the high-k material are detrapped via tunneling through the oxide layer into the silicon substrate. However, there have been few reports for the trapping properties of the bulk traps in high-k material.

The possibility of enhancing the performance limits of the conventional floating-gate device is being explored. In conventional floating-gate flash memory, if there is one defect chain across the tunnel oxide, all of the charges stored on the floating-gate will leak back to either the channel or the source/drain though the defect chain. Thus the floating gate memory requires a thick tunnel oxide to prevent charge loss through the defect chain. The serious leakage problem during retention can be eliminated by utilizing a semiconductor nanocrystal memory structure (shown in Figure 10.6). Only the electrons stored on the nanocrystal directly above the defect chain will be affected since the nanocrystals are separated from each other within the gate oxide dielectric. Hence the tunnel oxide thickness in the nanocrystal memory device can be reduced to allow faster programming and lower voltage operation. To alleviate the tunnel-oxide design trade-off for floating-gate memory devices, a single-transistor memory-cell structure with discrete nanocrystal charge-storage sites embedded within the gate dielectric has been proposed.
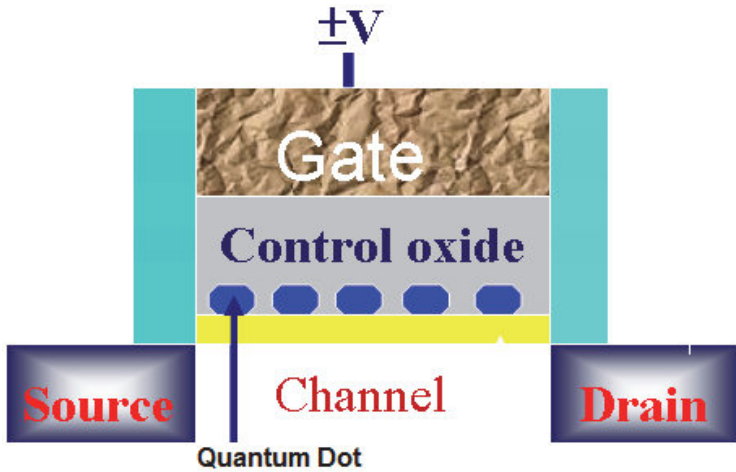
Figure 10.6. The cross section of Nanocrystal memory device.

SONOS: Programming and Erasing

Programming
- Electrons tunnel into nitride from channel during programming
- Peak shifted away from oxide/nitride interface
- Fowler–Nordheim tunneling
- Tunneling into nitride traps

Erasing
- Holes tunnel into nitride from channel during erasing.
- Peak shifted away from oxide/nitride interface
- Fowler–Nordheim tunneling

10.4 Simulation of SONOS Memory Devices

In the following, design of a 0.1μm SONOS single transistor memory cell via TCAD simulation is presented. The program and erase states of the device are simulated for several cycles to achieve a steady-state charge density in the device and then $I_d – V_g$ sweep was applied to determine the threshold voltage at the particular state of the device. Important considerations for embedded SONOS memories are; the lower

program and erase voltages compared to floating gate flash due to the thin bottom oxide. However, with low voltages (combined with high endurance), the data retention becomes an issue. The thickness of the top oxide is important to prevent Fowler–Nordheim tunneling of electrons from the gate during erase. To obtain the best compromise (optimization) between P/E voltage, endurance and retention, a detailed study has been taken up. Finally, a 10-year simulation is used to characterize the long-term charge retention capabilities of the SONOS structures.

The structure of the SONOS device is created by Synopsys TCAD tools and is based on a typical 100nm gate length NMOS device. The stack gate consists of an 18Å oxide layer, an 80Å silicon nitride (Si_3N_4) layer, as charge storage layer and 40Å oxide layer between the channel and poly (4.7 eV) gate. In simulation, models based on the Shockley–Read–Hall for trapping and recombination of the traps was used. Figure 10.7 shows an overall view of the doping distribution in the SONOS device. Figure 10.8 shows an overall view of the SRH recombination in the SONOS device.
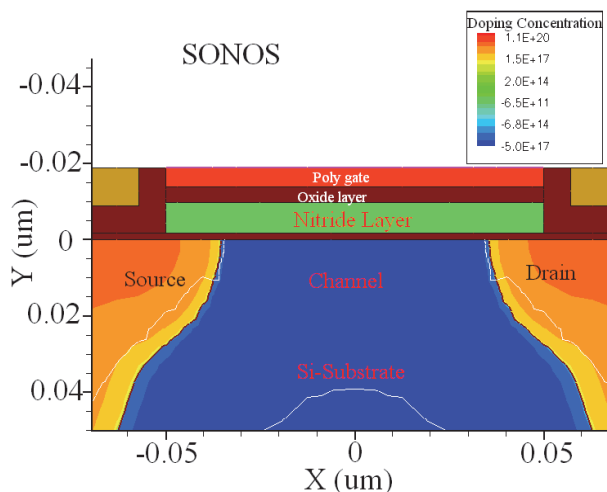
Figure 10.7. Flash memory device obtained from Sentaurus process simulation. After P. Chakraborty, Modeling and characterization of non-volatile flash memory devices, PhD Thesis, Jadavpur University, 2009.
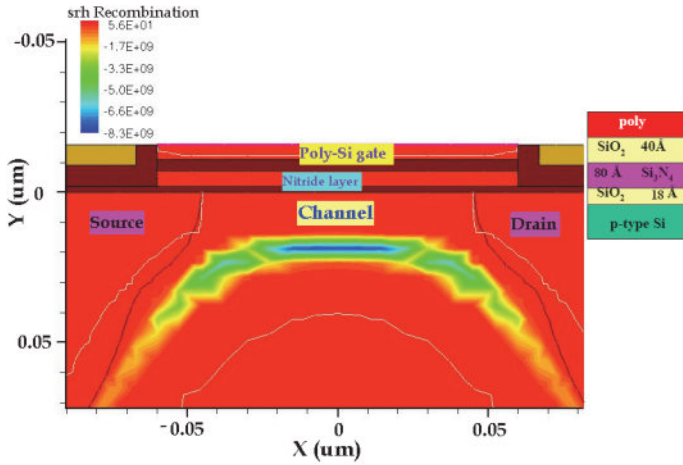
Figure 10.8. Distribution of SRH recombination in a SONOS device with gate length 100nm, nitride thickness 80Å. After P. Chakraborty, Modeling and characterization of non-volatile flash memory devices, PhD Thesis, Jadavpur University, 2009.

10.5 Programming/Erasing

The analysis of the erase/write charge transport and storage in SONOS devices is presented in this section for n-channel SONOS transistor with a negative gate potential (erase) and a positive gate potential (write). The latter is in the program mode since this operation requires a channel to be formed for tunneling of carriers through the ultra-thin tunnel oxide. The program (erase) pulse of 9 V, (−8V) for 2.5s (7.5s) as shown in Figures 10.9 (a–d) show the plot of trapped charge density in the nitride layer as a function of time. It is seen that it takes approximately 4-5 cycles before the charge in the nitride reaches a steady-state in the program and erase states. Throughout the program and erase cycle, the electrons and holes can tunnel between the nitride and channel and between the poly-Si gate and the nitride.
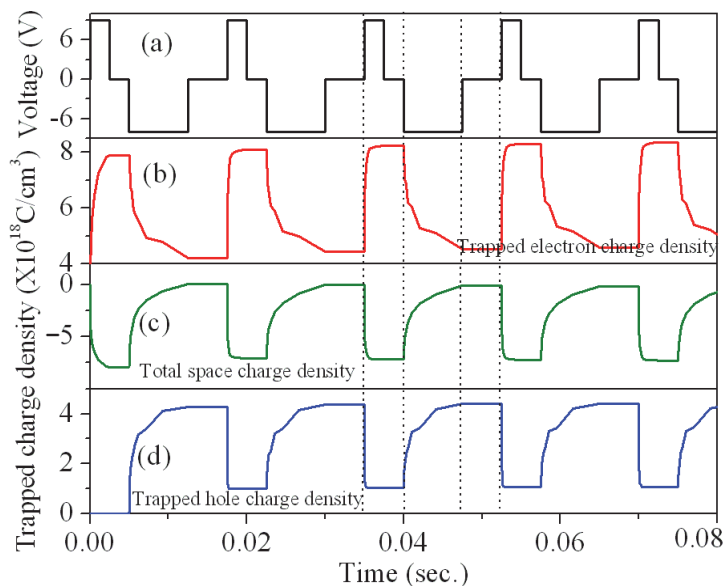
Figure 10.9. Trapped charge density in nitride region of a SONOS device as a function of time during the program and erase cycles. After P. Chakraborty, Modeling and characterization of non-volatile flash memory devices, PhD Thesis, Jadavpur University, 2009.

During the program state, the electron charge distribution (Figure 10.9b) established via Direct Tunneling (DT) and Fowler-Nordheim tunneling of electron into the nitride from the channel. Initially programming mechanism is direct band to band tunneling. For high-speed SONOS devices with good long term retention characteristics, as for the device presented here, the initial programming mechanism is DT. As charge storage occurs in the nitride layer, the tunnel oxide electric field relaxes and the tunneling mechanism becomes modified Fowler-Nordheim tunneling (MFN). During the erase cycle when a negative bias is applied to the gate, some electrons will tunnel out of the nitride and get into the channel, but the majority of the tunneling current comes from holes that tunnel into the nitride from the channel. Inside the nitride layer, they recombine with trapped electrons and remove the overall charge from the nitride. A plot of $I_d - V_g$ is shown in Figure 10.10 for both the program and erases states. Figure 10.10 illustrates that the threshold voltage is 2.428V and 0.49V at drain voltage of 0.05V during program and erase, respectively.
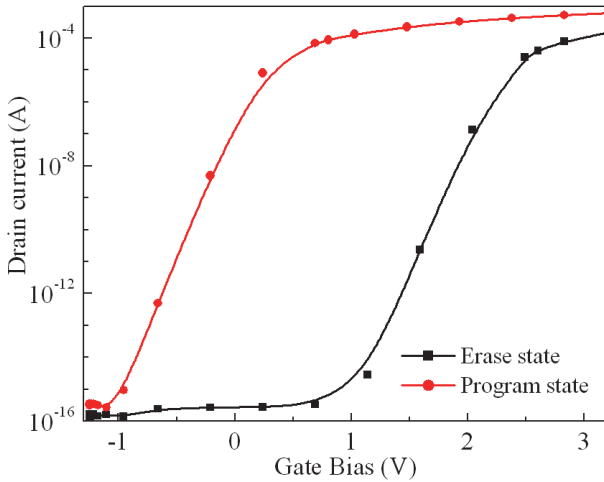
Figure 10.10. $I_d - V_g$ characteristics for the program and erase states of SONOS device.

10.6 Retention

Device and product reliability, such as electromigration and hot electron degradation of the transconductance play a crucial role in NVMs. Data retention is the ability of a NVM to retain data without power over a long period of time. Since the data of SONOS memory devices are represented as electron stored in the silicon nitride layer, the data may be lost if electrons leak away and the lateral migration of trapped charges in the silicon nitride layer. A high density of traps in silicon nitride is responsible for charge storage and memory action in SONOS devices. Trapped electron and hole charge density as a function of time are shown in Figures 10.11 and 10.12, respectively.

Various discharge mechanisms may be responsible for time and temperature dependent retention behavior of SONOS devices. These include trap-to-band tunneling, trap-to-trap tunneling, band-to-trap tunneling, thermal excitation and Poole–Frenkel emission retention loss mechanisms and may be classified into two categories. The first category contains tunneling processes that are not temperature sensitive. During retention, trapped electrons can 'back-tunnel' to the conduction band of

Figure 10.11. Trapped electron and hole charge densities during program state.
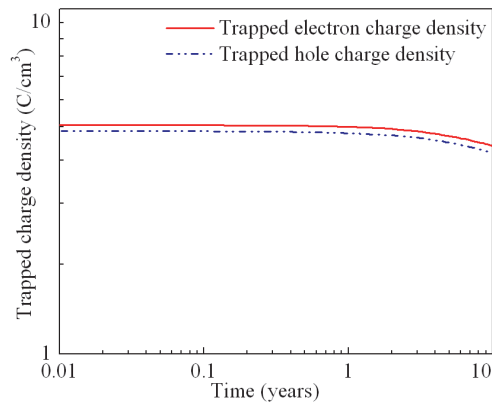
Figure 10.12. Trapped electron and hole charge densities during erase state.

the silicon substrate (trap-to-band tunneling), or to the Si–SiO$_2$ interface traps (trap-to-trap tunneling), under the influence of an internal self-built electric field. Also, holes from the substrate may tunnel through the thin tunneling oxide and become trapped in the nitride (band-to-trap tunneling).

For earlier MNOS devices, the trapped charge redistributed toward the gate electrode inside a relatively thick nitride layer (about 100nm). Also, the redistribution of charge storage within the nitride due to

Poole–Frenkel emission is an important retention loss mechanism. In advanced scaled SONOS devices, the movement of charge is 'blocked' by a 'blocking oxide', which results in the storage of charge in a much thinner (5–10nm) nitride film. Considering a nitride trap density on the order of 10^{19}/cm^3 (approximately equivalent to having one trap in a cube that measures 5nm on a side), the probability of charge vertically 'hopping' from one trap to another is limited. Also, barrier lowing due to the Poole–Frenkel effect diminishes during the later stage of retention, as the internal electric field relaxes after electron discharge. Therefore, one can ignore electron retention loss caused by a redistribution of trapped charge due to the Poole–Frenkel effect in scaled SONOS devices. It is observed that approximately one-third of the electron charge is lost over the 10 year period.

10.7 Simulation of SOHOS Memory Devices

A schematic cross section of the SOHOS device based on a reported experimental NMOS structure used in simulation is shown in Figure 10.13. The gate stack consists of an 18Å tunnel oxide layer, an 80Å charge trap/storage layer [high-k (hafnium oxide)], and a 40Å control oxide layer between the channel and the polysilicon (n$^+$ poly gate) gate material of work function of 4.05 eV.

c
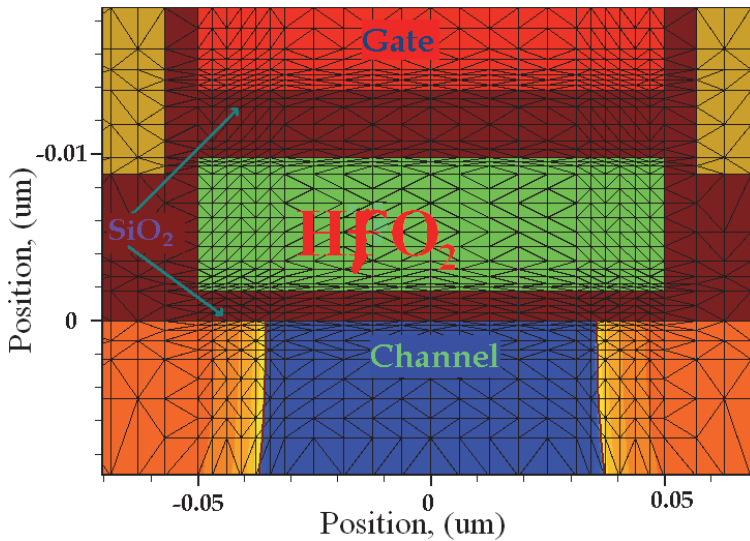
Figure 10.13. Schematic cross section of the simulated SOHOS devices with gate length 100nm, hafnium oxide thickness of the charge trapping layer (HfO$_2$) is 80Å in SiO$_2$ matrix. After P. Chakraborty, Modeling and characterization of non-volatile flash memory devices, PhD Thesis, Jadavpur University, 2009.

During simulation, trapping and recombination of the traps were modeled using the Shockley–Read–Hall mechanism. For programming and erasing, Fowler–Nordheim tunneling mechanism was used. In simulation, a trap density of 1×10^{19} cm^{-3}, substrate doping of 5×10^{17} cm^{-3} and the drain/source doping of 1×10^{20} cm^{-3} were used. High-k dielectrics are known to possess deep level traps. In simulation, material properties, such as, the band gap, permittivity and the electron affinity of the high-k dielectric (HfO$_2$) were used. During simulation, three partial differential equations viz., electrostatic potential, and electron and hole concentrations were solved self-consistently using Poisson's and the continuity equations.

10.8   Programming/Erasing Characteristics

The simulation results on the erase/write, retention and endurance electrical characteristics of the non-volatile flash memory device with HfO$_2$ high-k dielectric are discussed. A pulse of 9.0V for 2ns is applied

to the polysilicon gate for programming and an erase pulse of −8.0V for 7.0 ns was applied and is shown in Figure 10.14a. Figures 10.14b–d show the plot of trapped charge density in the high-k (hafnium oxide) layer as a function of time. It is seen that it takes approximately 4–5 cycles before the charge in the high-k reaches a steady-state during the program and erase states. Throughout the program and erase cycles, the electrons and holes are found to tunnel between the high-k and the channel and between the poly-Si gate and the high-k layers. The program and erase cycle is characterized by amount of charge stored in the high-k traps in the gate stack.

Tunneling into and out of the high-k region occurs by enabling barrier tunneling at the interface of the oxide layers. During the program state, the electron charge distribution (Figure 10.14c) is established via direct tunneling and by Fowler-Nordheim tunneling of electron into the high-k from the channel. Initially, during programming direct band to band tunneling occurs. As charge storage occurs in the high-k layer, the electric field in the tunnel oxide is relaxed and the tunneling takes place via Fowler-Nordheim mechanism. During erasing, the electrons have to
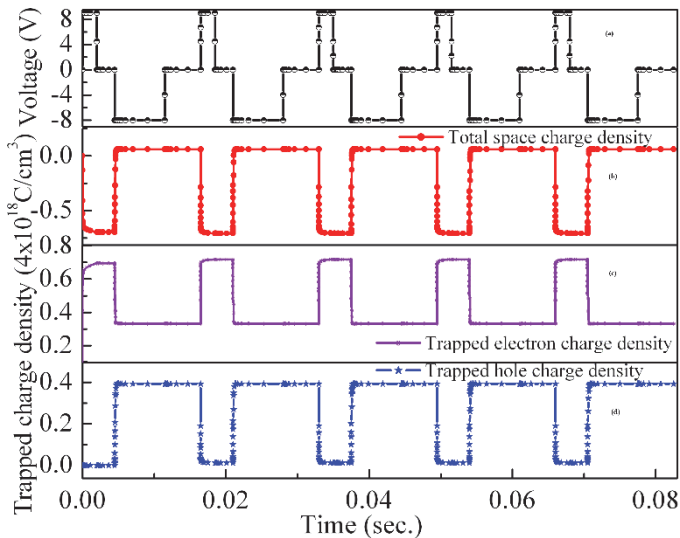
Figure 10.14. Trapped charge density in high-k region of SOHOS device as a function of time during the program and erase cycles. After P. Chakraborty, Modeling and characterization of non-volatile flash memory devices, PhD Thesis, Jadavpur University, 2009.

tunnel through the tunnel oxide layer to the silicon substrate by direct tunneling, which, however, is greatly affected by the tunnel oxide thickness. A number of interesting features may be observed from these plots.

Since the write and erase mechanism is Fowler–Nordheim tunneling, or field assisted tunneling, larger write and erase voltages result in higher tunneling probabilities due to the reduction of the effective barrier. A decrease in writing and erasing times with increased applied tunneling voltages is observed. During the erase cycle, when a negative bias is applied to the gate, some electrons will tunnel out of the high-k and enter into the channel, but the majority of the tunneling current comes from holes that tunnel into the high-k from the channel.

Inside the high-k layer, holes recombine with trapped electrons and remove the overall charge from the high-k resulting in a change in threshold voltage. A plot of $I_d - V_g$ is shown in Figure 10.15 for both the program and erases states. Since an important function of the gate dielectric is to isolate the gate terminal from the current-carrying channel region, it needs to be a good insulator.
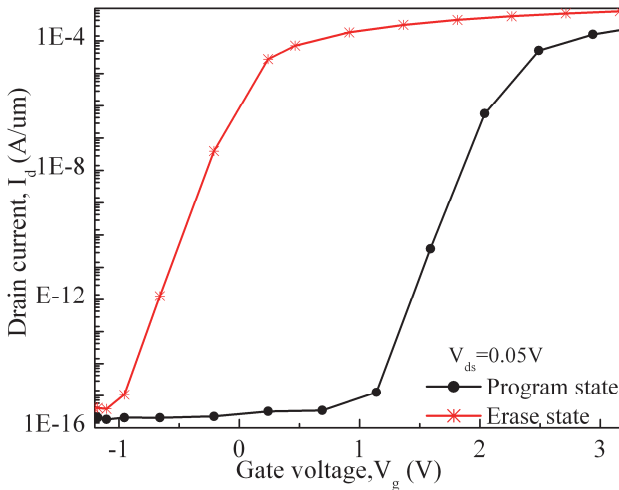
Figure 10.15. I_d–V_g characteristics for program and erase states of SOHOS device.

10.9 Retention Characteristics

A 10-year transient simulation (with all electrodes at zero bias) is used to model the charge retention of the in the high-k dielectric. The initial trap distribution is taken from the program or erases state. The trapped electron and hole charges, as a function of time, is shown in Figure 10.16.

It may be seen that one-third of the electron charge is lost over a 10-year period. The simulated retention characteristics are shown in Figures 10.16 and 10.17. It is observed that the HfO_2 charge trap layers provide a better retention for a wide range of trap energy levels because HfO_2 offers a large barrier offset which results in a longer retention time.
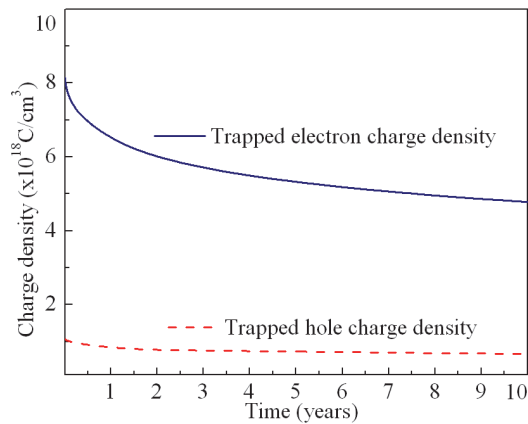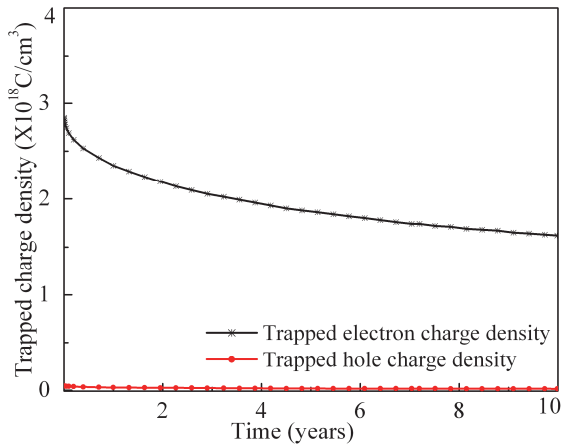

Figure 10.16. Trapped electron and hole charge densities during programming state.

C
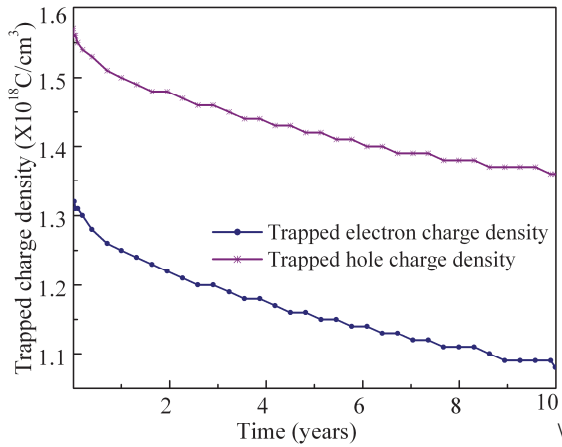
Figure 10.17. Trapped electron and hole charge densities during erasing state.

Summary

In this chapter, we have reviewed the status of nonvolatile flash memory technology. The basic characteristics such as transient programming and erasing characteristics, retention, and endurance, as well as the basic models for nonvolatile Flash memory have been discussed. Instead of continuous floating gate, use of "discrete" charge storage layer i.e., nanocrystal memory is also discussed.

SONOS memory structures with ultra-thin tunneling oxide and nitride layers have been studied. The operational characteristics of ultra-short SONOS memories down to 100nm effective gate length have been simulated. Effects of the oxide thickness between the nitride layer and the channel on charge retention and its kinetics are described. SONOS nonvolatile memory devices display a 10 year data retention after 10^6 erase/write cycles at 30°C, with 1ms +9/-8V program/erase voltages. Simulation results show that SONOS devices are promising candidates for low-voltage high-density flash memory applications. Use of HfO_2 and AlN as charge trap/storage layer to improve the SONOS-type flash memory performance is demonstrated. SONOS flash memory with HfO_2 and AlN trapping layer show enhanced endurance and programming properties much faster than the conventional SONOS memory while achieving good retention.

Chapter 11

Power Devices

Smart power integrated circuits, which monolithically integrate low-loss power devices and control circuitry, have attracted much attention in a wide variety of applications. Commonly used smart power devices are the lateral double diffused MOS transistors. The main design issues in the development of power devices are to obtain the best trade-off between specific ON-resistance and breakdown voltage, and to shrink the feature size without degrading device characteristics. Major concern is related to the origin of special phenomena like, for instance, quasi-saturation and explanation of the capacitive behavior of the HV device as a function of the biasing conditions. Another important issue is the planar integration of the HV LDMOSFETs with the low voltage CMOS devices. As lot of different devices need to be integrated onto a single chip for smart power applications, it is desirable to have as many shared process steps as possible to reduce costs [3.4]. While many process steps like N+ source and drain implant can be easily shared between digital, analog, and power MOSFETs, it is often impossible to share the same thermally grown gate oxide between digital MOSFETs and power lateral double-diffused MOS.

Electrical isolation scheme between the power devices as well as between the power devices and low-voltage CMOS devices is obtained using a deep trench to reduce the isolation distance between the high voltage devices considerably and hence to reduce the total chip area compared to technologies with a standard junction isolation scheme. In recent years, three-dimensional integration is becoming a reality. This technology uses through-silicon-via (TSV) and re-distribution layers to interconnect multiple active circuit layers. 3D integration offers significant improvements over two-dimensional integrated circuits on performance, functionality, and integration density.

In this chapter, we shall discuss novel power semiconductor concepts for smart power applications. New device structures are suggested and studied to improve the device characteristics related to traditional power devices. Two- and three-dimensional device simulations are performed to study new device concepts. Silvaco TCAD tools, general purpose device simulators, are used for the simulation of power semiconductor devices and integration schemes. Simulations are performed to highlight the physical phenomena that take place inside the HV LDMOSFETs. An extended analysis including DC and AC characteristics as well as TCAD simulations is performed revealing important effects. Then we consider the REduced SURface Field (RESURF) LDMOSFETs. A general purpose RESURF-LDMOS structure is designed based 2D and 3D numerical TCAD simulations to verify the RESURF concept. The key device/process parameters are optimized to get the best trade-off between ON-resistance and breakdown voltage.

11.1 LD-MOSFETs

An LDMOS transistor is one of the most important power devices in power ICs. A well-designed LDMOS should have a high breakdown voltage, low on-state resistance, and a small gate to drain capacitance. This section focuses mostly on how to use a TCAD tool to improve the device breakdown voltage without degrading the on-resistance. Most of the design efforts in power IC technology are concentrated on how to improve the performance of LDMOS devices. Like a vertical double-diffused MOS (VDMOS) transistor, the channel region is created by the difference between diffusions of body and source N^+ (or P^+) implants. This is a self-aligned process so that large lithographic feature sizes can be tolerated.

