Gate Leakage Variability in Nano-CMOS Transistors

Stanislav Markov

Submitted in fulfilment of the requirements for
the Degree of *Doctor of Philosophy*

Department of Electronics & Electrical Engineering
Faculty of Engineering
University of Glasgow

Dedicated to my parents

Abstract

Gate leakage variability in nano-scale CMOS devices is investigated through advanced modelling and simulations of planar, bulk-type MOSFETs.

The motivation for the work stems from the two of the most challenging issues in front of the semiconductor industry - excessive leakage power, and device variability - both being brought about with the aggressive downscaling of device dimensions to the nanometer scale. The aim is to deliver a comprehensive tool for the assessment of gate leakage variability in realistic nano-scale CMOS transistors.

We adopt a 3D drift-diffusion device simulation approach with density-gradient quantum corrections, as the most established framework for the study of device variability. The simulator is first extended to model the direct tunnelling of electrons through the gate dielectric, by means of an improved WKB approximation.

A study of a 25 nm square gate n-type MOSFET demonstrates that combined effect of discrete random dopants and oxide thickness variation lead to starndard deviation of up to 50 % (10 %) of the mean gate leakage current in OFF(ON)-state of the transistor. There is also a 5 to 6 times increase of the magnitude of the gate current, compared to that simulated of a uniform device.

A significant part of the research is dedicated to the analysis of the non-abrupt band-gap and permittivity transition at the Si/SiO_2 interface. One dimensional simulation of a MOS inversion layer with a 1 nm SiO_2 insulator and realistic band-gap transition reveals a strong impact on subband quantisation (over 50 mV reduction in the Δ-valley splitting and over 20 % redistribution of carriers from the $\Delta2$ to the $\Delta4$ valleys), and enhancement of capacitance (over 10 %) and leakage (about 10 times), relative to simulations with an abrupt band-edge transition at the interface.

If you are surprised one day - then you are learning!
Because, otherwise you are just absorbing
whatever is in your brain already.

— Antonio Martinez

Acknowledgements

Completing a PhD thesis is a personal achievement, and reflects a milestone in ones intellectual development, which is typically a long and twisted line. Writing acknowledgements is like scrolling back in time, along this line, thinking who were the people that influenced me all along. And I go back to 1996, to thank Robert Manolov, who was the first to trust my creative capacity and engage me in real-life electronics engineering, while I was still a student. I often recall his optimism and inventiveness, which helps me keep a bright outlook on any technical context.

The research reported in this thesis was performed with my best diligence and effort (both mental and physical). The right attitude for this was cultivated in me by Ventsy Manoev, and Mitio Mitev (my M.Sc. degree advisors, in The Technical University of Sofia), who I regard as my “parents”, in respect to the career path I took. I am still to meet wizards (in electronics) of their scale. They taught me to respect the details and to persist.

Why I faced the challenge of a doctorate is another story. But for being accepted as a PhD student, and for completing my dissertation, I am more than indebted to my supervisor, Prof. Asen Asenov – a wise, resourceful and enthusiastic leader, a visionary! Prof. Asenov and Dr. Scott Roy (my second supervisor) are inspirational not only because of their expertise, but also because of the commitment and exceptional promptness in their guidance and communication to me. They were always there – at any time of the clock, at any place of the globe – able to review my work, and give me their comments. It is this attitude of them that allowed me to submit in time reports and articles, and even this thesis.

I am thankful to Prof. Enrico Sangiorgi, for inviting me to work in the ARCES Research Centre at the University of Bologna, as a Marie Curie research fellow. It is there, where I really advanced in the study of the Si/SiO$_2$ interface transition layer (see Chapter 5). Looking back at this period, I am filled with gratitude to Prof. Claudio Fiegna, whose support at conferences, and dedicated advice throughout the research, have been elemental to the successful outcome of the fellowship – several publications, the IEEE Nicollian award, and the many things I learned!

There are two people, amongst my closest friends, whom I wish to say a special thank you – Dr. Antonio Martinez, and Dr. Giulio Ferrari, for the enlightening discussions; for the frequent exchange of books; for dragging me to the swimming pool (when I didn't feel ready for another challenge); for sharing moments away from the PhD studies; for taking a genuine interest in my work, and constantly questioning (only to improve it) my understanding of the solid state, quantum mechanics, and physics and maths in general. They are largely responsible for awaking my curiosity about nature, hence the re-activated interest in physics!

This is a good place to mention other device modellers from my immediate environment – Andrew Brown, and Gareth Roy, for sharing the 3D-'atomistic' code, and for their advice on simulations on a statistical scale; K. Kalna, C. Millar, B. Cheng, C. Alexander, C. Riddet and J. Ayubi-Moak, for occasional help on device architecture, modelling and simulation, data visualization, circuits, etc. and for their encouragement; and J. Wattling, for bringing to my attention Register's paper (1).

I am also obliged to my examiners, Dr. P.A. Tony Childs and Prof. John Davies, for carefully reviewing the thesis, and for their constructive feedback.

Finally, if you are still reading this, you probably do not have much interest in the contents of the thesis itself, do you! To better spend your time, why not visit www.ted.com – plenty of wholemeal, organic food for the brain: ideas worth spreading! Listening to the TED talks was what kept me safe from falling in despondency, while writing up. For bringing the TED conference to my attention, and for the eyes-opening course "Into the world of business", thanks goes to Patrick Andrews. Thank You All!

Contents

List of Figures

Nomenclature

Physical constants

$\hbar$	reduced Planck's constant ($1.054571628\times10^{-34}$ Js)
κ	dielectric constant
$\varepsilon = \kappa\varepsilon_0$	dielectric permittivity
ε_0	vacuum permittivity (8.85418782×10^{-14} F/cm)
k_B	Boltzmann's constant (1.38×10^{-23} J/K)
m_0	Electron rest mass (9.11×10^{-31} kg)
q	Elementary (unit) charge (1.60×10^{-19} C)

Other Symbols

δE_c	Conduction band offset at interfaces
δE_v	Valence band offset at interfaces
C_{ox}	MOS oxide capacitance
E_G	Electronic band-gap
EOT	MOS effective oxide thickness
F_{ox}	Electric field in the oxide
$I_{D,off}$	MOSFET sub-threshold drain current
$I_{D,sat}$	MOSFET drain current in saturation
I_D	MOSFET drain current

I_G	Gate current
J_G	Gate current density
L_G	MOSFET physical gate length
m	MOSFET ideality factor (when used without a sub-script)
N_b	MOS substrate impurity concentration
T	Temperature
T	Tunnelling probability
t_{ox}	MOSFET physical gate insulator thickness
V_D	MOSFET drain voltage, relative to the source contact
V_G	MOSFET gate voltage, relative to the source contact
V_T	MOS threshold voltage
V_{DD}	supply voltage
W_G	MOSFET physical gate width
W_d	MOS depletion layer width
X_j	MOSFET source/drain junction depth

Acronyms

(m)WKB	(modified) Wentzel-Kramers-Brillouin approximation
(S)TEM	(Scanning) Transmission electron microscope
(X)PS	X-ray Photoemission spectroscopy
ADF	Annular dark-field imaging
AFM	Atomic force microscope
ALD	Atomic layer deposition
BC	Boundary condition

BG	Band-gap
BS	Bound states
BTBT	Band-to-band tunnelling
BTE	Boltzmann transport equation
CMOS	complementary metal-oxide-semiconductor
CTL	Compositional transition layer
DD	drift-diffusion
DFT	Density functional theory
DG	density-gradient
DT	Direct tunnelling
EELS	Electron energy spectroscopy
EMA	Effective-mass approximation
ES	Extended states
FET	field effect transistor
FNT	Fowler-Nordheim tunnelling
HKGS	High-κ dielectric gate stack
HP	High performance (technology)
HRBS	High-resolution Rutherford back-scattering spectroscopy
HRTEM	High-resolution transmission electron microscope
IL	Interfacial layer
ITRS	International Technology Roadmap for Semiconductors
LER	Line-edge roughness
LOP	Low operating power (technology)

MC	Monte Carlo
MOS	metal-oxide-semiconductor
NEGF	Non-equilibrium Green's functions
OTV	Oxide thickness variation
PE	Poisson equation
PECT	Photo-emission current threshold
PGB	Poly-Si grain boundary
PS	Poisson-Schrödinger
QBS	Quasi-bound states
RDF	Random dopant fluctuations
RMS	Root-mean-square
RTA	Rapid thermal anneal
SE	Schrödingerequation
SILC	Stress-induced leakage current
SOI	Silicon-on-insulator device
SRAM	Static Random Access Memory
STL	Structural transition layer
TB	Tight-binding approximation
UTB	Ultra-thin-body device
UTB-DG	Ultra-thin-body double-gate device
VLSI	Very large scale integration
WF	Wave-function

Chapter 1

Introduction

We endeavour to analyse the gate leakage variability in nano-scale transistors, through advanced device modelling and simulation. A brief overview of the problems of leakage and variability in the context of ultra-scaled semiconductor devices reflects our motivation for the study. We summarise the objectives of this work, in line with the demands of the semiconductor industry and complete the introduction with an outline of the subsequent, more technical chapters.

1.1 Leakage and variability in ultra scaled devices

The miniaturisation (scaling) of complementary metal-oxide-semiconductor (CMOS) field-effect transistors (FETs), providing improved chip performance at a reduced cost through the simultaneous increase of transistor density and transistor switching speed, has been the main driver of silicon technology for over three decades (2). Subsequent to the 90 nm technology node, the semiconductor industry faced two formidable issues that challenged, and continue to challenge the value of traditional CMOS device scaling. These two issues are excessive leakage power, and device variability.

Ideally, the energy dissipation in CMOS circuits is due entirely to the switching between logic levels, referred to as dynamic, or active, power. This is true, under the assumption that once the input and output levels on the logic gates stabilise, current ceases to flow from the supply. In reality, leakage currents flow at all times, even when the voltage at the transistor terminals is stable, and these contribute to the so called static, or leakage, power. Static power is due to the manifestation of quantum-mechanical phenomena, including gate oxide and band-to-band tunnelling currents, and over-the-barrier carrier injection in the off-state of the transistor, resulting in non-negligible sub-threshold drain-source current. These phenomena become more pronounced with technology scaling, due to the exponential sensitivity of the gate tunnelling on oxide thickness, increased channel doping, and the much enhanced influence of the drain contact over the electrostatics of the device. All this leads to the rapid increase and dominance of static power dissipation in integrated circuits based on technologies with a smaller than 130 nm lithographic half-pitch, as reported for a variety of products (system-on-chip, desktop processors, field programmable gate arrays), marketed or prototyped by leading manufacturers (2; 3; 4; 5). An example of the increasing trend in the three most significant contributors to the total chip power is illustrated in Fig 1.1, for desktop processor technologies, with the gate leakage component being the highest. With the aggressive reduction of the gate oxide thickness beyond 2 nm, gate tunnelling current becomes in the order of, or higher than the sub-threshold current of a MOSFET, especially for low operating power, high-V_T devices (3; 6; 7; 8; 9; 10).

The introduction of alternative gate oxides with high dielectric constant (high-κ) mitigates the problem of gate tunnelling leakage to a significant extent (11; 12; 13), but entails numerous other challenges (14; 15; 16; 17), so that a solution for the ultimate scaling of the gate insulator down to the equivalent oxide thickness of 0.5 nm is as yet unknown.

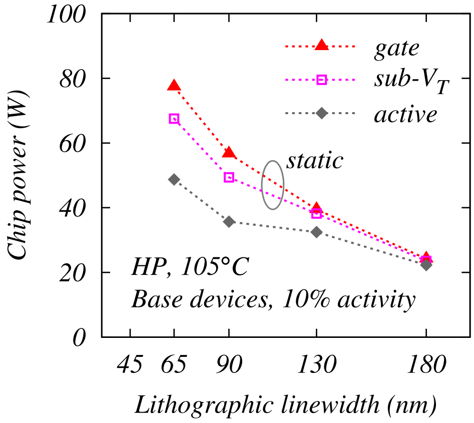


Figure 1.1: An example breakdown of the three most significant power components (*gate* leakage, *sub*-V_T drain leakage, and *active* power at 10% activity of the resources) of a 10 mm^2 chip, for a set of four different desktop processor, high-performance (HP), technology generations, referenced by their minimum lithographic line width. Data is from Ref. (7).

Statistical device variability is the dispersion in the electrical characteristics of identical MOSFETs. It is typically characterised by a statistical distribution of the threshold voltage, V_T, on which both the *on* and *off* (sub-V_T) drain currents depend. Device variability implies variability in the delay and leakage power of circuits and systems. Therefore it adversely affects functionality, yield and reliability, and has become a crucial obstacle in a power-constrained search for performance enhancement through device scaling (18; 19; 20; 21). Variability stems from two simple facts. First, the fast approach of critical device and interconnect dimensions to the nano-metre scale requires atomic level precision in their processing, and that is extremely challenging to maintain, if at all possible, by mass production fabrication facilities. Second, certain aspects of silicon technology and MOSFET design imply the existence of uncontrollable, intrinsic fluctuations in a number of *microscopic* features in devices with otherwise equivalent *macroscopic* parameters (i.e. layout and geometry). Here, we are concerned with intrinsic parameter fluctuations that are associated with the fundamental atomicity of charge and matter, and are stochastic in nature. Some well known sources of intrinsic variability are the different number and spacial configuration of discrete doping atoms, the oxide thickness variation induced by the roughness of the oxide interface, and the gate-line edge roughness (22; 23; 24; 25; 26; 27; 28; 29).

Gate leakage current has an exponential sensitivity to oxide thickness and oxide field, so that gate leakage variability is expected from the above mentioned intrinsic parameter fluctuation sources. A few works briefly address this problem, but do not provide a systematic approach for the characterisation of gate leakage variability in current and future technology generations (30; 31; 32; 33). At the same time, there is a growing effort to account for gate leakage variability in chip-power simulators, used for the power-constrained evaluation of different architectures and performance factors (4; 34; 35). It is suggested that ignoring the variability in leakage current leads to the underestimation of the total leakage power by as much as 30 %, in the 65 nm technology (4). Note however, that the underlying gate leakage models used in the chip-power simulator for making the above prediction are physically, and statistically oversimplified, because gate leakage variability is very difficult to measure and because no comprehensive device modelling studies on the subject exist (18). The aim of our work is to bridge this gap, by establishing a reliable framework for device level simulation of gate leakage variability in nano-scale MOSFETs, on a statistical scale. The specific objectives in this direction are summarised below.

1.2 Aims and objectives of the study

The main goal of this research is to develop a simulation framework that allows the study of gate leakage variability in nano-scale CMOS devices. This includes the following objectives:

- *selection of a direct tunnelling model*, with an emphasis on computational efficiency, sufficient accuracy, and viability for extension of the approach from a one-dimensional (1D) to a three dimensional (3D) simulation domain;

- *1D proof of concept* - implementation of the model in a 1D simulator and simulation of 1D gate leakage, comparison with relevant data, calibration of phenomenological parameters;

- *3D proof of concept* - incorporation of the tunnelling model in the *Glasgow 3D atomistic* drift-diffusion simulator with density-gradient quantum corrections, and simulation of a uniform device with continuous doping;

- *study of gate leakage variability* - statistical simulation of gate leakage variability due to the individual and combined sources of intrinsic parameter fluctuations (random dopant fluctuations and oxide thickness fluctuation) on a test sub-30 nm n-type MOSFET.

An additional aim of this work is to investigate the impact of the structural and compositional transition at the Si/SiO$_2$ interface, on the inversion layer and gate leakage characteristics of a MOS structure. The objectives towards this goal are:

- *problem formulation* - understanding how the atomic transition from Si to SiO$_2$ affects the change of electronic and dielectric properties at the Si/SiO$_2$ interface;

- *1D simulator development* - development of a 1D Poisson-Schrödinger (PS) solver that accounts for a non-abrupt change of conduction and valence band edges, dielectric constant and effective mass;

- *implementation of 1D quantum-mechanical tunnelling model* - selection of a quantum-mechanical tunnelling model, and its incorporation in the PS solver;

- *study of the physical consequences of the Si/SiO$_2$ transition* - simulation and analysis of the impact of gradual interface band-gap, permittivity and effective mass transition on the inversion layer quantisation, capacitance and gate leakage characteristics

1.3 Thesis outline

The exposition of this work is organised as follows.

In Chapter 2 we first review the implications of CMOS device scaling on leakage power, and gate leakage in particular. Then we elaborate the context of device variability, and anticipate which of the usual sources of intrinsic parameter fluctuations can affect gate leakage current. This is followed by a review of the prior art on gate leakage variability.

Chapter 3 comprises a review of the gate leakage mechanisms, with emphasis on direct tunnelling and its geometrical partitioning in a MOSFET, and a survey of the most common direct tunnelling models. The incorporation of a semi-classical, analytical tunnelling model in a Poisson-density-gradient solver is discussed, and results from 1D gate leakage simulations, used for the calibration of the model, are finally presented, showing a very good agreement with experimental data.

Chapter 4 is devoted to the 3D modelling and simulation of gate leakage and gate leakage variability. A summary of the most commonly used techniques for 3D device simulation is followed by a description of the *Glasgow 3D-atomistic* simulator (chosen as the main vehicle for the study of statistical gate leakage variability), with an emphasis on the modelling of random dopant fluctuations and oxide interface roughness, and the implementation of the direct tunnelling gate current model. Then we present the 3D gate current simulations of a uniform device, thus validating the established gate leakage modelling framework. The approach is finally applied to the study of gate leakage in an ensemble of 25 nm gate length n-type MOSFETs, subject to the individual, as well as the combined influence of random dopant fluctuations and oxide roughness. Hence, it is shown that for the biases corresponding to the stable points of a CMOS inverter, there is a large gate leakage variability, and the factors that lead to it are thoroughly investigated.

Chapter 5 is an extensive study of the non-abrupt transition of electronic properties at the Si/SiO$_2$ interface, and its impact on the characteristics of the MOS inversion layer. The issue is introduced by a review of experimental and first-principles studies of the Si/SiO$_2$ interface. This is followed by a systematic comparison of 1D quantum-mechanical simulations (in the envelope wave function, effective mass approximation) of the inversion layer with an *abrupt* (traditional), and a *linear* interface barrier models, showing a significant increase in leakage current, enhancement of gate capacitance, and reduction in subband splitting, due to the gradual band-gap transition. The analysis is extended to the non-abrupt change of dielectric constant and effective mass at the interface, and to the case of a *realistic* barrier model (obtained *ab initio*), and is applied to devices with high-κ dielectric stacks. The development of the 1D Poisson-Schrödinger solver used in this work is presented in Appendix A, while the procedure of obtaining a realistic band-edge transition profiles from first-principles calculations of the interface is described in Appendix B.

We conclude, in Chapter 6, with a summary of the major findings of this work, and a brief discussion of their implications in the context of present and future semiconductor technology, thus suggesting directions for future research.

Chapter 2

CMOS scaling, leakage power, and variability

This chapter starts with an overview of scaling trends in modern CMOS transistors, with an emphasis on the impact of scaling on leakage power, and a summary of the dominant leakage mechanisms. Then we elaborate the context of device variability and its impact on leakage power. Subsequently, we anticipate the sources of gate leakage variability, and finish with a review of existing studies on statistical simulations of gate leakage variability.

2.1 Impact of CMOS scaling on leakage power

2.1.1 Device scaling and power dissipation

The primary goal of CMOS scaling is reduction of the cost per functional power, by increasing the integration density of on-chip components. The elaboration of *constant field scaling* rules entails concomitant performance and power consumption improvements, which have shaped the evolution of silicon technology (36; 37). The concept of device scaling is illustrated in Fig. 2.1.