Typical Smart Power IC Process Flow includes the following steps:

- P Substrate
- N+ Buried Layer
- N Epitaxial Layer #1

- Deep N Layer #1
- N Epitaxial Layer #2
- Deep N Layer #2
- HV Twin-Well
- P body for n LDMOS
- Active/STI
- N/P Twin-Well
- LV Twin-Well
- Thick Oxide (TGOX)
- Thin Oxide (GOX)
- Poly Gate
- NLDD/PLDD
- NSD/PSD
- Contact

LDMOS with a breakdown voltage less than 150V are popular for power ICs in consumer electronics and transportation applications. A standard LDMOS structure without STI in the drain-drift region is shown in Figure 11.1. The starting wafer is p-type with a donor concentration of boron 1×10^{14}cm^{-3}. It represents the left half of our LDMOS device, measuring 10μm wide and 5μm tall. The final device will be 20μm wide and 5μm tall. ATHENA simulation starts by defining an asymmetric grid with finer grid in the drift region. In preparation for the gate oxide growth, a sacrificial oxide layer is grown and etched away. To insure adequate grid spacing in the growing oxide, the method "grid.ox" statement sets the grid to 100 A. The next step is a low-energy and low-dose boron V_t-adjust implant. Poly deposition, poly-gate definition and drift region implant are then performed. The drift region is masked and n+ phosphorus is implanted into the opened source and drain areas. For the LDMOS, an additional oxide layer is grown known as the FOX enhancement. FOX enhancement is a layer of oxidation beneath the poly-silicon and gate oxide that serves to boost performance parameters. It creates a deeper channel for current to travel, while still minimizing the length of the channel. It helps control the breakdown voltage and the R_{on} of the device. It is done at the same time as the LOCOS because it requires much growth time. For a power IC device, a high breakdown voltage is desired. A short channel length greatly reduces the on-resistance of the device. A minimal on-resistance is desired, it will allow for much higher current applications. FOX enhancement provides a

solution to the breakdown voltage and on-resistance trade-off issue. The next step is the poly-Si deposition and doping to improve conductivity. Figure 11.2 shows the active boron dose in the device. Figure 11.3 shows the LDMOS device prior to the masking of the p-well drive. Finally the contact is made. Figure 11.4 shows the final LDMOS structure.
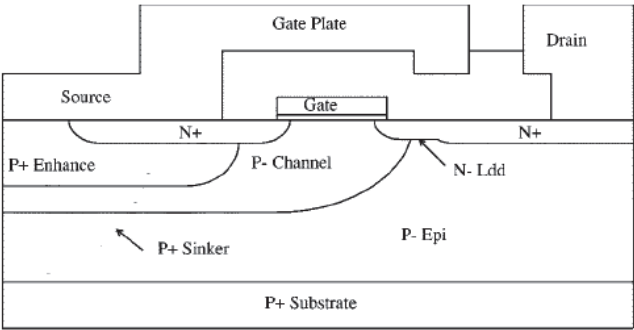
Figure 11.1. Typical standard LDMOS structure.
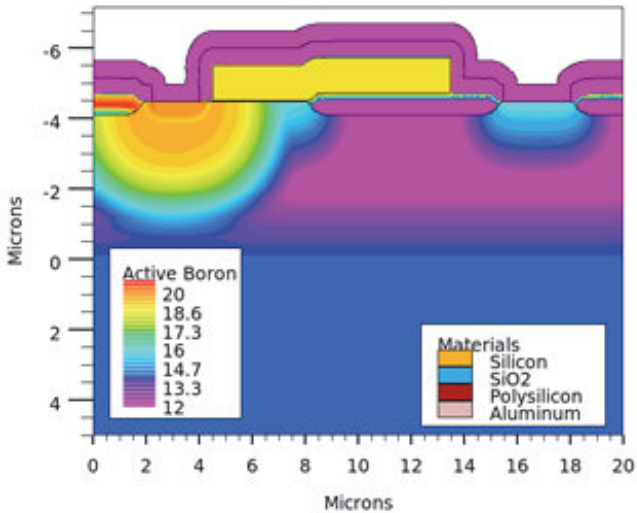
Figure 11.2. Boron doping in a LDMOS device.
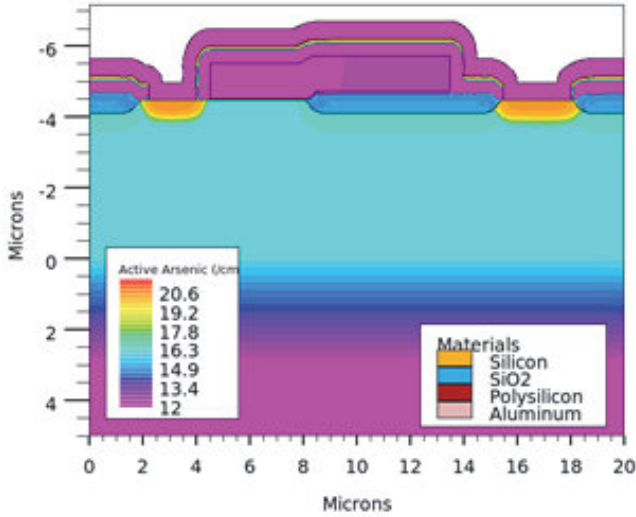
Figure 11.3. LDMOS device structure prior to the masking of the p-well drive.
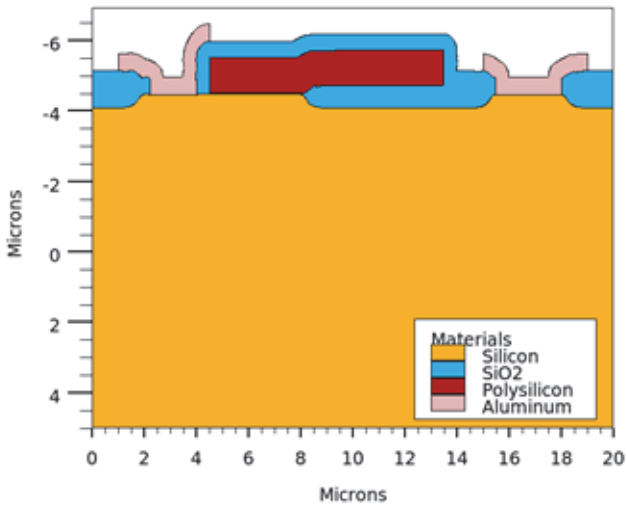
Figure 11.4. Final LDMOS structure.

The device simulated is an asymmetric LDMOS power device structure. The structure created by ATHENA is loaded into ATLAS when the

command "go atlas" is used. In ATLAS simulation begins by specifying the physical models to be used. For this example, they are CVT for transverse field dependent mobility and srh Shockley–Read–Hall recombination. The Selberherr impact ionization model is enabled using the impact "selb" statement. The full breakdown analysis is performed by solving for impact ionization using a two carrier approach. Figure 11.5 shows the impact ionization rate in the device. The "contact" statement sets the gate work function to that of degenerately doped n-type polysilicon. The sequence of solve statements ramp the drain voltage from zero to 100V. The breakdown simulation is performed with source/body contact and gate contact grounded while the potential at the drain contact is increased until leakage current increases rapidly. For this particular device, the breakdown voltage is about 50V. A plot of the I_d vs. V_d characteristic clearly shows the breakdown characteristics. The simulation results after device breakdown are shown in Figure 11.6. If we take a further look at the simulation, we find that in the electric field plot of Figure 11.7, the peak field is right underneath the gate edge, which also corresponds to a peak of the impact ionization rate in Figure 11.5. This strongly suggests that the critical field close to the gate edge is triggering impact ionization and reducing the breakdown voltage of the device. It is thus important to divert the peak electric field from the silicon to the oxide is an effective way for improving the breakdown voltage.
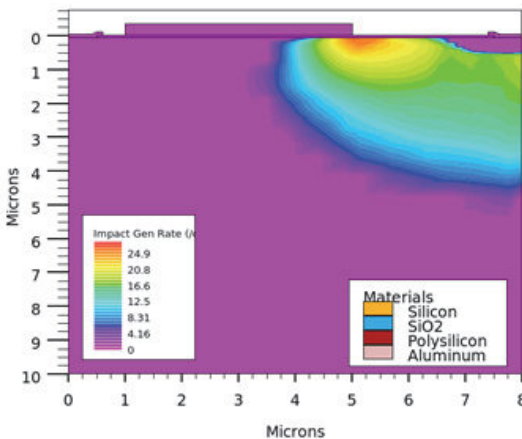
Figure 11.5. Impact ionization rate in the device.
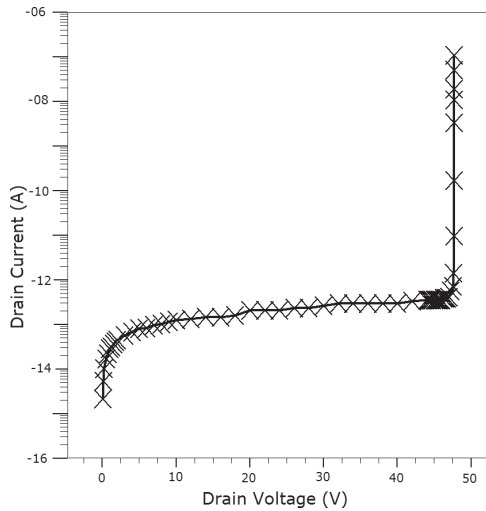
Figure 11.6. Breakdown characteristics of the device.
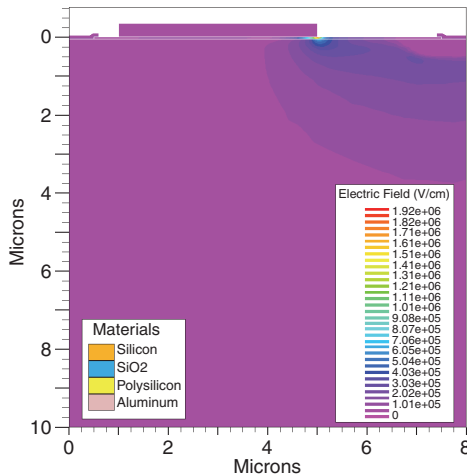
Figure 11.7. Peak electric field in the device.

11.2 Super Junction LDMOS RESURF Technology

For smart power applications, the requirements on the devices and the choice of technology depend on the intended application. For example, cell phones require a breakdown voltage for LDMOS of less than 20V but for automotive application, values as high as 80V may be required.

The trade-off between breakdown voltage and on-state resistance is a major issue for power semiconductor devices. In the past many ingenious designs have been proposed and implemented. One of the most promising design concepts is called the super junction, which is similar to the RESURF effect. In power electronics, N-type LDMOS is often preferred over P-type LDMOS because electrons have higher mobility than holes in silicon; this means the n-LDMOS will have a lower specific on-state resistance. For LDMOS, breakdown voltage and specific on-state resistance are almost always a trade-off, and meeting both goals simultaneously is not easy. By increasing the N-well doping, we are able to reduce the on-resistance, but the breakdown voltage is degraded.

Over the past decades both RESURF and Super Junction concepts were widely used, especially for high voltage ICs (>100V for RESURF, >300V for Super Junction). Both RESURF and Super Junction use the same charge coupling effect; RESURF is about two-dimensional optimization that can be applied multiple times (e.g., multiple-RESURF), whereas Super Junction is more about three-dimensional effects and usually consists of alternating N and P-regions in a direction perpendicular to the channel. We use process modeling for LDMOS device and include a brief overview for RESURF. In comparison with LDMOS, super junction LDMOS has a higher breakdown voltage with the same implant energy and dose for n-well. On the other hand, with the same breakdown voltage, super junction LDMOS can achieve higher dose and thus lower on-state resistance. Both approaches can greatly enhance the device figure of merit.

The schematic cross-section of a general-purpose RESURF LDMOS is shown in Figure 11.8. In this structure, the effective channel length LCH is defined by the difference in the lateral diffusions of the P-body and the N^+-source regions (lateral double-diffused MOS transistor). So, the doping in the channel is laterally graded along the channel length from the N^+-source to the K-point, which is called the intrinsic drain (the metallurgical junction between the P-body and the N-epitaxy regions). The LDMOS channel length depends only on the process (the vertical junction depth and doping concentration) and not on the device layout.
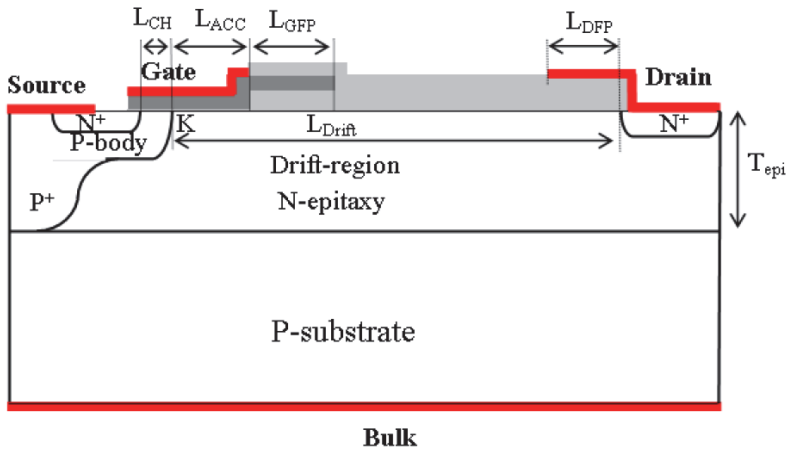
Figure 11.8. The general-purpose LDMOS structure optimized using the RESURF principle. After M. A. E. A. Ebrahim, Designing smart power integrated circuits with LDMOS devices compatible standard BiCMOS technology; Substrate coupling, PhD Thesis, INSA de Lyon, France, 2010.

In the following, simulation of a buffered super junction LDMOS is considered. Super junctions are used in LDMOS structures to greatly increase the breakdown voltage of small geometry devices by allowing the drain depletion region to spread in two dimensions instead of one at higher drain voltages (the additional direction being laterally across the super junction). The net result is that the drain voltage can now be spread across a much greater total distance than would otherwise be the case, which greatly lowers the field at the drain and therefore increases the breakdown voltage. Another advantage of super junction technology is that the n-doped stripes in the super junction can be doped to higher concentrations because the important depletion distance is now the lateral distance between the super junction stripes (determined largely by the technology node) rather than the full depletion width of the drain. Higher doping yields lower on resistance, increasing current drive over a standard construction design with a similar breakdown voltage.

The buffered super junction device takes this technology one step further by including an additional n-doped buffer layer under the super junction, which increases the breakdown voltage to even higher values by expanding the depletion region in the remaining third dimension which is down into the depth of the substrate. In this simulation we examine the effectiveness of this approach as it has an electrical gate length of only

2.5μm, but it shows a breakdown voltage of 85V. The device structure and net doping of a simple super junction LDMOS is shown in Figure 11.9.
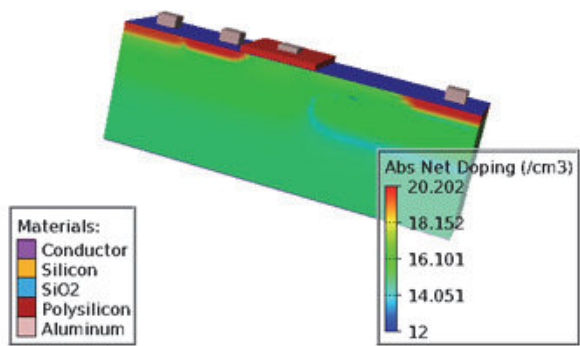
Figure 11.9. Net doping in the LDMOS structure.
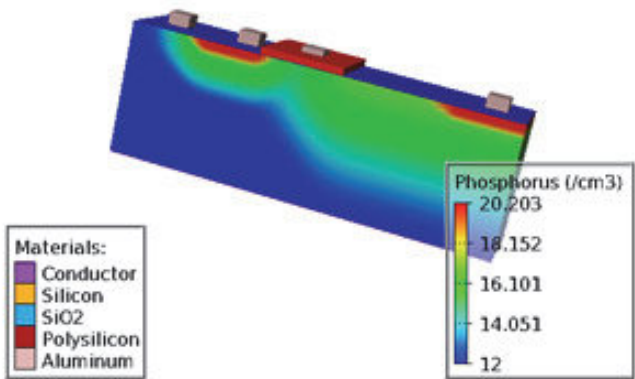
Figure 11.10. Doping profile in the simulated buffered Super Junction LDMOS.
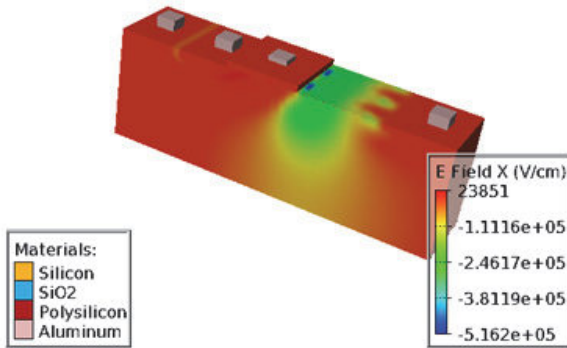
Figure 11.11. Electric field distribution with 80V applied to the drain for the simulated buffered Super Junction LDMOS. Impact ionization rate distribution at 80V drain voltage.
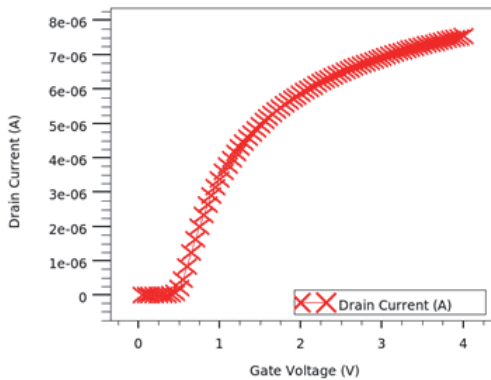
Figure 11.12. I_d–V_g characteristics of the simulated buffered Super Junction LDMOS.

The simulation of super junction RESURF LDMOS is similar to buffered super junction LDMOS but uses a low doped substrate and longer super junctions to increase the off state breakdown voltage to approximately 700V (see Figures 11.10 and 11.11). The 3D device is passed to Device3D to simulate the breakdown voltage, I_d–V_d, (see Figure 11.12) and I_d–V_g characteristics (see Figure 11.13). By careful optimization of the super junction widths and lengths together with the doping concentrations of the super-junctions and the related n-well, a 700V breakdown device is realized when coupled with a high resistivity substrate. The breakdown characteristics is shown in Figure 11.14.
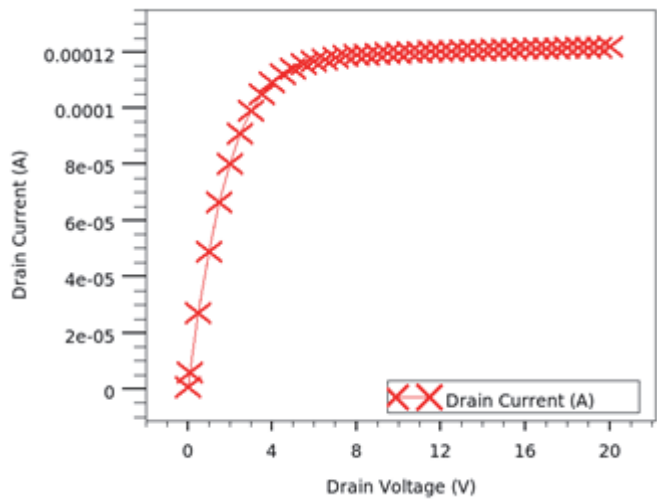
Figure 11.13. I_d–V_d characteristics of the simulated buffered Super Junction LDMOS.
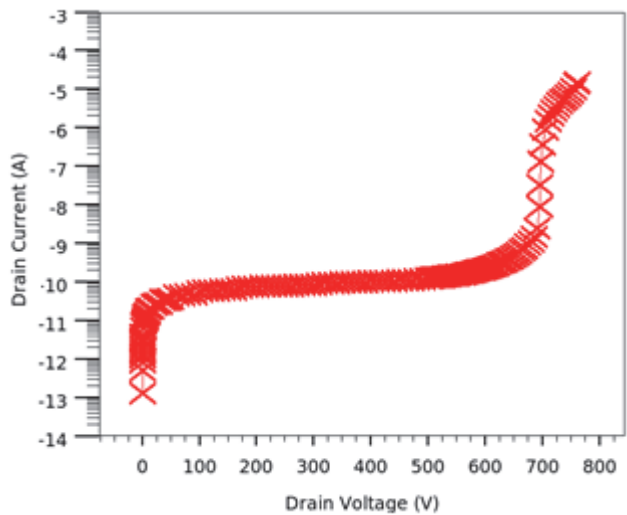
Figure 11.14. Simulation of breakdown voltage in RESURF LDMOS. The device shows a breakdown voltage of 700V.

11.3 SiC-Based Power Devices

Silicon based power devices are dominant today in power electronics. Wide bandgap semiconductors like SiC are now more and more used for high power, high-temperature applications because of superior thermal conductivity, lower intrinsic carrier concentration and better on-resistance compared to Silicon, which is a key figure of merit in power switching applications.

In the following, we demonstrate 3D trench SiC IGBT simulation using VictoryCell and VictoryDevice. In this case study, 3D trench shape effect on I–V and breakdown voltage (BV) characteristics are studied. The 3D layout-driven structure is created using VictoryProcess. 3D device simulation is done using VictoryDevice with built-in extended precision numeric. The 3D structure is saved using a full 3D Delaunay mesh using EXPORT victory (Delaunay) statement in VictoryProcess. The 3D structure is then loaded in VictoryDevice. The default 3D Delaunay discretization method combined with a parallelized iterative solver set by PAM.MPI statement in VictoryDevice which allows good convergence, speed and accurate results.

The structure considered in simulation has a rounded bottom trench as well as a floating P region underneath the trench for improved blocking capability and thus increased breakdown voltage. We will then show how one can optimize the breakdown voltage as a function of the trench shape. We compare 3 different structures; one fully Manhattan (i.e., 90 degree layout and trench) vs. 2 structures one having rounded trench edge and another one having rounded edge and angled trench. Finally, Aluminum implantation is simulated in 1D (to save simulation time) in ATHENA using Monte Carlo implantation before being imported in the 3D structure using the PROFILE statement in VictoryProcess. Figure 11.15 shows the net doping in the 3D structure after process simulation. The effect of the trench shape on the I_d–V_g and breakdown voltage characteristics has been simulated. Figure 11.16 shows the trench shape dependence of I_d–V_d characteristics. BV simulation reveals that the breakdown voltage can be increased by 30% using rounded layout and angled trench due to impact ionization not occurring any more only at the

corner of the trench. Figure 11.17 shows the breakdown voltage characteristics.

The anisotropic impact ionization model for materials with a hexagonal crystal structure for 4H-SiC has been used in simulation. The ionization coefficients in this model depend on the strength of the electric field and on the orientation of the field with respect to the optical axis of the crystal. One can specify the orientation of the optical axis of the crystal in VictoryDevice by supplying values for ZETA and THETA parameters on a MATERIAL statement. Figures 11.18 and 11.19 show the impact generation rate and electric field in the device. A comparison between 2D and 3D breakdown voltage simulation is shown in Figure 11.20. As expected the breakdown voltage is the same in 2D and 3D for the same structure. It is interesting to note that due to the presence of the P-type region underneath the trench, making a p-n junction, the maximum electric field and thus impact ionization rate is maximum at this junction location and not at the trench corner.
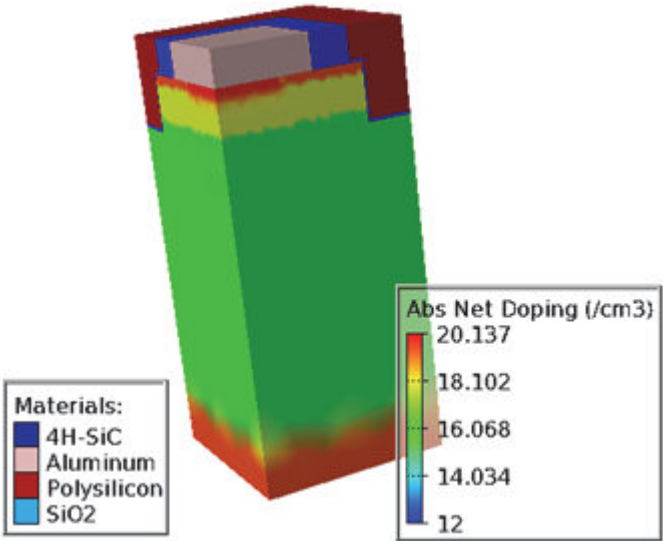
Figure 11.15. Net doping in the 3D structure after process simulation.
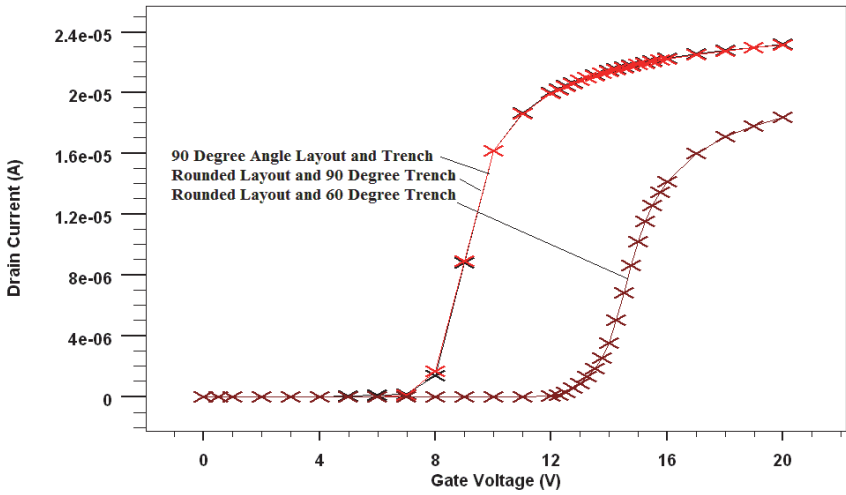
Figure 11.16. Trench shape dependence of I_d–V_d characteristics.
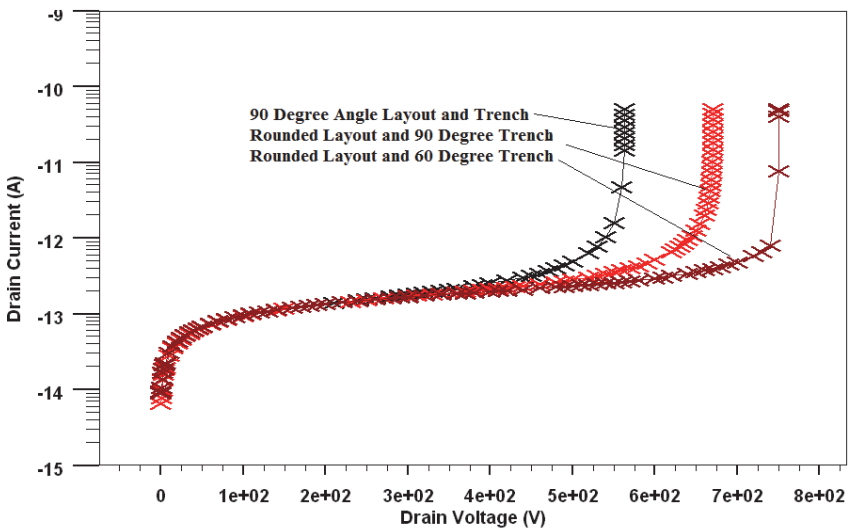
Figure 11.17. Breakdown voltage can be increased by 30% using rounded layout and angled trench.
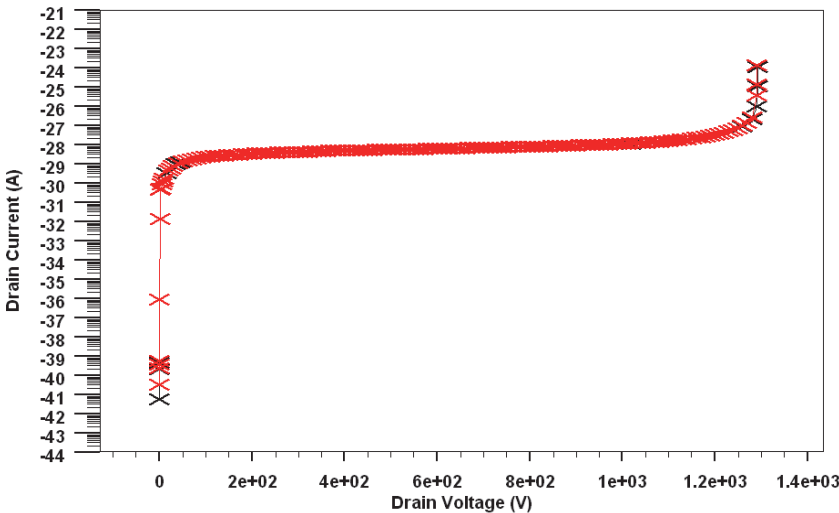
Figure 11.18. Comparison between 2D and 3D breakdown voltage simulation.
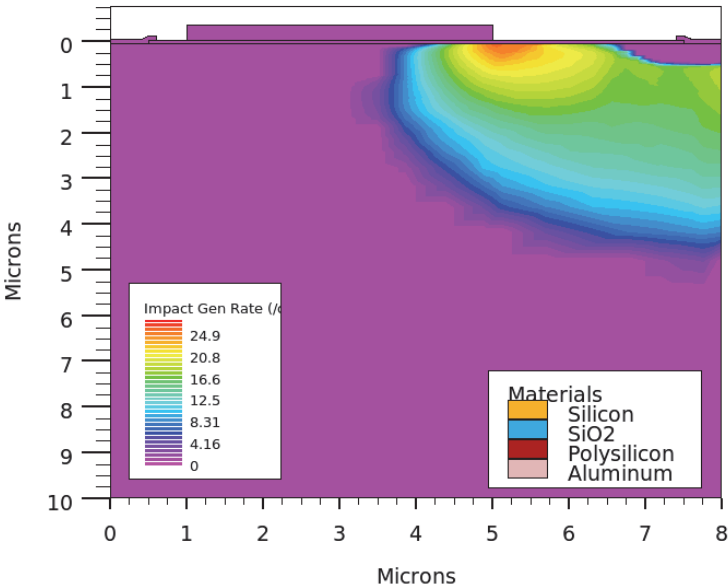
Figure 11.19. Impact generation rate 2D distribution in the device.
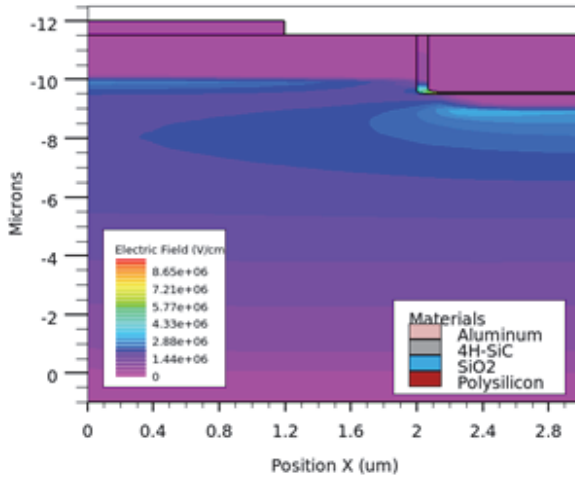
Figure 11.20. Electric field 2D distribution in the device.