In *constant field scaling*, the physical dimensions of the device (gate length L_G and width W_G, oxide thickness t_{ox}, and junction depth X_j), and the supply and threshold voltages (V_{DD} and V_T respectively), are reduced by the same factor, $\alpha > 1$, so that the two-dimensional pattern of the electric field is maintained constant, while circuit density increases by $\sim \alpha^2$. This implies that the depletion width (W_d) must also be reduced by the same amount, which is achieved by increasing the substrate doping N_B by α. Consequently, both the gate capacitance ($C = L_G W_G \varepsilon_{ox}/t_{ox}$), and the drain saturation current ($I_{D,sat}$) are scaled down by α. The

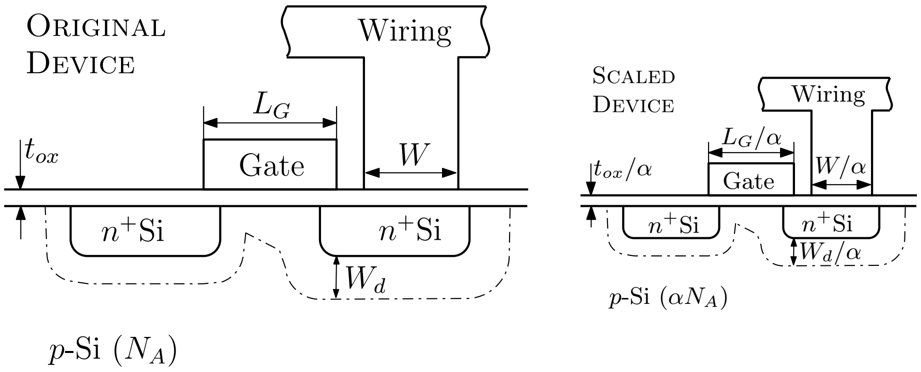


Figure 2.1: Conceptual schematic diagram of device scaling (38). Both device and wiring dimensions are required to scale by the same factor $1/\alpha$, in order to increase integration density by α^2. Scaling of the supply voltage by the same factor ($1/\alpha$) maintains the same 2D electric field pattern, subject to an equivalent scaling of the depletion width W_d.

saturation current determines the transistor intrinsic switching delay $\tau \sim CV_{DD}/I_{D,sat}$, which is thus reduced by α, leading to a performance improvement. At the same time, the power dissipation ($P \sim I_{D,sat}V_{DD}$) is reduced by α^2, so that the power density ($P/(L_G W_G)$) remains unchanged.

Table 2.1 shows the gate-length, supply voltage, and oxide thickness figures for several recent, and future technology generations, as projected by the ITRS. It is obvious that present day device scaling does not adhere to the constant field scaling rules. This is because of several fundamental, non-scaling factors, and practical considerations (37). Instead, the *generalised scaling* rules are followed, where the physical dimensions of the transistor are still reduced by a factor of α, providing the desired circuit density increase (α^2) and performance improvement (α), but the supply voltage is scaled by β/α, leading to an increase in the magnitude of the

Table 2.1: ITRS-projected L_G, *EOT* and V_{DD}.

Year	L_G (nm)	*EOT* (nm)	V_{DD} (V)
2003	45	1.3	1.2
2005	32	1.2	1.1
2007	25	1.1	1.1
2009	20	0.9	1.0
2011	16	0.6	1.0
2013	13	0.5	0.9
2015	10	0.5	0.9

Trends in two of the critical device dimensions, and power supply, from the ITRS 2003, 2005 and 2007 editions. Clearly, gate length scales much faster than supply voltage, resulting in an increase in the source-drain electric field, and a departure from *constant field* scaling.

electrical field by $1 \leq \beta \leq \alpha$ (37; 39). A full list of MOSFET physical parameters, and their scaling factors, is given for *constant field scaling*, and *generalised scaling* in Table 2.2. The last column in this table shows the rules for *selective scaling*, which relaxes one more constrain – the fixed ratio between gate length and width (38). Such relaxation is driven mostly by the slower scaling pace of on-chip interconnect lines.

Table 2.2: Device parameters and their scaling factors (38)

Physical parameter	Scaling Rules Factors		
	Constant Field	Generalized	Selective
Gate length (L_G), Oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	$1/\alpha_D$
Wiring width, channel width (W_G)	$1/\alpha$	$1/\alpha$	$1/\alpha_W$
Voltages (V_{DD},V_T)	$1/\alpha$	β/α	β/α_D
Substrate Doping (N_B)	α	$\beta\alpha$	β/α_W
Electric field	1	β	β
Gate capacitance ($C = L_G W_G \varepsilon_{ox}/t_{ox}$)	$1/\alpha$	$1/\alpha$	$1/\alpha_W$
Drive current ($I_{D,sat}$)	$1/\alpha$	β/α	β/α_W
Intrinsic delay ($\tau \sim CV_{DD}/I_{D,sat}$)	$1/\alpha$	$1/\alpha$	$1/\alpha_D$
Area ($A \propto L_G W_G$, or $\propto W_G^2$)	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_W^2$
Power dissipation ($P \sim I_{D,sat} V_{DD}$)	$1/\alpha^2$	β^2/α^2	$\beta^2/(\alpha_W \alpha_D)$
Power density (P/A)	1	β^2	$\beta^2(\alpha_W/\alpha_D)$

In *constant field* and *generalised* scaling, the same factor, α, is applied to the dimensional and voltage parameters. In *generalised* and *selective* scaling, β is the electric field scaling parameter. In *selective* scaling, vertical dimensions and gate length scaling is governed by α_D, while gate width and wiring scaling – by $\alpha_W < \alpha_D$.

As can be seen in Table 2.2, under *generalised* and *selective* scaling scenarios the dissipated power scales by β^2/α^2, while the power density increases by β^2. The power density is of paramount importance for chip packaging and systems design, and its increase imposes a practical limit for the exploitation of device scaling. It is worth noting that in the most recent technologies, and in the projections of the ITRS, supply voltage hardly scales, i.e. $\beta \approx \alpha$ (2; 10). This already suggests that the power density grows at the same rate as the integration density. Actually, the issue is even worse than it appears from considering the scaling rules alone, because of non-scaling factors, leading to the increase and dominance of the static, leakage power. This is clearly demonstrated by the crossing lines of Fig. 2.2, illustrating the trends in dynamic and leakage power densities with shrinking gate length. The actual measurements for devices with gate length between 1 µm and 65 nm are shown with symbols (2). The rapid escalation of leakage has ultimately led to power-constrained, application-specific evaluation of scaling scenarios, since different applications can tolerate different power densities. (e.g. 8 orders of magnitude difference between the allowed power density of a high-performance processor and an ultra-low power SRAM) (40).

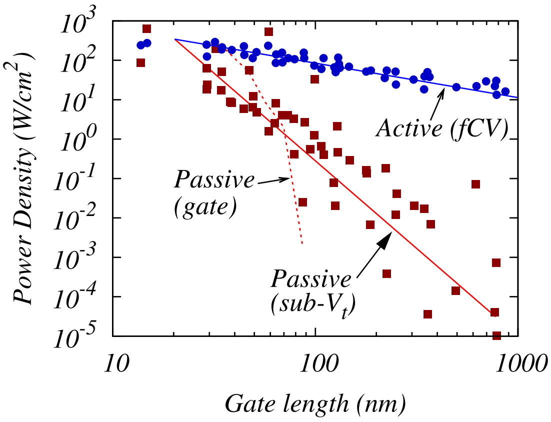


Figure 2.2: Comparison of measured active power, and leakage (passive) power in devices with gate lengths ranging from 1 µm to 20 nm (symbols). Lines indicate the trend of a particular power component, as indicated. (Data from (2).)

Of the two components of consumed power, the dynamic power, dissipated during a switching between logic states, can be ameliorated by limiting the switching frequency f, to which it is proportional ($P_{dyn} \approx CV_{DD}^2 f$, per device). The other component – static, leakage power – dissipated whilst maintaining a logic state, is exponentially sensitive to some of the device parameters and their variability, as well as to non-scaling factors. Leakage power dissipation is time-invariant (except for its comparatively weak dependence on the logic state of the transistors) and now poses the most significant scaling limit (2; 35).

2.1.2 CMOS leakage power

CMOS leakage power is due to sub-threshold drain current, and various tunnelling mechanisms through the potential barriers in the transistor (41).

2.1.2.1 Sub-threshold source-drain leakage

Unlike the saturation drain current, which scales down by β/α, as already discussed, the sub-threshold drain current actually increases with scaling. This is best understood from the approximate expression of the off current $I_{D,off}$ (37)

$$I_{D,off} = I_{DS}(V_G = 0, V_{DS} = V_{DD}) = \mu_{eff} C_{ox} \frac{W_G}{L_G} (m-1) \big(\frac{k_B T}{q}\big)^2 e^{-qV_T/mk_BT}, \tag{2.1}$$

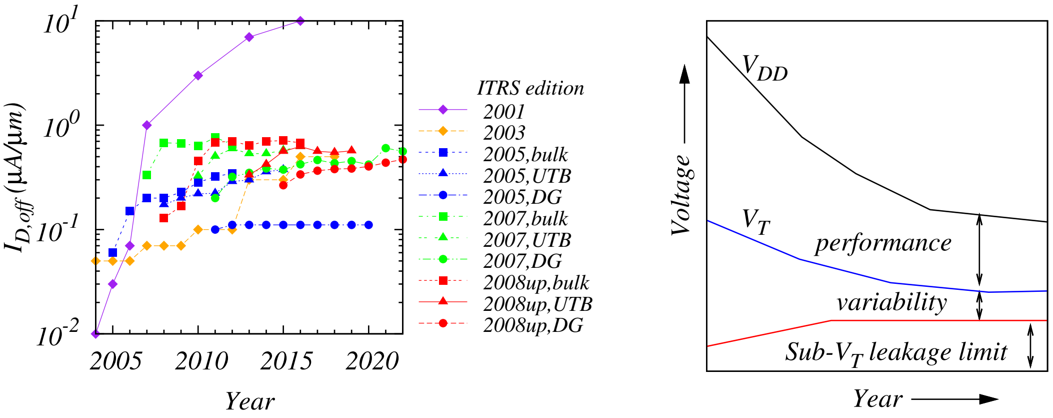


Figure 2.3: ITRS projections of the sub-threshold source-drain current (left) showing continuous relaxation of the sub-V_T leakage constrain to allow V_T down-scaling (42). The sub-threshold current implies a scaling limit for V_T, as shown schematically on the right. This limit is higher if variability is accounted for. The required performance margin ($V_{DD} - V_T$) stalls the power supply scaling too, in the absence of performance boosters.

where μ_{eff} is the effective mobility (relatively independent of scaling), and $(m-1) \sim 1$ is the body-effect coefficient. Note that the oxide sheet capacitance $C_{ox} = \varepsilon / t_{ox}$ scales by α, while V_T is supposed to scale by β/α, to avoid performance degradation. This latter scaling condition is related to the intrinsic switching delay of the transistor, which is determined by the saturation drain current, $I_{D,sat}$. For short channel MOSFETs, the saturation current may be expressed as

$$I_{D,sat} = I_{DS}(V_G = V_{DD}, V_{DS} = 0) = v_{sat} C_{ox} W_G (V_{DD} - V_T), \tag{2.2}$$

where v_{sat} is the saturation velocity (37). From this expression is clear that reducing V_T, at a given supply voltage, is beneficial to device performance. But this V_T reduction leads to an exponential increase in the off current, as is evident from Eq. 2.1. These contradictory requirements on V_T scaling have led to the continuous relaxation of the power dissipation constraints and the slowing of supply voltage scaling, as clearly illustrated in Fig. 2.3 (42). Still, V_{DD} scaling is desirable, because leakage power is proportional to it, while dynamic power is proportional to its square. In this regard, equation 2.2 also explains the importance of performance boosters (e.g. high-mobility channel materials, or strained channel) that aim to improve the saturation velocity v_{sat} (a material property) and alleviate the need to scale V_T in proportion to V_{DD}, without degrading performance.

2.1.2.2 Band-to-band tunnelling leakage

There are several quantum mechanical tunnelling mechanisms in short-channel MOSFETs, two of which result in leakage currents that are comparable to, or exceed, the sub-V_T leakage. One of these mechanisms is band-to-band (BTB) tunnelling through the reverse-biased pn-junction formed between the substrate and the drain region of the transistor. The tunnelling current is due to electrons tunnelling from the valence band of the p-region, to the conduction band of the n-region, as is schematically illustrated in Fig. 2.4. This current is much stronger than the diffusion current of a reverse biased pn-junction, and is a consequence of the retarded field scaling, and the abruptness of the pn-junction (37). The latter is a direct consequence of the former factor, because in a *generalised scaling* scenario, the substrate doping N_B must increase by the product $\alpha\beta$, i.e. even faster than if constant field was maintained. What additionally exacerbates the problem is that the control of short channel effects in modern, bulk-MOSFETs is accomplished by the use of a halo-doping implantation, aiming to minimise the drain-induced depletion in the substrate (41). As a result, the depletion region on each side of the drain pn-junction is sufficiently narrow that the voltage drop over the junction (due to the built-in potential, and the drain to body bias) exceeds the Si band-gap value of 1.12 eV (which cannot scale!), thus allowing band-to-band tunnelling. Deep traps in the depletion region additionally increase BTB tunnelling (40).

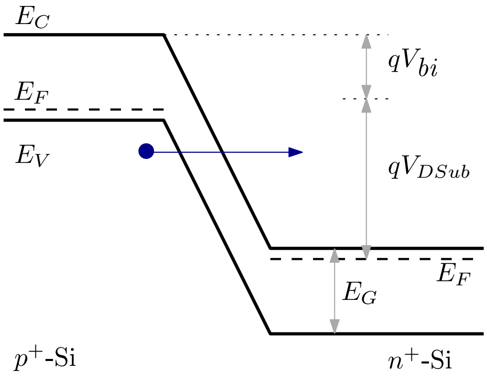


Figure 2.4: Schematic energy-band diagram representation of band-to-band (BTB) tunnelling, at the drain-substrate pn-junction of a MOSFET. Due to high doping levels, the space charge region is narrow, reducing the width of the tunnelling barrier, and $qV_{bi} \sim E_G$. High drain bias makes V_{DSub} large enough to align valence states from the p-Si to conduction band of the n-Si.

2.1.2.3 Gate oxide tunnelling leakage

The other critical tunnelling current is due to carriers tunnelling through the gate oxide of the MOSFET. It is a consequence of the continuous down-scaling of the SiO_2 gate dielectric, as required to guarantee gate control over the channel, due to device electrostatics. Several tunnelling processes give rise to gate leakage, and are discussed in Chapter 3. It will become clear

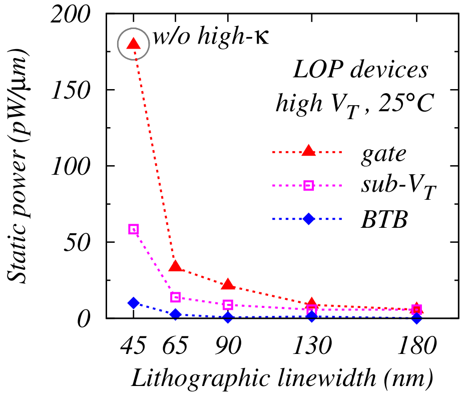


Figure 2.5: An example breakdown of the three most significant passive power components (*gate* tunnelling, sub-threshold (*sub*-V_T) leakage, and band-to-band (*BTB*) tunnelling) in devices with high threshold voltage (*high*-V_T), for a set of handheld, low operating power (LOP) technology generations, referenced by their lithographic line width. Data is from Ref. (7).

that gate tunnelling is exponentially sensitive to the oxide thickness, and the oxide perpendicular field, while it is only linearly dependent on the gate width, and source/drain extension overlaps. Therefore it also increases exponentially with device scaling. At this stage we must emphasise that gate leakage is the dominant tunnelling mechanism in devices with SiO_2 or with SiON dielectric gate insulators used in recent technologies. Fig. 2.5 compares the magnitude of leakage power due to leakage currents discusses so far.

The intolerable growth of gate tunnelling leakage has triggered a radical change in the silicon technology, which aims to introduce dielectric permittivity scaling through material engineering. The goal is to scale the oxide sheet capacitance C_{ox}, which is required to maintain good electrostatic control by the gate, but avoid decreasing the physical thickness of the gate insulator. This may be achieved by increasing the gate dielectric constant κ, of the insulator, and is realised in the SiON gate oxides that are in use today. With respect to scaling, the relevant metric is now the equivalent oxide thickness (EOT), defined through the relation $\varepsilon_{HK}/t_{HK} = \varepsilon_{ox}/EOT = C_{ox}$, where ε_{HK} and t_{HK} are the permittivity and physical thickness of the high-κ material, and ε_{ox} is the SiO_2 permittivity. Therefore, the introduction of high-κ dielectric materials is seen as the only way to realise the projections of the ITRS for effective oxide thicknesses reaching 0.5 nm.

However, the transition to high-k gate dielectric also entails a replacement of the poly-Si gate with a metal gate, which is another formidable technological challenge, in addition to the difficulties of controlled growth of ultra-thin high-κ oxide films (14). These complications have delayed the adoption of alternative gate dielectric stacks, and have lead to the continuous relaxation of EOT scaling requirements, as can be seen in Fig. 2.6.

Concluding this subsection we note that even with viable high-κ dielectric stack solutions, which are based on hafnia or hafnia-silicate (43), the gate leakage power is anticipated to remain

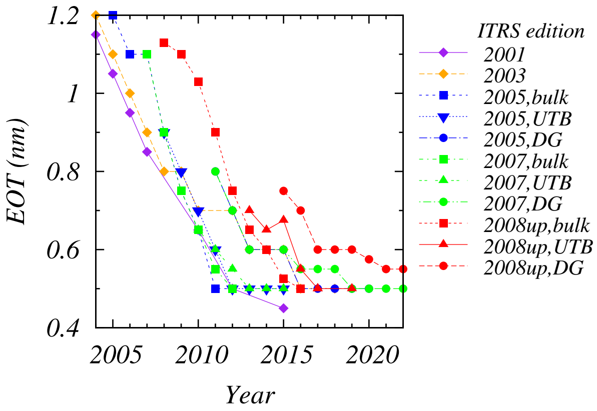


Figure 2.6: Projected trends in EOT scaling, from different editions and updates of the ITRS, and for different architectures (bulk, UTB-SOI, DG) of high-performance devices (42). The requirement for aggressive thinning has been successively relaxed over the last three editions of the roadmap (2005, 2007 and 2008-update), reflecting the delayed adoption of high-κ dielectric stacks.

comparable to the sub-V_T leakage power. This is illustrated in Fig. 2.7, showing the allowed dissipated power per device, the projected total leakage power, and the contribution from gate leakage. Note that the ITRS-projected limit of gate leakage density, J_G is given in terms of the target sub-V_T off-current at 300 K, $I_{sd,leak,TGT} = 0.2\ \mu$A/$\mu$m, as $J_G L_G = I_{sd,leak,TGT}$. At the same time, the actual limit of the sub-V_T leakage is between 0.13 and 0.69 μA/μm, for the high-performance bulk-MOSFET (10).

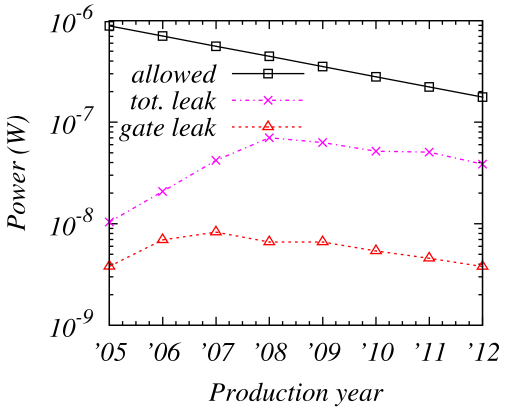


Figure 2.7: ITRS-projected short term trends in the allowed power per device (high-performance), its total leakage power, and the contribution of gate leakage. The trend in the allowed power is dictated by the package-constrained maximum power density. The negative slope in the gate leakage and total leakage curves assumes successful deployment of high-κ dielectric gate stacks.

From the above discussion it is clear that gate leakage continues to be a major component of leakage power, and its accurate evaluation remains of paramount importance.

2.2 Variability in ultra-scaled devices

Device variability has emerged as the most significant challenge to continuous device scaling. This is because the tolerance in the electrical characteristics of devices, induced by the various sources of variability, cannot scale in proportion to the nominal values of the electrical parameters. Therefore, design margins are reduced at a circuit and system level, and this impacts yield. The taxonomy of device variability differs according to the adopted viewpoint (18; 19), but to understand the importance of characterisation and modelling of variability at different levels of the design and optimisation process, one must keep in mind the following classifications.

2.2.1 Classification

From a device perspective, variability is *intrinsic*, arising from fundamental phenomena linked to the granularity of charge and matter, or *extrinsic*, triggered by unavoidable tolerances in the process conditions. Typical sources of intrinsic device variability are random dopant fluctuations (RDF) (both in terms of number and spacial distribution), line edge roughness (LER), microscopic oxide thickness variation (due to Si/SiO$_2$ interface roughness), all of which result in device-to-device variability (29). Examples of extrinsic variability are the die-to-die, or wafer-to-wafer oxide thickness fluctuation, or implantation dose uncertainty (19).

An alternative classification, according to the causes that effect parameter fluctuations, distinguishes *systematic* from *random* variability. The former includes fluctuations of device characteristics that are identifiable with the specifics of the device layout and neighbourhood, e.g. L_G variation due to optical proximity corrections or transistor orientation. The latter concerns parameter fluctuations in transistors with identical layout or environment. Systematic variation can be anticipated, and dealt with by technology optimisation, or in a deterministic way during circuit design, while random variability simply widens the design margins needed to guarantee functionality and performance despite the stochastic contribution to timing and leakage (18; 19).