Summary

In this chapter, we have presented the process design, modeling, and analysis of laterally diffused metal oxide semiconductor transistors (LDMOS). Technology CAD challenges for developing an integrated circuit process is described. As a case study, process recipe for a medium voltage (40–50V) LDMOS device has been developed and optimized through the use of a design of experiment approach based on TCAD simulations. Silicon based power devices are dominant today in power electronics. However, wide bandgap semiconductors like SiC are being more and more used for high power, high-temperature applications because of superior thermal conductivity, lower intrinsic carrier concentration. We have addressed the process to device simulation of SiC devices in 2D and 3D and optimized this promising technology.

Chapter 12

Solar Cells

The discovery of the photovoltaic effect dates back to 1839 when Becquerel found the photo-voltage between the two electrodes in electrolyte solutions. The first working solar cells were constructed by Charles Fritts in 1883. Despite holding great promise, solar energy, especially PV, is still considered uneconomical due to its relatively low electricity-generating capacity and high production and installation costs. It currently contributes less than 1% to the global energy supply. The solar energy industry has thus been focusing on cost reduction to add more competitive advantages to the various photo-voltaic (PV) technologies. Alternative material photovoltaic have started gaining more attention recently. Successful developments of efficient solar cells using alternative absorbers will significantly enrich the PV industry and reduce the market gap with other energy sources. This will make the PV technologies a more economically viable alternative energy option.

To continue this growth trend, certain long-term requirements have been identified which any photovoltaic technology shall need to satisfy. These requirements are:

- High efficiency
- Non-toxic
- Based upon highly abundant materials
- Durable
- Suitable for Thin-Film Technology

The predominance of c-Si in the market is a result of numerous factors. Firstly, silicon fulfils the physical requirements for application in solar cells due to its natural semiconductor properties, e.g., the suitable band

gap etc. Secondly, c-Si modules have also exhibited excellent stability with very little degradation and non-toxicity. Thirdly, silicon is the second most abundant element on the earth's crust. The combination of these factors makes c-Si modules the most economical technology today for large-scale systems. However, the production cost of crystalline silicon solar cells is still high, and this inhibits it from widespread terrestrial application and replacing the fossil fuels.

Besides c-Si solar cells, thin-film solar cells, based on amorphous silicon (a-Si:H), cadmium telluride (CdTe) and copper indium gallium diselenide (CIGS) are available or on the verge of commercialization. Thin film solar cells require less material, and they can also be deposited on cheap substrates, including various flexible plastics and/or stainless steel, and hence it is possible to reduce manufacturing costs. However, they also have some disadvantages. Amorphous silicon technology suffers from degradation and low efficiencies and the module efficiency is less than 10%. Substantial efforts have been made to compensate for these disadvantages by introducing sophisticated double and triple junction solar cells. Nevertheless, the stabilized efficiencies still stay below 13%. Additionally, the more sophisticated the solar cell structures, the smaller the margin of reducing cell production costs. Non-silicon based solar cells, such as CIGS and CdTe, have better device performance than amorphous Si. The efficiencies of CdTe and CIGS modules are around 11% and 13%, respectively. But they suffer from two major disadvantages: (i) potential toxicity and (ii) scarcity of some of the material components. Both of them limit the global application and become critical issues when photovoltaic is considered as a technology to supply energy worldwide on a GW scale. Hence, although the market share of thin-film solar cells has increased in recent years, it remains below crystalline silicon.

Current photovoltaic cell production is overwhelmingly based around bulk crystalline silicon wafers, so-called 'first generation' cells. While this is a proven and well established technology compatible with standard industrial techniques, and for the most part satisfying the first four of the above criteria, the high cost of silicon wafers results in a limited cost reduction potential. An important drawback of the silicon solar cells is the large amount of energy that is required in the production of high-quality silicon. Thin-film solar cells, or 'second generation' solar cells, offer the potential of considerably reduced 'cost-per-watt'

production than bulk crystalline technologies due to a) the ability to produce the solar cells on inexpensive large-area substrates such as glass; b) reduced use of semiconductor material; c) to monolithically series connect the cells. Thin-film PV therefore offers the potential of considerably reduced 'cost-per-watt' production than bulk crystalline technologies.

At present, however, most thin-film PV technologies are unable to compete with the higher efficiencies of bulk crystalline solar cells. Especially attempts are being made to develop photovoltaic devices and materials that can deliver higher energy conversion efficiency in a cheaper way to meet this need. To address the energy needs and climate challenges at a broad front, a wide variety of photovoltaic technologies are being explored by scientists all over the world. This includes conventional cells made of materials such as gallium arsenide, cadmium telluride and copper indium gallium diselenide as well as more exotic technologies like dye sensitized solar cells, polymer based solar cells, tandem cells, hot carrier cells and intermediate band solar cells. Today the photovoltaic market is dominated by crystalline silicon cells – a first generation technology.

Cells that show the most promise for reaching higher efficiencies, namely, the CIGS and CdTe based cells, are non-silicon based technologies and are largely incompatible with the fabrication infrastructure that the electronics industry has invested over the last few decades, driving up the cost of these cells. Furthermore, CIGS and CdTe based cells use expensive and toxic materials and do not satisfy criteria (ii) and (iii) above. Hence, there is a natural demand to advance thin-film technologies that are compatible with current main stream semiconductor fabrication, such as, silicon based materials which exhibit competitive efficiencies compared to first generation bulk crystalline technologies, and use materials that are both abundant and environmentally friendly semiconductors. Until recently, the development of photovoltaic technology has mainly centered to inorganic semiconductors, especially the polycrystalline silicon. The upper limit of silicon solar cell efficiency is 31%, which is substantially higher than the best laboratory (25%) and large-area commercial (24%) cells. Cell efficiencies above 25% appear to be feasible in both a laboratory and production environment.

While commercial silicon cells typically has an efficiency in the range 13 to 20%, the current world record for photovoltaic energy conversion is 41.6% and was reported by Spectrolab in 2009. The record was set using a triple-junction tandem cell. Main focus in this chapter is on the design of several types of advanced solar cell structures which will include the GaAs, InAs/GaAs Quantum Dot Solar Cells

12.1   The Basics

The basic principle of a photovoltaic cell is that electrons are excited from electron states with low energy to electron states with higher energy by absorption of photons. If the electrons can be extracted from the high energy states in some way, some of their increased energy can be harvested before they are re-inserted in the low-energy states. For simplicity, and with reference to the semiconductor implementation of the photovoltaic cell, the high-energy states will be referred to as the conduction band (CB) and the low-energy states as the valence band (VB), as shown in Figure 12.1. The working principle of all solar cells is based on the photovoltaic effect. In general, the photovoltaic effect means the generation of electrical energy in the form of current and voltage at the junction of two different materials in response to electromagnetic (i.e., light, including infrared, visible, and ultraviolet) energy. The basic processes behind the photovoltaic energy conversion are (see Figure 12.2):

- Absorption of photons in the materials (that form a junction) and the same causes charge carrier pair (free negative- and a free positive-charge) generation.

- Photo generated charge carriers are then separated at the junction due to transport mechanism (built-in electric field).

- Collection of photo generated charge carriers at the terminals of the junction.
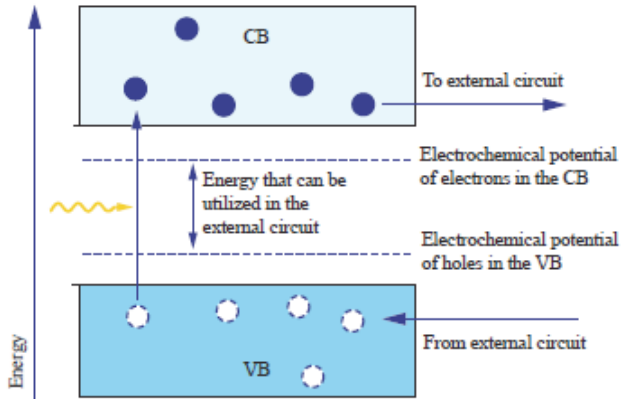
Figure 12.1. After being excited from the valence band to the conduction band, the electrons can be extracted to an external electrical circuit. The energy that can be delivered to the circuit by an electron is given by the difference in electrochemical potential between electrons and holes in the CB and VB.
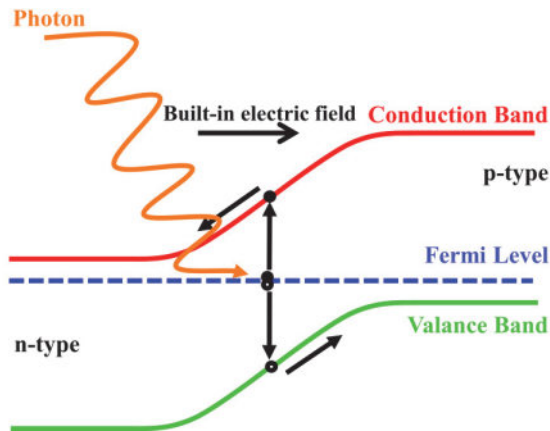
Figure 12.2. Schematic diagram of the basic working principle of solar cell.

Thus, any solar cell structure needs a photovoltaic absorber layer that is not only able to absorb the incident radiation efficiently but also to create electron-hole pairs, which are separated at the terminals of the device without significant loss of energy. After light absorption electrons and holes are present in the absorber and must be directed towards the two

different contacts to the absorber, that is, the final charge carrier separation step. Typical solar cell structure (in cross-sectional view) is shown in Figure 12.3.

The band gap energy of semiconductor, acting as a photovoltaic absorber is the most important consideration for the fabrication of solar cells. The parameter internal quantum efficiency determines how many charge carriers are generated from solar photons with energy $E \geq E_g$. Large band gap material is preferred for high open circuit voltage, due to less recombination and higher potential barrier, while small band gap material is preferred for high short circuit current, due to more absorption. Therefore, one intuitively expects that an optimum band gap energy exists between $E_g = 0$, maximizing the generated electron–hole pairs, and $E_g \to \infty$, maximizing the generated energy contained in a single electron-hole pair. There is clearly a trade-off where the optimal band gap is near 1.1 eV. For a given semiconductor, increasing optical absorbance is another consideration since it is related to short circuit current. In practice, a thicker active region is preferred for higher optical absorbance, although this is accompanied by lower charge collection efficiency and high cost.
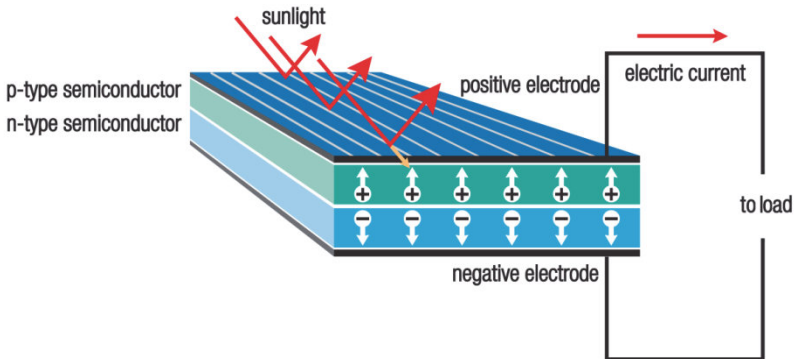
Figure 12.3. Cross-section of a typical solar cell. The collecting electrodes (cathode and anode) are shown with the top electrode being transparent.

After the charge carriers are transported to the active layer (i.e., absorber)/electrode interface, they must be extracted from the active layers to the electrodes. To achieve high efficiency in charge extraction, the potential barrier at the active layer/electrode interfaces have to be minimized. An ohmic contact is needed to pass the current without any voltage drop. As a general rule of thumb, the large work function metal are used as the contact to p-type semiconductor materials, and low work function materials are used as the contact of n-type semiconductor to form ohmic contacts.

For the generation of electron-hole pair, light must enter the solar cell absorber (i.e., active) layers. Thus transparent contacts are needed to allow light. The material to be utilized must have high conductivity, high transparency, band edge position, and proper work function to serve as ohmic contacts. There is also a trade-off between resistance and transmission with film thickness. Transparent conductive oxide (TCO) materials meet these requirements. TCOs that have been examined for transparent contact purposes include indium oxide, tin oxide, indium tin oxide, zinc oxide, aluminum zinc oxide (AZO), gallium indium oxide, indium zinc oxide, gallium indium tin oxide, and zinc indium tin oxide. Among these, tin oxide, ITO and zinc oxide have found most widespread use.

The energy supply for a solar cell is the photon coming from the sun. This input energy is distributed, in ways that depend on variables like latitude, time of day, atmospheric conditions, and over different wavelengths. The various distributions that are possible are called solar spectra. The product of this light energy input, in the case of a solar cell, is usable electrical energy in the form of current and voltage. Some common "standard" energy supplies from the sun, which are available on the earth, are shown in Figure 12.4. The spectrum of available energy from the sun at the outer edge of the earth's atmosphere is referred to as the air mass zero, AM0 spectrum. This spectrum has an integrated power density of 1366.1 W/m^2. The two spectrums that are relevant on the earth's surface are referred to as the AM1.5G and AM1.5D spectrum. The AM1.5G spectrum includes direct and diffuse radiation and has an integrated power density of 1000 W/m^2, which is relevant for flat panel photovoltaic, while the AM1.5D spectrum consists of only direct
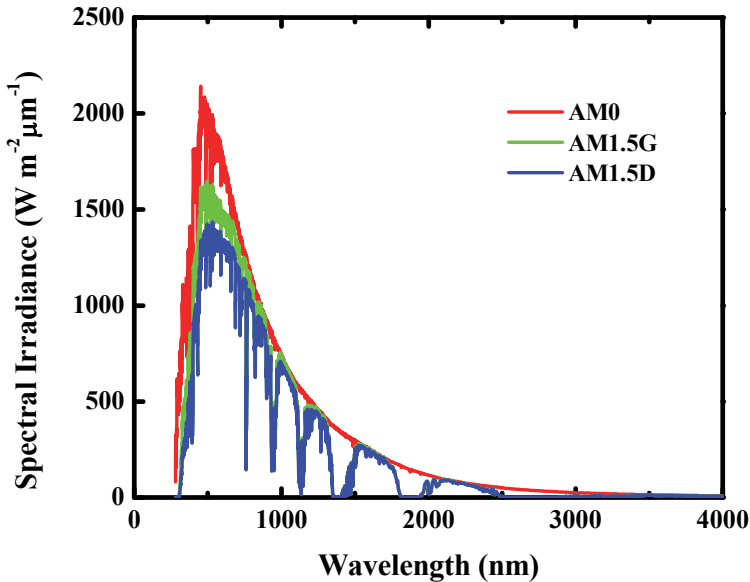
Figure 12.4. The AM0, AM1.5G and AM1.5D spectrums.

radiation from the sun and is applicable to concentrator photovoltaic. It has an integrated power density of 900 W/m^2.

As we have discussed earlier, photon absorption in a solar cell depends on active materials and its band gap. Also it may be noted that input solar energy depends on various factors. There must be a trade-off between the two factors to decide the optimal band gap for efficiencies of solar cells. Based on a detailed calculation, Shockley and Queisser determined the maximum theoretical efficiency for solar cells in the 1960s. Using the AM1.5G spectrum and assuming no light concentration, the limit for a single junction solar cell is 33%. The main limitation of such cells is the thermalization of hot carriers generated by photons with energies greater than the band gap energy and the non-absorption of photons with energy smaller than the energy gap (see Figure 12.5). Figure 12.5 shows the main loss mechanisms in a solar cell (spectral losses). It shows the maximum achievable energy of a silicon solar cell in relation to the sun spectrum. Photons carrying a specific energy can only generate one electron-hole pair even if their energy is higher. The excess energy greater than band gap energy is lost in thermalization of hot carriers i.e.,

heat (see upper grey part in Figure 12.5). Photons with energies lower than the band gap cannot generate an electron-hole pair (non-absorption, right grey part in Figure 12.5). These two losses account for about 50% of the power carried in the sun spectrum. Thus, it is important to choose the right band gap energy in order to balance and minimize these losses. Fortunately, semiconductors like silicon and GaAs are very close to the optimum band gap energy (see Figure 12.6).
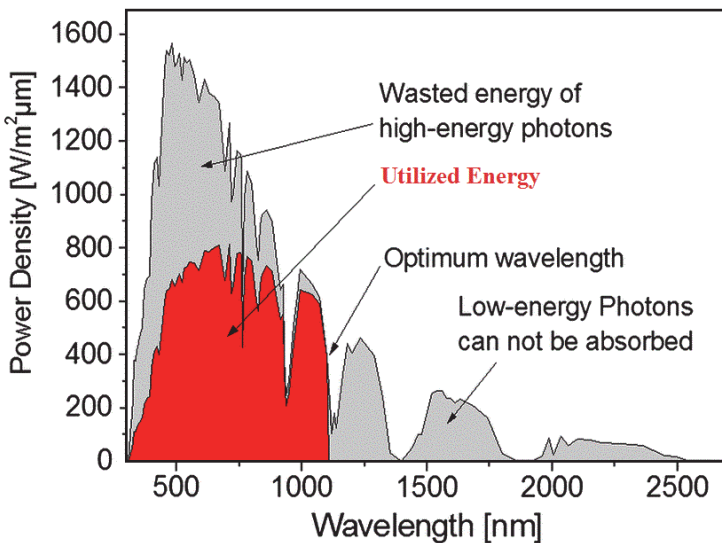
Figure 12.5 Spectral losses in a solar cell. The figure shows the maximum achievable energy of a silicon solar cell in relation to sun spectrum (AM 1.5G). After A. Bag, Fabrication and characterization of iron disilicide heterojunction solar cells, PhD Thesis, IIT Kharagpur, 2014.
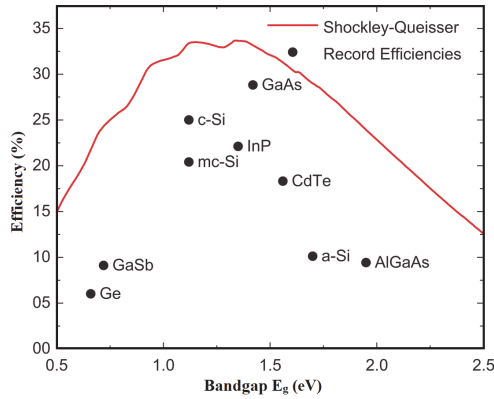
Figure 12.6. Band gaps of various solar cell materials and conversion efficiencies. After A. Bag, Fabrication and characterization of iron disilicide heterojunction solar cells, PhD Thesis, IIT Kharagpur, 2014.
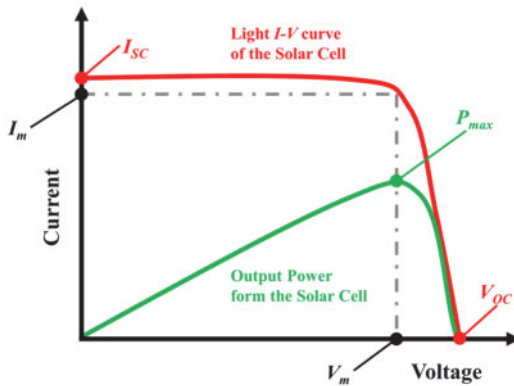
Figure 12.7. Illustration of I–V, P–V characteristic and some parameters of a solar cell under illumination.

12.2 Solar Cell Parameters

There are several important parameters of solar cells that can be derived from I–V characteristics: the short-circuit current (I_{SC}), the open-circuit voltage (V_{OC}), the fill factor (FF), and the efficiency (η). They are shown on the I–V characteristics in Figure 12.7.

Short-Circuit Current:

The short-circuit current is the current flowing through the solar cell when the voltage across the cell is zero, i.e., when the solar cell is short circuited. It is due to the generation and collection of light-generated carriers. So ideally, it is equal to the light-generated current and is the largest current may be drawn from the solar cell.

Open-circuit Voltage:

The open-circuit voltage V_{OC} is the maximum voltage available from a solar cell, which occurs at zero current. This is the voltage of which the current in the external circuit is zero. By setting the net current to zero, we can get the ideal value of V_{OC}:

$$V_{OC} = \frac{k_B T}{q} \ln\left(\frac{I_{Ph}}{I_0} + 1\right)$$

The above equation shows that V_{OC} depends on the saturation current I_0 and the photo generated current I_{Ph} of the solar cell. Since I_{Ph} typically has a small variation, the key effect on V_{OC} is the saturation current I_0, which may vary by orders of magnitude for different solar cells.

Fill Factor:

Although I_{SC} and V_{OC} are the maximum current and voltage for a solar cell, the output power of the cell at these two extreme points is zero. Hence they are not suitable for operation. The power output of a solar cell for any operating point can be expressed as the rectangle area in the I–V characteristic, indicated in Figure 12.7. There exists one particular operating point (V_m, I_m), which maximizes this power output. "Fill factor" (FF), is a parameter related to the maximum power from a solar cell. It is a measure of the "squareness" of the I–V characteristics, and is defined as the ratio of the maximum power from the solar cell to the product of V_{OC} and I_{SC} as following equation:

$$FF = \left(\frac{I_m \times V_m}{I_{sc} \times V_{oc}}\right)$$

The theoretical FF from a solar cell can be found as the derivative of the power for a solar cell with respect to voltage equals to zero. By setting:

$$\frac{d(I \times V)}{d(V)} = 0,$$

it gives:

$$V_m = V_{OC} - \frac{k_B T}{q} \ln\left(\frac{V_m}{k_B T / q} + 1\right)$$

Theoretically, once V_m is obtained from the above equation, I_m is available and the FF can be obtained. However, the above technique does not yield a simple or closed form. A more commonly used expression for the FF is given by:

$$FF = \frac{v_{oc} - \ln(v_{oc} + 0.72)}{v_{oc} + 1}$$

where v_{OC} is defined as a normalized V_{OC}:

$$v_{oc} = \frac{q}{k_B T} V_{OC}$$

Efficiency:

Efficiency (η) is defined as the ratio of energy output from the solar cell to the input energy from the sunlight:

$$\eta = \frac{P_{max}}{P_{in}} = \frac{V_{OC} \times I_{SC} \times FF}{P_{in}}$$

It is the most important and commonly used parameter to evaluate the performance of one solar cell. Besides that, the cell efficiency also depends on the spectrum and intensity of the incident sunlight and the operation temperature.

12.3 Solar Cell Characterization

The performance of solar cells which convert incident photon energy to electrical energy can be quantitatively analyzed by measuring the current-voltage (I–V) characteristics under illumination. In the simplest case with infinite shunt resistance and zero series resistance, the I–V characteristics of the solar cell without illumination follow diode characteristics. If the diode is ideal and diffusion is the dominant carrier transport mechanism at the junction interface, its I–V characteristics can be described by the Shockley diode equation,

$$I = I_0 \left[\exp\left(\frac{qV}{k_B T} \right) - 1 \right]$$

where, I_0 is the "dark saturation current", q is the electronic charge, k_B is Boltzmann's constant, and T is the absolute temperature (K). The dark saturation current I_0 is an extremely important parameter for diode, and it reflects the recombination in a device.

When the solar cell is illuminated, electron-hole pairs are generated and a photo current flows in the opposite direction of current flow under bias. Then I–V characteristics of solar cell become the superposition of the I–V characteristics of a normal p-n junction diode in the dark plus a light-generated current. Qualitatively, the light has the effect of shifting the p-n junction's dark I–V characteristics down into the fourth quadrant where power can be extracted from the diode. Under illumination the ideal diode law becomes:

$$I = I_0 \left[\exp\left(\frac{qV}{k_B T} \right) - 1 \right] - I_{Ph}$$

where I_{Ph} is the light generated current or photo current. The effect of light on the I–V characteristics of a p-n junction, together with their equivalent electrical symbol, is shown in Figure 12.8.
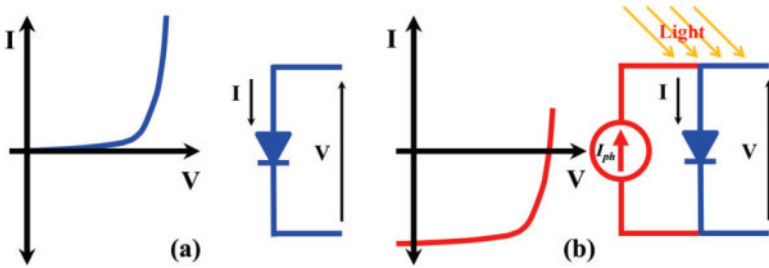
Figure 12.8. I–V characteristics of a p-n junction and their equivalent electrical equivalent circuit model: (a) in dark, (b) with light illumination.

Electrical Characterization

For characterization aiming the actual conditions under which a solar cell operates, standardized spectra based on empirical data are available. Of these, the AM1.5 spectrum is widely used. "Air mass" AM1.5 tells us that the spectrum is based on the solar spectrum after it has passed through air corresponding to 1.5 times the thickness of a standardized atmosphere of the earth. Similarly, the AM0 spectrum is based on the spectrum of the sunlight outside the atmosphere, resembling the conditions experienced by satellites orbiting the earth. Various spectra relevant for photovoltaic technology are shown in Figure 12.9. The AM1.5G and AM0 data are available from the National Renewable Energy Laboratories (NREL), Colorado.

The most fundamental and important characterization of solar cell performance is the measurement of current density-voltage (J–V) characteristics. Figure 12.10 shows the schematic setup of a J–V measurement system. An artificial light source that simulates the sunlight shines on the solar cell. As scanning the voltage, the current between the probes is recorded for each applied voltage.
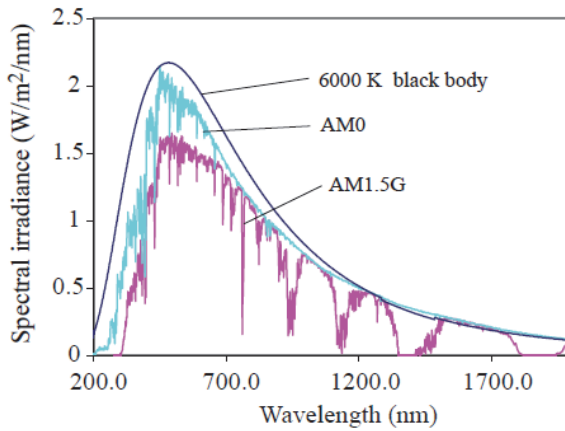
Figure 12.9. Various spectra relevant for PV-technology. After R. Strandberg, Theoretical studies of the intermediate band solar cell, PhD Thesis, Norwegian University of Science and Technology, 2010.
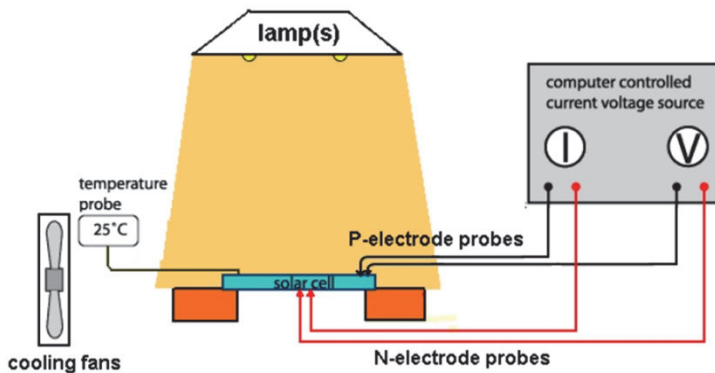
Figure 12.10. Schematic setup of I-V measurement system. After A. Bag, Fabrication and characterization of iron disilicide heterojunction solar cells, PhD Thesis, IIT Kharagpur, 2014.

From measurement of J–V characteristic, we can extract some basic parameters that determine the solar cell performance. Figure 12.11 shows a typical J–V characteristic. It can be seen that open circuit voltage (V_{OC}) and short circuit current density (J_{SC}) can be obtained directly from J–V characteristics. The J–V characteristics both under dark and AM 1.5
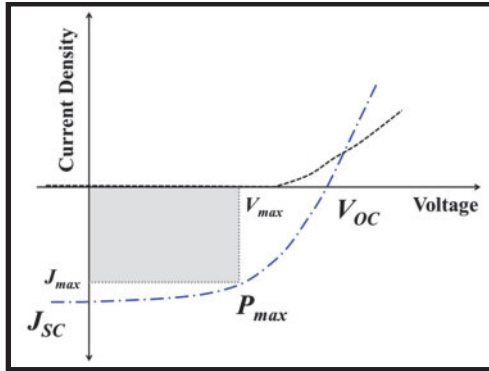
Figure 12.11. Typical J–V characteristics of a solar cell.

illumination (using Model XES-151S) with a power density of 100 mW/cm^2 may be measured with Agilent B1500A semiconductor device analyzer (see Figure 12.12).

Solar Cell Characterization Setup

In order to characterize the solar cells, devices are placed into a solar simulator (see Figure 12.12), consisting of a xenon arc lamp. Solar simulator was used to illuminate the solar cells with the broadband AM1.5 spectrum (100 mW·cm^{-2}). The AM1.5 spectrum was generated using Model XES-151S. The incident power was adjusted to 100 mW/cm^2 using neutral density filters. The solar cells' I–V data were collected and recorded using a semiconductor parameter analyzer (Agilent Technologies, Model B1500A).

For the temperature dependent I–V characteristics, the data are collected as a function of temperature between 120 K and 320 K in the dark in a cryogenic microprobe station. The experimental setup used for I–V analysis is shown in Figure 12.12. The temperature of the cryogenic platen is measured at two locations and controlled using a combination of continuous liquid nitrogen flow and a heater. The desired temperature can be set with the temperature controller, which alters the current flow through the heating element in order to achieve thermal equilibrium of the sample stage at the set temperature. The temperature range of the kit is ~80 K up to ~400 K.
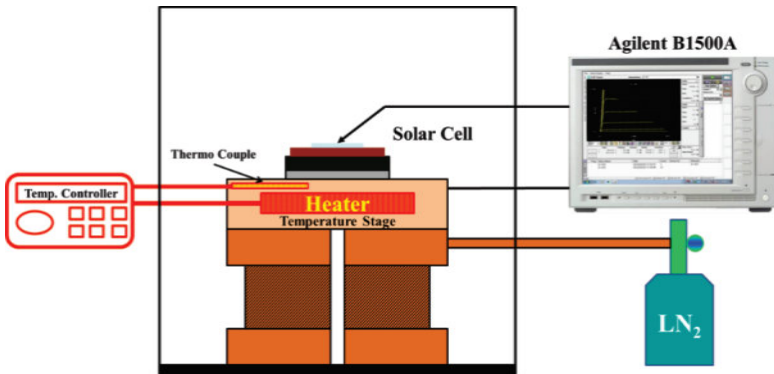
Figure 12.12. Schematic experimental setup for measuring the temperature dependent I–V characteristics of solar cells. After A. Bag, Fabrication and characterization of iron disilicide heterojunction solar cells, PhD Thesis, IIT Kharagpur, 2014.

To measure the solar cell temperature, an additional calibrated temperature sensor is used. It was attached directly on the surface of a second solar cell placed next to the solar cell being tested. The two solar cells need to be clamped in the same way to the cryogenic chuck so that monitoring the temperature of the second solar cell gives an accurate measure of the temperature of the solar cell being studied. Once the solar cell is stabilized at the desired temperature, a voltage sweep is applied to it, and the corresponding current flow is recorded. This data is then multiplied by the appropriate factors in order to obtain the quantity in units of $mA \cdot cm^{-2}$.

12.4 Solar Cell Modeling

Device modeling of solar cells can be done to various degrees of complexity. A complete treatment should incorporate carrier generation and all recombination mechanisms as well as carrier transport due to drift and diffusion in all regions of the cell. A widely used approximation is the depletion approximation. This implies assuming that there are no carriers in the space charge region (SCR). The mathematics involved in solar cell is rigorous, but simplifications can be made which yields analytical expressions. Several simplified assumptions are; (a) the

number of photo-generated carriers will be small compared to the number of carriers introduced by the doping, so called low-injection conditions, and (b) n-region (or the p-region) is so thin that only a negligible fraction of the generation and recombination processes occur there. Combining these approximations, it is possible to construct a simple device model of the solar cell. Use of SILVACO TCAD software will be made for modeling and simulations of both standard GaAs solar cell and InAs/GaAs p-i-n quantum dot solar cells.

In simulating the limits of a solar cell, it is common to take the temperature of the sun to be 6000 K and that of the device to be 300 K. It is possible to analyze the maximum possible performance of a photovoltaic device. By providing each individual material's bandgap into the detailed balance model, the fundamental limiting performance is determined. The performance of several of the elemental and binary semiconductors is plotted in Figure 12.13 as a set of I–V characteristics.
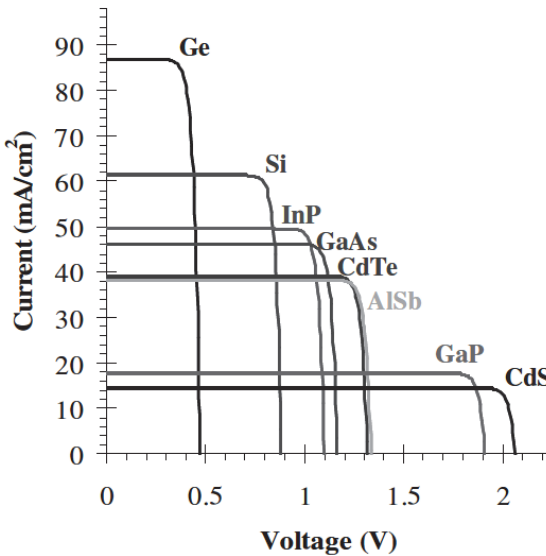
Figure 12.13. Electrical characteristics of several semiconductors operating at the detailed balance limit. Larger bandgap materials exhibit larger open circuit voltages but smaller short circuit currents. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaic based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.

Figure 12.13 helps to understand the trade-offs inherent in materials selection. The smaller bandgap semiconductors allow for a larger short-circuit current because they are able to absorb a larger portion of the solar spectrum; however, their small bandgap places a fundamental limit on the open-circuit voltage. Therefore it makes sense that there will be an optimum bandgap such that the corresponding I–V characteristic gives the maximum attainable efficiency.

GaAs Solar Cell

The structure of a GaAs standard solar cell is a p-n junction and is shown in Figure 12.14. Then other layers are deposited around p-n junction, such as window layer, antireflection layer, back surface field layer, and metal electrodes. P-n junction is used to set up the built-in electric field and space charge region, which are the key factors to help photo-generated carriers apart and generate electricity. Photons propagate with exponential decay in semiconductors. In addition, most photo-generated carriers which are useful for photo current located at the diffusion region and drifting region. Hence, p-n junction should be close to the top surface of solar cell. The structure of the p-n junction is usually designed to have a thinner upper layer and thicker lower layer as shown in Figure 12.14. Since the mobility of holes is much lower than mobility of electrons, the thinner upper side is usually designed as p-type region and the thicker lower side is usually designed as n-type region. As a result, the holes with low mobility will have shorter drifting distance, and electrons with high mobility will have longer drifting distance.

Figure 12.14. Structure of standard solar cell. After B. Dong, Modeling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.

Figure 12.15. Effects of (a) window layer and (b) back surface field layer. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.

Window layer is mainly used to reduce the front surface recombination. For materials that can be chosen as window layer, there are several requirements. First, it should have better transmittance, and most of the light should not be absorbed in this layer. Thus, semiconductor materials with a high band gap can be used as window layer. Also, the layer should be very thin to reduce absorption of light. Second, for good contact with metal electrode, the concentration of window layer should be higher than the concentration of p-type region. Figure 12.15a is the energy band diagram of window layer and p-type region, which can help explain the positive effect of window layer.

In Figure 12.15, the window layer is made of materials with higher band gap than that of p-type region. Also, since it has higher concentration than p-type region, it can be treated as P^+ layer. We can find out that a potential barrier for electrons is built at the junction of window layer and p-type region. The electric field generated at the p^+-p junction will help push those electrons which are minority carriers at p-type region back to

depletion region. After being pushed back to depletion region, these electrons will drift to n-type region with the help of built-in electric field in space charge region. In other words, these electrons come back to the right path under the effect of window layer, which also can be regarded as reduce the probability of top surface recombination.

The effect of back surface field layer is similar to window layer. As shown in Figure 12.15b, n-N^+ junction is built by increasing the concentration of back surface field layer. Similarly, an electric field, as well as the potential barrier for holes, is generated at the junction of n-type region and back surface field layer. When holes which are minority carriers at n-type region come, the electric field will have reflecting effect on these holes, and the potential barrier will push them back to depletion region. Then, with the help of built-in electric field in depletion region, these holes will drift to p-type region. In other words, BSF (back surface field) layer can help block minority carrier holes from diffusing to the bottom and contact layer of solar cell, and thus is helpful to reduce the bottom surface recombination.

The structure of GaAs solar cell used in simulation is shown in Figure 12.16. As shown in Figure 12.16, the structure used in this simulation is a p-i-n solar cell, which has six layers in total. Two red layers in

Figure 12.16. Structure of a typical p-i-n solar cell. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.

Figure 12.16 are made of InGaP, which are called window layer and back surface field layer, respectively. They are used to reduce the surface recombination at top and bottom of the cell, and to block minority carriers. The other four green layers are made of GaAs, which are p-type emitter region, intrinsic region, n-type base region, n-type epitaxial seed layer, respectively.

Silvaco ATLAS is chosen as simulator to build the structure of solar cell and to simulate. The group of statements used to build solar cell structure are: definitions of meshes, regions, electrodes and doping concentration. For solar simulation, we need to define the illumination. In ATLAS 2D device simulation, illumination is activated by a light beam lying above the device. Command BEAM is used to set illumination. After defining the illumination, next step is to define the characteristics under different wavelength to be used in simulation. For obtaining the I–V characteristics, the light density should be a constant. Command "power.file" is used to load files with spectral information, which indicates the relationship between wavelength and density. Next step is to apply bias voltage on electrodes of solar cell, and change their values. We use the "sweep" function in ATLAS, to change values of parameters automatically. Figure 12.17 shows the I-V characteristics of standard p-i-n solar cell. The red solid line is from simulation while the blue dotted line is for comparison with the measured data.
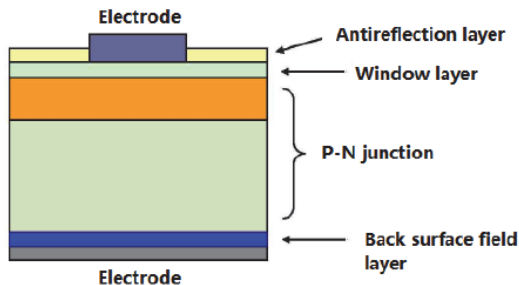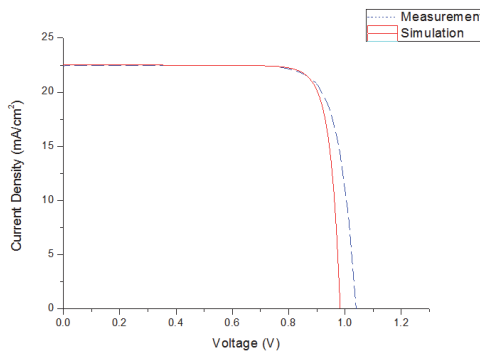

Figure 12.17. Comparison of I–V characteristics of standard p-i-n solar cell. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.
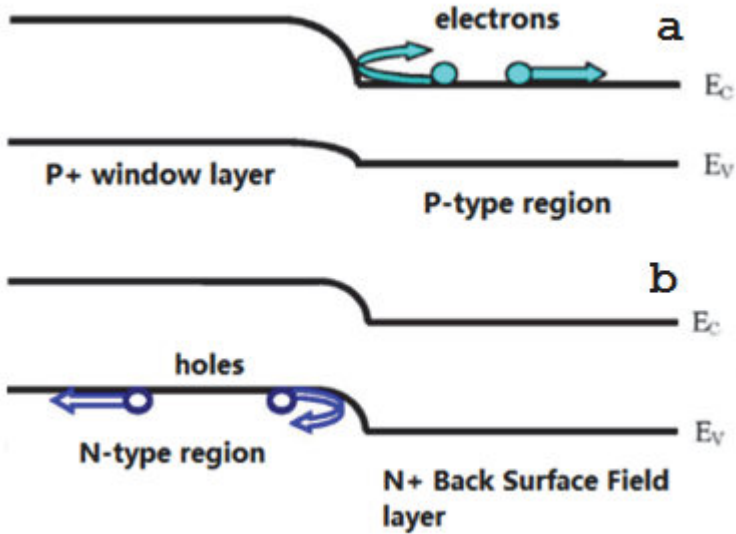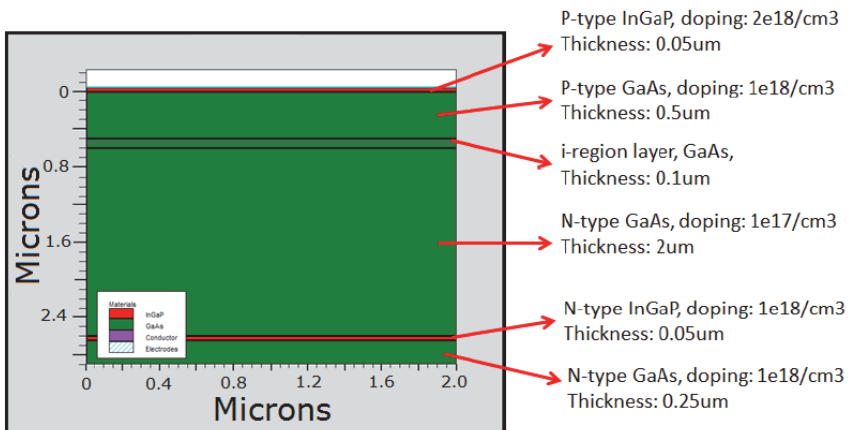
Quantum Dot Solar Cells

Conventional solar cells can absorb photons with different wavelength ranges by using different materials, however, none of them can absorb infrared with longer wavelength. In quantum dot solar cells this can be easily achieved by simply changing the size of the quantum dots. Smaller quantum dots absorb photons with higher energy, and larger quantum dots absorb photons with lower energy. For the third generation solar cells, quantum dots can improve the conversion efficiency via two effects; that a photon with high energy can produce multiple exciton and that intermediate band can help absorb photons with lower energy.

Quantum dot is a nanostructure whose sizes in three dimensions are all smaller than De Broglie wavelength of exciton. In these structures wave property is applicable, which leads to various quantum effects, such as size effect, surface effect, resonant tunneling effect, and multiple exciton effects which are beneficial for the design of quantum dot solar cells. Theoretical conversion efficiency of as high as 60% has been predicted for quantum dot solar cells. Compared to current highest solar cell conversion efficiency which is around 30%, it is a large improvement. Many research results have been presented about different quantum dot solar cells with different materials and structures. Up to now, the highest reported conversion efficiency of fabricated quantum dot solar cells has reached 19.4% which is under two-sun illumination.

For conventional solar cells, the high absorption region of photons ranges from 400nm to 875nm and multiple electron-hole pair generation by collision ionization can rarely occur, and thus cannot help much to increase the conversion efficiency. Photons with higher energies responsible for collision ionization have shorter wavelength and are almost located in the ultraviolet region from 200nm to 400nm. Also the rate of collision ionization must be close to rate of energy relaxation caused by electron-phonon scattering. Due to quantized energy levels in both the conduction and valence bands, electrons and holes move from higher energy level to lower energy level, and the extra energy is used to generate the multiple exciton.

Multiple exciton generation has been found in several nanocrystals. For example, the reported PbSe quantum dots having a diameter 3.9nm (energy of photons is as four times as the band gap), the quantum efficiency is 300%. Although the multiple exciton generation in quantum dots can bring such a high quantum efficiency, fabricated quantum dot solar cells have demonstrated a highest quantum efficiency of only 100%.

In the following, we present p-i-n quantum dot solar cell simulation using ATLAS. P-i-N structured InAs/GaAs quantum dot solar cell used in the simulation is shown in Figure 12.18. The quantum dots are inserted inside the intrinsic layer in between the p- and n-type regions. The advantage of this structure is to improve the drifting action of carriers inside quantum dots and intrinsic region by making use of the built-in electric field of p-n junction.

There are six layers in the structure in Figure 12.18. All of them have the same materials and sizes as the structure used in standard p-i-n solar cell simulations (as shown in Figure 12.16), except the intrinsic region. After inserting the quantum dots, the size of intrinsic region has been changed and depends on the number of quantum dot layers, as shown in Figures 12.19a and b.
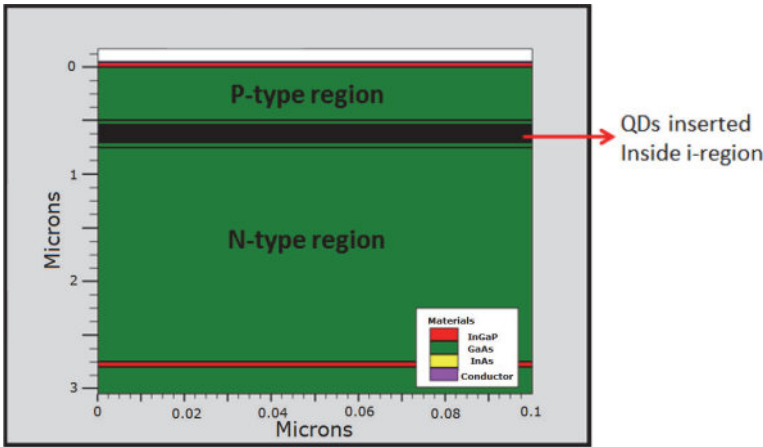
Figure 12.18. Structure of InAs/GaAs quantum dot solar cell with 10-layer QDs embedded in intrinsic region. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.
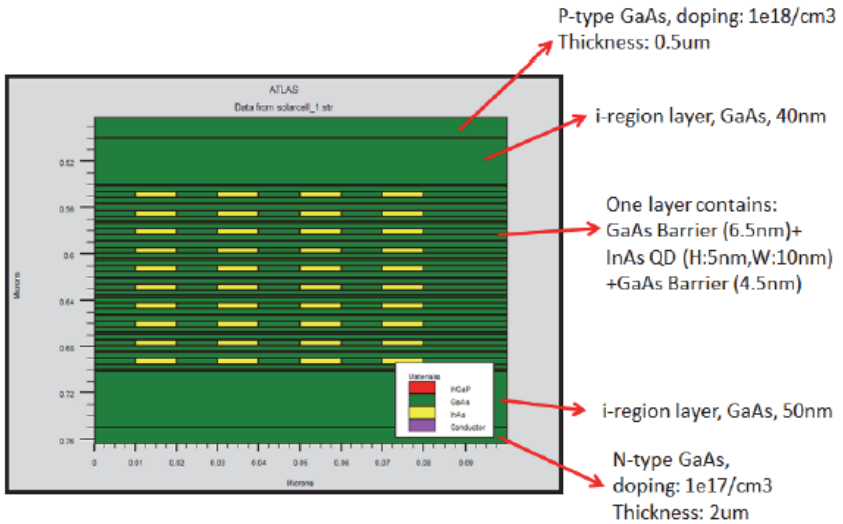
Figure 12.19a. Zoomed-in view of 10 layers quantum dots inserted in intrinsic region. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.
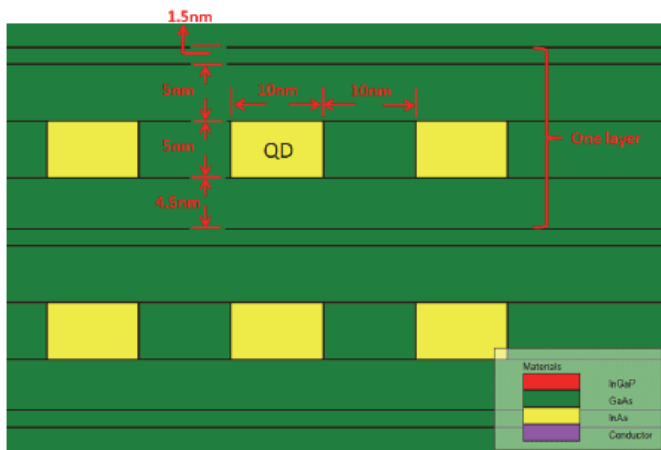
Figure 12.19b. Sizes of quantum dots and barriers in one layer. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.

Figure 12.19a shows the intrinsic region with zoomed view. In this structure, there are ten layers of quantum dots being embedded. Each layer contains several InAs quantum dots and GaAs barrier layers. When simulating quantum dot solar cells, in addition to build the structure of multiple-layer quantum dots, we also need to choose a proper physical model to apply the effect of quantum dots. However, there is no built-in physical model in ATLAS library for simulating quantum dots. However, there are several other quantum effect models in ATLAS which can be used to simulate various effects of quantum confinement. One of them is general quantum well physical model. This model can be used to apply the quantum confinement effects in those quantum well light emitting devices, including quantum well solar cells. In this model, the Schrodinger equation is solved to calculate the bound state energies and wave functions, which can be used in modeling optoelectronic gain, radiative recombination and absorption.

Quantum well and quantum dot are similar nanostructures which both have quantum confinement effects. However, for quantum well, in one dimension, the size is in nanometer scale. For quantum dot, its sizes in three dimensions are all in nanometer scale. As a result, quantum well has one-dimension quantum confinement effect (in the vertical direction), but quantum dots have three-dimension quantum confinement effect. This is one of the differences between the two. In other words, in quantum well, carriers (electrons and holes) are in two-dimension. Thus, one needs to use an approximation to achieve the quantum effects in of the simulation of quantum dots. In ATLAS, the Schrodinger equation is solved along discrete slices of calculation grid in the quantization direction. Here, the quantization direction can be controlled, and the 2D orientation can be set by the flag "2DXY.SCHRO" on the "MODELS" statement. When this flag is on, it tells ATLAS that the Schrodinger equation need be solved in XY plane. In other words, the Schrodinger equation will be solved along discrete slices both in X-axis direction and Y-axis direction. As a result, the carriers are no longer confined in only one direction (the vertical direction), but in both X-axis and Y-axis directions. Since quantum dot is in nanometer scale, its size is much smaller than other normal layers in a solar cell. Thus, the mesh density around quantum dots set by command "MESH" might be coarse, and decrease the accuracy of simulation results. Important parameters of solar cells, such as short circuit current, open circuit voltage, fill factor

and conversion efficiency, are computed using ATLAS. The results are helpful for analyzing the effect of embedded quantum dots, especially the spectral response of solar cell.

Figure 12.20 shows the simulation results of four different solar cells. One is the standard p-i-n solar cell without quantum dots. The other three are all quantum dot solar cells with different number of quantum dot layers, which are 10-layer, 20-layer and 40-layer, respectively. Figure 12.20 shows the I–V characteristics of both standard solar cell and quantum dot solar cell, which are plotted from the simulation results. It is observed that short circuit current increases when more quantum dots are inserted, and open circuit voltage almost remains the same. As a result, the conversion efficiency increases relatively 31.5% when comparing 40-layer quantum dot solar cell with standard solar cells.