The selection of appropriate design margins is crucial for the realisation of a profitable chip design, and requires quantitative knowledge of the impact of different sources of parameter fluctuation, and the correlation between them. Concerning random variability, statistical device and circuit modelling and simulation is the only way to obtain such insight.

In our study, we focus on *random*, *intrinsic* device variability of gate leakage current. The different sources of variability and their relevance to gate leakage will be discussed in the next section (2.3).

2.2.2 Impact on leakage power and yield

Device variability adversely affects leakage power. This can be inferred directly from Eq. 2.1, showing an exponential dependence of the sub-threshold drain current (I_{off}) on V_T, the variability of which is extensively studied (27; 29; 44). The same sources of variability that induce threshold voltage fluctuations induce a log-normal distribution in I_{off} (i.e. $log(I_{off})$ has a normal distribution) (19). The inverse correlation between V_T and I_{off} means that faster devices have excessive sub-V_T leakage. Fig. 2.8 shows the measured frequency-leakage dispersion for microprocessors from a single wafer. The ratio between the maximum and minimum measured stand-by leakage (I_{sb}) can be over 20, while the corresponding spread in operating frequency is limited to 30 % (45). In this case, the spread in leakage current has been correlated to gate length variation and V_T variation. In a power constrained design, this may eventually render a large number of devices outside specification, and reduce yield.

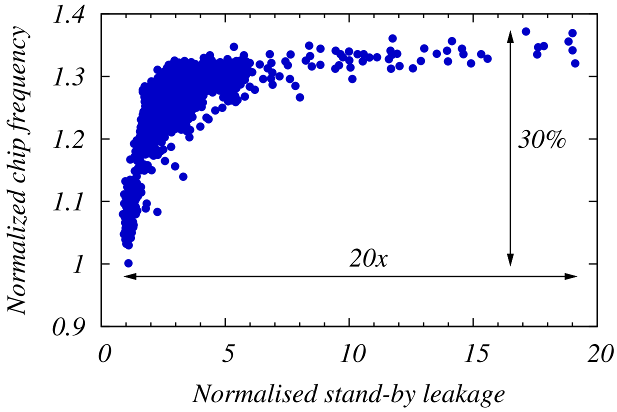


Figure 2.8: Spread in frequency and stand-by leakage, measured in microprocessors from a single wafer (45). Leakage is inversely correlated to chip performance, and suffers much wider fluctuation.

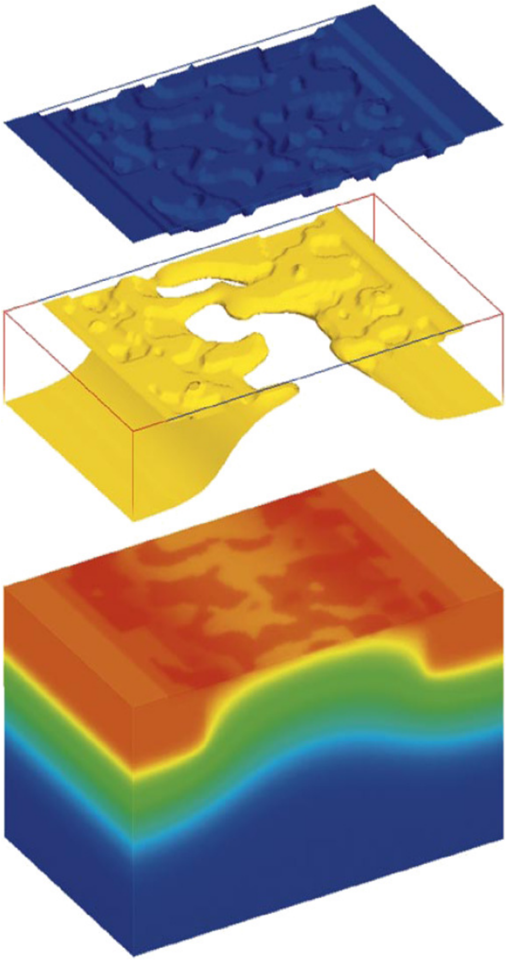
Knowing the sub-V_T leakage dependence on the sources of variability, and more importantly, knowing the distribution of V_T itself, allows performance and power constrained design optimisation, (35), and yield-prediction modelling (46), including I_{off} tolerances. However, the gate leakage, despite its comparable magnitude, and exponential sensitivity to the same sources of variability, is typically ignored, or accounted for on a simplified, unphysical basis (e.g. assuming $I_G = I_{G,nom} \exp(f(\delta t_{ox}))$, where f is a linear function of the oxide thickness fluctuation δt_{ox}) (46).

The fact that gate leakage variability is typically neglected stems from the difficulties in its characterisation (18), and from the lack of comprehensive simulation studies on the subject, as will become clear from the following section. Furthermore, there is no simple correlation between gate leakage and any of the performance-related parameters.

2.3 Statistical gate leakage variability - prior art

One reason to anticipate gate leakage variability is the exponential dependence of the gate tunnelling current on the gate oxide thickness, which is known to vary on a lateral scale of hundreds of nm, as well as on a much smaller lateral scale of 1 to 30 nm (47; 48). The thickness fluctuations on a smaller lateral scale are associated with Si/SiO_2 interface roughness, leading to a deviation from the nominal oxide thickness by one Si(001) inter-atomic plane distance (0.28 nm) (49; 50). For a $1-1.5$ nm SiO_2 gate dielectric, such microscopic thickness fluctuations lead to local tunnelling current density fluctuations of orders of magnitude, and increases the mean of the total tunnelling current, relative to that of a uniform device (30; 51; 52). In Ref. (32), this issue is studied on a statistical scale, concluding that as the linear dimensions of a device are reduced to the order of the interface roughness characteristic length, the fluctuations in the tunnelling current density cannot self-average and translate to a statistical variation of the total tunnelling current. However, the modelling approach disregards actual device electrostatics, and cannot be used to predict the statistical distribution of gate leakage variability in realistic devices. Fig. 2.9 shows a reconstruction of the interface surface profile and its impact on the potential, and on the inversion carrier population distributions in a 30 nm gate length n-channel MOSFET, at low drain and high gate bias. It is expected that the tunnelling current density fluctuations due to the thickness variation itself are further emphasised by the correlation of higher local electron density to the thinner regions of the oxide.

Random dopant fluctuations are also considered a source of intrinsic gate leakage variability, because they locally modulate the electric field and electron density, to which the direct tunnelling current is very sensitive. This is visualised in Fig. 2.10, where a typical distribution of discrete dopant atoms, resulting from a three-dimensional process simulation is shown at the top of the figure. The bottom part of the figure shows the effect of these impurities on the surface potential and electron density distribution in a 30 nm MOSFET. This effect is sufficient to induce local tunnelling current density fluctuations, which may not average over the tunnelling area, and lead to gate current variations from device to device. However, a statistical simulation of a 50-device sample with random dopant fluctuations, reported in Ref. (31), suggests that in a 30 nm gate length MOSFET, the fluctuations in gate current are associated with the uncertainty of the source and drain junction positions, rather than to the dopant-induced fluctuations in the oxide field. Unfortunately, the study was limited to low-drain/high-gate bias of the transistor, when ionized impurities are screened by the inversion charge carriers.



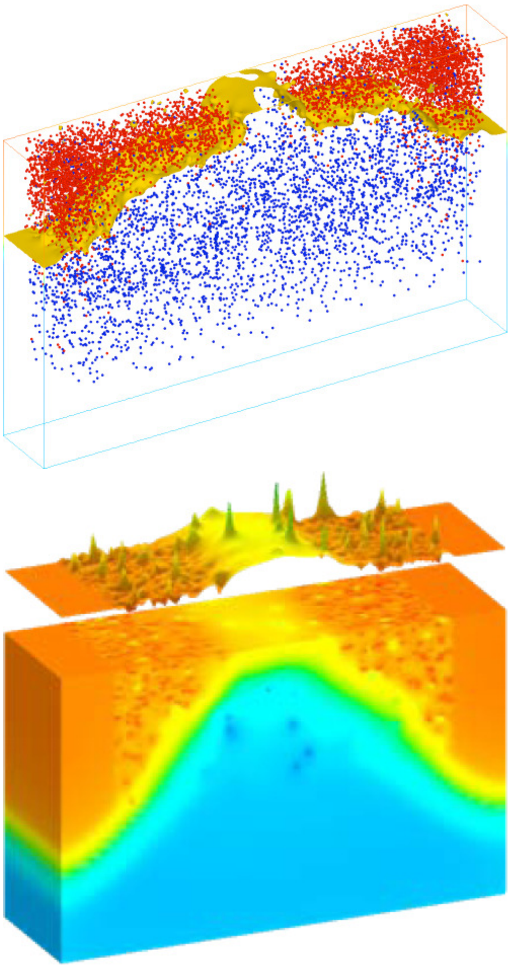
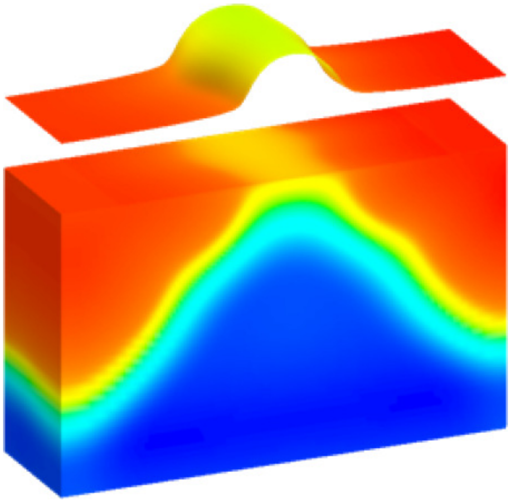


Figure 2.9: *(Left)* Si/SiO$_2$ interface profile (top), electron iso-concentration surface, and electrostatic potential, simulated in a 30 nm gate length MOSFET at high V_G and low V_D (27) (figure courtesy of A. R. Brown).

Figure 2.10: *(Right)* Discrete impurity distribution, resulting from a 3D process simulation, and electron concentration iso-surface, obtained from drift-diffusion/density-gradient simulations (53) (top) (figure courtesy of A. R. Brown). Effects of the discrete impurities on the surface potential (surface plot) and electron density distribution in a 30 nm MOSFET (3D slab).

An additional limitation of the approach concerns the purely classical treatment of the device electrostatics, and the modelling of the discrete impurities through their long range potential only. Therefore, discarding any correlation between oxide field fluctuations and gate leakage variability is not definitive.

Another possible source of intrinsic gate leakage variability is line-edge roughness (LER).



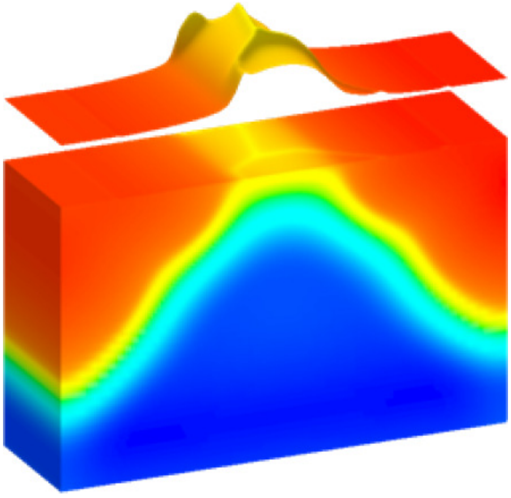


Figure 2.11: *(Left)* Micrograph of parallel lines with sub-100 nm width, showing uncertainty in the line edge, at a length scale comparable to contemporary MOSFET gate length (top). Effect of line edge roughness on the electrostatic potential of a 30 nm MOSFET (bottom) (figure courtesy of A. R. Brown).

Figure 2.12: *(Right)* Micrograph of a large poly-Si area, showing the random poly-grain size, shape and orientation, and the very well defined grain boundaries (top). Effect of the poly-Si granularity on the device electrostatic potential, for grain boundaries across the gate (figure courtesy of A. R. Brown).

Line edge roughness is caused by the tendencies of the lithographic photoresist to aggregate in polymer chains. These aggregates are large enough to locally affect the speed of the resist development process, and eventually translate to loss of resolution and low fidelity of the line edge, as illustrated in Fig. 2.11. This is of particular importance for the formation of the gate pattern, and translates to an uncertainty of the gate length along the width of the device. The impact on the device electrostatic potential is also shown in Fig. 2.11. Although the gate current is linearly proportional to the gate dimensions, it must be kept in mind that the distribution of random impurities forming the source and drain extension is correlated to the gate line edge roughness.

In ultra-scaled MOSFETs, the poly-Si gate grain boundaries also become a source of gate leakage dispersion between devices (54). This is explained with the segregation of ions at the grain boundaries which sharply alter the surface potential along those grain boundaries. This is shown in Fig. 2.12. One may expect that similar effect are to be found due to grains in high-κ gate dielectric stacks, but none of these effects is further investigated.

It appears therefore, that prior to our study, there was no attempt to thoroughly address the intrinsic gate leakage variability. This is in spite of the fact that gate leakage variability has been suggested as a scaling roadblock (due to apparent V_T fluctuations induced by the gate tunnelling current fluctuations) (55), and despite the need to account for it in circuit and system design and optimisation (34), and parametric yield modelling (46).

2.4 Summary

While device scaling aims to reduce device dimensions, to achieve improvements in integration density and performance, leakage power grows in an exponential fashion. The design and optimisation process becomes performance and power constrained. In this paradigm, the accurate evaluation of transistor variability in both performance, and power is of paramount importance, as this information critically impacts design margins, and hence yield. Of the two dominant leakage mechanisms, only the dependences of sub-V_T drain leakage variability are understood, while gate leakage variability, despite its importance, has not been studied in detail.

Chapter 3

Leakage through ultra-thin gate oxides

A brief overview of the known tunnelling mechanisms through ultra-thin oxides and their relevance to gate leakage in nano-scale CMOS transistors is followed by a survey of direct tunnelling models, and gate current simulation approaches. A semi-classical model, based on a modified WKB approximation, and its incorporation in a 1D *Poisson-density gradient* solver are presented in greater detail. A single value of the oxide tunnelling effective mass provides a good fit to experimental data for gate leakage from both inversion and accumulation layers, for a broad range of oxide thicknesses.

3.1 Gate oxide leakage mechanisms

Tunnelling in solids comprises a range of phenomena (56), a few of which give rise to gate leakage current in the context of a MOSFET. In a semiconductor-oxide-semiconductor structure, different tunnelling processes take place depending on the applied bias, oxide thickness, and the conditions of electrical and thermal stress. One may distinguish Fowler-Nordheim tunnelling (FNT), direct tunnelling (DT), and trap-assisted tunnelling (TAT), as is schematically illustrated in Fig. 3.1. This classification is based on the analytical, semi-classical models that have been established to explain the tunnelling current-voltage characteristics of the structure under different experimental conditions (37; 57; 58; 59; 60). Each of these processes contribute to a different degree to the gate current of a CMOS transistor, and the reasons are discussed below.

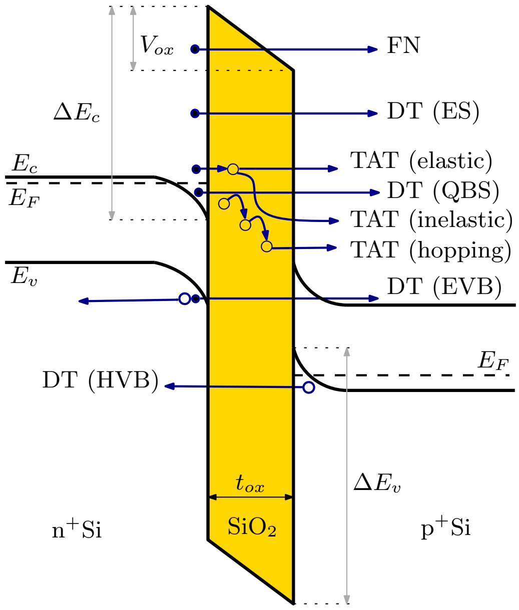


Figure 3.1: Conceptual diagram of tunnelling processes through an ultra-thin gate oxide: Fowler-Nordheim tunnelling (FNT) to the oxide conduction band, trap-assisted tunnelling (TAT) through the oxide band-gap, direct tunnelling (DT) through the oxide band-gap. DT may be due to conduction band electrons in extended (ES) or quasi-bound states (QBS), or from valence band electrons. Holes may tunnel in the DT regime in an identical fashion to electrons.

3.1.1 Fowler-Nordheim tunnelling

FN tunnelling is associated with electrons reaching the SiO_2 conduction band through a triangular potential barrier. In nano-scale MOSFETs, where the supply voltage is in the order of 1 V, one expects relatively few electrons to possess enough energy to do so. This is because the Si-SiO_2 conduction band discontinuity of 3.15 eV is large (37), while at the same time the voltage drop due to the built-in potential and the applied bias (both in the order of 1 V) is distributed across the oxide and the space charge layers, resulting in an oxide field of about 8 MV/cm, i.e. 0.8 V across 1 nm oxide. Hence electrons need an excess of 2 eV of incident energy. Quantitative insight is obtained from the simplest expression for the FN tunnelling current $J_{FN} = F_{ox}^2 C \exp(-\beta/F_{ox})$ (61), which in good agreement with experiment (61; 62; 63). Here, F_{ox} is the oxide field, while C and β are constants, involving the potential barrier height ϕ_B and oxide effective mass m_{ox}, so that J_{FN} is not explicitly dependent on the oxide thickness. [1] Assuming $\phi_B = 3.15$ eV, and $m_{ox} = 0.5m_0$ as in Ref. (63), we obtain J_{FN} of 1.0×10^{-12} and 2×10^{-4} A/cm^2, for F_{ox} of 6 and 10 MV/cm, respectively. These values appear to be orders of magnitude lower than the measured tunnelling current, even in oxides as thick as 3.2 nm at similar F_{ox} (59; 64; 65; 66). Figure 3.2 shows the vast difference between measured gate leakage from sub-5 nm oxides, and the calculated current J_{FN}, based on the expression

[1] Experimental results are typically presented on a plot of $log(J_{FN}/F_{ox})$ versus $1/Fox$, to which a straight line can be fitted, according to the quoted relation, over a wide range of F_{ox} value.

above. This discrepancy indicates that there is a dominant transport mechanism, different from FNT, taking place in ultra-thin oxides (i.e. t_{ox} below 4 - 5 nm). The data shown in Fig. 3.2, and from various other studies, has been reproduced to a good agreement by modelling the gate current as direct tunnelling current of electrons (57; 65; 67; 68).

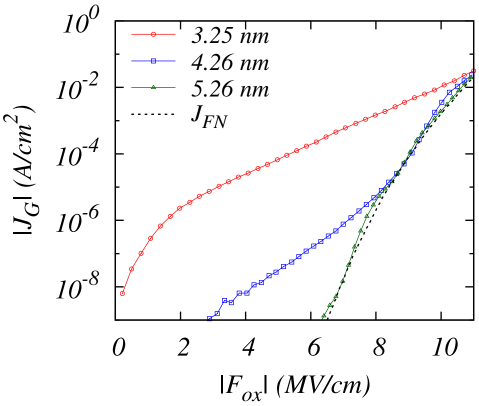


Figure 3.2: Measured (65) gate leakage current from n$^+$poly-Si gate into a p-Si(001), for three different oxide thicknesses, as a function of the oxide field. Significant disagreement with the calculation of J_{FN} using the relation quoted earlier in the text, shows there is another, more important tunnelling mechanism at sub-5 nm oxide thickness, which was successfully modelled as a direct tunnelling process (65; 68).

3.1.2 Direct tunnelling

In the direct tunnelling regime, electrons traverse the entire width of the potential barrier that is imposed by the conduction band discontinuity (56). The most important feature of the DT regime is the experimentally established exponential sensitivity with respect to oxide thickness (64; 65; 69). [1] As the barrier gets thinner, the tunnelling probability dramatically increases even for electrons near the bottom of the conduction band (i.e. with a very small incident energy, compared to the top of the barrier). Moreover, the density of these electrons is much higher, than the ones near the top of the barrier, and as a consequence the DT electron flux, dominates the gate leakage through sub-5 nm SiO$_2$ films, as was shown in Fig. 3.2.

Direct tunnelling in MOS structures exhibits a weak temperature dependence (associated with the temperature dependence of the Fermi level on each side of the oxide) (70; 71), allowing successful interpretation of experimental data in the framework of a single particle, elastic tunnelling (56). A survey of models for the simulation of DT gate current in MOS structures, and the associated quantitative results, are deferred until the next two sections. Here, it suffices to stress, that a two-fold reduction of the oxide thickness t_{ox} from 3 nm to 1.5 nm increases the DT gate current by more than seven orders of magnitude (69; 72), with a current density at F_{ox} = 8 MV/cm greater than 10 A/cm^2, for the thinner oxide.