As the most important impact of quantum dots is to help solar cells absorb those photons with lower energy, which also means long wavelength. The maximum wavelength of photons that can be absorbed by a typical standard GaAs solar cell is around 875nm. After inserting
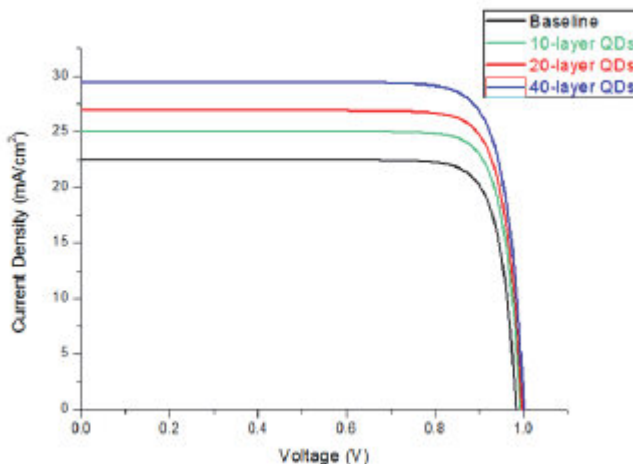
Figure 12.20. I-V characteristics of standard solar cell and quantum dot solar cell with different number of QD layers. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.
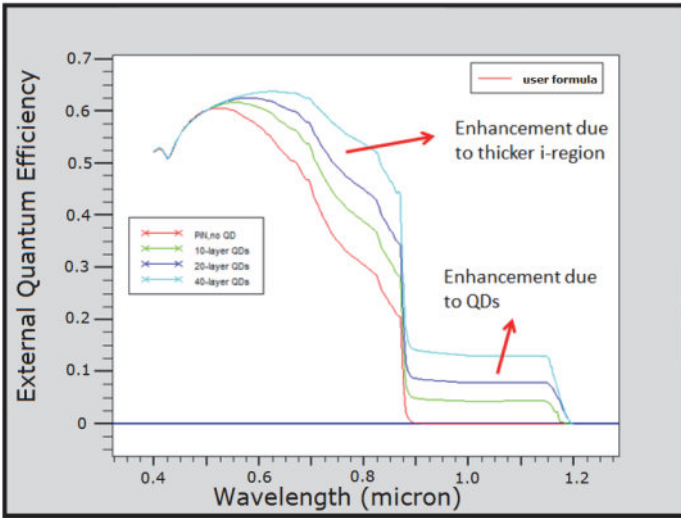
Figure 12.21. External quantum efficiency of standard solar cell and quantum dot solar cell with different number of QD layers. After B. Dong, Modelling and simulation of InAs/GaAs quantum dot solar cells in Silvaco TCAD, MS Thesis, The George Washington University, 2014.

quantum dots, photons with wavelength between 875nm to 1200nm are also able to be absorbed.

As shown in Figure 12.21, the external quantum efficiency of standard p-i-n solar cell without quantum dots indeed turns to zero around wavelength of 875nm. For the solar cells with 10-layer, 20-layer and 40-layer quantum dots, the external quantum efficiency from wavelength of 875nm to 1200nm varied from 5% to 15%.

12.5   Intermediate Band Solar Cell

The concept of intermediate band solar cell (IBSC) which has a theoretical efficiency of 63.2% at full light concentration was proposed by Wolf. In an IBSC there is a band of electron states, an intermediate band, in the main band gap of the solar cell material as shown in Figure 12.22. This band allows electrons to be exited from the valence band

(VB) to the conduction band (CB) in a two-step process. Ideally, the interband (IB) should only be optically coupled to the VB and the CB, meaning that electrons should only be allowed to make transitions to or from the IB while absorbing or emitting a photon. The increased efficiency of the IBSC compared to ordinary solar cells can be attributed to a reduction of both absorption losses and thermalization losses. The main band gap E_{cv} can now be larger than in conventional cells, providing a better utilization of the high energy photons. Photons with energy lower than the main band gap, can still be absorbed while exciting electrons over the two sub-band gaps E_{ci} and E_{iv}. The feature that distinguishes the IBSC from other photovoltaic concepts is that it allows electrons to be excited from low energy electron states to high energy states in a two-step process via an intermediate band (IB).

Despite the promising efficiency limit of 63.2%, which is more than 50% higher than the efficiency limit for conventional solar cells, the IBSC will be just a good idea unless ways are found to implement the concept in reality. The first proposed implementation of the IBSC concept was the use of quantum dot superlattices. Quantum dots (QDs) are nanosized particles of a particular material. Due to quantum mechanical effects,
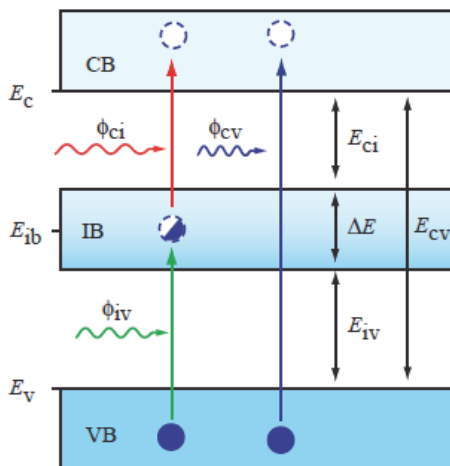
Figure 12.22. The intermediate band solar cell has an intermediate band situated in the main band gap. This allows for a two-step excitation of electrons from the VB to the CB. After R. Strandberg, Theoretical studies of the intermediate band solar cell, PhD Thesis, Norwegian University of Science and Technology. 2010.
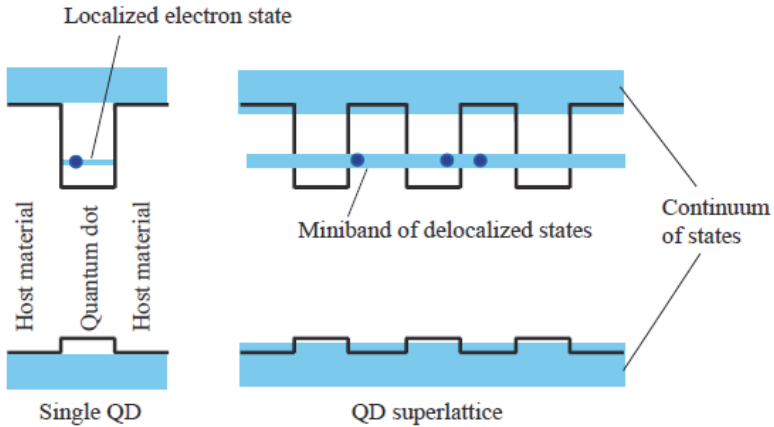
Figure 12.23. In quantum dots quantum effects give rise to new electron levels. After R. Strandberg, Theoretical studies of the intermediate band solar cell, PhD Thesis, Norwegian University of Science and Technology, 2010.

particles of this size will have discrete energy levels. The position of the energy levels can be tuned by varying the size of the QDs. When many QDs are closely spaced, typically a few nanometers between them, the electrons can tunnel from QD to QD. The electron wave functions have become delocalized and have formed a miniband as illustrated in Figure 12.23.

Quantum dot intermediate band solar cell material triads (quantum-dot/barrier/substrate) have been proposed that yield thermodynamic conversion efficiencies of over 60% at maximum concentration. These material triads are selected based upon several design rules. The valence band offsets between the quantum dot material and barrier material are negligible, thus mitigating efficiency loss due to miniband formed by hole states. The conduction band offsets between the quantum dot materials and barrier materials are at least as large as the energy gap (between the conduction band and intermediate band) required for high efficiency conversion. Direct bandgap are preferred as they have larger absorption coefficients, with respect to indirect bandgap, thus permitting the epitaxial growth of thinner solar cells. The lattice constants of the substrates are either lattice matched with their barriers or are in-between the lattice constants of their quantum dot materials and their barriers.

Although a very high theoretical conversion efficiency have been predicted for intermediate band quantum dot solar cells, experimental solar cell performance is found to be much lower. This is due to the quantized energy levels inserted in the band gaps produce unexpected non-radiative recombination centers, which leads to the decrease of solar cell performance. To improve performance, quantum dot solar cell with intermediate band requires tightly arrangement of quantum dots in order to get consistent energy level. In addition, a high density of quantum dots is required for absorbing more photons.

12.6 Multi-junction Solar Cell

In the last few decades, ever increasing demand for alternative renewable clean energy sources have stimulated scientific research in solar energy harvest. It is clear that the material bandgap plays an essential role in determining the performance of a solar cell and any single device could be made more efficient simply by extending the I–V characteristic such that a larger open-circuit voltage is achieved which is the basic idea behind the multi-junction solar cell. The device considered in this section is the dual-junction, InGaP-GaAs, tandem solar cell as diagram in Figure 12.24. This design places an InGaP solar cell atop the GaAs solar cell. From top to bottom, the InGaP design is a 100nm GaAs contact layer doped at 1×10^{19} acceptors/cm^3, 30nm AlAs window doped at 1×10^{18} acceptors/cm^3, 100nm InGaP emitter doped at 1×10^{18} acceptors/cm3, 1 μm InGaP base doped at 1×10^{17} donors/cm^3, and a 100nm InGaP BSF doped at 1×10^{18} donors/cm^3. In between the InGaP and GaAs solar cells is an n-on-p GaAs tunnel junction; 5×10^{19} cm^{-3} donor concentration and 2×10^{19} cm^{-3} acceptor concentration. This layer allows for electrons in the top cell to recombine with holes in the bottom cell by means of internal field emission thus driving the current; this is easily seen in the device's band diagram (Figure 12.25). Similar to the design of the bottom cell, the AlAs window and InGaP BSF of the top cell serve to reflect and accelerate minority carriers towards the top cell junction. The results of simulating the tandem cell with the discussed modifications are shown in Figure 12.26.
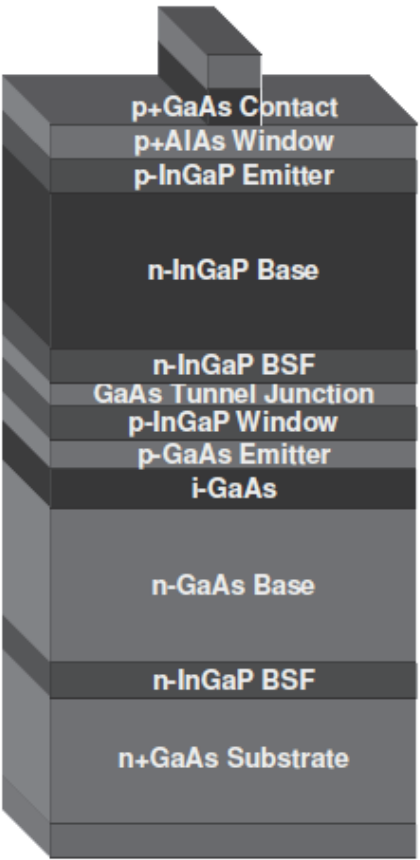
Figure 12.24. Device structure of the InGaP-GaAs tandem solar cell considered in this work. The InGaP top cell more efficiently collects short-wavelength light while remaining transparent to the light more efficiently collected by the GaAs bottom cell. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaic based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.

Figure 12.25 clearly shows how the window and BSF layers act as potential barriers that reflect minority carriers towards their respective junctions. Also observable from this diagram, at ~1.35μm, is the tunneling region that drives current between the two sub-cells. The simulated I–V characteristic of the InGaP-GaAs tandem cell is shown in Figure 12.26.
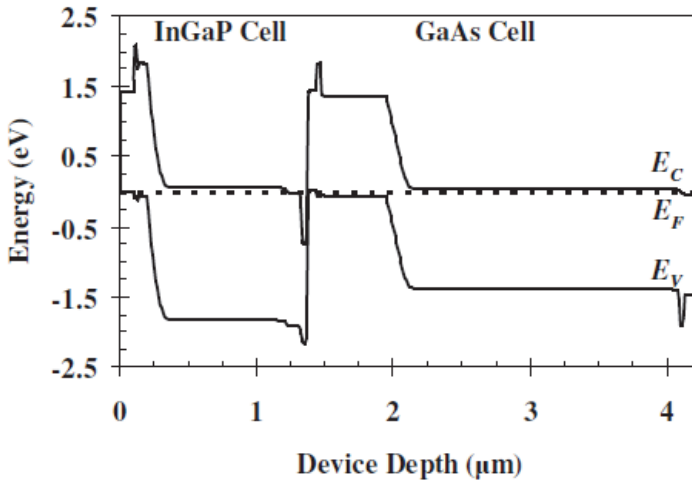
Figure 12.25. Band diagram of the InGaP-GaAs solar cell. After R. Aguinaldo, Modelling solutions and simulations for advanced III–V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.
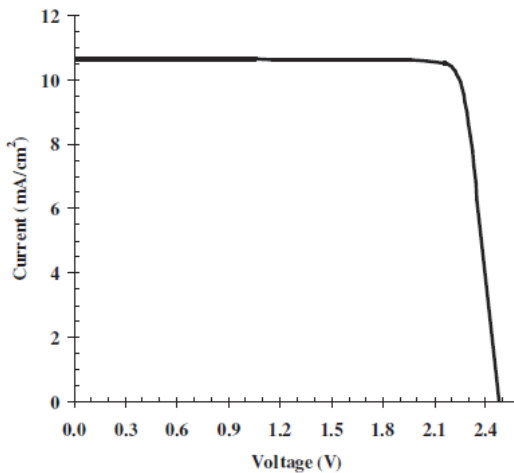
Figure 12.26. Simulated I–V characteristic of the InGaP-GaAs tandem cell. After R. Aguinaldo, Modelling solutions and simulations for advanced III–V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.
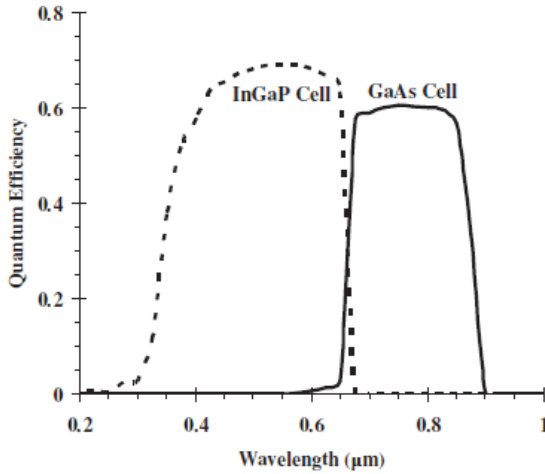
Figure 12.27. Spectral response of the InGaP-GaAs tandem cell showing the contributions of each sub-cell. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.
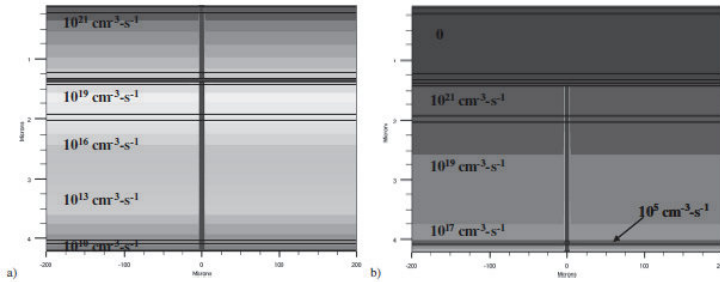
Figure 12.28. Photo generation rate throughout the device for a) a 550nm and b) 800nm light source. For the 550nm illumination, the photo generation occurs primarily in the InGaP sub-cell; no photo generation occurs in that cell at 800nm. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.

The spectral response of the tandem cell is shown in Figure 12.27. It is clear that the tandem cell is better suited to perform photo conversion throughout a wider range of wavelengths. The photo generation throughout the device is displayed in Figure 12.28 for a 550nm light

source. This indicates that, at this wavelength, the vast majority of photo generation occurs due to the InGaP cell.

The intermediate band solar cell is a particularly interesting photovoltaic concept because it has a theoretical efficiency limit of comparable size to the triple junction tandem cell, but can, if suitable materials are found, be implemented in a less complicated way. However, a multi-junction device, increases solar efficiency by enabling additional and more efficient absorption throughout the solar spectrum. While the triple junction tandem cell consists of three cells stacked on top of each other, the intermediate band solar cell consists of a single cell. Triple-junction tandem cells are structures with three solar cells with different optical properties stacked on top of each other. As a specific case of the multi-junction solar cell, a triple-junction solar cell is shown in Figure 12.29a. InGaP-GaAs-Ge triple junction cell and represents the state-of-the-art in basic multi-junction approaches. The materials are chosen to conform to the constraint of lattice matching. In this device, there are actually three separate p-n junctions, each composed of a different material. This constitutes an equivalent circuit of three solar cells connected in series as seen in Figure 12.29b.
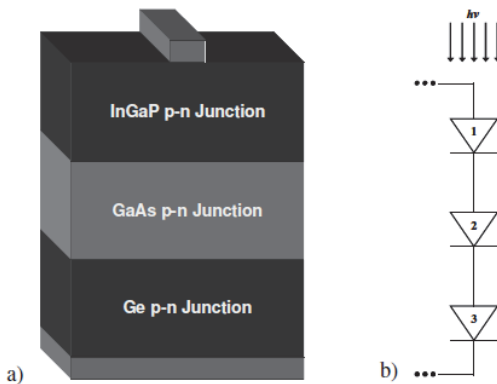
Figure 12.29. InGaP-GaAs-Ge triple-junction solar cell. a) Diagram to show the proper placement of each layer such that the smallest bandgap material is placed at the bottom with increasing bandgap towards the top of the stack. The device is b) diagram as an equivalent circuit with three individual solar cells connected in series representing each of the individual junctions. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.

The materials are placed such that the one with the largest bandgap is located at the top of the stack; the remaining materials follow the trend of decreasing bandgap towards the bottom of the stack. This design allows for the most efficient conversion of higher energy photons by the top cell. The top cell is transparent to sub-bandgap light; this light is then absorbed by one of the remaining cells underneath. This allows for the splitting of the absorption of the solar spectrum into more efficient means; this is shown in Figure 12.30.

Efficiency of this device is enhanced from the single-junction cell due to the photo generation occurring over several junctions at once. In general, a separate photo voltage will be dropped across each junction at any given time; these voltages add giving the total voltage across the multi-junction device. This is the mechanism which leads to an increased open-circuit voltage, i.e., the I–V characteristic is stretched along the voltage axis. The band gaps of InGaP, GaAs, and Ge are 1.89 eV, 1.42 eV, and 0.66 eV, respectively. Using these parameters in the model the efficiency limit is 33.5 % for the triple-junction cell. By decreasing only the middle cell bandgap to 1.20 eV, the limit is increased to the maximum point at 47.5 %. A quantum well or quantum dot array placed in the GaAs middle
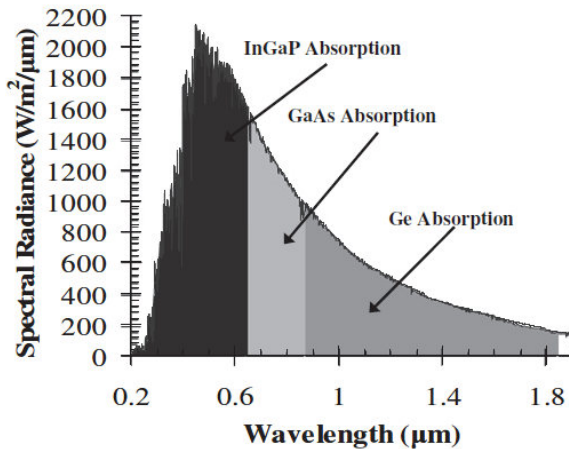
Figure 12.30. AM0 solar spectrum split into separate absorption regions corresponding each of the sub-cells in the triple-junction InGaP-GaAs-Ge stack. After R. Aguinaldo, Modelling solutions and simulations for advanced III-V photovoltaics based on nanostructures, MS Thesis, Rochester Institute of Technology, 2008.

cell may therefore induce an effective bandgap lowering such that the overall triple-junction efficiency limit increases from 33.5% to 47.5%.

Summary

In this chapter, the design and simulation of various advanced solar cell structures have been discussed. Experimental setup and characterization of several important parameters for a solar cell, such as short circuit current, open circuit voltage, fill factor and conversion efficiency have been described in detail. First, the simulation of a standard p-i-n GaAs solar cell without quantum dot has been taken up. Then simulations have been performed with quantum dot layers of 10-, 20-, and 40. The 40-layer quantum dot solar cell showed an improved conversion efficiency relative to the standard solar cell by 31.5%. Then analyses of advanced III–V photovoltaic devices and their enhancement by the incorporation of semiconductor nanostructures have been performed.

Chapter 13

Heterojunction Solar Cells

Contributed Chapter by Dr. Atanu Bag

The tremendous rise in fossil fuel consumption, along with its greenhouse effects, and other negative environmental impacts have resulted in research efforts that focus on renewable energy sources like hydropower, biomass, solar, wind, geothermal and ocean energy. Among these, the solar is the most promising and reliable source of energy. In semiconductor industry metal silicides have gained increased interest as contact. Among various metal silicides semiconducting iron disilicide (β-FeSi$_2$) has attracted intense research interest recently due to its remarkable optical and electrical properties. In this chapter, we focus on iron disilicide based heterojunction solar cells. The semiconducting iron disilicide (β-FeSi$_2$) has recently attracted considerable attention from both the scientists and engineers due to its remarkable optical and electrical properties. Semiconducting β-FeSi$_2$, made from ubiquitous and nontoxic materials, is expected to be used in optical devices including thin film solar cells as it has a high absorption coefficient ($\alpha > 10^5$ cm^{-1}, which is 200 times larger than that of crystalline silicon at 1.5 eV), a direct energy bandgap of 0.85 eV, chemical stability at temperatures as high as 937°C, high thermoelectric power (Seebeck coefficient of k $\sim 10^{-4}$/K) and high resistance against the humidity, chemical attacks and oxidization. It is also compatible with silicon technology.

Therefore, β-FeSi$_2$ is one of the most promising materials for various applications such as light-emitting diodes, infrared sensors, and solar cells. It is a novel photovoltaic material with theoretical energy conversion efficiency of ~16–23%. β-FeSi$_2$/Si heterojunction solar cell enables the absorption of solar light over a wide wavelength range from ultraviolet to near-infrared and therefore, a high photovoltaic efficiency

is expected. Theoretical energy conversion efficiency, as high as 23%, may be achieved in heterojunction β-FeSi$_2$/Si solar cells. However, until now, there have been limited reports on β-FeSi$_2$ solar cells with poor conversion efficiency (~3%). In this chapter, we investigate the photovoltaic properties of sputter-deposited β-FeSi$_2$ solar cell via simulation.

To understand the reasons of low conversion efficiency for β-FeSi$_2$/c-Si heterojunction solar cells, the effects of surface recombination and interface states on the cell performance has been investigated by numerical simulation. It is well known that surface quality of β-FeSi$_2$ film plays an important role in limiting the cell performance as β-FeSi$_2$ is highly sensitive to surface recombination due to its high optical absorption coefficient. As the surface quality, defects and traps, interface states between β-FeSi$_2$ film and crystalline silicon, work function, are critical to solar cell performance, numerical simulation has been taken up to study the effects of the above parameters using SILVACO ATLAS and the simulations results are presented in this chapter.

For simulation, SILVACO ATLAS is chosen to build the models of solar cell. SILVACO TCAD tools contain many different modules, such as interactive tools DeckBuild and TonyPlot, process simulator ATHENA, device simulator ATLAS, structure and mesh editor DevEdit , and so on. Figure 13.1 describes the flow diagram when simulating through SILVACO TCAD. All these parts are managed through DeckBuild interface. ATLAS is used as it is powerful device design and simulation software [2.17].
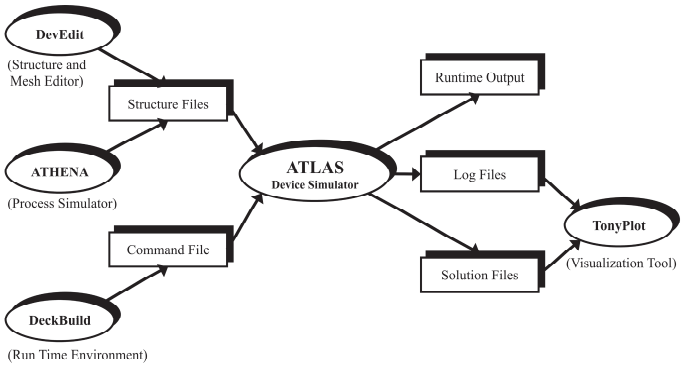

Figure 13.1. Simulation flow diagram of SILVACO ATLAS.

From Figure 13.1, we can see that device simulator ATLAS is at the core of SILVACO TCAD tools. It can simulate semiconductor device characteristics, including electrical behavior, optical behavior, thermal behavior, and so on. ATLAS provides an easy-to-use, modular and extensible platform, which is based on physics. In addition, ATLAS affords a large number of physical models for users to choose through C-interpreter function [2.17].

General Feature of ATLAS

- It is capable of simulating AC, DC, and transient simulation.
- Available transport models are drift-diffusion, energy balance and hydrodynamic.
- Available carrier statistics models are Fermi-Dirac and Boltzmann statistics.
- It can simulate graded and abrupt heterojunction.
- Different recombination models available are SRH, radiative, Auger, and Surface recombination.
- Carrier mobility models available are concentration dependent e and h mobility, Klaassen's mobility.

As a device simulator, ATLAS owns a full set of functions from building models to calculating output results. All these functions are achieved through statement description. ATLAS statements can be classified into different groups by five different functions, which are structure specification, material models specification, numerical method selection, solution specification, results analysis, respectively.

Structure specification: Including the division of meshes, definition of regions with different materials, definition of electrodes, and setup of doping concentration.

Material models specification: Including setup of material parameters (band gap, absorption coefficient, minority lifetime, electron affinity etc.), choose of physical models, specification of contact characteristics, and definition of interface properties.

Numerical method selection: Computer simulation is based on numerical calculation. To make sure that the calculation can converge, numerical method should be selected properly. In ATLAS, there are Newton iteration method, Gummel iteration method, Block iteration method, and combination iteration method.

Solution specification: Obtain device working characteristics by setting up voltages, currents, illumination, and so on.

Results analysis: To help analyze, specified data or values can be extracted from simulation results obtained in previous steps.

13.1   Design of FeSi$_2$ Heterojunction Solar Cells

Two-dimensional (2D) device simulator, SILVACO ATLAS, which can handle complicated device structures, has been used for the simulation of solar cell. ATLAS tool is used to analyze carrier generation-recombination, traps and current voltage characteristics. 1μm×1μm area is considered for simulation. Solar cell structures with different layers is shown in Figure 13.2. Material parameters of Si and β-FeSi$_2$ for simulation purpose are listed in Table 13.1. ITO is used as front contact material and back contact is considered ohmic. Indium tin oxide is usually used as a high work function electrode since it combines good conductivity and transparency in the visible range.
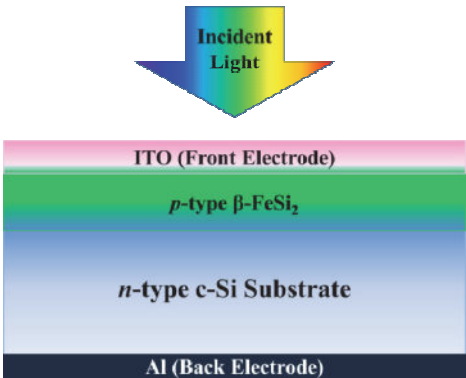

Figure 13.2. Schematic structure of β-FeSi$_2$/c-Si heterojunction solar cell used in the simulation.

13.2 Simulation and Modeling

Heterojunction effects were considered because materials used have different bandgap, electron affinity, density of states and mobility. Dominant recombination mechanism in Si is Shockley–Read–Hall and Auger. It is considered to be the same for FeSi₂. Matrix methods are fast and accurate way to simulate EM wave propagation through layered materials. Propagation of EM wave in device structure is modeled by transmission matrix method. It directly relates the electric filed amplitude of reflected and transmitted wave to the amplitude of incident wave. Wavelength dependent complex index of refraction ($\tilde{n} = n + ik$) is used in simulation for ITO, FeSi₂, and Si. The optical source considered was the standard AM1.5 spectrum and the solar cells were assumed illuminated using normal incidence of light.

Table 13.1. Material parameters and device dimensions of β-FeSi₂ and c-Si layer. *Data source*: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and Properties". http://www.ioffe.ru/.

Parameters and units	β FeSi₂ (p)	c-Si (n)
Thickness (μm)	0.05 to 0.08	100
Dielectric constant	31.0	11.8
Electron mobility (cm²V⁻¹s⁻¹)	70	1350
Hole mobility (cm²V⁻¹s⁻¹)	20	450
Acceptor/Donor conc. (cm⁻³)	1×10^{19}	1×10^{18}
Bandgap (eV)	0.87	1.12
Effective conduction band density (cm⁻³)	2.5×10^{19}	2.8×10^{19}
Effective valence band density (cm⁻³)	2.5×10^{19}	1.04×10^{19}
Electron affinity	3.86	4.05

Defects modeling in bulk FeSi$_2$

Disordered or amorphous material contains a number of defects states within bandgap of material. Total density of states of mid-gap states, $g(E)$ is modeled by a combination two Gaussian distributions and two exponential decaying band tail states.

$$g(E) = g_{GA}(E) + g_{GD}(E) + g_{TA}(E) + g_{TD}(E)$$

$$g_{GA}(E) = NGA \exp\left[-\left(\frac{EGA - E}{WGA}\right)^2\right]$$

$$g_{GD}(E) = NGD \exp\left[-\left(\frac{E - EGD}{WGD}\right)^2\right]$$

$$g_{TA}(E) = NTA \exp\left[\frac{E - E_C}{WTA}\right]$$

$$g_{TD}(E) = NTD \exp\left[\frac{E_V - E}{WTD}\right]$$

Here, E is the trap energy, E_C is the conduction band energy, E_V is the valence band energy and the subscripts (T, G, A, D) stand for Tail, Gaussian (deep level), Acceptor and Donor states respectively.

Interface modeling at Fesi$_2$/Si interface

Mid band traps/defects are always present at interfaces of two semiconductors. For analysis of interface defects/traps, 5nm thick interface is considered with uniform trap density. This region is placed intentionally to analyze its effects on carrier generation and recombination. Two trap levels (acceptor like traps and donor like traps) are used in simulation. Parameters used to model interface traps are from Tables 13.2 and 13.3.

Table 13.2. Parameters for defects modeling in amorphous FeSi$_2$. *Data source*: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and Properties". http://www.ioffe.ru/.

Density of acceptor, donor like states for tail distribution (NTA, NTD) (cm^{-3}eV^{-1})	3×10^{18}, 3×10^{18}
Characteristics decay energy for acceptor, donor like states for tail distribution (WTA,WTD) (eV)	0.01, 0.01
Capture cross section for electrons, holes for tail distribution of acceptor like states (SIGTAE, SIGTAH) (cm^2)	1×10^{-16},1×10^{-15}
Capture cross section for electrons, holes for tail distribution of donor like states (SIGTDE, SIGTDH) (cm^2)	1×10^{-15},1×10^{-16}
Density of acceptor, donor like states for Gaussian distribution (NGA, NGD) (cm^{-3}eV^{-1})	1×10^{15},1×10^{15}
Characteristic decay energy for acceptor, donor like states for Gaussian distribution (WGA, WGD) (eV)	0.1, 0.1
Gaussian distribution peak energy for acceptors, donors (EGA, EGD) (eV)	0.4, 0.4
Capture cross section for electrons, holes for Gaussian distribution of acceptor like states (SIGGAE, SIGGAH) (cm^2)	1×10^{-15},1×10^{-16}
Capture cross section for electrons, holes for Gaussian distribution of donors like states (SIGGDE, SIGGDH) (cm^2)	1×10^{-16}, 1×10^{-15}

Table 13.3. Parameters for defects modeling at FeSi$_2$/Si interface. *Data source*: Ioffe Physico-Technical Institute. "New Semiconductor Materials Characteristics and

Properties". http://www.ioffe.ru/.

Energy of discrete trap level for acceptors, donors like trap (E.LEVEL) (eV)	0.4
Capture cross section of traps for electrons, holes for acceptor type trap (SIGN, SIGP) (cm^2)	$1\times10^{-12}, 1\times10^{-14}$
Capture cross section of traps for electrons, holes for donor type trap (SIGN, SIGP) (cm^2)	$1\times10^{-14}, 1\times10^{-12}$

13.3 Simulation Results

Effect of χ of FeSi₂

From semiconductor heterostructure theory, the conduction (ΔE_C) and valance (ΔE_V) band offset of β-FeSi$_2$/Si are obtained as.

$$\Delta E_C = \chi_1 - \chi_2$$

$$\Delta E_V = E_{g_1} - E_{g_2} + \Delta E_C$$

where χ_1 and χ_2 are electron affinities of FeSi$_2$ and Si, E_{g_1} and E_{g_2} are bandgap of FeSi$_2$ and Si. Both conduction and valance band offsets increase with affinity of FeSi$_2$.
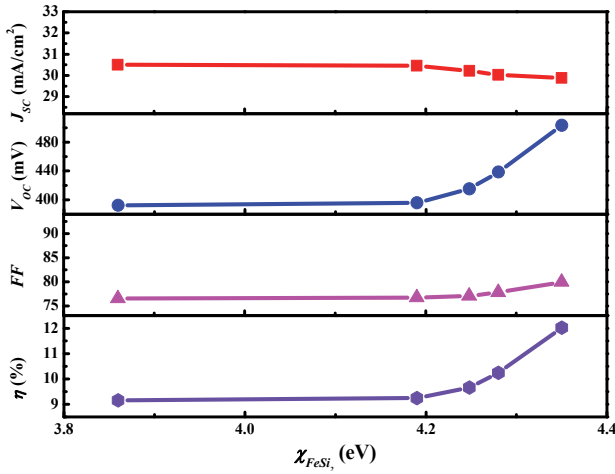
Figure 13.3. Photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cell against χ FeSi$_2$.

The data from Table 13.1 is used to investigate the effects of χ_1. Figure 13.3 shows the photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cell against χ FeSi$_2$. Ohmic contact is considered for ITO and FeSi$_2$. The drift of photo generated electrons from p-β-FeSi$_2$ to n-Si is the main carrier collection process. Therefore, the conduction band offset of p-β-FeSi$_2$/n-Si at the interface affects the movement of photo generated electrons. A spike in conduction band can be observed in Figure 13.4 for χ_1 = 4.35 eV. This spike acts as electron barrier and the reason behind decrease in short circuit current. Open circuit voltage of p-n junction is proportional to the work-function differences (qϕ_1−qϕ_2) of materials. Work-function of FeSi$_2$ (qϕ_1) increases with increase in affinity of FeSi$_2$, so increase in V_{OC} with affinity of FeSi$_2$ is observed. Short current density (J_{SC}) decreases by slight margin and V_{OC} increases significantly with increase in χ_1 of FeSi$_2$ as a result efficiency increases with χ_1 of FeSi$_2$.
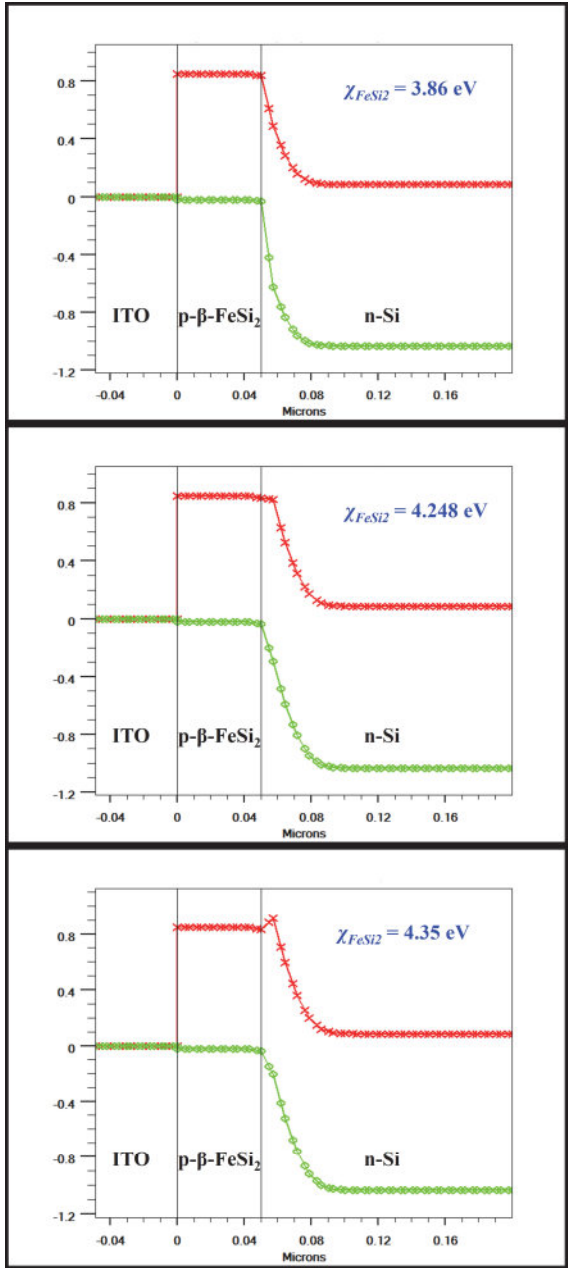
Figure 13.4. Band diagram variation with different values of χ_{FeSi_2}.

Fill Factor follows efficiency, increases with affinity of FeSi$_2$. Maximum η of 12.02 % is observed at an affinity of 4.35 and it further increases to 15.73% at an affinity of 4.5.

Effect of work-function of ITO

Indium tin oxide (ITO) is used as transparent contact material for p-β-FeSi$_2$. High work function contact material is required for a better contact on p type material. Parameters used for simulation are listed in Table 13.1, affinity of FeSi$_2$ is taken as 4.35 and work functions of ITO are taken from literature. Bandgap diagram for different work function of ITO is shown in Figure 13.5. The bandgap at the interface of ITO/FeSi$_2$ changes with the change in work-function of ITO. The rise in conduction band energy at this interface surface helps photo generated electrons in p-β-FeSi$_2$ to reach n-Si region, and rise in valance band energy helps photo generated holes in p-β-FeSi$_2$ to move towards ITO (anode).

This kind of band diagram helps in electrons and holes separation, as a result J$_{SC}$ increases with increase in ITO work function. The rise of conduction band at ITO/p-β-FeSi$_2$ interface can be also considered as increase in work function of FeSi$_2$ near ITO interface region, which is responsible for increase in V_{OC}. As V_{OC} and J$_{SC}$ increases with increase in work function of ITO, efficiency increases. Fill factor increases as efficiency increases. Maximum η of 15.5% (without AR 14.64) is obtained at ITO work function of 5.75. Henceforth in all simulations affinity of FeSi$_2$ and work function of ITO as 4.35 and 5.75, respectively, were used. Figure 13.6 shows the photovoltaic properties of p-β-FeSi$_2$/ n-Si heterojunction solar cells as a function of W$_{ITO}$.
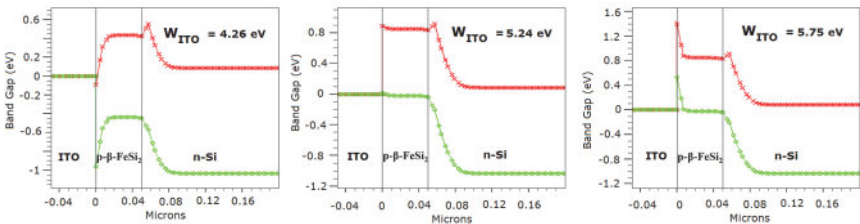

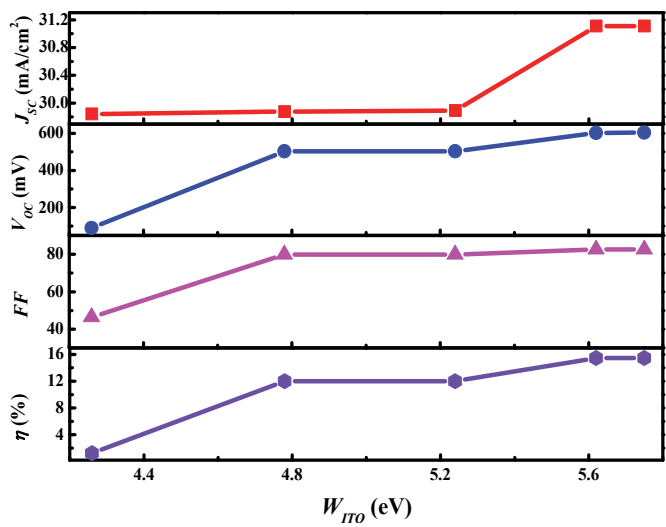Figure 13.5. Band diagram variation with different values of W$_{ITO}$.

Figure 13.6. Photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cell against W$_{ITO}$.

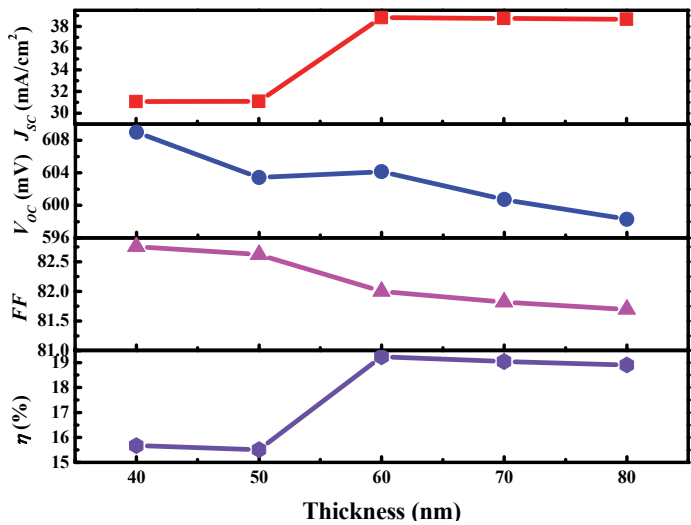## Effect of thickness variation of FeSi$_2$

Figure 13.7. Photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cell against FeSi$_2$ thickness.

Solar structure is tuned for maximum efficiency, all the parameters used are from Table 13.1 except thickness of p-β-FeSi$_2$ and, affinity of FeSi$_2$ and work function of ITO used are 4.35 and 5.75, respectively. The variation of different solar parameters with respect to p-β-FeSi$_2$ thickness is shown is the Figure 13.7. Maximum efficiency of 19.23% is obtained at a thickness of 60nm. Tuning of FeSi$_2$ thickness maximizes the absorption and minimizes reflection of light in the solar cell.

Effects of Surface Recombination

Effects of front surface recombination velocity (SRV) is shown in Figure 13.8. All the parameters used in simulation are from Table 13.1 except β-FeSi$_2$ thickness (0.06μm), for max efficiency. For simulation purpose SRV of electron and hole are considered equal. SRV influences the important solar parameter V_{OC} and J_{SC}. Increase in SRV from 10^3 to 10^6 cm/s reduces J$_{SC}$ from 38.40 mA/cm^2 to 19.18 mA/cm^2, and reduces V_{OC} from 588.3 mV to 495.04 mV. Efficiency decreases from 19.23 % to 6.26 % when front SRV is varied from 1 to 10^7cm/sec. This suggests that front SRV is one of the key parameter that can significantly deteriorate cell performance. Due to very high optical absorbance of FeSi$_2$ ($\alpha > 10^5$ cm^{-1}), large amount of photo carriers are generated in thin film of FeSi$_2$.
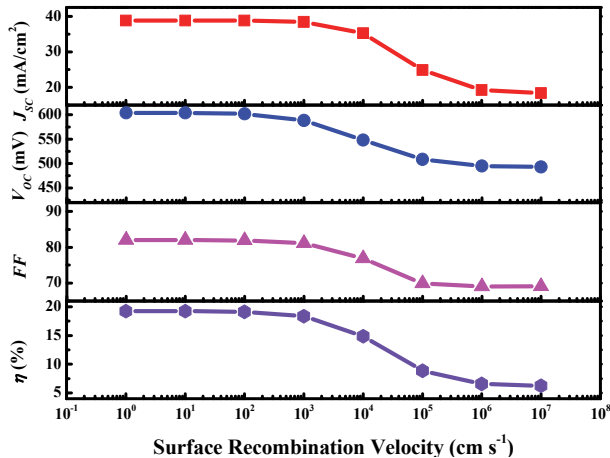

Figure 13.8. Effects of SRV on photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cell.

The poorer the surface quality, the larger the value of SRV. The net surface recombination rate R_{surf} increases with the incremental surface recombination velocity.

$$R_{surf} = \frac{np - n_{ie}^{2}}{\tau_p^{eff}\left(n + n_{ie}\exp\left(\dfrac{ETRAP}{k_B T_L}\right)\right) + \tau_n^{eff}\left(p + n_{ie}\exp\left(\dfrac{-ETRAP}{k_B T_L}\right)\right)}$$

$$\frac{1}{\tau_n^{eff}} = \frac{1}{\tau_n^{i}} + \frac{d_i}{A_i}\text{S.N}$$

$$\frac{1}{\tau_p^{eff}} = \frac{1}{\tau_p^{i}} + \frac{d_i}{A_i}\text{S.P}$$

τ_n^{i} is the bulk lifetime of electrons calculated at node i along the interface. τ_n^{eff} is the effective lifetime for electrons. The d_i and A_i parameters are the length and area of the interface for node i. P in subscript represents the similar parameters for holes. n_{ie} is intrinsic carrier concentration, ETRAP is trap level energy, k_B is Boltzmann constant and T_L is lattice temperature. The S.N and S.P parameters are the recombination velocities for electrons and holes respectively. The increase in recombination reduces collection efficiency of photo carriers and as a result J_{SC} decreases. Dark saturation current density also increases with increase in surface recombination velocity

$$J_0 \cong \frac{q n_i^2}{N_A}\text{S.P}$$

$$V_{OC} = \frac{k_B T}{q}\left[\ln\left(\frac{J_{SC}}{J_0}\right) + 1\right]$$

So V_{OC} decreases with increase in surface recombination velocity. As V_{OC} and J_{SC} decreases with increase in surface recombination velocity, η decreases. As a consequence, to obtain higher conversion efficiency, an effective passivation of surface defects and cleaning of the surface for β-FeSi$_2$ film should be carried out to reduce the surface recombination velocity. As far as the back surface recombination velocity of the cell is concerned, its influence is found to be little for the cell because there are

few photo generated minority carriers at back side of the cell as incident light is almost completely absorbed by β-FeSi$_2$ layer.

Effects of interface traps

Due to lattice mismatch between semiconductor materials, dangling bonds are present at semiconductor interfaces. These dangling bonds acts as defects or traps sites and these may influence electrical characteristics of devices. Energy of traps centers lies in forbidden gap and exchange charge with the conduction and valence bands through the emission and capture of electrons. Space charge density in bulk semiconductor and recombination rates is influenced by the presence of trap states. Charges due to interface traps and bulk traps are added to the space charge term in Poisson's equation. Standard SRH recombination term is modified to consider trap effects as follows:

$$R = R_D + R_A$$

$$R_D = \frac{np - n_{ie}^2}{Taun\left[p + DEGEN.FAC.n_{ie}\exp\left(\dfrac{E_i - E_t}{k_B T_L}\right)\right] + Taup\left[n + \dfrac{1}{DEGEN.FAC}.n_{ie}\exp\left(\dfrac{E_t - E_i}{k_B T_L}\right)\right]}$$

$$R_A = \frac{np - n_{ie}^2}{Taun\left[p + \dfrac{1}{DEGEN.FAC}.n_{ie}\exp\left(\dfrac{E_i - E_t}{k_B T_L}\right)\right] + Taup\left[n + DEGEN.FAC.n_{ie}\exp\left(\dfrac{E_t - E_i}{k_B T_L}\right)\right]}$$

where R is total recombination rate term of carrier continuity equation, R_D and R_A are recombination rate term for donor like states and acceptor like states respectively. n, p, n_{ie} are electron hole and intrinsic carrier concentration. E_i and E_t are intrinsic Fermi level energy and trap energy level. τ_n and τ_p are carrier life times for electrons and holes, related to capture cross section of carriers.
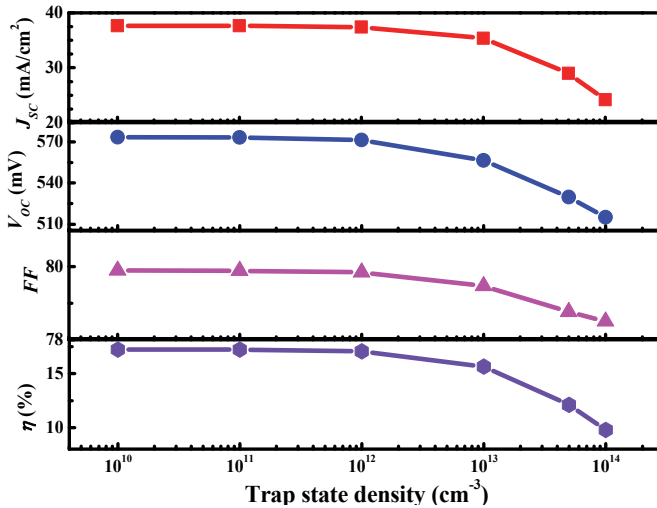
Figure 13.9. Effects of trap state density on photovoltaic properties of p-β-FeSi$_2$/n-Si heterojunction solar cells.

In the simulation of this section, all the parameters used are the same as previous section. Front SRV is set at 2000 cm/sec. Two level traps is used to model interface traps between β-FeSi$_2$/Si, which is discussed in section 2 and density of acceptor like traps is considered equal to density donor like traps for simulation purpose. Figure 13.9 depicts that V_{OC} and J_{SC} is reduced by significantly from 571.42 mV to 515.02 mV and 37.41 to 24.17 mA/cm^2, respectively, with increase in interface trap density of states from 1×10^{12} per cm^3 to 1×10^{14} per cm^3. Overall η decreases from 17.25 to 9.77 % for interface trap density variation from 1×10^{10} per cm^3 to 1×10^{14} per cm^3. The results indicate that interface states have also large impacts on solar cell efficiency.

These traps sites acts as recombination sites in solar cell. Due to increase in trap density the carrier life times (τ_n and τ_p) decreases, and the recombination parameter increases, which is responsible for decrease in J_{SC} with increase in trap density. Dark current also increases with defects as discussed in previous section, which is responsible for decrease in V_{OC} with increase in trap density. As J_{SC} and V_{OC} decreases FF and η decreases with increase in trap density.

Effects of bulk traps in FeSi$_2$

Both amorphous and crystalline form of FeSi$_2$ have high photo absorption coefficient, so it is important to investigate the effects of defects in amorphous FeSi$_2$ solar cell. Figure 13.10 shows the density of state distributions of FeSi$_2$.
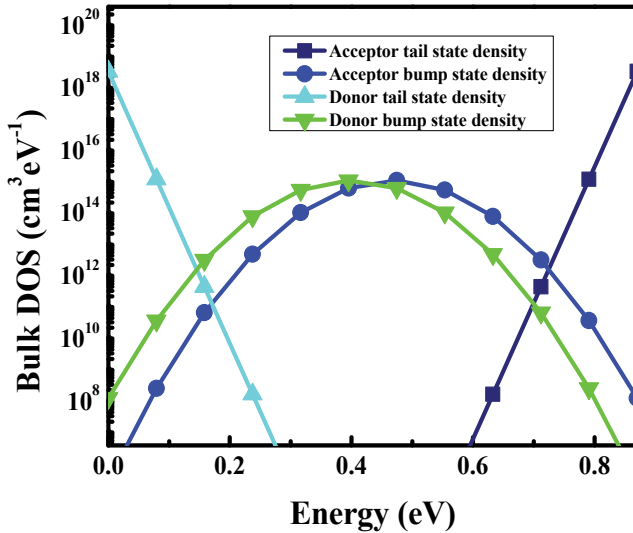

Figure 13.10. Density of state distributions of FeSi$_2$.

The distributed defects are modeled similar to amorphous Si, and defects parameters are used from Table 13.2. All the parameters in simulation are considered same as previous section and interface trap density of donor and acceptor like states are set at 1×10^{13} per cm^3. Effects of defects are shown in Figure 13.11. The decrease in J_{SC}, V_{OC}, FF and η while considering bulk trap in FeSi$_2$ can be explained by the similar type of reasons as disued in previous section.
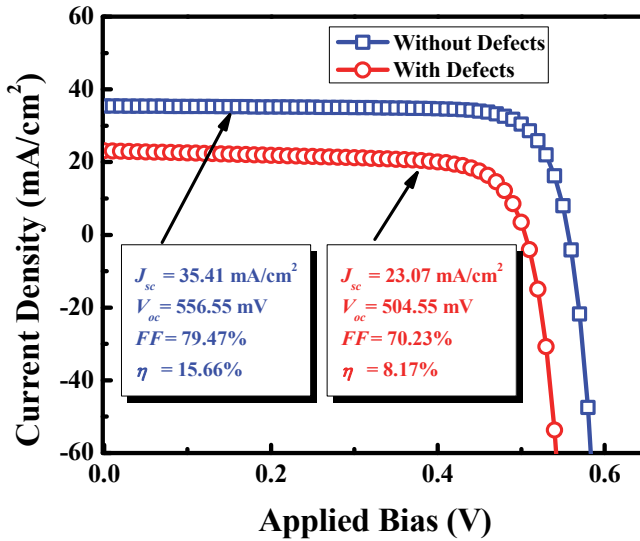
Figure 13.11. A typical solar cell *J–V* characteristics showing the effects of bulk trap.

Summary

In this chapter, numerical simulation has been taken up to study the effects of various solar cell parameters such as surface quality, defects and traps, interface states between β-FeSi$_2$ film and crystalline silicon, and work function using SILVACO ATLAS. Effects of electron affinity of β-FeSi$_2$ and different defects are also investigated. It is shown that the work function of ITO boosts the efficiency of heterojunction solar cells. It is shown that the electron affinity of β-FeSi$_2$ and valence band offset influence the performances of p-β-FeSi$_2$/n-Si heterojunction solar cells. A spike in the conduction band offset for p-β-FeSi$_2$/n-Si heterojunction solar cells causes a degradation of short circuit current. The simulation results show that large surface recombination velocity for p-β-FeSi$_2$ films and high interface states between the p-β-FeSi$_2$ film and n-Si can lead to low conversion efficiency for the cell.

Chapter 14

SPICE Parameter Extraction

Current TCAD process and device simulation tools enable one to predict the effects of process changes on device performances. Changes in process parameters and mask variation may be described using the Virtual Wafer Fabrication (VWF) framework. The ATHENA process simulator, the ATLAS device simulator and the UTMOST SPICE parameter extractor are included in this framework. The SPICE models extracted by UTMOST may be imported into SPAYN for a detailed statistical analysis and model parameter sets corresponding to various process corner are derived. SPAYN, in turn, can be linked to circuit simulator SMARTSPICE. This allows circuit performance to be analyzed using the corner models so that worst-case circuit performance for any particular circuit application can be studied.

In order to successfully design chips with higher integration and higher transmission speed, the work of modeling engineers to develop accurate device models up to tens of Gigahertz becomes more and more challenging. An absolute prerequisite for achieving this goal are accurate measurements, checked for data consistency, accurate instrument calibration, and correct de-embedding. For a general transistor modeling as an example, the measurement of the DC performance with respect to its input and output characteristics as well as its transfer function is done first, followed by the so-called CV modeling, i.e., the characterization of the depletion capacitances at 1MHz. Finally, the S-parameters of the transistor as well as of the dummy devices (contact pad capacitances and inductances) are measured. The dummy devices consist of an OPEN dummy (representing the contact pads and the open connection lines to the device-under-test (DUT), a SHORT (the device is replaced by a metal plane, thus shorting both ports to ground at the location of the device), and a THRU (the device is replaced by a strip-line between

port1 and port2 of the network analyzer). After the de-embedding process has been verified (by modeling the THRU dummy device for example), the S-parameters of the inner device-under-test are de-embedded (OPEN and SHORT dummy de-embedding) and the device modeling can be applied. However, one should keep in mind that the HF application of a device is not necessarily linear. While a S-parameter measurements by a network analyzer is usually performed at HF-signal levels below −30dBm (<1μW), and while therefore modeling is based on these data, it must be checked if the model can also predict correctly the device performance above these low HF power levels.