[1] Exponential sensitivity of the gate current J_G on oxide thickness t_{ox} is evident from a plot of $log(J_G t_{ox}^2)$ versus t_{ox}, where the data of all devices for a given bias falls on a straight line (69).

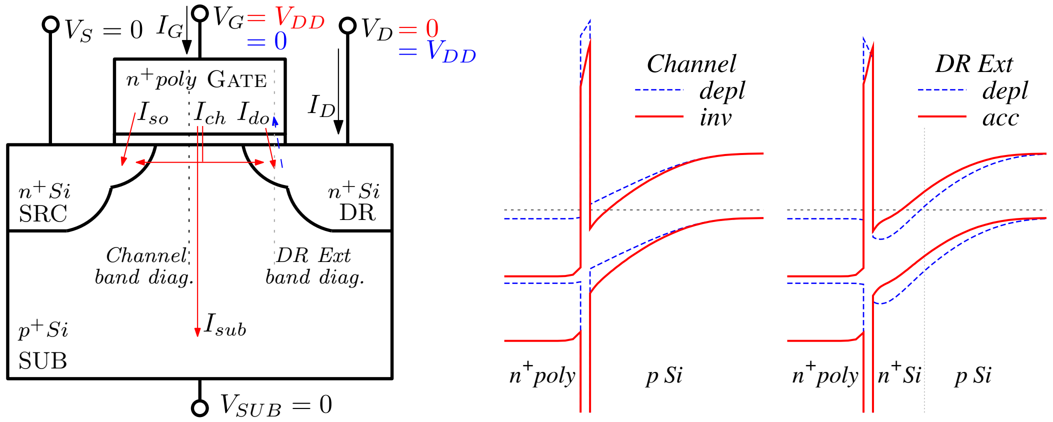


Figure 3.3: Schematic, 2D diagram of an nMOSFET, and simulated energy band-diagrams at the middle of the channel (*Channel*), and at the end of the overlapped drain extension (*DR Ext*). Two extreme biases are shown: $V_G = V_{DD}, V_D \sim 0$ (red, solid), and $V_G = 0, V_D = V_{DD}$ (blue, dash). The former case implicates electrons tunnelling from substrate, while the latter one - mainly from the gate, to the drain extension. The same energy reference is used, coincident with the equilibrium Fermi level deep in the Si substrate.

As shown in Fig. 3.1, DT may involve electrons from the conduction band (commonly termed ECB), or from the valence band (EVB). *Electrons in the conduction band* can be conceptually split in two groups - quasi-bound state (QBS) electrons, and extended state (ES) electrons. The former are confined to the narrow accumulation or inversion quantum well next to the oxide interface, and have the properties of a 2D electron gas (73). The term *quasi-bound* reflects their ability to tunnel through the oxide to unoccupied extended states in the gate, although the energy levels that these electrons occupy correspond to the stationary (bound) states of the well. [1] The ES electrons occupy the continuum of energy levels above the surface quantum well, and have the properties of a 3D electron gas. The presence of electrons with different properties bears great relevance to the tunnelling models, since accurate determination of the available amount of charge for tunnelling is of crucial importance.

Figure 3.3 shows a schematic diagram of an n-channel MOSFET and band diagrams in the direction normal to the oxide at two different geometrical positions, and two different bias conditions. In the case of high gate voltage and low drain voltage, carriers in the substrate are confined, and occupy QBS (refer to the red, solid lines in the band-digrams). Tunnelling from ES may be ignored due to their negligible occupancy. Moreover ES electrons have to overcome

[1] Assuming a weak coupling between the cathode (the inversion/accumulation layer) and the anode, so that equilibrium distribution in the cathode is insignificantly disturbed from leakage.

the depletion layer in the substrate before reaching the oxide interface and tunnel. This is a relatively slow process compared to the tunnelling rate through an ultra-thin oxide, due to the wide depletion layer. ES electrons must be considered if electrons are the majority carriers in the tunnelling cathode (e.g. the gate in accumulation), since at a low confining potential, ES electrons form a significant fraction of the total electron population (refer to the blue, dashed line of the *DR ext* band-diagram, in the case of low gate voltage and high drain voltage).

Tunnelling of electrons from the valence band (TEVB) results in a hole left behind, at the cathode, as illustrated in Fig. 3.1. In the context of an n-channel transistor in a CMOS logic circuit, TEVB could happen only under sufficiently strong inversion of the substrate (i.e. high gate bias), in which case substrate valence band electrons tunnel to available states in the conduction band of the gate, contributing to the gate current I_G, while the holes left in the substrate contribute to the substrate current I_B (74; 75). TEVB requires that the *valence* band edge at the cathode interface is higher than the *conduction* band edge at the anode interface. Otherwise, valence band electrons meet a wider tunnelling distance than conduction band electrons, due to the band bending in each of the Si electrodes, as shown in Fig. 3.1, and TEVB is comparatively insignificant. Ignoring band-gap narrowing, and assuming the intrinsic Fermi level to be in the middle of the Si band gap, it is easy to translate this requirement into a condition for the oxide field strength, namely $|F_{ox}|t_{ox} > E_g/q$. If this condition is satisfied, the flux of VBE tunnelling electrons is limited only by the tunnelling probability, since electron density in the valence band is high. However, at this stage the Fermi level in the cathode is above the conduction band edge, and conduction band electrons have high density too. Moreover, conduction band electrons have higher tunnelling probability, because the potential barrier for them is lower by $E_g = 1.12$ eV, and TECB remains the dominant component of the gate current ($I_B/I_G < 0.2$) (76). A voltage drop above $E_g = 1.12$ eV is in fact not possible if the power supply V_{DD} is in the order of 1 V, and therefore TEVB may be disregarded for the study of CMOS gate leakage variability. It should be kept in mind for tunnelling model evaluations and calibration however, since oxide characterisation data is often available up to high-voltages where TEVB is not negligible.

3.1.3 Trap-assisted tunnelling

Point-like defects in the insulator contribute electronic states that are energetically localised in the oxide band-gap. These states can trap charge, due to their spacial localisation, or facilitate tunnelling through the oxide - the process being called trap-assisted tunnelling (TAT). There is a finite probability of an electron from the cathode (the emitting electrode) tunnelling to the trap, and this probability is higher than the one for tunnelling through the entire oxide at the energetic level of the trap. There is of course a finite probability of tunnelling out of this trap to the anode at opposite side. This simplified description of the process renders TAT as a two-step tunnelling phenomenon, and forms a basic model for quantitative understanding through the relation (77) $J_{TAT} = q \int N_T/(\tau_c + \tau_e)dx$, where the $N_T(x)$ is the trap density at a distance x away from the tunnelling cathode, τ_c and τ_e are the characteristic times for carrier *capture* and *emission* by the trap (dependent on the energy and space location of the trap, and the oxide field), and integration is performed over the oxide distance. TAT models of various complexity are reviewed in Ref. (78), differing in the way that *capture* and *emission* times are calculated, and in their account for hopping (through a sequence of traps) and inelastic TAT (79; 80; 81; 82; 83).

TAT modelling is of greatest importance for the accurate simulations of stress-induced leakage current (SILC) – the increase in gate leakage of a device subjected to an electrical stress, relative to the leakage of the unstressed device.. Since SILC is attributable to the generation (due to stress) of traps in the oxide, it is used to extract their density, which is linked to oxide degradation and breakdown (84; 85; 86).

In good quality sub-3 nm thick oxides, the trap concentration is less than 10^{17} cm^{-3} (87; 88; 89), and 10 times lower than the concentration in bulk oxides (90). Note that the corresponding areal density of traps in a 1 nm oxide is 10^{10} cm^{-2}, which is in the order of the electron sheet density in a MOS tructure with highly doped substrate at zero gate voltage. Therefore J_{TAT} is limited by N_T, and should affect gate current mainly at low bias, where the DT or FNT components are relatively small. This hypothesis is confirmed by experiment – a comparison between $I_G - V_G$ characteristics of unstressed and stressed devices with identical structure shows a characteristic flaring of the IV curves of the stressed device at low V_G, and a corresponding increase in the tunnelling current (64; 85).

Experimentally, TAT in sub-2 nm SiO$_2$ appears to be negligible (91), and remains a concern only for non-volatile memory devices (78), and MOSFETs with high-permittivity dielectric gate stacks (92; 93).

3.1.4 Hole tunnelling

The tunnelling mechanisms discussed so far for electrons are available for holes too. However, tunnelling for holes happens at a lower rate, compared to electrons for two main reasons – *i)* holes have higher tunnelling effective mass (94), and *ii)* the potential discontinuity for holes is very significantly larger than that for electrons (compare $\Delta E_v \sim 4.8$ V, and $\Delta E_c \sim 3.1$ V for SiO_2(37)). This is why p-channel MOSFETs typically exhibit lower gate leakage current than n-channel MOSFETs at a give channel length, as hole tunnelling current dominates gate leakage in p-channel MOSFETs with p^+poly-Si gate (94; 95).

In line with our objectives to study gate leakage in variability in n-channel nano-CMOS FETs, we focus, in the subsequent exposition, entirely on the primary tunnelling process – direct tunnelling of conduction band electrons.

3.2 Direct tunnelling models

Here we give an overview of the models describing direct tunnelling of electrons emitted from a semiconductor cathode, adopting the independent-particle point of view (96). Tunnelling is viewed as a one dimensional process, with the customary approximations of total particle energy conservation, effective mass approximation in all areas of the tunnelling junction, translational invariance in lateral direction, and a phenomenological treatment of the barrier (56). The last aspect was implicit in the energy band-diagrams presented in the previous section, and concerns the barrier shape, height $\Delta E_{c,v}$ (conduction/valence band discontinuity), and width t_{ox} (oxide thickness), which are determined from tunnelling-independent experiments. In addition, tunnelling is assumed to be sufficiently weak, as to not disturb the equilibrium in the cathode – a condition typically satisfied in MOSFETs under normal operating conditions.

3.2.1 Generalised expression

The expression for the electron tunnelling current density from quasi-bound (QBS) and extended states (ES) can be written as (30; 56; 78; 97)

$$J_G = q \sum_{\nu,\iota} \frac{n_{\nu,\iota}}{\tau_{\nu,\iota}} + q \int_{q\psi_s}^{\Delta E_c} T(E_\perp) N(E_\perp)\, dE_\perp. \tag{3.1}$$

The sum in Eq 3.1 accounts for tunnelling from QBS, where the index ι identifies a subband, in a valley ν, of the quantised layer in the semiconductor, and $n_{\nu,\iota}$ is the corresponding sheet density of electrons. Each electron tunnels with a rate $1/\tau_{\nu,\iota}$, where $\tau_{\nu,\iota}$ is the characteristic lifetime of the electron QBS.

The integral accounts for tunnelling from ES with a higher energy than the depth of the potential well, $q\psi_s$ (assuming the bottom of the potential well as an energy reference, and ψ_s

being the surface potential). The integrand functions T and N are transmission probability and supply function respectively (56), and $E_\perp$ is the incident kinetic energy of a particle with a total energy $E = E_\perp + E_\parallel$. The supply function itself is (98)

$$N(E_\perp) = \frac{4\pi m^*_{3D}}{h^3} \int_0^\infty (f(\xi_c, E) - f(\xi_a, E))\, dE_\parallel, \tag{3.2}$$

where m^*_{3D} is the 3D density of states mass in the electrodes, $f(E)$ is the occupational probability function (i.e. Fermi-Dirac or Maxwell-Boltzmann distribution at thermal equilibrium),[1] and $\xi_{c,a}$ is the Fermi energy on each side of the insulator, i.e. the cathode and the anode.[2]

Finding the transmission probability is the major challenge associated with the expression for the tunnelling from ES. Moreover, $T(E_\perp)$ is characteristic for the barrier penetrability at a given energy, independent of the supply function, and some models for tunnelling from QBS also make use of it. Next we elaborate on the methods of its calculation.

3.2.2 Transmission Probability

The transmission probability $T(E_\perp)$ is defined as the ratio of the transmitted and incident probability currents (100). Note that it implies itinerant states on each side of the barrier, i.e. a plane wave, incident on one side of the potential barrier, connected (through a decaying wave function in the barrier) to another plane wave, on the other side of the barrier. The plane waves are not normalizable, but as we are interested in the ratio of their densities, transmission is a well defined concept. Finding T, given a trapezoidal potential barrier, is a typical quantum-mechanics textbook problem, which is solved *i)* semi-analytically, deploying *Airy functions* (101; 102), *ii)* numerically, using the *transfer-matrix* (TM) method (103), or *iii)* approximately, using the *Wentzel-Kramers-Brillouin* (WKB) approximation (98).

The semi-analytical solution is computationally intensive, but not extensible to stacked dielectrics, which is why it has gained limited popularity (102). The TM method is also computationally demanding, but can be applied for multi-layered dielectric stacks and barriers of arbitrary shape. It is typically embedded in self-consistent Poisson-Schrödinger solvers (104; 105; 106; 107; 108), and the method is elaborated in Appendix. A. The dependence of the accuracy and stability of the TM method on the grid density is a disadvantage, particularly with a 3D simulation framework in mind. The grid resolution dependence is circumvented by

[1] By selecting an appropriate non-equilibrium distribution functions for f, it is possible to more accurately account for hot-carrier tunnelling (78; 99).

[2] Equation 3.2 assumes electrodes are described by the same parabolic band-structure (i.e. the same m^*_{3D}), and when applied to the integral of Eq. 3.1, accounts for the bidirectional flux through the barrier.

the development of the *quantum transmitting boundary* (QTB) method (109), which requires however, the inversion of a complex matrix with the dimensionality of the number of grid points. The TM and the QTB approaches are therefore not considered for the study of gate leakage variability at this stage, because of their implementation complexity and computational cost.

Here we focus on an improvement of the WKB approximation, which, for the DT regime, yields results in excellent agreement to the semi-analytical approach involving Airy functions. The transmission probability in this case is (1; 110)

$$T = T_R \times T_{WKB} \tag{3.3}$$

with the standard WKB transmission probability being (100)

$$T_{WKB} = exp(-2\int_0^{-t_{ox}} \kappa(x)\, dx) \tag{3.4}$$

and the correction factor, in a band-structure-independent form, being (1; 110)

$$T_R = \frac{4v_k(0)v_\kappa(0)}{v_k^2(0) + v_\kappa^2(0)} \times \frac{4v_k(-t_{ox})v_\kappa(-t_{ox})}{v_k^2(-t_{ox}) + v_\kappa^2(-t_{ox})}, \tag{3.5}$$

where $k(x)$ and $i\kappa(x)$ are the real and imaginary electron wave vectors, outside, and inside the barrier, respectively, while v_k and v_κ are the corresponding real and imaginary group velocities ($dE/d(\hbar k)$ and $dE/d(\hbar\kappa)$, respectively). Hereafter, we refer to the transmission probability given by the above set of equations as *modified WKB* (m-WKB) approximation.

The quality of the "correction" is demonstrated in Fig. 3.4, which compares the transmission probability through a trapezoidal barrier, calculated with Airy functions (i.e. the exact solution), with the conventional WKB approximation (i.e. Eq. 3.4), and with the m-WKB approximation (Eqs. 3.3 to 3.5).

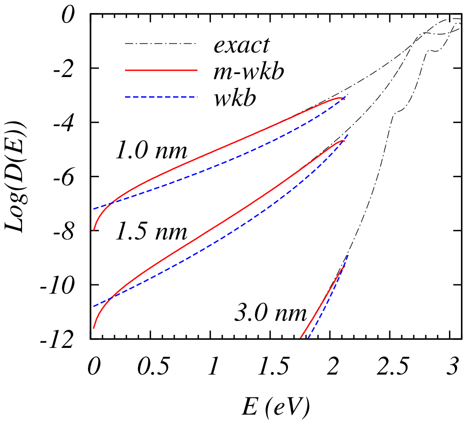


Figure 3.4: Transmission probability as a function of the incident energy, for a trapezoidal barrier of 3.15 eV height, and width as indicated. Voltage drop accross the barrier is 1 V, and potential is flat in the electrodes. In the DT regime, there is an excellent agreement between the modified-WKB (m-WKB) and the exact calculation based on Airy functions. The discrepancy with the traditional WKB is quite obvious.

To elucidate the physical significance of the correction term T_R, we first recall that the derivation of the conventional expression of T_{WKB} is done for smoothly varying potential barrier, using the WKB approximation for the wave function in all regions of its validity (i.e. where $|dk/dx| << |k^2(x)|$), and with the help of connection formulas (100). As the connection formulas are invalid at potential discontinuities, Ref. (100) suggests that in such cases, the exact wave function solutions for each region are used, and matched smoothly at the points of discontinuity. This procedure is followed in the derivation of Eq. 3.3 (110).[1] Plane waves are assumed on each side of the barrier, and connected to real exponents at the inner (for the barrier) side of the potential discontinuities. The matching conditions ensure wave continuity and probability current conservation, and give rise to the correction term T_R. Wave propagation only inside the barrier is treated with the WKB theory, reflected in the T_{WKB} term. Therefore the correction term T_R accounts for the wave reflections at the abrupt potential discontinuities (1). Note that in the limit of a square barrier and a single effective mass throughout, the m-WKB expression for T reduces to the exact analytical solution for low incident energy (or $\kappa t_{ox} >> 1$) (103)

$$T_{sq} = \frac{16k(-t_{ox})k(0)\kappa^2}{(k^2(-t_{ox}) + \kappa^2)(k^2(0) + \kappa^2)} exp(-2\kappa t_{ox}). \tag{3.6}$$

As a final comment we show that the requirement for the validity of the WKB approximation, $|d\kappa/dx| << |\kappa^2(x)|$, is satisfied. Considering 1 V voltage drop over 1 nm thick oxide, i.e. F_{ox} = 10 MV/cm, and an electron with $0.5m_0$ effective mass (hence parabolic band structure in the oxide) tunnelling from 3 eV below the top of the trapezoidal barrier, we evaluate

$$|\kappa^2(x)|\left|\frac{d\kappa(x)}{dx}\right|^{-1} = \kappa^2(0)\frac{t_{ox}}{(\kappa(0) - \kappa(-t_{ox}))} = 9.44 \tag{3.7}$$

Therefore, the WKB approximation is acceptable for the purpose of our study. In addition, tunnelling calculations using the m-WKB expression for the transmission probabilities have already shown good agreement with experiment (72; 91; 94; 95; 110; 112).

3.2.3 Tunnelling from quasi-bound states

A rigorous treatment of tunnelling from a quasi-bound state (QBS) requires the solution of the time dependent Schrödinger equation. The solution is of the form (113)

$$\Psi(x,t) = \psi(x)e^{-\lambda_n t/2}e^{-i(E_n+\delta E)t/\hbar}. \tag{3.8}$$

[1] Alternative derivations of the same expression for the transmission probability are obtained independently, following Bardeen's transition probability approach (111), and Harrison's independent-particle tunnelling model (96), in Refs. (95; 112).

This assumes the maximum probability of finding the particle in the quantum well initially, at $t = 0$, and hence at a corresponding bound state (BS) level E_n. The coupling of the QBS with the continuum of states in the gate leads to the following consequences. First, the probability of finding the particle in the well after time $\tau_n = 1/\lambda_n$ is $1/e$, i.e. τ_n is the characteristic lifetime of the QBS. Second, there is a small shift δE of the QBS energy, with respect to the BS level, which is due to the coupling of the QBS to the continuum of states in the gate.

One could follow the time dependent formulation to find the leakage current from a QBS, and obtain that the tunnelling current exponentially decreases in time (114), similarly to an α-particle decay (113). The case with QBS in a MOSFET is different, since for sufficiently weak coupling through the gate oxide (assumed at the onset of this section), the tunnelled particle is instantly replaced through the contacts, and the steady state tunnelling current is constant. This instantaneous refill of the QBS allows one to ignore its time dependence altogether, and determine the characteristic lifetime τ_n, from the properties of the corresponding stationary state E_n (114). This is by far the most common approach in device modelling (115). Therefore, the challenge in modelling QBS leakage current is in the calculation of the QBS lifetime, while the subband sheet density is usually obtained from a self-consistent solution of the Poisson and time-independent Schrödinger equations, or an analytical approximation of the inversion layer. Subsequently we elaborate on the methods of calculating QBS lifetimes.