All aspects of semiconductor device design, process technology development, and compact model generation can be illustrated by using Technology CAD [2.2]. TCAD tools not only provide modeling of real processes, but allow predictive simulation during technology development. A robust circuit design requires knowledge about the variation in key device parameters and TCAD is finding applications in the generation of statistical compact models, which are useful for yield and performance improvement. The use of professional tools is a prerequisite to develop accurate device models and simulation circuits, from DC to GHz, including noise modeling and nonlinear HF effects, within a reasonable time.

This extraction process is usually a combination of direct parameter determination out of the measured data, followed interactively by parameter fine-tuning with an optimizer. For a given device, an adequate model is selected first. This can be a single model (like for transistors), or a composed sub-circuit consisting of standard devices. Next, the model equations, which are solved for the model parameters during model parameter extraction, give a clear indication about what kind of measurements and what type of stimulus sweeps are required for characterization. Then, during the parameter extraction process, more and more model parameters are extracted, and the selected model will fit more and more precisely the measured device. The purpose of this chapter is to demonstrate the method of SPICE parameter extraction using ICCAP. The method of extraction is discussed.

14.1 UTMOST

The most important test structures in an IC manufacturing process are the devices themselves. It is imperative that these devices are accurately characterized so that important information regarding device performance, model performance, and the best model parameter set for the device under test can be extracted. UTMOST is a data acquisition and parameter extraction tool [2.24] with applications in the areas of device characterization and modeling. UTMOST can accept data from process and device simulation and direct measurements on devices. UTMOST also supports the simulation of DC, transient, capacitance and s-parameter characteristics. UTMOST supports commercial device models, user-defined device models and macro models. Direct communication with the models in most popular commercial circuit simulators is also supported. The flow diagram used in UTMOST is shown in Figure 14.1a.

UTMOST includes routines for the extraction of many DC MOSFET parameters including length reduction, width reduction, threshold voltage, low-field mobility, body effect, velocity saturation, resistance, breakdown, and sub threshold slope parameters. In addition, UTMOST supports the extraction of overlap and junction capacitances from capacitance measurements. For bipolar devices there are extraction routines supplied to extract resistance, breakdown, saturation, temperature, and leakage, forward and reverse gain, early voltage, knee current, and basic Gummel-Poon parameters from DC measured characteristics. Extraction routines for cut off frequency, forward and reverse transit time, base resistance, and excess phase parameters are
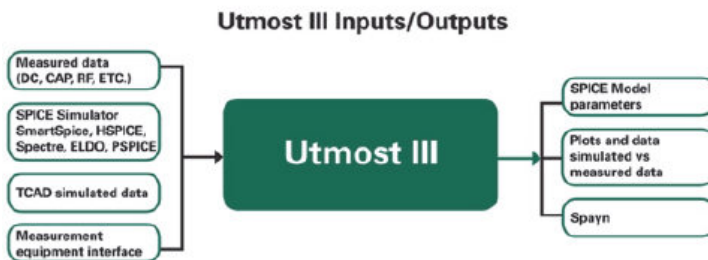
Figure 14.1a. UTMOST flow diagram. Adapted from UTMOST User's manual.
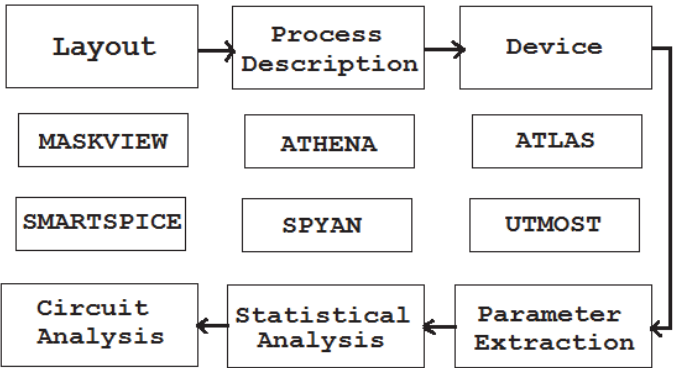
Figure 14.1b. UTMOST parameter extraction flow diagram. Adapted from UTMOST User's manual.
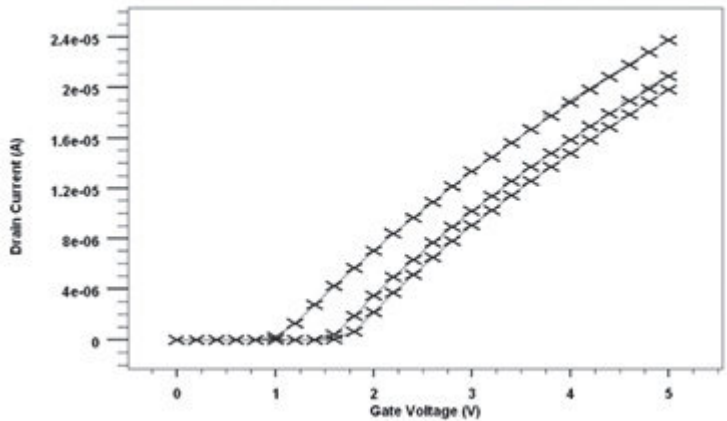

Figure 14.2. ATLAS run generates $I_d - V_g$ characteristics at three substrate biases.

implemented for s-parameter measurements. The interrelationship of UTMOST with other Silvaco tools for parameter extraction is shown in Figure 14.1b.

In the following example, we demonstrate the extraction of BSIM3 SPICE model parameters for the process-simulated MOSFET device described in Section 4.2. In fact, this example shows the integrated device design; viz., process to device simulation and SPICE model parameter extraction.
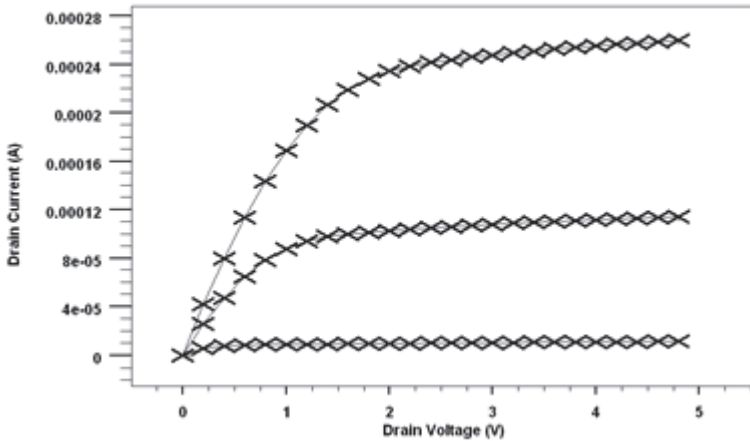
Figure 14.3. ATLAS run generated $I_d - V_d$ characteristics at three substrate biases.

The process simulation used for this example is an NMOS process flow in ATHENA. During process simulation, **extract** statements are used to extract important process parameters. For example, in terms of the SPICE model extraction it is vital to extract the gate oxide thickness in meter for use in UTMOST. The process-simulated MOSFET structure is interfaced with device simulator ATLAS. The first ATLAS run generates $I_d - V_g$ characteristics (Figure 14.2) at three substrate biases. The sequence of **solve** statements is to first save three solutions at each back bias with V_d=0.1V and V_g=0.0V. Then in turn each of these three files is loaded using **load** and the gate voltage ramped up to the supply voltage defined as 'vdsmax'. All the three characteristics are saved to a single log file. The second ATLAS run simulates $I_d - V_d$ characteristics (Figure 14.3) at three different gate voltages. A similar technique to the first run is used. Three solution files are saved at each gate bias with $V_d = 0V$. These are then loaded in turn and the drain voltage ramped to 'vdsmax'. All the characteristics are saved to a single log file.

The final stage of the example is to run UTMOST to extract the SPICE model. Important information about the structure such as gate oxide thickness and gate length is transferred to UTMOST using the results of **extract** statements in the ATHENA simulation. The log files from ATLAS are loaded and appended together. UTMOST then fits the

SPICE model to the complete $I_d - V_g - V_{bs}$ and $I_d - V_d - V_{gs}$ data sets. All the UTMOST parameters are stored to a file and then extract is used to print out the parameters of interest (Figure 14.4).

******* Extracted BSIM3 SPICE parameters from UTMOST *******

```
VGS 2 0
VDS 1 0
VSS 3 0 0
VBS 4 0    0.00E+00
M1 1 2 3 4 MMOD W=1.0000E-06 L=1.0000E-06
      NRD=4.00E-02 NRS=4.00E-02
.DC VDS 0 5.000 0.200 VGS 1.500 4.500 1.500
.PRINT DC I (VDS)
.OPTIONS TNOM= 27.00
.TEMP    27.00
.OPTIONS DCGMIN=1.00E-18 RELTOL=1.00E-03
      VNTOL=1.00E-06 ABSTOL=1.00E-12
.OPTIONS NUMDGT=5
.MODEL MMOD NMOS (LEVEL = 8
+TNOM = 27 TOX = 1E-8 XJ = 1.5E-7
+NCH = 1.7E17 NSUB = 6E16 VTH0 = 0.7973586
+K1 = 0.6170425 K2 = -0.05 K3 = 80
+K3B = 0 W0 = 2.5E-6 NLX = 1.74E-7
+DVT0W = 0 DVT1W = 0 DVT2W = -0.032
+DVT0 = 2.2 DVT1 = 0.53 DVT2 = -0.032
+VBM = -5 U0 = 393.7633007 UA = 2.153867E-9
+UB = 1E-21 UC = 5.9097E-11 VSAT = 8.2304E4
+A0 = 0.6152605 AGS = 0 B0 = 0
+B1 = 0 KETA = -0.047 A1 = 0
+A2 = 1 RDSW = 0 PRWG = 0
+PRWB = 0 WR = 1 WINT = 0
+LINT = 0 DWG = 0 DWB = 0
+VOFF = -0.15 NFACTOR = 1.5015015 CIT = 0
+CDSC = 2.4E-4 CDSCD = 0 CDSCB = 0
+ETA0 = 0.08 ETAB = -0.07 DSUB = 0.56
+PCLM = 1.6683 PDIBLC1 = 0.39 PDIBLC2 = 8.6E-3
+PDIBLCB = 0 DROUT = 0.9761121 PSCBE1 = 4.24E8
+PSCBE2 = 5E-5 PVAG = 0 DELTA = 0.01
```

```
+MOBMOD = 1 PRT = 0 UTE = -1.5
+KT1 = 0 KT1L = 0 KT2 = 0
+UA1 = 4.31E-9 UB1 = -7.61E-18 UC1 = -5.6E-11
+AT = 3.3E4 NQSMOD = 0 WL = 0
+WLN = 1 WW = 0 WWN = 1
+WWL = 0 LL = 0 LLN = 1
+LW = 0 LWN = 1 LWL = 0
+CAPMOD = 1)
.END
```

Figure 14.4. UTMOST extracted BSIM3 SPICE parameters for the n-MOSFET.

The results presented here show how worst-case SPICE model parameters can be derived by the simulation of variations in the CMOS manufacturing process. The whole process can be automated as an experiment in the Virtual Wafer Fabrication. Statistical fluctuations inherent in any IC manufacturing process cause variations in device and circuit performance. Product yield and manufacturing problems necessitate costly redesign cycles. SPAYN can be used to determine the underlying statistical nature of IC manufacturing processes and their impact on device performance. Resulting variations in circuit performance caused by process fluctuations can then be predicted early in the design cycle.

14.2 ICCAP

ICCAP is a model parameter extraction and simulation software package from Agilent [14.1]. It serves **as** a platform for measurement, simulation and data analysis. It has an interface for the user through its Main Window and Status Window. The Status Window updates the ICCAP outputs and errors/warnings for the user. ICCAP allows the user two methods of parameter extraction: direct and optimized. Direct parameter extraction involves extracting a particular parameter or set of parameters directly from the measured characteristics. Optimization may be used after direct parameter extraction, and is used to optimize the fit between the measured characteristics and the simulated characteristics produced from the extracted model parameters.

The 'Optimize' function in ICCAP is used to do the optimization by setting the 'X' and 'Y' data ranges. The parameters to be used for the optimization are also specified in the ICCAP. The optimizer iteratively solves for a set of model parameters which produce simulated data that optimally matches with the measured data. There are different algorithms used for optimizations in ICCAP. The suitable algorithm is chosen depending upon the goal of optimization and nature of the model equations involved. If the model is well behaved and local minima are not a problem, the Levenberg-Marquardt optimization is used. The Levenberg-Marquardt method is a non-linear, least squares fit algorithm. It is a combination of steepest descent and Gauss-Newton methods. It calculates the specified model parameters in each iteration until the RMS error between measured and simulated data is minimized.

The fit is optimized by varying one or more model parameters from their extracted values and checking the fit. The range over which the parameters may be varied is set by the user, as are the parameters being optimized. Additionally, the order in which parameters are optimized and possibly re-optimized, and the number of possible methods of optimization is infinite.

Additionally, the three transistor sizes required for MOS extraction using ICCAP (large, short channel, and narrow channel). Using ICCAP, a macro is written, which utilize the auto-probing capabilities of the Semi-Automated Wafer Prober. The macro controlled automated movement of the prober so that each die location on the wafer could be tested. Initial alignment of the wafer and probe placement on a reference position was required before automated probing could begin. At each location measurements are made on each of the three transistor sizes for the type of devices being measured. Once the devices are measured, the various current-voltage plots are displayed and the operator is prompted to extract SPICE parameters, save the results and optionally print the plots. The parameter extraction process required the operator to make a judgment call of whether or not the current-voltage plots represented a "good" or a "bad" device. A "bad" device consist of any characteristic which represented non-operational devices, and also included characteristics where the operation of the device was poor enough to cause it to be judged a "bad" device. Once parameter extraction is

completed using ICCAP, one can compare the simulated I-V characteristics from the device simulation.

The ICCAP Modeling System is used to measure semiconductor device and circuit modeling characteristics and analyze the resulting data. The general functional flow of a typical ICCAP system is shown in Figure 14.5. There are various simulators like Spectre, ADS, HSpice etc. that can invoked from ICCAP during simulations. Once the simulation is done using the model parameters and the appropriate input conditions the next step is to optimize the response so as to match with the measured response. The optimization procedure followed by ICCAP is as shown in Figure 14.6. Random optimization is used in the initial phase before a Levenberg-Marquardt optimization to reduce the error to an acceptable limit. Random optimization makes random guesses of the parameter values until the specified RMS error value is obtained.

To use ICCAP, one needs:

- A workstation
- Instruments that perform DC, Capacitance, AC, and Time-Domain measurements
- A test fixture
- A test device
- The ICCAP software

The main window provides the graphical interface for the user. ICCAP is utilized for the following tasks. For using the ICCAP Modeling System for modeling purposes, one needs to follow:

- Start the program and load the model file into memory.
- Measure the device characteristics.
- View the measured data.
- Perform a simulation.
- Extract the model parameters from the measured data.
- Simulate with extracted parameters.
- Compare measured data and simulated data.
- Optimize model parameters as needed.
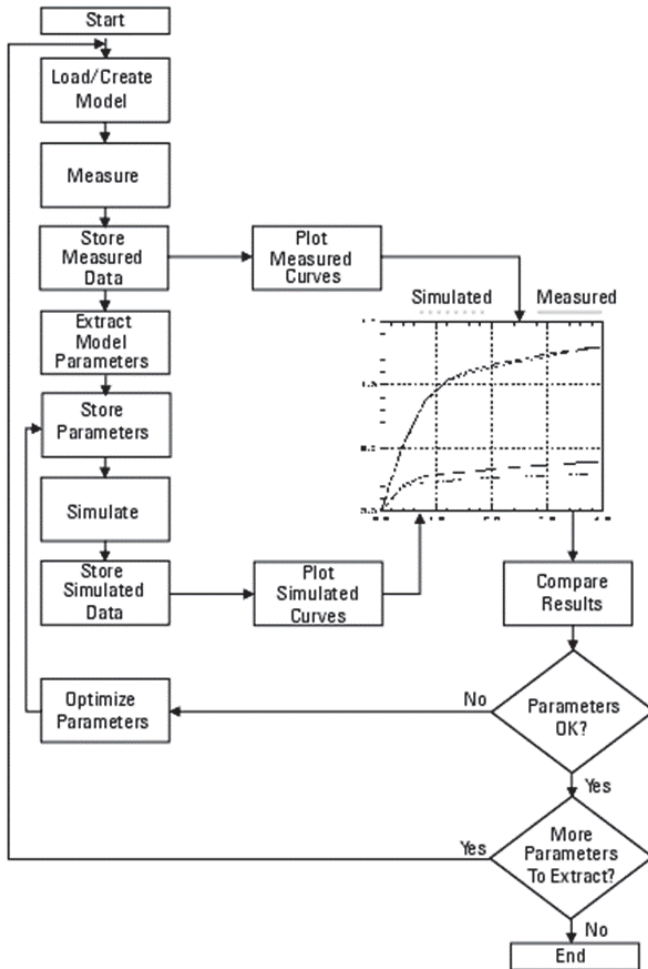
Figure 14.5. General functionality of ICCAP system. After A. T. Karingada, Estimation of thermal impedance parameters of silicon germanium heterojunction bipolar transistors, MS Thesis, The University of Texas at Arlington, 2011.
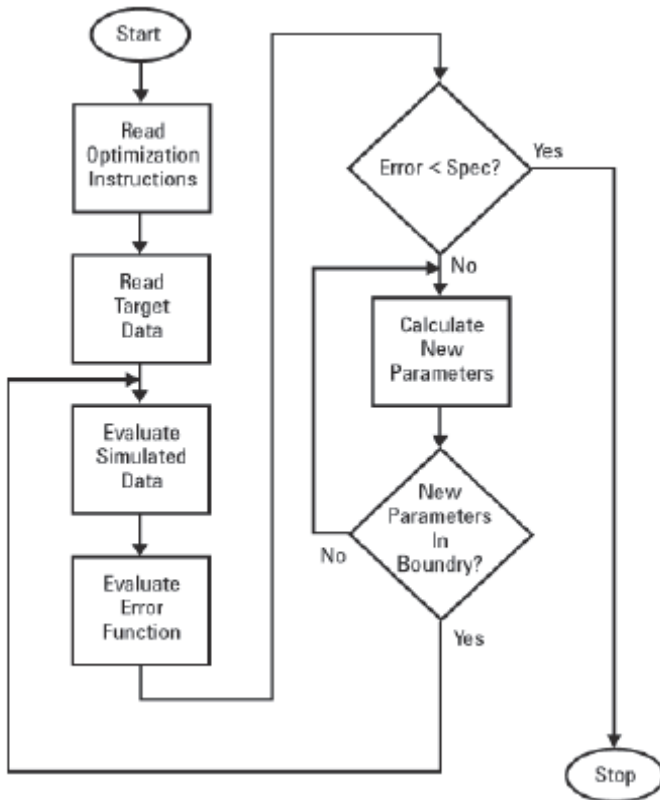
Figure 14.6. Optimization flow diagram. After A. T. Karingada, Estimation of thermal impedance parameters of silicon germanium heterojunction bipolar transistors, MS Thesis, The University of Texas at Arlington, 2011.

Controlling Measurement:

To measure the device characteristics, one can use a semiconductor parameter analyzer such as HP 4145B. It is connected to a computer using an interface card. It is connected for ICCAP by selecting it in the hardware list prior to using it for measurement. Proper configuration is applied for the SMUs of the HP 4145B.The ICCAP's "Device under Test" set-up is used for specifying the voltages and currents to be applied at a particular terminal of the device. After specifying the compliance, the measurement was carried out by ICCAP commands.

Simulation for the particular device set-up:

ICCAP provides the facility for selecting the simulator. ICCAP supplies a variety of model extractions in model files that one can load and work with immediately. Proper links and paths should be provided to activate the simulator. The "Circuit" field of ICCAP accepts the model parameter for the device. This should have the format required by the respective simulator. After the proper circuit has been described, the "Model Parameter" field is updated with the parameter listed in the "Circuit" field. The simulator uses these parameters while simulating. Simulation is performed by the commands of ICCAP.

Data analysis using graphs:

ICCAP provides the graphical presentation for measured as well as simulated data. It also allows mathematical operations on the data and simulation while presenting in graphical format. It has various options which enriches the graphical presentations. Multiple parameters can be presented in same graphs with various options.

Optimizing using ICCAP tools:

Various mathematical functions and program s are available for data manipulation. ICCAP provides the "Macro" and other mathematical functions for curve fitting. The "Optimization" option compares the measured data and simulated data. It adjusts the model parameters to give minimum specified error between them. Various options are available while using the optimization function. This facility was widely used for the thermal resistance measurement and the optimization procedure.

In the following, we provide an example of a typical ICCAP session using a supplied model file. Associated Figures (screenshots: a-t) show the procedures need to be followed when one plan to extract SPICE parameters starting from real device measurement to optimization.
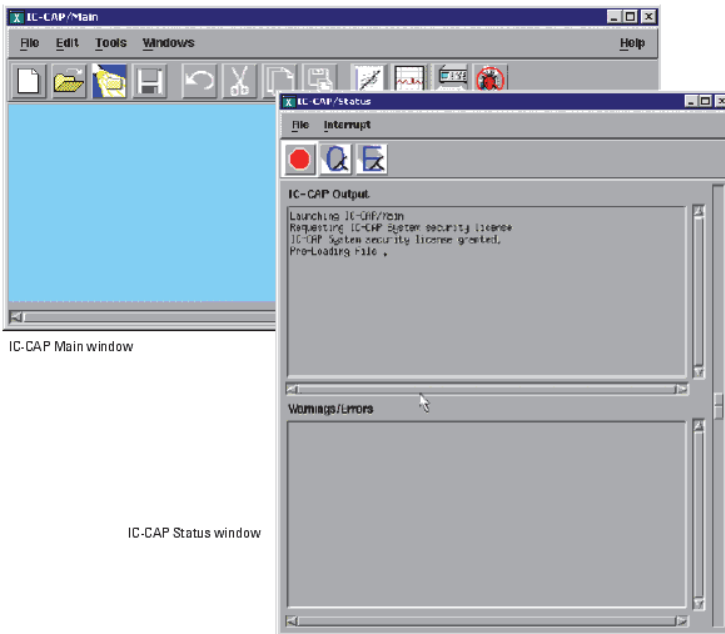
IC-CAP Main window

IC-CAP Status window

- In the ICCAP/Main window, click Examples on the toolbar.
- The directory dialog box displays the directories of examples. Double-click model files to see the models that are provided with the program. Double-click the MOSFET directory. Then double-click nmos2.mdl

An icon of the model file displays in the ICCAP/Main window and the Model window opens. The nmos2.mdl example extracts parameters for the Level 2 model (N-channel UCB MOSFET transistor model).
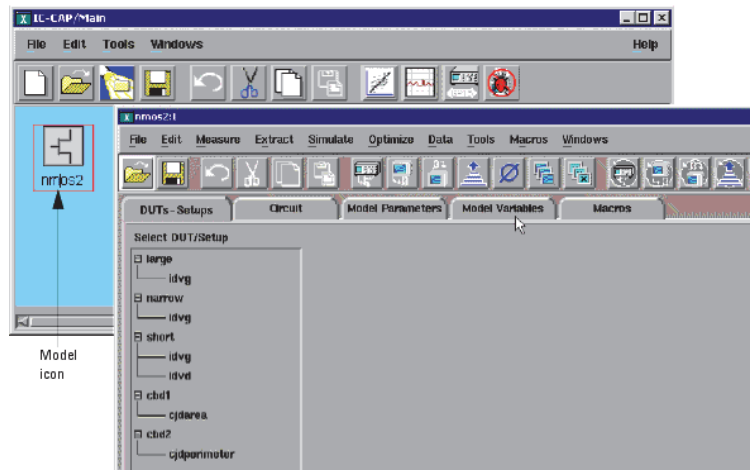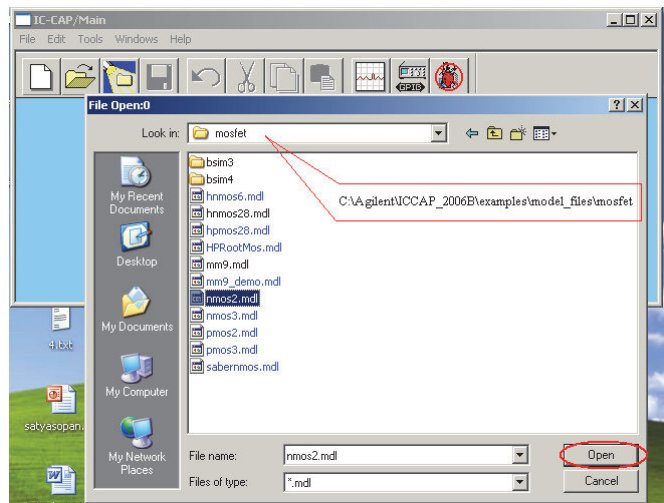
Viewing the Circuit Description

The Circuit description defines the device topology and parameters to be extracted. To view the default values of the parameters, select Circuit.

The circuit definition begins with the Options definition (.OPTION). The definition lists the device to be measured, the device nodes, the name of the model card that contains the model parameters used by the device, and the device parameters.

Circuit Options definition

```
┌─Device ID              ┌─Name of model card
    Device nodes                     Device parameters

M1 1=D 2=G 3=S 4=B MOSMOD L=2u W=3u AD=9p PD=12u AS=9p PS=12u
```
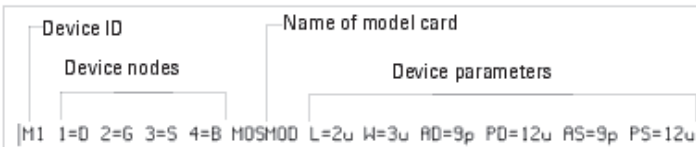
The next section is the model card definition (.MODEL). The definition lists the name of the model card, the model type, and the contents of the model card. The contents of the model card consists of the parameters that are to be extracted (model parameters) and their assigned initial values.

Circuit Model definition

```
                                            ┌─Model card name
                                              ┌─Model type
Begin model definition ────    .MODEL MOSMOD NMOS
                            ┌─ + LEVEL        = 2
                               + UO           = 600
                               + VTO          = 0
                               + NFS          = 0
                               + TOX          = 100n
Model parameters               + NSUB         = 1e15
                               + UCRIT        = 10.00K
                               + UEXP         = 0
                               + VMAX         = 1MEG
                               + RS           = 0
                            └─ + RD           = 0
```
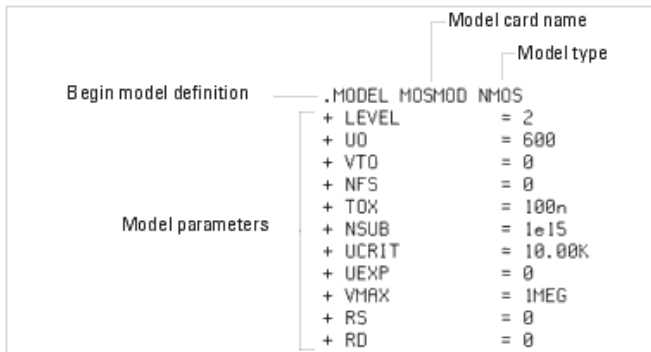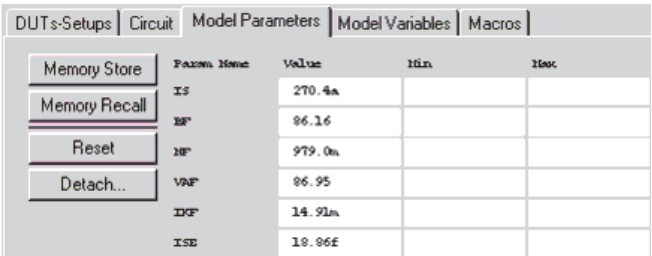
The Model Parameters folder displays the current parameter values. These parameters are common to all DUTs for that model and are used in a simulation. When a new model is read, these values are set to the defaults defined in the circuit. The values change after extractions and optimization are performed.

To define the model parameters, select Model Parameters.

Click Model Parameters
to view current parameter values

DUTs–Setups	Circuit	Model Parameters	Model Variables	Macros
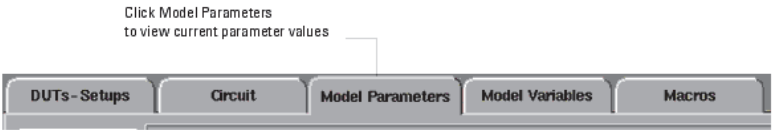
The Parameters table contains the same names as those specified in the MODEL definition of the Circuit. However, by editing the table, the values can be set independently of those in the circuit definition and can be specified as a real number or an equation. In addition, one can set limits by entering minimum and maximum values for each parameter. Parameter values that are outside their limits are clamped to their minimum or maximum values.

DUTs-Setups | Circuit | Model Parameters | Model Variables | Macros

Control value set selection

	Param Name	Value	Min	Max
Memory Store	IS	270.4a		
Memory Recall	BF	96.16		
Reset	NF	979.0m		
Detach...	VAF	96.95		
	IKF	14.91m		
	ISE	18.86f		

After editing the defaults, one can store the current set of parameter values in a temporary buffer and recall these values at any time during the ICCAP session (Memory Recall). If one prefers, one can restore the default values (Reset).

Defining Model Variables

Variables can be defined at the top, or system, level or at the Model, DUT and Setup levels. One can define as many variables as one needs and use these variables for other functions, such as extractions. Model variables can be accessed by any of the components of a model, including the DUTs and setups, while DUT and Setup variables (such as inputs or outputs) can be accessed only by the components of that setup. A useful example of a variable defined at the model level is the compliance limit for measurements. Typically, compliance values are defined for each input in a setup. One can refer all of the compliance values in a model to one variable defined at the model level by creating a compliance variable for the model. This compliance variable can be

modified to change the compliance values for all inputs in the model. The Model Variables folder contains the names and values of all variables that are global to the model. To view the model variables, select Model Variables.

Click Model Variables to view current values

System Variables	Name	Value
Detach...	WD	512.6n
Print	AreaCap1	10n
	PerimCap1	400u
	AreaCap2	300p
	PerimCap2	206u
	POLARITY	NMOS
	OPEN_RES	1.0E12

DUTs-Setups | Circuit | Model Parameters | Model Variables

Add new variable in empty field
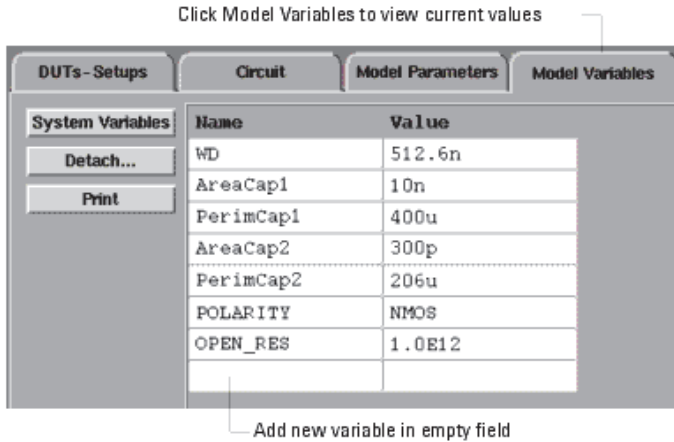
To define a new variable:

- Type the variable name in the first empty Name field row.
- Press Tab to move the cursor to the Value field. Type a value for the variable.
- Press Enter to accept the value and open a new row.

ICCAP models usually contain several DUTs. DUTs contain groups of setups that have a similar physical connection to the device. Each DUT contains its own DUT parameter set and test circuit. If two setups require differences in either of these areas, one must define a different DUT for each. Before one can measure a device, one must make any needed changes to the DUT options. The DUTs-Setups folder displays the DUTs and setups in the model. Three folders are available for defining DUT options. To view the DUT options and make a device active, select DUTs-Setups. Then select the DUT. If necessary, display the list by clicking the arrow button.

Click DUTs-Setups to select the DUT

Click the arrow
to see the setups
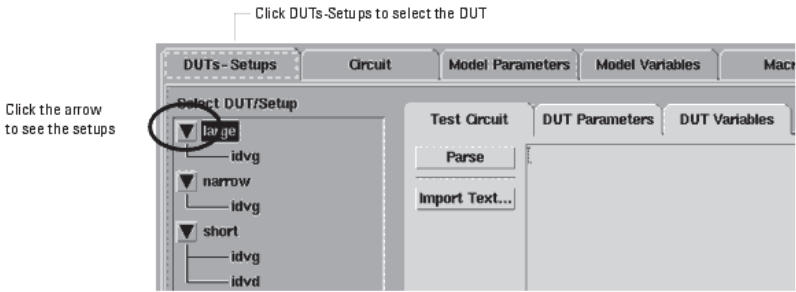
One can edit the values of parameters that change from one DUT to another (such as, channel length and width for a MOSFET). In addition, one can set limits by defining minimum and maximum values for each parameter. If a parameter value is outside its limits, it will be clamped to its minimum or maximum value. After editing the values, one can temporarily store the current set of parameter values (Memory Store), recall stored parameter values (Memory Recall), apply the changes to the circuit definition (Update Circuit), and reset the default values specified in the circuit definition (Reset). To view the DUT parameters, select DUT Parameters.

Click DUT Parameters to view current values

For frequent editing,
detach table
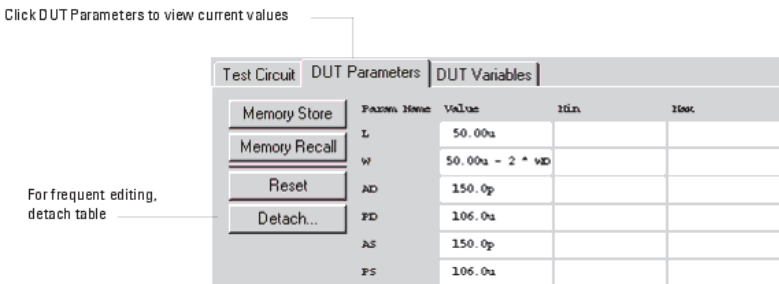
The DUT Variables folder contains the names and values of all variables that are global to the DUT. To view the DUT variables, select DUT Variables

Click DUT Variables to view current values

Click to view system variables

Add new variable in empty field
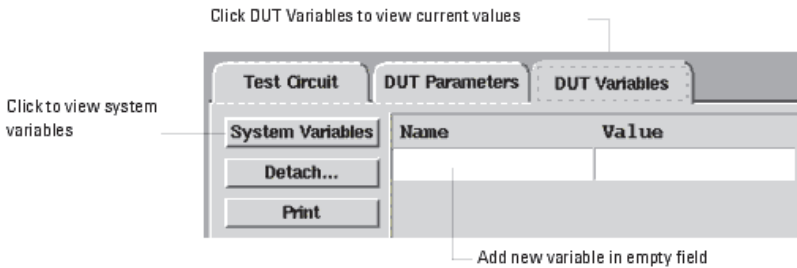
One can define setup options from the DUTs-Setups panel. A setup contains definitions for inputs, outputs, instrument options, measuring and simulating, and results. To view and edit a setup, select a setup from the DUTs-Setups panel.
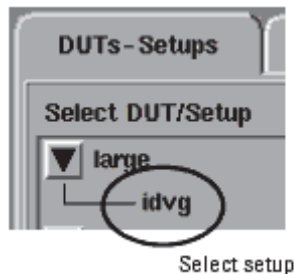
Select setup

The Measure/Simulate folder opens so one can define the inputs and outputs for a particular setup.
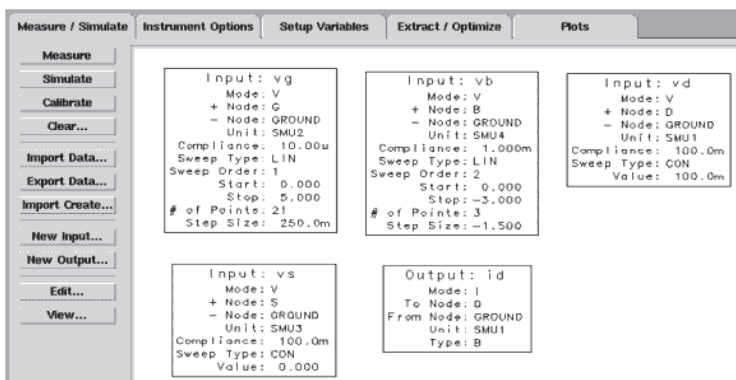
Setup input and output data are listed in a table (sometimes called a tile). One can edit input and output in one of two ways:

- Edit directly in the table.
- Click or double-click a field and type the new data.

Edit directly in the table

Input: vs
Mode: V
+ Node: S
− Node: GROUND
Unit: SMU3
Compliance: 100.0m
Sweep Type: CON
Value: 0.000

Use the Input or Output Editor. To open the Input or Output Editor, first select the table and then click Edit. Edit the fields and click OK.

Edit...   Click to open the Input or Output Editor

Input Editor:1

Input: vg

Mode: Voltage   Click arrow for list of values

Voltage
Current
Frequency
Time
Parameter
User

+ Node: G

− Node: GR

Unit: SM

Compliance: 10

Sweep Type: Linear

Sweep Order: 2

Start: 0.000

Stop: 5.000

# of Points: 21

Step Size: 250.0m

OK   Cancel   Help

Output parameters can be derived from measured data, simulated data, or both. The Setup Variables folder contains the names and values of all variables that are global to the Setup.

To view the Setup variables, select Setup Variables.

Click Setup Variables to view current values

Click to view system variables

Measure / Simulate	Instrument Options	Setup Variables	Ext

System Variables	Name	Value
Detach...		
Print		

Add new variable in empty field
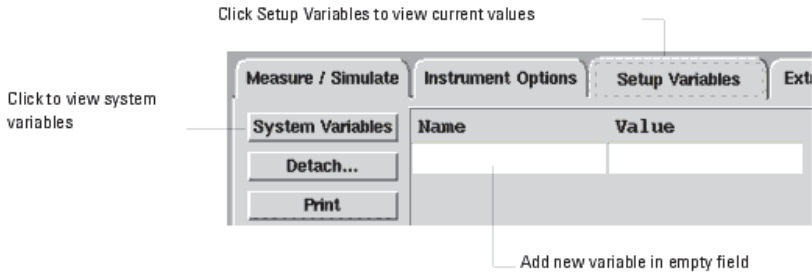
This section describes the measurement procedure in detail. The measurement on the devices provides sufficient information for extracting a full set of model parameters. To avoid errors, measuring devices, and extracting parameters must be performed in a specified order as follows:

- Log-in to the computer and start ICCAP. Two windows will appear on screen, status window and main interface window.

- Turn on the HP 4145B. It will follow the self-calibration cycle. After its calibration is over, it will indicate "0" in its display. Now, it is ready for interface with ICCAP.

- Using the ICCAP main window, include HP 4145B into its Hardware set-up list. After ICCAP recognizes the HP 4145B in its hardware list, configure its Source/Monitor units.

Now it is ready for measurements.

- Open the ICCAP model file for the transistor under measurement. Verify that it has the correct circuit description for the selected simulator. Open the appropriate "Device under Test"

set-up. Specify the voltage/current to be applied at the specific terminal of the device.

- With appropriate model parameter, simulate the set-up and observe if the simulation result is according to the expectation. If it is, the measurement can be performed.

- Select the designated transistor. From the map of the IC, find out its location on the wafer. Place the wafer on the probe station. With the help of the microscope, connect the probes to the appropriate pads on the IC. Verify that the probe makes proper contact with the pads.

- Turn off the microscope light and the light in the measurement room to minimize photo-generation.

- The vibration produced by the walking in the room can create the misalignment of the probe. Hence, the vibrations must be minimized by careful behavior in the measurement room.

- Check the compliance specified on the measurement set-up and verifies if they are correct.
- After the set-up is ready, measure the set-up using the "Measure" command of ICCAP.

- ICCAP will indicate, "Measure in Progress" till the measurement is finished.

- Observe the graph for the current-voltage relationship. If the noise is observed at the lower voltage/current range, change the integral time in the measurement set-up to a longer time period to minimize the noise effect.

- Repeat the measurement until the noise is within the tolerable range.

- After the proper data have been taken, save the file.

- Perform the measurement using all the five specified set-ups.

- Now, the data are ready for the optimization. After optimization of all the parameters is over, the result is ready for the analysis.

- After finishing the measurement, disconnect the probes from the transistor pads; turn off the ICCAP section and the HP 4145B.

- One measurement cycle is completed.

• To initiate the measurement using the active setup, select the setup and click Measure Setup.

• To initiate the measurement using all setups in the active DUT, select the DUT and click Measure DUT.

One can determine the accuracy of current parameter values by performing a simulation and comparing the results to the measured data.

• To perform a simulation on the active setup, select the setup and click Simulate Setup.

• To perform a simulation on all setups in the active DUT, select the DUT and click Simulate DUT.

14.3 Extracting Model Parameters

Agreement between two sets of data is not always good prior to extraction. ICCAP modules provide standard device extractions, as well as macros for performing an extraction. If the model one is using does not have macros available, one can set the criteria in the Transforms folder. One can choose also from a large variety of functions to use for extracting model parameters in a Transform.

To view or edit transforms:

- Select a setup.
- In the Extract/Optimize folder, choose the extract transform.
- Click Browse to view the Function list.

- Select a Function from the list.

The initial values of the process-dependent parameters for the extraction are entered using a macro. To run a macro:

In the series of prompts, enter the initial values.

In the example, the initial values are:
TOX = 40n
XJ = 0.2u
LD = 0.4U
RS = RD = 1 0

- In the Model folder, select Macros.
- In the Select Macro list, choose init_parameters and click Execute.
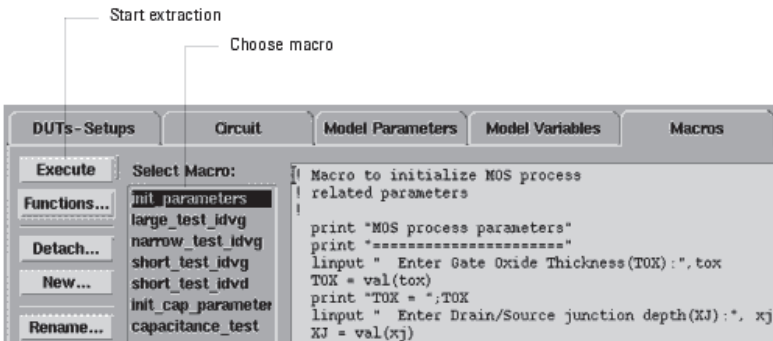
One can extract all extraction transforms in the active setup or extract all setups in the active DUT.

   • To perform an extraction on all extraction transforms in the active setup, select the setup and click Extract Setup.
   • To perform an extraction on all extraction transforms in all setups in the active DUT, select the DUT and click Extract DUT.

When one extracts all setups in the active DUT, the extractions are performed in the left-to-right order listed in the setup. This order is usually critical to proper extraction performance. Typically, extractions are completed instantly. The newly extracted model parameter values are listed in the ICCAP Status window and are placed in Model Parameters.

```
MOSDC_lev2_lin_large Extraction Results:
        VTO = -862.7m
        NSUB =  38.49T
        UO =  1.056MEG
        UEXP =  2.289
        VMAX =  1.000MEG
        NFS =  1.000T
```

After extracting the model parameters, one can perform a simulation with the extracted parameters and compare the results to the measured data.

- To perform a simulation on the active setup, select the setup and click Simulate Setup.
- To perform a simulation on all setups in the active DUT, select the DUT and click Simulate DUT.

Repeat the simulation procedure for each DUT in the model. One can observe the differences in output that result from changes to various model parameters. For example, change the value of one or more parameters. Then run a simulation and view the changes in the plot.

14.4 Optimizing the Model Parameters

To achieve greater accuracy, one can optimize the model parameters. This optional step obtains the best possible fit between measured and simulated data by altering the parameter values iteratively until the difference between the data sets is minimized. Since repeated simulations are required, optimization is more time-consuming than an extraction.

To perform an optimization:

- Select a setup.
- In the Extract/Optimize folder, select the optimize transform.
- Define the Inputs, Parameters, and Options for the selected transform.
- Execute the optimization:
- To perform an optimization on the active setup, select the setup and click Optimize Setup.
- To perform an optimization on all setups in the active DUT, select the DUT and click Optimize DUT.

One can view the results of the measured data for a setup by selecting the Plots folder in the Setup window. The Plot Finder list contains all of the plots available for a Setup. If there are multiple plots, select the desired one to view. One can view multiple plots by pressing the Ctrl key while selecting the plot names in the list. To see the plot, click Display Plot or Display All. The ICCAP/Plot window opens.

Figures 14.7 and 14.8 show examples of optimization results obtained. The comparison between simulation and measurement: drain-source current vs. drain-source voltage is shown. The BSIM3 model parameters for devices are extracted using ICCAP. The simulation results are compared with the experimental data as shown in Figures 14.7 and Figure 14.8. Good agreement between the model predictions and measurement for transistors is observed.

Length Scaling: group device extraction (I_d vs. V_d @ V_{bs} = 0V at Different V_g).
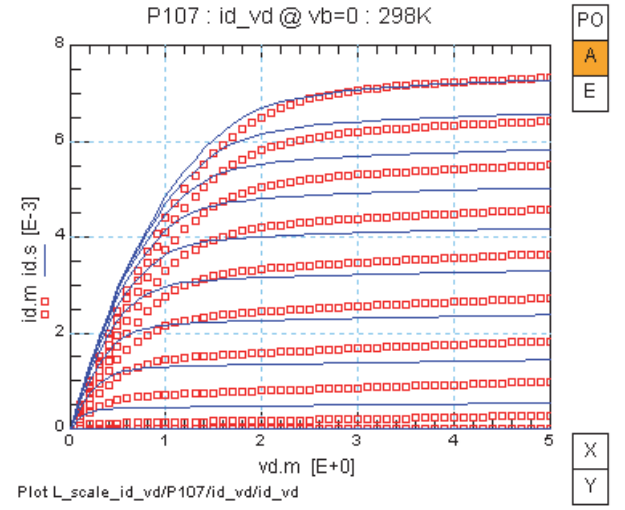
Figure 14.7. Group-Fitting of I_d vs. V_d @ V_{bs} = 0V for different V_g for one set of devices (Large and Fixed W and Different L). Fixed W = 20μm. L: (a) 1.25μm, (b) 1.0μm, (c) 0.875μm and (d) 0.75μm, respectively.
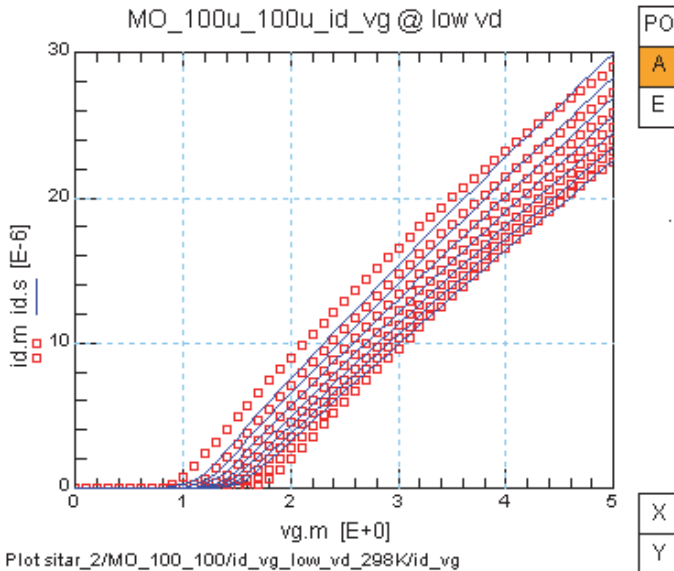
Figure 14.8. Fitting of I_d vs. V_g @ V_d = 0.1V at different V_b of Large Size Device [Large W and L: (W × L) = (100 μm × 100 μm)].

Table 14.1. Locally Extracted Parameters**.**

Parameter	Extracted Value	Unit
VTH0	0.7529	V
K1	0.0469	$V^{1/2}$
K2	-918.3×10^{-15}	$-$
U0	687.5	cm^2/Vs
UA	1.4×10^{-9}	m/V
UB	4.289×10^{-18}	$(m/V)^2$
UC	63.97×10^{-12}	V^{-1}

Summary

As modern circuits are usually very complex, the performance of such circuits is difficult to predict without accurate device models. The focus of this chapter is on device compact modeling methodologies and SPICE parameter extraction. An N-Well CMOS was process simulated and BSIM3 model parameters were extracted using UTMOST. Prediction of statistical circuit performance variation is becoming critical for 22nm technology node and beyond. The predictive capability of TCAD to estimate the worst-case SPICE models for a process, based on the physical variations of process parameters has been examined. ICCAP parameter extraction procedures are discussed in detail. The integrated circuit design tools, such as the Cadence, H-SPICE, Agilent Advanced Design Systems (ADS), may be used for compact models to predict the behavior of a design.

Bibliography

1.1. Maiti, C. K. and Armstrong, G. A. (2001) *Applications of Silicon-Germanium Heterostructure Devices* (Institute of Physics Publishing (IOP), UK).

1.2. Maiti, C. K. Chakrabarti, N. B. and Ray, S. K. (2001) *Silicon Heterostructures: Materials and Devices* (Institute of Electrical Engineers (IEE), UK).

1.3. Maiti, C. K. Chattopadhyay, S. and Bera, L. K. (2007) *Strained-Si Heterostructure Field-Effect Devices* (CRC Press (Taylor & Francis), USA).

1.4. Maiti, C. K. Editor, (2008) *Selected Works of Professor Herbert Kroemer* (World Scientific, Singapore).

2.1. Dutton, R. W. and Yu, Z. (1993) *Technology CAD Computer Simulation of IC Processes and Devices* (Kluwer Academic Publishers, USA).

2.2. Armstrong, G. A. and Maiti, C. K. (2008) *TCAD for Si, SiGe and GaAs Integrated Circuits* (The Institution of Engineering and Technology (IET), UK).

2.3. Synopsys Inc, Sentaurus Workbench User Guide, September 2011.

2.4. Synopsys Inc, Sentaurus Structure Editor User Guide, September 2014.

2.5. Synopsys Inc, Sentaurus Mesh User Guide, September 2014.

2.6. Synopsys Inc, Sentaurus Device User Guide, September 2011.

2.7. Synopsys Inc, Sentaurus Process User Guide, September 2011.

2.8. Synopsys Inc, Sentaurus Visual User Guide, September 2014.

2.9. Synopsys Inc, Inspect User Guide, December 2010.

2.10. Synopsys Inc, TecPlot SV User Guide, September 2011.

2.11. TSuprem4 User Manuals, Synopsys Inc., Mountain View, Calif., 2006.

2.12. Taurus Process User Manuals, Synopsys Inc., Mountain View, Calif., 2006.

2.13. Medici User Manuals, Synopsys Inc., Mountain View, Calif., 2006.

2.14. Silvaco Inc, VWF Interactive Tools, User's Manual, 2009.

2.15. Silvaco Inc, DeckBuild User's Manual, 2013.

2.16. Silvaco Inc, DevEdit User's Manual, 2013.

2.17. Silvaco Inc, ATLAS User's Manual, 2013.

2.18. Silvaco Inc, ATHENA User's Manual, 2013.

2.19. Silvaco Inc, TonyPlot User's Manual, 2014.

2.20. Silvaco Inc, VictoryProcess User's Manual, 2014.

2.21. Silvaco Inc, VictoryDevice User's Manual, 2014.

2.22. Silvaco Inc, VictoryStress User's Manual, 2014.

2.23. Silvaco Inc, VictoryCell User's Manual, 2014.

2.24. Silvaco Inc, UTMOST User's Manual, 2013.

3.1. Vasileska, D. Goodnick, S. M. and Klimeck, G. (2010) *Computational Electronics: Semiclassical and Quantum Device Modeling and Simulation* (CRC Press, Taylor & Francis Group, Boca Raton).

3.2. Sun, Y., Thompson, S. E. and Nishida, T. (2010) *Strain Effect in Semiconductors: Theory and Device Applications* (Springer Science+Business Media, LLC, New York).

3.3. Sverdlov, V. (2011) *Strain-Induced Effects in Advanced MOSFETs* (Springer-Verlag/Wien).

3.4. Li S. and Fu, Y. (2012) *3D TCAD Simulation for Semiconductor Processes, Devices and Optoelectronics* (Springer Science+Business Media, LLC, New York).

3.5. Wang, X. (2010) *Simulation study of scaling design, performance characterization, statistical variability and reliability of decananometer MOSFETs*, PhD Thesis, University of Glasgow.

7.1. Maiti, C. K. and Maiti, T. K. (2012) *Strain-Engineered MOSFETs* (CRC Press (Taylor & Francis), USA).

7.2. Jankovic, N. D. and O'Neill, A. (2004). 2D device-level simulation study of strained-Si pnp heterojunction bipolar transistors on virtual substrates, *Solid-State Electronics*, 48, pp. 225–230.

7.3. Mukherjee, C. (2010) *Strain engineering in heterojunction bipolar transistors*, MTech Thesis, IIT Kharagpur.

7.4. Al-Sadi, M. (2011) *TCAD based SiGe HBT advanced architecture exploration*, PhD Thesis, L'Universite Bordeaux I.

8.1. Colinge, J. P. Editor (2008) *FinFETs and Other Multi-Gate Transistors* (Springer, Berlin).

8.2. Han, W. and Wang, Z. M. Editors (2013) *Toward Quantum FinFET* (Springer International Publishing, Switzerland).

8.3. Gaynor, B. (2014) *Simulation of FinFET electrical performance dependence on fin shape and TSV and back-gate noise coupling in 3-D integrated circuits*, PhD Thesis, Tufts University.

8.4. Mohd Zain, A. S. (2013) *Scaling and variability in ultra-thin body silicon on insulator (UTB SOI) MOSFETs*, PhD thesis, University of Glasgow.

9.1. Deleonibus, S. Editor, (2009) *Electronic Device Architectures for the Nano-CMOS Era: From Ultimate CMOS Scaling to Beyond CMOS Devices* (Pan Stanford Publishing Pte. Ltd., Singapore).

9.2. Liu, K.-M. (2008) *Schrödinger equation Monte Carlo-3D for simulation of nanoscale MOSFETs*, PhD Thesis, University of Texas at Austin.

9.3. Ran, C. (2014) *Strain engineering for advanced silicon, germanium and germanium-tin transistors*, PhD Thesis, National University of Singapore.

14.1 ICCAP: (2006) *IC-CAP Parameter Extraction and Device Modeling Software* from Agilent Technologies. Web site: http://eesof.tm.agilent.com/products/iccap_main.html

Index

C

C

C

C