3.2.3.1 Complex eigen-states

The time dependent factor in Eq. 3.8 can be re-written in the form $exp(-i(E_n - i\lambda_n\hbar/2)t/h)$ (ignoring δE). Therefore, one can obtain $\Psi(x,t)$ from the stationary state Schrödinger equation for $\psi(x)$, allowing for complex eigen-energies of the form

$$\mathcal{E}_n = E_n - i\hbar/2\tau_n = E_n - i\Gamma_n. \tag{3.9}$$

This suggests a direct method of obtaining the QBS lifetimes, from the imaginary part of the complex eigen-states of the time-independent Hamiltonian of the system. However, the coupling of the QBS to the semi-infinite domain of the anode requires the application of open boundary conditions. This is realised in the quantum transmitting boundary method (QTBM), but the Hamiltonian becomes non-linear, because some of the coefficients of the augmented matrix (incorporating the BC) are energy dependent (109). Despite its complexity, the method has been successfully used for the simulation of gate current (66; 116; 117). A recently proposed alternative, the perfectly matched layer (PML) method, based on artificial absorbing boundary conditions, makes the Hamiltonian appear linear, non-Hermitian, for finding the eigen-states of which there are more efficient algorithms, and agrees well with the QTBM results (118).

The complex eigen-value problem can also be solved using an analogy to the transmission line theory, in a method referred to as the general impedance concept (119), or as the transfer-resonance method (120). It is successfully used in several works (120; 121; 122; 123).

3.2.3.2 Reflection time and transfer-matrix

There is a widely used alternative of calculating QBS lifetimes. It is based on a consideration of the reflection time t_r, associated with extended states, incident from the cathode, and reflected back to it (124). It is found that an incident wave with energy coinciding with a QBS suffer much longer reflection time, compared to incident waves at energies away from the QBS. Moreover, the energy dependence of t_r in the vicinity of a QBS is described by a Lorentzian, with a half width at half maximum (HWHM) $\Gamma_n = \hbar/2\tau_n$, τ_n being the QBS lifetime (124). Finding τ_n is thus reduced to the problem of finding Γ_n.

There are two ways of finding Γ_n, based on symmetry properties of the transfer-matrix. The first is to evaluate the derivative of the phase, $\Theta(E)$, of the complex reflection coefficient, since $d\Theta(E)/dE$ exhibits the same resonances as the reflection time (125). This approach requires a very fine step of energy scanning ($\sim$1 peV) to estimate numerically $d\Theta/dE$, and subsequently, the width of the Lorentzian peaks (105; 114). The second approach relies on the width of a Lorentzian peak in the energy dependence of $1/|W_{11}|^2$, W being the transfer matrix (115). Note that $1/|W_{11}|^2$ is the principal factor in the transmission coefficient of a resonant system (125). Although transmission is not a valid concept in the case of QBS, the lack of incident waves imposes that $W_{11}(\mathcal{E}_n) = 0$, with $\mathcal{E}_n = E_n + i\Gamma_n$. Consequently, assuming $W_{11} \propto (E - \mathcal{E}_n)$ in the vicinity of E_n, it is straight forward to show that $1/|W_{11}|^2$ has a Lorentzian form around E_n, with a HWHM of Γ_n. Eliminating the need for numerical evaluation of $d\Theta/dE$, this approach is easier to implement (107; 115).

We implement the second method for QBS lifetime calculation in the Poisson-Schrödinger solver described in Appendix A, and use it for the work presented in Chapter 5.

3.2.3.3 Wave-function logarithmic derivative

The QBS lifetime can also be obtained from the wave-function of the stationary state $\psi(x, E)$, following the approach recently presented in Ref. (126). For a real E, ψ can be real too (and normalizable, except in the anode, where the QBS leaks into extended states), and one can define a real function $G(x, E) = (1/\psi)(d\psi/dx)$ (i.e. the logarithmic derivative of ψ). G can be *analytically continued* to assume a complex argument (with a small imaginary component) of the form $\mathcal{E} = E - i\Gamma$, using first order Taylor series (implicitly dependent on x):

$$G(\mathcal{E}_n) = G(E_n) + \frac{dG}{dE}\Big|_{E=E_n} (\mathcal{E}_n - E_n). \tag{3.10}$$

Taking $x = 0$ to be the interface between the oxide and the anode (e.g. the gate, for inversion layer QBS), and an outgoing wave for $\psi \propto exp(-ikx)$, for $x < 0$, it follows that $G(x, \mathcal{E}) = -ik$. By implication, $G(0, E_n) = 0$, since for real energy, $x = 0$ is a classical turning point of $k = 0$. Therefore, Eq. 3.10 yields the following relation for Γ_n, and hence for the QBS lifetime τ_n (with the help of Eq. 3.9)

$$\Gamma_n = \frac{k}{dG(0, E)/dE} \Rightarrow \tau_n = \frac{\hbar}{2k} \frac{dG(0, E)}{dE}\Big|_{E=E_n}. \tag{3.11}$$

An expression for dG/dE can be obtained from the time-independent Schrödinger equation, written for two infinitesimally differing energies, E_1 and E_2, such that $E_2 - E_1 = dE$, and taking the difference of the two equations.[1] The result, obtained in the original reference (126) for real wave-functions, and assimilated in Eq. 3.11, reads for τ_n

$$\tau_n = \frac{m(0^-)}{\hbar k(0^-)} \frac{1}{\psi^2(0, E_n)} \int_0^\infty \psi^2(x, E_n)\, dx, \tag{3.12}$$

where $m(0^-)$ and $k(0^-)$ are the effective mass and wave number in the anode side of the barrier at $x = 0$.

The remarkable convenience of this approach originates from its dependence only on the wave-function, which is typically required in a self-consistent calculation anyway, and hence is known. This eliminates the need for additional scanning in energy and searching for resonance peaks, which we found to be numerically challenging, especially for relatively low confinement and small subband level differences. Consequently, we implemented this approach in the Poisson-Schrödinger solver described in Appendix A. The numerically obtained wave-functions are complex in this case, and we obtain the QBS lifetime (with $\psi_n(x) = \psi(x, E_n)$)

$$\tau_n = \frac{m(0^-)}{\hbar k(0^-)} \frac{1}{\psi_n(0)\psi_n^*(0)} \int_0^\infty \psi_n(x)\psi_n^*(x)\, dx. \tag{3.13}$$

Agreement with the lifetimes obtained from this equation, and from the HWHM of $1/|W_{11}|^2$ resonance, discussed in the previous subsection, is very good.

[1] The technique is similar to the one used in the derivation of the quantum-mechanical current operator (103).

3.2.3.4 Impact frequency and barrier transparency

The earliest conceptual model of tunnelling from QBS consists of a particle oscillating with a given energy E_n between two potential barriers, at least one of which has a finite transparency $T_n(E_n)$, thus giving rise to an escape rate

$$1/\tau_n = f_n T_n, \tag{3.14}$$

where τ_n is the QBS lifetime, as previously defined. Here, f_n is the number of incidences of the particle on the barrier, per unit time, hence referred to as impact (or attempt) frequency (127). For a semi-classical wave-packet representation of the particle, one can obtain the impact frequency from the reciprocal of the round-trip time between the classical turning points (97)

$$f_n^{-1} = 2 \int v^{-1}(x)\, dx \tag{3.15}$$

where $v(x) = \sqrt{(2/m)(E_n - U(x))}$ is the classical (group) velocity, integration is performed between the two classical turning points (where $E_n = U(x)$), and the factor of two accounts for the full opacity of one of the barriers. This definition of f is formally equated to the quantum-mechanical propagation time defined through the energy derivative of the phase change induced by a round trip of the wave function $\psi(E_n)$ (127).

The other ingredient of Eq. 3.14, T_n, can be found by any of the methods described earlier for the calculation of the tunnelling probability.

This model has the advantage of simplicity, since it does not explicitly require numerical evaluation to obtain τ_n, once E_n is known (e.g. from a self-consistent Poisson-Schrödinger solver), and has found very wide use (97; 106; 128). It can be coupled to approximative solutions of the quasi-levels (or even bound states) of the quantum well (1; 72), and forms the basis of the fully analytical tunnelling model described next.

3.2.4 Register's model

Here we link together the model of impact frequency and the modified-WKB expression for the tunnelling probability, to an approximation of the quantisation effects and sheet-density of the semiconductor cathode, and obtain an analytical oxide-field dependence of the gate current density. We refer to this tunnelling model as Register's model, after its first author (1).

The gate tunnelling current density in this model is

$$J_G = Q_s f(E_\perp) T_{mWKB}(E_\perp), \tag{3.16}$$

where Q_s is the sheet charge available for tunnelling, $f(E_\perp)$ is the impact frequency given by Eq. 3.15, T_{mWKB} is the tunnelling probability given by Eqs. 3.3 to 3.5. In this form, the model assumes that a single subband, at energy $E_\perp$, contains all inversion or accumulation charge in the quantised layer. Energy is referenced from the bottom of the conduction band at the oxide interface of the cathode, $E_c(x = 0)$. Next we present the analytical relations of these quantities to the oxide field F_{ox}.

The quantum well is modelled as triangular, with slope $dE_c/dx = q(\varepsilon_{ox}/\varepsilon_{Si})F_{ox}$, where ε_{ox} and ε_{Si} are the permittivity of $\mathrm{SiO_2}$ and Si, respectively. Assuming rigid walls for the quantum well, the wave-vector quantisation condition becomes

$$2\int_0^{x_c} k_\perp(x)\,dx = 2\pi, \tag{3.17}$$

where 0 and x_c are the classical turning points, and $k_\perp(x) = \sqrt{(2m_\perp/\hbar^2)(E_\perp - E_c(x))}$ reflects a parabolic $E(k)$ dispersion in Si, with an effective mass $m_\perp$. Solving Eq. 3.17 yields

$$E_\perp = 0.6\frac{(3\pi\hbar q m_\perp)^{2/3}}{2m_\perp}\left(\frac{\varepsilon_{ox}F_{ox}}{\varepsilon_{Si}}\right)^{2/3}, \tag{3.18}$$

where the factor of 0.6 compensates for the exaggerated confinement implicated by the assumed rigid walls, and is obtained by comparison to numerical simulations (1).[1] Knowing $E_\perp$, the interface-normal velocity $v_\perp(x) = \sqrt{(2/m_\perp)(E_\perp - E_c(x))}$ can be substituted in Eq. 3.15, and integrated from 0 to x_c to obtain the impact frequency

$$f = 0.6\frac{2q}{(3\pi\hbar q m_\perp)^{1/3}}\left(\frac{\varepsilon_{ox}F_{ox}}{\varepsilon_{Si}}\right)^{2/3}. \tag{3.19}$$

The expression for T_R is that of Eq. 3.5, independently of the $\mathrm{SiO_2}$ band-structure, while the expressions for T_{WKB} depend on the assumed oxide band-structure, i.e. parabolic or Franz-type ($k = \sqrt{(2m/\hbar^2)E(1 - E/E_g)}$, where E_g is the oxide band-gap (130)), and are stated in the original reference (1), and need not be repeated here.

The last quantity of interest is the amount of tunnelling charge Q_s. For accumulation, it is modelled as field induced, i.e. $Q_s = \varepsilon_{ox}F_{ox}$. For inversion, only the mobile inversion charge can tunnel, and therefore $Q_s = \varepsilon_{ox}F_{ox} - Q_D$, where Q_D is the depletion charge sheet-density (91).

The efficiency of this model is favourable for its inclusion, as a post-processing step, in any self-consistent field calculation, where F_{ox} and Q_D can easily be obtained from the solution of

[1] Recall that the text-book treatment, based on the WKB approximation, allows for wave penetration into the sloped barrier, changing the 2π in Eq. 3.17 to $3\pi/2$, thus leading to a prefactor of ~ 0.8 (129).

the Poisson's and charge-density equations (e.g. density-gradient (DG), or Schrödinger equations). The model has demonstrated sufficient accuracy, even in its simplest form presented here (1; 91), and is extensible to account for multiple subbands in the quantisation layer (72; 110). These qualities make it a suitable choice for the study of gate leakage variability, and its incorporation in a 1D Poisson-DG simulator, and in the 3D device simulator, will be elaborated upon in later parts of this work. To close the section of direct tunnelling models, we briefly address several aspects not mentioned so far.

3.2.5 On the underlying approximations

3.2.5.1 Effective mass approximation

If the true test for the validity of a theory is its relation to experiment, then the use of the effective mass approximation (EMA) in tunnelling models is justified at least for oxides down to 1.2 nm thickness, by the reasonably good fits to measured $J_G(V_G)$ data (55; 59; 106; 131). The answer is not so simple however, because of the number of phenomenological parameters that enter the equations – shape, thickness, and height, of the tunnelling barrier, and band-structure of the oxide.

More confidence is gained from microscopic calculations of tunnelling through ultra-thin oxides. In particular, Ref. (132) presents a tight-binding (TB) calculation of tunnelling in a Si/SiO$_2$/Si super-cell, accounting for the 3D atomic structure of the tunnelling junction, as well as the entire electronic structure of Si and SiO$_2$ up to several electron volts. Systematically comparing the TB and the EMA results, the authors conclude (in regard to electron tunnelling) that - *i)* transmission through oxide as thin as 0.4 nm can be qualitatively described with a bulk band structure, and depends exclusively on the incident (interface-normal) energy of the particle, up to 1.5 eV electron energy; *ii)* the EMA, with a single imaginary band in the oxide, remains adequate for modelling direct tunnelling in oxides as thin as 0.7 nm; *iii)* an energy dependent effective mass (e.g. Franz type $E(k)$ dispersion) gives qualitatively accurate voltage dependence of the transmission probability, although the quantitative overestimation by the EMA requires the band-edge effective mass, or the oxide thickness to be treated as a fitting parameter.

An additional support for the EMA applicability is the fact that Ohmic-conduction (i.e. $J_G \propto V_G$) at an infinitesimal bias, described to be a ground-state property in an *ab-initio* density-functional theory simulation (133), is demonstrated also in an EMA-based simulation (112).

3.2.5.2 Band-structure mismatch

The adoption of a single, imaginary band, to describe direct tunnelling through the oxide in the EMA, brings about the issue of momentum conservation, since at the same time, the Si cathode has 6-ellipsoidal band structure. The problem is understood by the following argument. Assuming that energy is referenced at the bottom of the Si conduction band at the Si/SiO$_2$ interface, and decomposing the total kinetic energy of the particle in normal ($\perp$), and parallel ($\|$), to the interface plane, the conservation of total energy requires that

$$\frac{\hbar^2 k_\perp^2}{2m_\perp} + \frac{\hbar^2 k_\|^2}{2m_\|} = \Delta E_c + \frac{\hbar^2 \kappa_\perp^2}{2m_{ox}} + \frac{\hbar^2 \kappa_\|^2}{2m_{ox}}, \tag{3.20}$$

where k and κ denote the wave vector in Si and in SiO$_2$, respectively. Supposing parallel wave-vector is conserved, i.e. $\kappa_\| = k_\|$, Eq. 3.20 is transformed into

$$\frac{\hbar^2 \kappa_\perp^2}{2m_{ox}} = E_\perp - \Delta E_c - \frac{\hbar^2 k_\|^2}{2m_{ox}}\Big(1 - \frac{m_{ox}}{m_\|}\Big), \tag{3.21}$$

where $(\hbar^2 k_\perp^2)/(2m_\perp)$ is substituted with $E_\perp$ – the incident, or subband energy. In Si, $k_\|$ and $m_\|$ depend on orientation of the substrate, with respect to the interface, and moreover, the ratio $m_{ox}/m_\|$ can be greater, or smaller than 1. Therefore, Eq. 3.21 means that the potential barrier for tunnelling, and hence the tunnelling current itself, is dependent on the substrate orientation (134). However, a systematic study of the issue showed that tunnelling current from Si substrate is essentially independent of the substrate orientation (135). This implies a violation of parallel wave-vector conservation, which is ascribed to interface roughness and the amorphous nature of SiO$_2$.

Nevertheless, parallel momentum conservation has been advocated (136), since short range order within the sub-stoichiometric region of the oxide has been observed. Once again more insight is provided from microscopic, tight-binding calculations, showing that above a certain critical oxide thickness (between 1 and 3 nm), $k_\|$-breaking is intrinsic to the Si/SiO$_2$ lattice mismatch, even in crystalline SiO$_2$ models with high degree of lateral periodicity (132). This readily explains the experimental findings of Ref. (135), where the thinnest oxide is much larger than 5 nm. However, some barrier height dependence on parallel momentum is observed in the TB calculations of Ref. (132) for oxide thickness below 2 nm.

Although the issue of band-structure mismatch remains debatable, in our work, we assume $k_\|$-conservation is violated, and that the tunnelling barrier is independent of the subband and valley from which electrons tunnel.

3.2.5.3 Image force

Another complex issue with tunnelling concerns the image force correction to the potential barrier, as discussed in Refs. (136; 137). Comparison of transmission probabilities calculated with and without accounting for the image force, indicates that the effect of barrier lowering due to the image-force can be absorbed to the fitting parameter m_{ox}, although a qualitative discrepancy becomes apparent at high energies (137; 138; 139). Since we are mostly concerned with tunnelling of electrons near the Si conduction band (i.e. the first two subbands, 0.1 - 0.3 eV above the conduction band), rather than close to the top of the barrier, we consider the effect of image force to be secondary, and have neglected it, as is common.

3.3 1D simulation of direct tunnelling gate current

The simulation of the gate tunnelling current dependence on gate voltage requires a self-consistent calculation of the field and charge distribution in the device for each gate bias. Here we give an overview of a 1D Poisson-density-gradient (P-DG) solver, to which we couple Register's tunnelling model (described in the previous section 3.2.4), to obtain $I_G - V_G$ characteristics from inversion and accumulation. This could be regarded as a 1D prototype of the 3D simulation framework, anticipated for the study of gate leakage variability. It is used to calibrate the only fitting parameter in the tunnelling model – the oxide effective mass, m_{ox}.

3.3.1 Simulation approach

The 1D P-DG solver provides a self-consistent solution for the potential and the carrier distribution in the direction normal to the interface of an n$^+$poly-Si/SiO$_2$/p-Si structure. The solver is based on a finite difference discretization, and an iterative solution of the *Poisson* and *density-gradient* (DG) equations, which take into account quantum confinement effects in the inversion or accumulation layer (140; 141). The equations are

$$\frac{d}{dx}(-\varepsilon\frac{d\psi}{dx}) = q\,(p - n + N_D - N_A) \tag{3.22}$$

$$\frac{d^2 S_n}{dx^2} = \frac{12qm_n^*}{\hbar^2}\,S_n\left(\frac{\phi_n - \psi}{2} + \phi_T \ln{(S_n)}\right) \tag{3.23}$$

$$n = n_i S_n^2 \tag{3.24}$$

$$\phi_p = -\,\phi_T \ln\left(\frac{N_A}{n_i}\right),\ \phi_T = \frac{k_B T}{q} \tag{3.25}$$

$$p = n_i e^{(\phi_p - \psi)/\phi_T}, \tag{3.26}$$

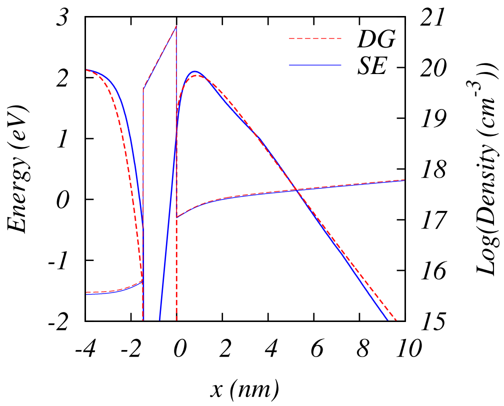


Figure 3.5: Typical agreement in the potential and substrate electron density distribution between the Poisson-DG and Poisson-Schrödinger solvers. The MOS structure has 1.46 nm oxide, 1×10^{20} cm^{-3} n^{+}-poly-gate, and 5×10^{17} cm^{-3} p-Si substrate. Gate voltage is 1.5 V.

The effective mass m_n^* in the DG equation for S_n 3.23, and m_{oxDG}^* – setting a Neumann boundary condition for S_n at the Si/SiO$_2$ interface,[1] are treated as phenomenological parameters. They are calibrated through fitting of the substrate carrier distribution against a more accurate, self-consistent Poisson-Schrödinger solver. Figure 3.5 shows the type of agreement obtained for $t_{ox} = 1.46$ nm at $V_G = 1.5V$, with $m_n^* = 0.15m_0$ and $m_{oxDG}^* = 0.14m_0$. We have verified that there is a very good agreement over the range of the confining fields of interest for the calibration of the tunnelling model for oxide thicknesses in the range of 1 to 2.5 nm, and gate biasing from 0 to 2.5 V.

The direct tunnelling calculation is based on REGISTER's model (1):

$$J_G = Q_s f(F_{ox}) T(F_{ox}). \tag{3.27}$$

Since we assume zero charge in the SiO$_2$ in the Poisson equation, the P-DG solution for the electrostatic potential, at any gate bias V_G, varies linearly across the oxide, allowing the oxide field F_{ox} to be calculated. This allows calculation of $J_G(V_G)$. The dependencies of the impact frequency $f(F_{ox})$ and tunnelling probability $T(F_{ox})$ remain the same as in the original work, and sheet charge in accumulation is still modelled as field-induced, i.e. $Q_s = \epsilon_{ox} F_{ox}$.

The novelty in our work is in the different way of calculating the charge available for tunnelling in inversion. The original model used $Q_s = \epsilon_{ox} F_{ox} - Q_d$, where Q_d is the depletion charge, accounting for the immobile portion of the induced charge (91). Although this approach is acceptable for 1D simulations, where Q_d can be calculated from the doping concentration of the substrate, it cannot be adapted to 3D MOSFET simulations with realistic doping profiles,

[1] The Neumann BC for S_n is derived on the assumption of an exponential decay of the electron density in the oxide, with a characteristic penetration length $x_p = \hbar/\sqrt{2m_{oxDG}\Delta E_c}$, where ΔE_c is the conduction band discontinuity at the interface (142).

particularly if discrete random dopants are to be modelled. The solution is to use the self-consistent electron concentration $n(x)$ from the P-DG solver and integrate it over the depth of the substrate, to obtain the inversion sheet charge

$$Q_s = Q_i = q \int_0^{t_{sub}} n(x)dx. \tag{3.28}$$

This is readily applicable to 3D simulations.

3.3.2 Tunnelling model calibration

Here we present 1D simulation results for the direct tunnelling gate current, based on the approach described above. Comparison is drawn with experimental data used for the calibration of the tunnelling effective mass, the only fitting parameter in the tunnelling calculations. The experimental oxide thickness remains unchanged and the conduction band discontinuity at the Si/SiO$_2$ interface is assumed to be 3.15 eV.

3.3.2.1 Tunnelling from inversion layer

Figure 3.6 shows measured gate current (lines) from n-channel MOSFETs with different oxide thickness under positive gate bias, with source and drain contacts grounded (data from Ref (72)). The large gate area (4.65 × 10 μm^2), and the uniform surface potential in the substrate at this bias allows for the interpretation of the data as one-dimensional tunnelling from the inverted channel, ignoring the minor contribution from the gate edge. The simulated current (symbols), shown in the same figure, agrees very well with the experimental results, for the entire range of oxide thickness. Throughout the entire study a non-parabolic, Franz-type band-gap dispersion relation is used for the SiO$_2$ (130), with the band-edge effective mass treated as

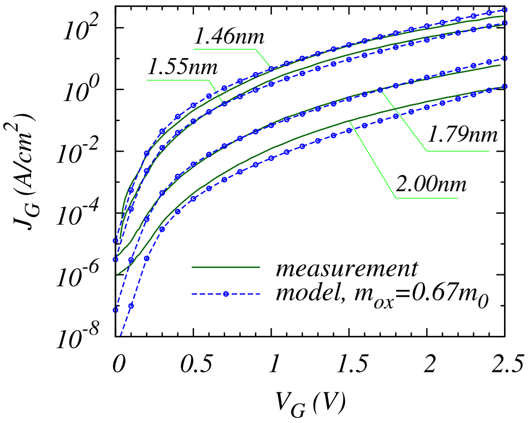


Figure 3.6: Simulated (symbol) and measured (line, after (72)) direct tunnelling gate current density from the inversion channel of large-area MOSFETs with different oxide thicknesses, as indicated. All structures have identical doping levels of $1 \times 10^{20} cm^{-3}$ for the n+poly-Si gate, and $5 \times 10^{17} cm^{-3}$ for the $p-$Si(100) substrate.

a fitting parameter. The best fit, shown in Fig. 3.6, is obtained by setting the tunnelling oxide effective mass m_{ox} to $0.67m_0$. This is the only fitting parameter with respect to tunnelling, and has a single value. For the oxide thicknesses considered here, the slight increase (from $0.61m_0$) of the oxide effective mass compared to Ref. (1) and (72) may be attributed to the different way that the tunnelling charge is calculated (non self-consistently with the potential, in the case of Refs. (1; 72)). In our case, the self-consistent charge from the Poisson-DG solver is used, as described in the previous subsection.

Since we perform numerical integration using a simple trapezoidal approximation, to obtain the charge available for tunnelling from the inversion layer, it is important to check if there is a strong dependence of the tunnelling current on the mesh size. The simulations presented so far were done using a uniform grid with a mesh size of 0.1 nm, which is too fine for 3D simulations. Fig. 3.7 compares the DT current density simulated with four different node spacings – 0.1, 0.2, 0.25 and 0.5 nm. As expected, there is a slight dependency, observable particularly at lower gate bias, which actually results from the solution of the discretized Poisson and density-gradient equations, rather than the numerical integration of the sheet charge. This is evidenced from the electron density profile away from the interface (at $x = 0$) shown on the same figure, for two gate voltages – 0.1, and 1.0 V. Clearly, the electron density is underestimated when using the coarsest grid (particularly for the low gate bias), which is propagated to the value of the electron sheet charge. However, the agreement at 0.25 nm mesh spacing is excellent, and we choose that grid size for the inversion layer discretization in the 3D simulations, to be presented in the next chapter.

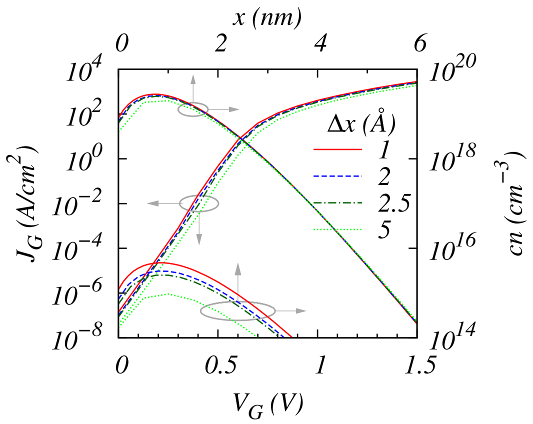


Figure 3.7: Mesh-size (Δx) dependence of the simulated gate current density and electron concentration (at $V_G =$ 0.1 V, and $V_G = 1$ V) for a MOS structure biased in inversion. The structure is n+poly-Si ($1 \times 10^{20} cm^{-3}$), 1 nm SiO$_2$ p-Si(100) ($5 \times 10^{18} cm^{-3}$). Oxide effective mass is 0.67m$_0$, as in the preceding simulations.

3.3.2.2 Tunnelling from accumulation layer

Figure 3.8 presents a comparison between experimental (solid, green lines) and simulated (connected symbols, blue lines) gate current from the accumulated gate of an n$^+$poly/SiO$_2$/p-Si(100) structure (measurements from Ref. (91)). Good agreement is achieved for oxide effective mass of 0.67m$_0$. Note that this is the same value of the tunnelling effective mass, used for the simulation of gate leakage from the inversion layer. Fig. 3.8 additionally shows (red dash) experimental data for a similar device structure and oxide thicknesses from a different source, Ref. (1). An appreciable difference is observed between the experimental data from the two sources, particularly at the thinner oxide of 2.2 nm, although the reported details of samples preparation are the same. The model matches the data from (1) when simulations are performed with a lower effective mass of 0.63m_0, but the results are omitted from the graph for clarity.

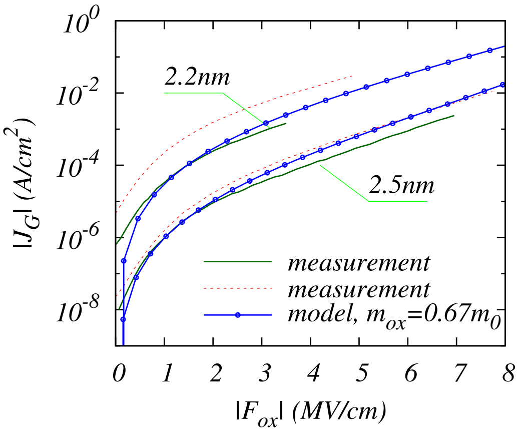


Figure 3.8: Simulated (symbol) and measured (line (91), and dash (1)) direct tunnelling current density from the accumulated poly-Si gate of large-area MOSFETs. Substrate is p-Si(100) doped to $5 \times 10^{17} cm^{-3}$. Oxide thickness is as indicated. The n$^+$poly gate is modelled as $1 \times 10^{20} cm^{-3}$-doped n$^+$Si(110), according to the experimentally observed predominant orientation of the poly grains.

There is commonly some disagreement between reported experimental data relating to devices with the same nominal oxide thickness, in the case of tunnelling from the accumulated poly-gate (1; 91; 112; 121). This may be attributed to the different poly-grain orientation and size, poly-Si doping concentration, or the quality of the poly-Si/SiO$_2$ and Si/SiO$_2$ interfaces for different processing conditions. (In the present simulations, the poly-Si band structure is approximated by that of a Si(110), as suggested by the predominant grain orientation reported in the original experimental works used for the comparison in Fig. 3.8 (1; 91).) On the other hand, there is better agreement between reported data for the case of tunnelling from inversion layers (6; 72; 91; 120). Because of this, we choose to keep the tunnelling effective mass set to $m_{ox} = 0.67m_0$, which provides the best fit for inversion, and a good fit to tunnelling from accumulation. All 3D simulations, to be presented in the next chapter are performed with this value of m_{ox}.

3.4 Summary

In this chapter we identified the direct tunnelling of electrons from quasi-bound states (QBS) to be the gate leakage mechanism of primary concern, and elaborated on the geometrical partitioning of the tunnelling fluxes under different bias conditions of an n-channel MOSFET. Subsequently we reviewed the principle models for direct tunnelling of electrons from a semiconductor cathode, explaining the fundamental concepts behind each of them, and clarifying the approximations involved.

We considered a modified-WKB expression for the transmission probability through the barrier, and demonstrated an excellent agreement with the exact evaluation based on Airy functions. Further emphasis was put on the different ways of calculating QBS lifetimes using the transmission and reflection coefficients in a transfer-matrix-based approach, and using the computed wave functions, in an approach by Price (126), which we extended to apply for complex wave-functions. These two methods are implemented in the Poisson-Schrödinger solver described in App. A and used for the study of the Si/SiO_2 transition layer (see Chapter 5).

For the study of gate leakage variability, we chose the model based on impact frequency and transmission probability, due to its efficiency, accuracy, and adaptability to any 1D or 3D device simulator. This approach of QBS lifetime estimation is linked to an approximation of the quantisation layer in the semiconductor, forming a fully analytical tunnelling model, referred to as REGISTER's model (1), in which we improve the tunnelling sheet charge evaluation.

In the last section, we elaborated on the implementation of this model in a 1D Poisson-density-gradent simulator used as a prototype of the 3D device simulator anticipated for the gate leakage variability simulations. We show a very good agreement between the simulated gate leakage current density and the reported experimental data for both accumulation and inversion bias. Notable in our simulations is the use of a single value of the effective mass in the oxide, $m_{ox}/m_0 = 0.67$, which is the only fitting parameter in the tunnelling model.

Chapter 4

MOSFET gate leakage variability

While the gate leakage current of a uniform transistor may be analysed as a two-dimensional problem, to account for microscopic fluctuations that affect the transistor characteristics requires 3D simulations. In this chapter, we summarise the techniques most commonly used for 3D device simulation and among all, select the density-gradient framework, for the study of statistical gate leakage variability. The methods used to account for random dopant fluctuations and oxide interface roughness are detailed, as is the implementation of the direct tunnelling model. Results from the simulations of a uniform device are presented. The approach is finally applied to the study of gate leakage variability in an ensemble of 25 nm gate length n-type MOSFETs, resulting from discrete random dopant fluctuations and oxide thickness variation.

4.1 Simulation of MOSFET variability

Initial attempts to model device variability have been based on analytical models with an *a priori* assumed statistical (usually Normal) distribution of the parameters, introduced primarily by the impact of discreteness of charge and granularity of matter on device structure (23; 143; 144; 145). It is now well accepted that only a three-dimensional (3D) simulation framework can fully capture the complex effects of microscopic fluctuations in real space, and provide an accurate estimate of the magnitude of variability in future, scaled technologies. 3D simulation techniques are numeric, and typically, computationally expensive. This latter fact is an additional challenge, since to obtain a statistically reliable distribution of a given transistor characteristic, e.g. gate current or V_T, one must simulate a sufficiently large ensemble of macroscopically identical, but microscopically different devices. In practice, the ensemble size would be in the order of 200, to obtain the standard deviation with no more than 5% error, but should exceed $10,000$, if the details of the tails of the distribution are required (146). In the following section, we summarise the most established 3D simulation approaches for studying nano-CMOS device variability, and discuss their applicability for gate leakage simulation.

4.1.1 3D simulation techniques

4.1.1.1 Drift-diffusion with density-gradient corrections

Within the drift-diffusion approximation, electric current is described as a sum of carrier drift (electric field-driven motion) and carrier diffusion (random motion resulting in a net flow only in the presence of charge concentration gradient) (147). The corresponding equation for the electron current density $\boldsymbol{J}_n$ is

$$\boldsymbol{J}_n = -qn\mu_n \boldsymbol{\nabla}\psi + qD_n \boldsymbol{\nabla} n. \tag{4.1}$$

The first and second terms on the right represent the drift and diffusion components respectively, where q is the elementary charge, n – the electron concentration, ψ – the electrostatic potential, μ_n – the electron mobility, and $D_n = \mu_n(k_B T/q)$ – the diffusion coefficient. Equation 4.1 reflects a macroscopic point of view, based on an estimate of the average carrier velocity and its linear dependence on the electric field, and assuming a constant carrier temperature throughout, equal to the lattice temperature. Its validity is therefore limited to slow field variation, near-equilibrium conditions, and to devices with dimensions larger than the carrier mean free path, as could be elucidated by a derivation of Eq. 4.1 from the Boltzmann transport equation (BTE) (148; 149). A great emphasis in the practical application of this equation (viz. Eq. 4.4) lies on the use of an appropriate mobility model.

The DD transport model is at the core of DD device simulators that consist of an iterative numerical procedure for the solution of the coupled *Poisson* and *current continuity* equations. [1] The method, first proposed by Scharfetter and Gummel (150), has a remarkable numerical stability and lends itself to very efficient implementations even in a 3D solution domain (148; 149; 151; 152). This quality has served as a major pull to extend the applicability of the DD approach to the simulation of nano-scale devices. Two aspects have been successfully addressed in this regard - *i*) modifications of mobility models to include dependency on the carrier concentration, and on the lateral and vertical electric fields (151; 153), and *ii*) quantum

[1] The full set of equations relevant for n-channel MOSFET is presented in Section 4.1.2.1.

corrections to the carrier density (via the Density Gradient (DG) generalisation of the drift-diffusion framework (140; 141) or the Quantum Drift Diffusion approach (154)), to account for carrier confinement effects in the inversion layer.

Even with these advancements, the DD framework cannot capture the non-equilibrium transport effects in a rapidly varying electric field, disregarding velocity overshoot effects (155). However, in an ultra-short channel MOSFET, the drain current is essentially defined by the average velocity of carriers near the source, and their concentration around the peak of the lateral potential barrier (156; 157; 158), so that non-equilibrium carrier dynamics at the drain end of the device have indirect and relatively minor impact (37). It has been demonstrated that close to the source, the extracted velocity profile from the DD approach (with suitably increased saturation velocity) is very similar to the carrier velocity profile obtained from a more accurate, Monte Carlo simulations of a 50 nm gate-length MOSFET (158). $I_D - V_G$ characteristics simulated with carefully calibrated density gradient (DG) simulators compare very well against experimental data, or full-band Monte Carlo simulations, of devices with gate length as small as 10 nm (159; 160). The agreement is particularly close in the sub-threshold region, suggesting appropriate account of the device electrostatics by the DG framework.

Due to the above, the DG simulation approach with quantum corrections has become the primary technique for the investigation of statistical variability in ultra-scale MOSFETs (27; 29; 161). One of the most advanced implementations using this approach is the *3D-atomistic* simulator developed at the University of Glasgow, incorporating a number of the known sources of intrinsic parameter fluctuations, including random dopant fluctuations, oxide interface roughness, line edge roughness, poly-Si granularity and Fermi-level pinning (27; 162). It has been successfully applied to the prediction of threshold voltage and drive current variability in n-channel MOSFETs, scaled according to the ITRS requirements (29).

The applicability of a DD framework for the study of gate leakage is exemplified by existing commercial and in-house developed 3D DD simulators (31; 151). These are based on a 1D semi-classical direct tunnelling model, coupled to the 3D drift-diffusion solution of carrier density and potential. Using a classical DD device simulation and such 1D tunnelling model, gate leakage variability due to random dopants has been briefly addressed in Ref. (31).

4.1.1.2 Monte Carlo technique

The Monte Carlo (MC) simulation of modern semiconductor devices is based on a particle description of the charge transport within the system (163; 164). It provides increasingly accurate approximation to the solution of the BTE when the number of MC particles is increased. Individual point-particle dynamics (drift-collision sequences under the influence of external force) in an ensemble of free carriers are modelled by a semi-classical equation of motion – the quantum mechanical interaction between free carriers and the crystal lattice is abstracted away by the introduction of band structure and the effective mass approximation, while scattering is modelled through perturbation theory, following the Fermi Golden rule. The Monte Carlo method, as a mathematical algorithm for the aggregation of a probabilistic result from a random sample of possible inputs, enters to the simulation framework in several ways – selection of an initial and after-scattering state for each particle, selection of the particle free flight duration, and selection of a scattering process at the end of the free flight. Throughout the simulated time for the evolution of the system, statistical information about the individual position of each carrier in the six-dimensional phase-space is collected. For steady-state simulation, the simulated time is large enough to ensure any effect of the initial state of the ensemble has eroded, so that the time average of the system states may be interpreted as a steady state ensemble average. Consequently, the phase-space distribution function of the carriers may be used to obtain their real-space distribution, as well as the terminal currents.[1] Typically, calculation of particle dynamics is self-consistently coupled to a solution of the Poisson equation through an iterative loop, to establish the self-consistent electrostatic potential that is the force field for carrier acceleration, in the presence of an applied bias to the system terminals (167).

The advantage of MC device simulations comes from the ability of the procedure to capture non-equilibrium and non-local effects, i.e. those of velocity overshoot carrier heating due to finite carrier relaxation characteristics. It could be modified for the modelling of devices with microscopic fluctuations, and non-uniformities in the structure. Successful implementation of *ab initio* impurity and surface roughness scattering has been demonstrated in a 3D simulation study of V_T and drive current variability in devices with random dopant fluctuations and interface roughness (168). A significant drawback of the MC approach however is its large computational cost and implementation complexity.

MC simulations have not gained popularity for the study of direct tunnelling gate leakage current. This is due to the fact that there is no model of particle dynamics for the electrons tunnelling with energy lower than the oxide conduction band minimum. Thus the MC method of solving the BTE within the oxide is applicable only to electrons with higher energy, and

[1] It has been shown that the (phase-space and time dependent) distribution function obtained through the described MC simulation technique satisfies the semi-classical Boltzmann transport equation (BTE) (165; 166).

with the adoption of idealised $\mathrm{SiO_2}$ band-structure and scattering mechanisms (136). MC device simulations have been combined with a method of estimating the tunnelling probability of carriers reaching the oxide interface, emphasizing on the importance of the electron energy distribution function on the gate leakage (30; 99; 136; 169; 170; 171; 172). [1] We note however, that for a short-channel n-type MOSFET with an ultra-thin oxide *i)* the maximum gate leakage current corresponds to the transistor ON state ($V_{GS} = V_{DD}$ and $V_{DS} = 0$, i.e. one of the two static points of a CMOS inverter), and *ii)* even at $V_{GS} = V_{DS} = V_{DD}$, the peak gate leakage density is linked to the maximum electron concentration near the source end, rather than to the carrier energy peak near the drain end (30).

4.1.1.3 Non-equilibrium Green's function formalism

The non-equilibrium Green's function (NEGF) approach to device simulations is based on quantum transport theory. The central quantity of interest, from which electron and current densities are obtained, is the density matrix describing the correlated manner in which energy states may be occupied (173; 174). The NEGF formalism is a prescription for obtaining the density matrix from the single electron Green's function $\boldsymbol{G}(E)$ given by the (simplified) matrix equation

$$\boldsymbol{G} = [E\boldsymbol{I} - \boldsymbol{H} - \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_S]^{-1}, \tag{4.2}$$

and signifying the response of the system to an impulse excitation within the device. The energy E here is not an eigen-state, but a contact excitation energy, over which one has to integrate, to obtain the density matrix. The Hamiltonian $\boldsymbol{H}$ represents the kinetic and potential energy of a single particle, and reflects the details of the device structure. The electrostatic potential is obtained self-consistently through a solution of the Poisson equation. $\boldsymbol{I}$ is the identity matrix. The crucial element in NEGF is the treatment of the contact boundary condition and scattering through self-energy matrices. In equation 4.2, $\boldsymbol{\Sigma}_{1,2}(E)$ couple the device to a semi-infinite contact leads in an analytically exact way, even if they have different Fermi levels. Typically, the source and drain contacts of a MOSFET are considered, but the gate may be treated on equal footing, to obtain the gate leakage current (175). The self-energy $\boldsymbol{\Sigma}_S(E)$ accounts for scattering. [2]

The strength of NEGF approach is in its ability to treat an open boundary, non-equilibrium system in a quantum-mechanically consistent way. Its applicability to device variability stems

[1] Note that these works use 2D MC simulators, except Ref. (30), which uses a quasi-3D approach

[2] A more elaborate description, and a derivation of a kinetic equation is found in Refs. (175; 176; 177).

from two sources. First, one obtains directly the details of the current density within the device. This is of great importance to understanding the physical origins of the effects that microscopic non-uniformities have on the device characteristics. (178). Second, one could use a tight-binding basis representation of the Hamiltonian, and capture the structural variations, e.g. oxide roughness, and band-structure effects in ultra-scaled devices in truly atomistic representation (179). This is conceptually attractive for gate leakage modelling if a crystalline model of the gate oxide is deployed (132). In addition, the device, including the gate contact, can be modelled as a single entity. These properties of the NEGF approach with respect to leakage is in contrast to the previously discussed simulation approaches, which require a different transport model for gate and drain current. Yet, only a few NEGF studies of MOSFET leakage are known, all with an effective mass Hamiltonian (175; 180; 181).

There are two limitations with present NEGF simulators. The first is large computational cost, due to the necessity to invert large matrices. This issue is actually prohibitive for using NEGF approach for variability study on a statistical scale. The second is due to the complexity in implementing scattering mechanisms. In particular, the scattering self-energy matrix $\mathbf{\Sigma}_S$ depends on the density matrix itself, and must be solved self-consistently, through iterations, much like the electrostatic potential (173). However, the dependence of $\mathbf{\Sigma}_S$ on the density matrix can be very involved. Reliable approximations have not yet been developed nor evaluated, since the technique is relatively new. This is why even the most advanced 3D real-space NEGF simulators models the ballistic limit of MOSFET operation (178).

4.1.1.4 Precision versus simulation time

We close the review of three-dimensional (3D) simulation techniques by looking at their relative accuracy and a consideration of their computational cost. Figure 4.1 shows the $I_D - V_G$ characteristics of a uniform, 22 nm gate length MOSFET, simulated with NEGF, MC, and DG. The characteristics at high drain bias ($V_D = 1.0$ V) are shown plotted on a logarithmic scale (left axis), while the low drain bias ($V_D = 0.1$ V) IV curves are on a linear scale (right axis). In both cases the current obtained from the NEGF simulations is highest, since no scattering model is included. The drift-diffusion current appears to be lowest, since it is limited by the saturation velocity, the effect being appreciably stronger at high-field (high drain bias). However, at low drain bias, or low electron density in the channel (high drain bias and lower gate bias), the agreement between the MC and DG simulations is very good. This is due to the fact that in such a bias regime, the current is mostly due to carrier diffusion, and is controlled by the

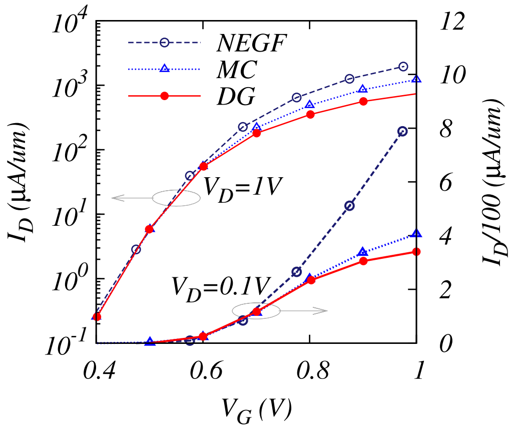


Figure 4.1: Comparison of $I_D - V_G$ characteristics simulated with 2D NEGF, 3D MC and 3D DG. The NEGF simulation is ballistic, and predictably yields the highest current. High-field current in the DG case is limited by velocity saturation and is lower, relative to the MC simulation, but at low field compares very well. (Data courtesy of C. Alexander, A. Martinez, A.R. Brown, unpublished.)

electrostatics of the device, which are captured with a sufficient accuracy, due to the DG corrections incorporated in the DD simulator. At the same time, since gate leakage is of a critical relevance to static power dissipation, the bias condition for achieving maximum drive current, i.e. high V_{GS} voltage and high V_{DS} voltage, is less important. This devalues the advantage of a MC simulator in providing the energy distributions of electrons at high-fields.

Regarding computational cost, the structure simulated above takes typically a few minutes per IV point on a modern CPU, for the DG simulator. The MC and the NEGF require the equivalent of about 10 CPU days per IV point. Efficient parallelization of the NEGF code makes the problem manageable for analysing uniquely selected configurations of non-uniform devices, but there is still a gap of over three orders of magnitude performance, relative to DG.

In conclusion, for the study of gate leakage variability on a statistical scale, the complexity and computational burden of a 3D MC or NEGF device simulator is not justified. The generalisation of the DD approach, via the DG theory to account for quantum confinement, provides the essential components for computing the direct tunnelling gate current. It models the electrostatic potential and electron distribution in the desired bias regime with sufficient accuracy and in a computationally efficient manner. The following subsection summarises the details of the three-dimensional atomistic DG simulator, at the core of the simulations presented later in this chapter.

4.1.2 *3D atomistic* device simulator

Here we give a brief account of the underlying models and details of the implementation of the Glasgow *3D-atomistic* Drift-Diffusion device simulator with Density-Gradient quantum corrections, the development of which is described in more details elsewhere (26; 27; 182; 183).

4.1.2.1 Mathematical Formulation

The following set of equations constitutes the drift diffusion framework with density gradient corrections (DD-DG).

The *Current continuity* equation is solved only for electrons (we model an n-channel MOSFET) in steady state:

$$\boldsymbol{\nabla} \cdot \boldsymbol{J}_n = qR \tag{4.3}$$

$$\boldsymbol{J}_n = -qn\mu_n \boldsymbol{\nabla}\psi + qD_n \boldsymbol{\nabla} n, \tag{4.4}$$

where the electron concentration n is unknown variable. Note that in our case the recombination rate R is zero. It is possible however to include the gate leakage current into the self-consistent drain current through R (31).

The Non-linear Poisson equation is

$$\boldsymbol{\nabla} \cdot (-\varepsilon \boldsymbol{\nabla}\psi) = q\left(p - n + N_D - N_A\right), \tag{4.5}$$

with free carrier concentration given by Boltzmann statistics according to [1]

$$n = n_i e^{(\psi-\phi_n)/\phi_T}, \ \phi_T = \frac{k_B T}{q} \tag{4.6}$$

$$p = n_i e^{(\phi_p-\psi)/\phi_T} \tag{4.7}$$

$$\phi_n = \psi - \phi_T \ln\left(\frac{n}{n_i}\right) \tag{4.8}$$

$$\phi_p = -\phi_T \ln\left(\frac{N_A}{n_i}\right). \tag{4.9}$$

The quasi-Fermi level for holes is fixed (everywhere in the device) to the quasi-Fermi level of the source contact, which is the potential reference, i.e. where $\psi = 0$. Impurities are treated as fully ionised.

To account for quantum confinement effects, Eq. 4.6 and Eq. 4.7 are replaced by the appropriate density gradient (DG) relations:

$$n = n_i S_n^2, \ \ p = n_i S_p^2 \tag{4.10}$$

$$\nabla^2 S_n = \frac{12qm_n^*}{\hbar^2} S_n \left(\frac{\phi_n - \psi}{2} + \phi_T \ln\left(S_n\right)\right) \tag{4.11}$$

$$\nabla^2 S_p = \frac{12qm_p^*}{\hbar^2} S_p \left(\frac{\psi - \phi_p}{2} + \phi_T \ln\left(S_p\right)\right), \tag{4.12}$$

[1] The electrostatic potential is implicitly defined in terms of the intrinsic Fermi level E_i: $\psi = -E_i/q$.

where m_n^* and m_p^* are anisotropic effective masses for electrons and holes (with normal and lateral components relative to the oxide interface), and are both treated as fitting parameters. [1] Accordingly, the quasi-Fermi level ϕ_n is modified to account for the *effective quantum* potential

$$\phi_n = \psi + \frac{2\hbar^2}{12qm_n^*}\frac{\nabla^2 S_n}{S_n} - \phi_T \ln\left(S_n^2\right), \tag{4.13}$$

where $S_n = \sqrt{(n/n_i)}$ is the solution of Eq. 4.11.

Ohmic contacts are assumed by ensuring charge neutrality using Dirichlet boundary conditions (BC) for the potential. The particulars are given in Table 4.1. Charge neutrality and carrier equilibrium define the BC for mobile carriers at the contacts. Non-physical device boundaries (e.g. side walls and top side between gate and source/drain contact) assume Neumann BC with zero normal-derivative of the potential and carrier concentration (148). The current normal to these boundaries is zero. The density gradient equations 4.11 and 4.12 are subject to Neumann BC at the oxide interface, which corresponds to a finite penetration of mobile carriers in the oxide, although such charge is not accounted for in the Poisson equation (the oxide is idealised, free of any charge). At Ohmic contact and non-physical boundaries, the BC for $S_{n,p}$ are zero normal-derivative. This latter fact implicates Neumann boundary conditions for the minority carriers at contacts as well (184).

Table 4.1: Boundary conditions.

Contact	ψ
Source	0
Drain	V_D
Substrate	$-\phi_T \ln\left(\left(N_A N_D\right)/n_i^2\right)$
Gate	$V_G + \phi_T \ln\left(N_P/N_D\right)$

N_P, N_D and N_A are the concentrations of poly-Si gate donors, source/drain donors, and substrate acceptors, respectively. V_G and V_D are relative to the source contact, which is connected to the substrate.

4.1.2.2 Implementation Aspects

In discretizing the Poisson, current continuity, and density gradient equations, their integral form is used, and the *control-volume* discretization method is applied (also known as *box integration* (148)). The basic idea can be visualised in Figure 4.2. It shows the abstract rectangular box (a cuboid) that surrounds a given grid node, and the six nearest grid nodes. With the help of the divergence theorem, an integral over the volume of the cuboid is reduced to an integral

[1] The R-parameter of the DG formalism equals 3 and is implicit in the constant preceding the S term in equations 4.11 and 4.12.

over the surface of the cuboid. The Poisson equation is taken as an example in the caption of the figure. Note that the computation of the surface integrals requires only discretization of first order partial derivatives, normal to the faces of the control-volume cuboid.

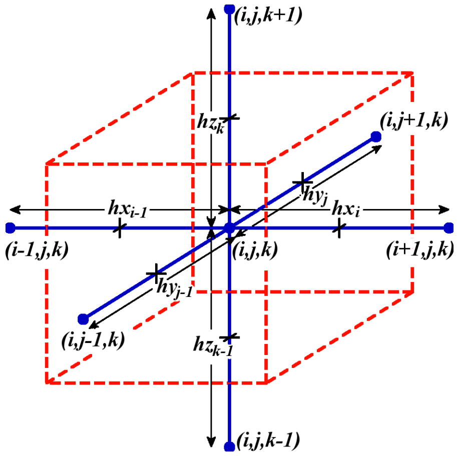


Figure 4.2: An abstract cuboid with a volume dv, and a surface ds, surrounds a grid node to which a net charge density ρ is assigned. Integrating the Poisson equation 4.5 over dv, we arrive at the integral form of Gauss law: $\int(\nabla\cdot\epsilon\nabla\psi)dv = \oint(\epsilon\nabla\psi)\cdot ds = \int\rho\, dv$. Assuming constant $\nabla_{\perp}\psi$ over each face of the cuboid, the surface integral is split into sums, and discretization does not require interpolation of ϵ (even if it changes across an interface), as would be the case in ordinary finite difference scheme.

The control-volume discretization dispenses with the necessity of interpolating elemental material parameters (constants within an element defined by four grid nodes) which are typically discontinuous across interfaces (e.g. dielectric constant). This is essential for simulating devices with microscopic non-uniformities like oxide roughness and discrete impurity atoms.

Discretization of the current continuity follows the Scharfetter-Gummel method, which assumes exponential, rather than linear, variation of free carrier density between neighbouring grid nodes. The Gummel cycle ensures convergence of the iterative, self-consistent solution by solving the Poisson and current continuity equations in turn (148; 150). Note that the solution of the linearised Poisson equation itself is done through a self-consistent, iterative loop with the solution of the density gradient equations, which is required due to the exponential dependence between free carrier concentration and electrostatic potential. Neumann boundary conditions are implemented using the method of image (or phantom) nodes, outside the solution domain, used to define the derivative at a boundary grid node via a central difference scheme (148; 149).

4.1.2.3 Modelling of random dopant fluctuations

Modelling of the discrete doping atoms consists conceptually of two parts – *i)* creating such a distribution of random impurities that is representative of a given continuous doping concentration profile, and *ii)* accounting for the discrete ionized, dopants charge of in the device electrostatics.

In the *3D atomistic* simulator, random dopant distribution is obtain following Ref. (185). A Si-crystal lattice is constructed (independently of the numerical discretization grid) within the simulation domain of the device. The probability for finding a dopant at a given lattice site is computed from the ratio between the desired local doping concentration, and the intrinsic Si concentration. Using rejection technique, a dopant replaces a Si atom if this probability exceeds a random number (in the range of 0 to 1) generated for the specific lattice site. It has been demonstrated that this procedure yields a Poisson distribution, centred around the average dopant number (of a certain type) per device, as expected (183).

Accounting for the discrete ionised impurities in the drift-diffusion framework requires the assignment of their charge to the discretization nodes, since it is the charge density that enters the Poisson equation 4.5, via N_D^+ and N_A^+, which are position dependent. Practically, each dopant is surrounded by an elemental volume dV of the device grid, so that the question is how to distribute the equivalent concentration $1/dV$ between the vertices of this volume. Three different methods – *cloud in a cell* (CIC, first-neighbour grid vertexes are assigned charge density inversely proportional to their distance from the dopant), *nearest grid point* (NGP, entire charge density is assigned to the nearest grid vertex), *Gaussian smearing* (the charge density is smeared even beyond the boundaries of the volume, with each grid vertex being assigned an amount proportional to a Gassian distribution centered at the dopant) – have been evaluated and result in negligible differences in the context of density-gradient corrected drift-diffusion simulator (183).

This deserves a clarification, since it is known that the Coulomb potential due to an ionised impurity is analytically singular (i.e. $\propto 1/|\boldsymbol{r}|$). In a classical simulation such as DD, the mobile carriers could become trapped in the sharply resolved Coulomb wells of the ionised impurities. In a numerical calculation this is further complicated by the fact that the magnitude of the Coulomb potential is mesh dependent, in addition to the dependence on charge assignment method mentioned above. This is a problem, particularly for majority carriers in the source/drain regions, where such unphysical trapping results in decreased conductance. The issue is resolved in a quantum mechanically consistent way in the *3D atomistic* simulator. The density-gradient corrections to the mobile charge, reflect the effect of quantisation within the potential well of an impurity, making the mobile carriers density less localised around the dopants and smoothing the potential. [1]

[1] Due to quantisation, shallow impurities have transport activation energies of about 50 meV (37).

In the present study, we used CIC charge assignment and DG corrections to both electrons and holes, so that donors and acceptors are treated on equal basis. This practically eliminates mesh sensitivity, and allows the adoption of a finer mesh in the inversion channel, as required for the accurate estimation of the charge available for tunnelling.

4.1.2.4 Modelling of oxide thickness fluctuation

The systematic variation in the oxide thickness due to process conditions is not considered here, and we focus instead on the stochastic oxide thickness fluctuation induced by the microscopic roughness of the Si/SiO$_2$ interface (47). The model is based on the assumption that the Si/SiO$_2$ interface is well characterised by a two parameter autocorrelation function (ACF), $C(r)$, of the actual material boundary distance $\Delta(r)$ from an ideal (001) interface plane (r is the correlation radius). [1] The ACF is usually approximated by a Gaussian or an exponential function (47; 48), having the form

$$C(r) \simeq \Delta^2 \exp\left(-r^2\lambda^{-2}\right) \quad (Gaussian) \tag{4.14}$$

$$C(r) \simeq \Delta^2 \exp\left(-\sqrt{2}r^2\lambda^{-1}\right) \quad (Exponential). \tag{4.15}$$

Here Δ is the root-mean-square (RMS) of $\Delta(r)$, and λ is the correlation length, describing the decay of the autocovariance in $\Delta(r)$.

The aim is then to create a rough surface that could be characterised by an *a priori* chosen set of ACF parameters Δ, and λ, and use it to define the position of the boundary between Si and SiO$_2$ in the simulator. This is achieved by creating a 2D matrix, representative of an interface plane, each element of which has a magnitude determined by the power spectrum obtained from a Fourier transform of the ACF, and a randomly selected phase (26). An inverse-Fourier transform is performed to yield a real function that can be interpreted as the distance $\Delta(r)$. This function is then quantised to one atomic layer of 0.28 nm, to obtain a physically meaningful interface, with roughness arising from atomic level steps (187).

A comment regarding the values of Δ and λ is worthwhile. For the Si(001)/SiO$_2$ interface, experimentally obtained values of Δ are in the range of 0.15 - 0.3 nm, depending on process conditions and oxide thickness, and HRTEM analysis show one atomic layer steps being at the origin of the fluctuations (47; 187; 188; 189). Rough surface reconstruction for device simulations is typically done with Δ = 0.3 nm (26; 47; 48; 190). However, there is over an order of magnitude disagreement in the reported range of λ. Values of 1 to 3 nm are

[1] Sharp structural transition is implied, ignoring the sub-stoichiometric oxide (186).

deduced from HRTEM interface image analysis (47; 187), and established from calculations, fitting mobility data limited by interface roughness scattering (190; 191). At the same time, AFM surface analysis report a correlation lengths between 15 and 55 nm (48; 188; 190). It is argued that AFM cannot resolve short range correlation features (190). More importantly, smaller correlation length and RMS fluctuations were deduced in Ref. (47) after removing a background non-uniformity of the oxide thickness, appearing on a scale larger than 20 nm, on the assumption that such a trend is systematic. The issue remains controversial, particularly in view of the post-AFM self-affine fractal analysis in Ref. (48), suggesting that stochastic fluctuations on a larger scale (tens of nm) are intrinsic to the oxidation kinetics.

4.2 MOSFET gate leakage

4.2.1 Direct tunnelling model

We adopt the following approach to calculating the direct tunnelling gate leakage in a 3D MOSFET. A one-dimensional (1D) semi-classical tunnelling model is coupled to the *3D-atomistic* device simulator, described in the preceding section, at a post-processing stage. In doing so, the following approximations are involved. Gate leakage current is not self-consistently calculated, and its effect on the drain current is ignored. This is justified by the fact that typically, the direct tunnelling gate leakage in well scaled MOSFETs does not affect significantly the equilibrium carrier concentration in the tunnelling cathode. [1] A finite surface element discretization of the gate oxide interface is performed, and the direct tunnelling current density J_G, through each surface element with area dS, is obtained by 1D calculation. Numerical integration over the gate area yields the total gate leakage current, I_G. In practical terms, and with n being the number of surface elements,

$$I_G = \sum_{i}^{n} J_{Gi} dS_i. \tag{4.16}$$

Hence, there is an implicit assumption that dS_i is sufficiently small and a constant leakage current density is a good approximation of the actual or average tunnelling flux through a segment. Lateral coupling between the different segments is not considered. Subsequent simulations are performed using a uniform lateral grid with 1 nm node spacing.

[1] This approximation could be dispensed of however, by the inclusion of leakage contribution through the Generation/Recombination rate term in the current continuity equation, which would allow one to account of the gate leakage impact on the drain current.

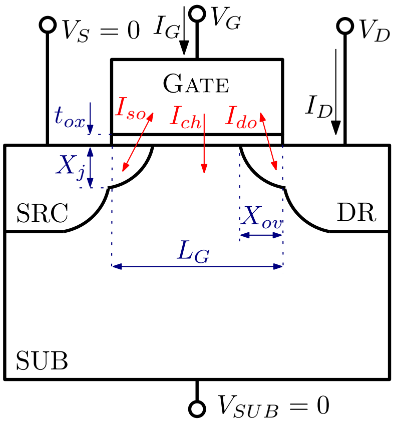
The calculation of the elemental direct tunnelling current density J_G is based on the oxide-field-dependent, analytical model described in Section 3.3, according to

$$J_G = Q_s f(F_{ox\perp}) T(F_{ox\perp}) \tag{4.17}$$

The equivalent sheet-charge density Q_s, and the normal component of the oxide field, $F_{ox\perp}$, on which the semi-classical impact frequency f and tunnelling probability T depend, are derived from the three-dimensional electrostatic potential and carrier concentration distributions obtained from the *3D atomistic* simulator, at each bias point. In inversion, Q_s equals the electron charge concentration, integrated over the depth of the substrate, along a line normal to the interface at the given lateral position of the gate. This advancement of the original model was already exploited in our 1D simulations, and is of a critical importance when discrete random dopants are introduced in the device, for 3D simulations of gate leakage variability. In accumulation, Q_s is modelled as field induced, by $Q_s = \epsilon_{ox} F_{ox,\perp}$. $F_{ox,\perp}$ is found from the normal-derivative of the electrostatic potential in the oxide. Since the approach is identical with the 1D case, we do not need to re-calibrate the tunnelling effective mass, which remains $0.67 m_0$ in our 3D simulations.

4.2.2 Geometrical partitioning

Before we move on to study gate leakage variability, we analyse a uniform device with nominal geometry and doping related parameters specified in Fig. 4.3 and the adjacent table. The aim is



Parameter	
Dimensions	
Gate length, L_g (nm)	25
Gate width, W_g (nm)	25
Gate-S/DE overlap, X_{ov} (nm)	5
S/D Junction depth, X_j (nm)	6
SiO_2 thickness, t_{ox} (nm)	1
Doping	
Substrate p-Si, N_A (cm^{-3})	5×10^{18}
Gate n+poly-Si, N_D (cm^{-3})	1×10^{20}
Source/Drain n+Si, N_D (cm^{-3})	1×10^{20}

Figure 4.3: 2D schematic diagram of the simulated MOSFET, and its geometrical and doping related parameters. Nominal values are given in the table. Substrate, source, and drain are uniformly doped. The red arrows represent geometrical partitioning of the gate current.

to understand the influence of gate and drain bias on gate leakage, and to revise the importance of the different geometric regions under the gate, i.e. the source and drain extension overlaps, and the channel. The red arrows in Fig. 4.3 partition the gate current according to these three distinct regions. The bi-directional arrows associated with the S/DE overlaps anticipate the bias dependence of the direction of the electron flux through the oxide. In particular, since the source and drain regions are degenerately doped to the same level as the poly-gate (1×10^{20} cm^{-3}), the direction of the current through the oxide over the drain extension, I_{do}, depends on the voltage difference $V_G - V_D$. Regarding the channel component I_{ch} of the direct tunnelling current, it is always in the indicated direction (for normal CMOS-logic bias of the transistor). This is clarified with the help of the band-diagrams in Fig. 4.4. At high gate voltage and low drain voltage, the channel is strongly inverted, while the contact extension overlaps are accumulated. The electric field is facilitating electron tunnelling from the substrate into the gate. At low gate voltage and high drain voltage, the poly-gate over the drain extension is weakly accumulated, facilitating electron tunnelling from the gate into the drain extension. Due to the built-in potential, the channel region is in depletion, hence the field remains in the same direction as for high gate voltage.

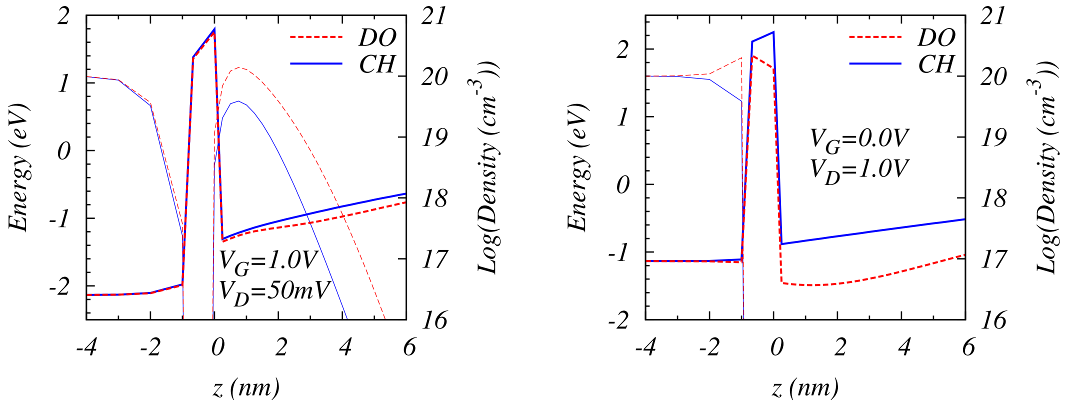


Figure 4.4: Conduction band diagrams from two different lateral positions – middle of the channel (CH), and middle of drain extension overlap (DO). Two different bias conditions are shown – $V_G = V_{DD}, V_D \sim 0$ (left), and $V_G \sim 0, V_D = V_{DD}$ (right).

Note that the bias conditions reflected in these two figures correspond to the two stable points of a CMOS inverter. Hereafter, the condition of $V_G = V_{DD}, V_D \sim 0$ is referred to as the On-state of the MOSFET, while the condition of $V_G \sim 0, V_D = V_{DD}$ is referred to as the Off-state. The power supply voltage V_{DD} is 1.0 V throughout this work. Under any of these bias conditions, the transistor dissipates only static power, hence the relevance of gate leakage and its variability is greatest. Next, we separately analyse the influence of V_G and V_D, and show the actual results from the simulation of gate leakage.

4.2.3 Gate voltage dependence

Figure 4.5 shows the simulated $I_G - V_G$ (red), and $I_D - V_G$ (blue) characteristics of the uniform device for two different drain voltages – $V_D = 50$ mV (thick, solid lines), and $V_D = 1.0$ V (thin, dash+ lines). Note the difference in measuring units for the drain (μA) and gate (nA) current. For both drain voltages, the drain current is between 1 (at low V_G) and 5 (at high V_G) orders of magnitude higher than the gate leakage. The obvious qualitative difference in the $I_G - V_G$ curves for low and for high drain voltage requires further elaboration, and is best understood with the help of Fig. 4.6, showing the lateral distribution (along the channel) of the direct tunnelling current density.

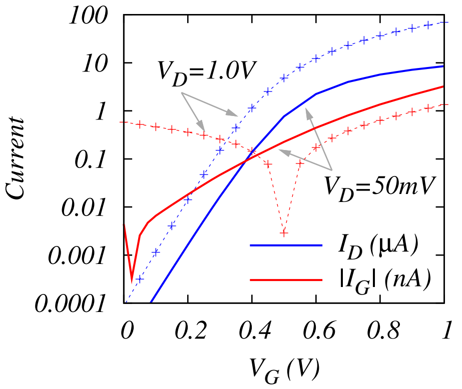


Figure 4.5: $I_D - V_G$ (blue) and $I_G - V_G$ (red) characteristics at low (solid) and high (dashed,symbol) V_D. Note that I_G and I_D are in nA and μA, respectively. The dip in the I_G curve indicates a change in current direction at the corresponding V_D.

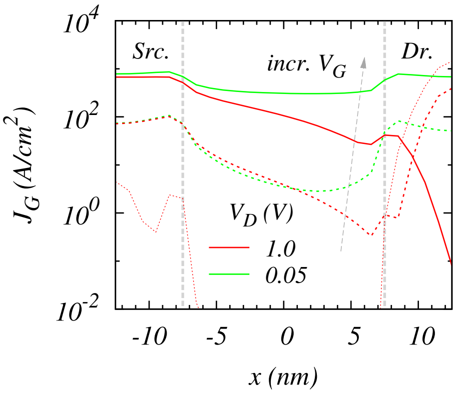


Figure 4.6: Distribution of the gate current density J_G along the channel direction. Two sets of curves are shown, at high (red) and low (green) V_D. The arrow indicates upper curves are for higher V_G, i.e. 0 (dot), 0.5 (dash), and 1.0 (solid) volts.

At low drain voltage, the increase of V_G inverts the channel and accumulates the source and drain extension overlaps, so that the charge available for tunnelling from the substrate constantly increases. At the same time, the oxide potential barrier becomes more transparent due to the stronger confinement that elevates the subbands in the inversion or accumulation layer in the substrate. This leads to the steady exponential growth of the gate current density

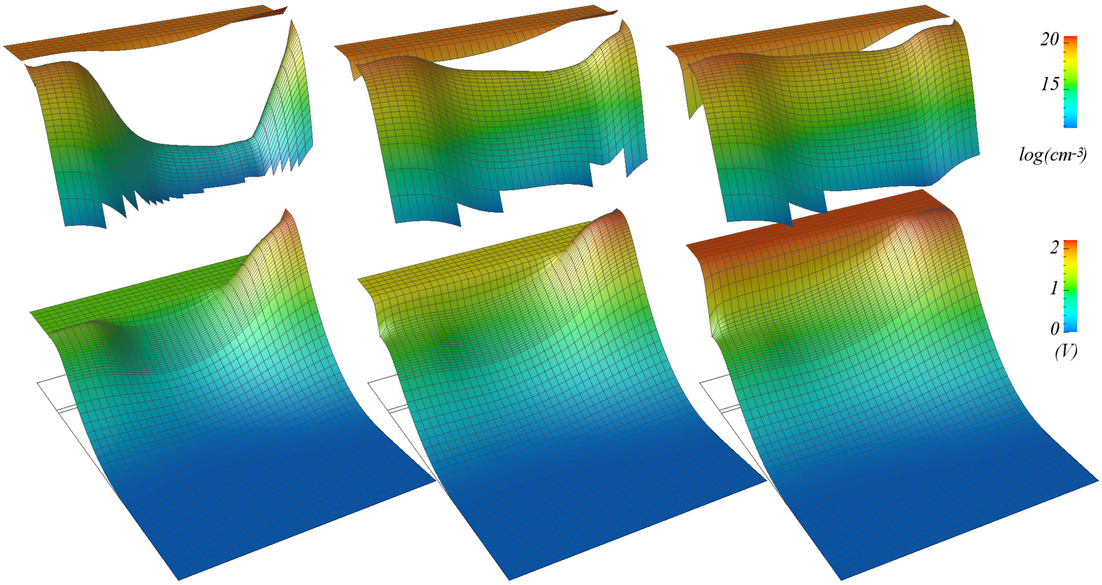


Figure 4.7: Surface plots of the 2D electron density (top) and electrostatic potential (bottom) for V_G = 0, 0.5, and 1.0 V (left to right), and V_D =1.0 V. The gate-drain overlap is the most distant corner in each plot. Electron concentration below 10^{11} cm^{-3} is not shown.

in all three geometrical regions (refer to Fig. 4.3 and the green curves on Fig. 4.6), and the corresponding increase of the total leakage I_G with the increase of V_G.

At high drain voltage, there are two competing trends, which determine the $I_G - V_G$ curve, and the fact that the tunnelling current changes its sign around the middle of the simulated gate voltage range. Consider the following three gate bias points – 0, 0.5 and 1.0 V. The 2D electrostatic potentials and electron concentration distributions at these bias points are shown (as surface plots) in Fig. 4.7. At low gate voltage, the substrate is depleted, while the gate regions overlapping the source and drain extensions are accumulated. The electrostatic field at the drain end is conductive for a significant flux of electrons from this region of the gate, which explains the J_G distribution at $V_G = 0$ in Fig. 4.6 (red, dotted curve). As gate voltage increases, so does the population of electrons in the source extension and the adjacent part of the channel, as seen in the middle plots of Fig. 4.7. Tunnelling from the substrate becomes appreciable, and competes with the tunnelling from the gate, which is meanwhile reduced due to the diminished drain-gate voltage difference (see Fig. 4.6, red, dashed curve). Eventually, at $V_G \sim 0.5$ V, the two tunnelling fluxes cancel each other, leading to the lowest point in the $I_G - V_G$ curve. Further V_G increase makes the substrate tunnelling component dominate the gate current, while tunnelling into the drain extension is minimised (see Fig. 4.6, red solid line).

4.2.4 Drain voltage dependence

The drain voltage dependence of the direct tunnelling gate current is shown in Fig. 4.8, for fixed gate voltage of 0 V(solid, red line), and for 1.0 V (dashed, symbol). The $I_D - V_D$ characteristic at low V_G is also shown for a comparison (solid, blue line). In this case, the gate leakage current is larger in magnitude than the simulated drain current. This is in apparent contradiction to previous studies (175; 192), and requires further consideration. We first look at the limiting case of $V_G = 0$ and $V_D = 1.0$ V, as it was already discussed in the preceding subsection. Gate current is almost entirely composed of electrons tunnelling from the gate into the drain extension. These electrons would form part of the drain terminal current if gate leakage was self-consistently included (unlike the simulations presented here), and would raise the sub-threshold drain leakage to the magnitude of the gate leakage, in agreement with Ref. (175; 192). This only stresses the importance of including gate leakage in a self-consistent manner for accurate prediction of static power dissipation, since this limiting bias condition corresponds to a stable point of a CMOS inverter, the OFF-state of the n-type MOSFET.

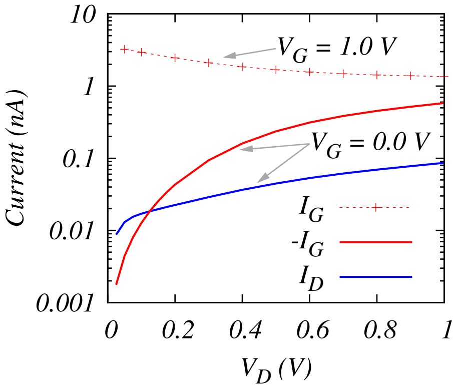


Figure 4.8: $I_D - V_D$ (blue) and $I_G - V_D$ (red) characteristics at low (solid) and high (dashed,symbol) V_G. Note that the I_G at low V_G is larger in magnitude than the corresponding I_D, and opposite in sign (i.e. direction) to the I_G for high V_G.

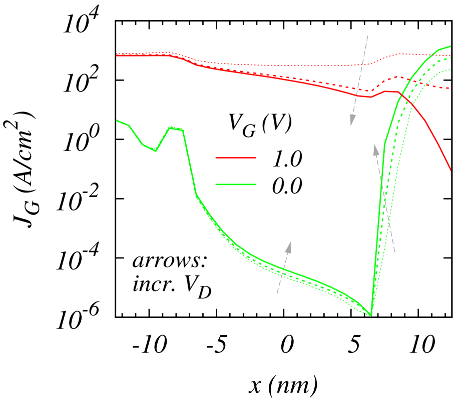


Figure 4.9: Distribution of the gate current density J_G from source to drain. Two sets of curves are shown, at high (red) and low (green) V_G. The arrow indicates the trend of J_G with increasing V_D (∼ 0 (dot), 0.5 (dash), 1.0 (solid)) at a fixed V_D.

The shape of the $I_G - V_D$ characteristics is easy to explain following the exposition so far, and referring to Fig. 4.9 for the 2D distribution of the direct tunnelling current density. At low V_G, a reduction of V_D from 1.0 V steadily reduces the drain-overlap gate leakage component as shown in Fig. 4.9 (green curves), but leaves the channel and source-end components almost unaffected, since nearly the entire voltage drop between the drain and the source happens at the drain junction. These latter components are minor, since the substrate is depleted of electrons. At high V_G of 1.0 V the substrate has very high electron concentration, and tunnelling from the source extension and the channel is dominant. If drain voltage is also high, i.e. $V_D = V_G$, the contribution from the drain-overlap region is minimised, as discussed in the previous subsection. However, reducing V_D leads to the accumulation of electrons in the drain extension itself, and effectively increases the direct tunnelling flux from the substrate. This is clearly seen in Fig. 4.9 and explains why the decrease of V_D, at high V_G, actually increases the gate leakage.

4.2.5 Critical evaluation

Our results and conclusions presented thus far agree qualitatively and in order of magnitude with previous investigations of gate leakage in sub-100 nm gate length, sub-2 nm gate oxide MOSFETs, studied with different techniques – e.g. experimental ((192; 193)), 2D Monte Carlo ((30)), and 2D NEGF ((175)) simulations. Direct comparison against measurements of gate leakage from an ultra-scaled MOSFET requires a more realistic doping profile to be considered, outside the immediate objectives of this work. Before applying our approach to study gate leakage variability however, it is worth briefly evaluating its shortcomings.

The disregard of hot-carrier phenomena has already been discussed in association with the choice of a 3D simulation methodology, and is known to have little effect on direct tunnelling gate current, in the devices of interest (30). A more significant issue is the dependence of tunnelling current only on the perpendicular component of the electric field, effectively employing a 1D picture of the tunnelling process. This is acceptable for simulations at low drain voltage, since the lateral component of the field is much smaller than the perpendicular one. There is also an experimental evidence that the gate-drain extension tunnelling at high drain voltage and low gate voltage also depends only on the vertical electric field (192) (the measured $I_D = I_G$ regardless of the substrate negative bias). We could therefore limit our simulations to low drain voltage, for $I_G - V_G$ characteristics, and to low gate voltage, for $I_G - V_D$ characteristics, which are also the most important regimes relevant to static power dissipation (i.e. the On- and Off-states of the MOSFET, explained earlier).

Three other aspects remain disregarded, currently, in our framework. First, the self-consistency of the drain and gate current. This could be implemented at a later stage. Second, the effects of electron diffraction around the gate edge (electrons having indirect tunnelling path through the oxide spacers) has been recently shown to increase the gate leakage in both On (over 10-times) and Off-state (about 3 times) (194). To address this, a 3D quantum mechanical approach would be required. Finally, the effect of gate width scaling, also shown to increase the tunnelling current by increasing the perpendicular field near the shallow trench isolations, is also disregarded (195). This would require the implementation of different boundary conditions, in the DG simulator.

4.3 Gate leakage variability

The 3D simulation approach described above is applied here on a statistical scale to study the direct tunnelling gate leakage variability in realistic nano-scale MOSFETs. Two prime sources for local gate current density variation are taken into account - discrete random dopant fluctuations (RDF) and oxide thickness variation (OTV).

4.3.1 Nominal device and microscopic fluctuations

An ensemble of 230 macroscopically identical, but microscopically different n-channel MOSFETs is simulated. The nominal (macroscopic) parameters of the devices are the same as the uniform device presented in Fig. 4.3 in the preceding section – a 25 nm gate length n-channel MOSFET with 1.0 nm SiO_2 dielectric is taken as an example, with a simplified uniform impurity profile and geometry, but similar to a 'well-behaved' realistic bulk MOSFET at the given gate-length (29). In choosing a square gate device we aim to obtain the upper limit in gate leakage variability, since in a wider device, the intrinsic fluctuations in local tunnelling density will self-average over a wider area, and reduce the anticipated spread in the gate current-voltage characteristics. To illustrate the impact of each of the two sources of statistical variability individually, as well as their combined effect, the following three sets of devices are simulated:

1. RDF set: devices vary only in the number and spacial configuration of the discrete doping atoms introduced in the domain under the gate oxide, i.e. in the substrate, and the source and drain extensions. The poly-Si gate has a uniform, continuous doping level;

2. OTV set: devices vary only in the roughness features of the Si/SiO_2 interface, implicating local oxide thickness variation. Interface roughness of one atomic layer of Si (0.3 nm) is generated over a 1 nm nominal oxide thickness and a correlation length of 1.8 nm;

3. Combined set: the microscopic variations corresponding to the devices of the previous two sets are simultaneously introduced.

4.3.2 Random dopant fluctuations

4.3.2.1 $I_G - V_G$ characteristics

Figure 4.10 shows (light-blue lines) the $I_G - V_G$ characteristics of a sub-sample of 50 simulated devices with randomly distributed impurities in the substrate, including the source and drain contacts, at a low drain voltage of 50 mV. The $I_G - V_G$ curve of the uniform device with the same nominal doping is shown in black. The curves in red, labelled *worst* and *best*, correspond to the devices with highest and lowest leakage, respectively, at $V_G = 1.0$ V. Such a bias condition (high gate/source and low drain/source voltages), is important for digital CMOS technology, and corresponds to the On-state of the transistor. Notably in this case, there is a relatively minor spread of the gate leakage magnitude, but there is an appreciable increase of the mean value, with respect to the gate tunnelling current of the uniform device. The increase in magnitude could be tentatively associated with the exponential sensitivity of the direct tunnelling current density on the oxide field, which is strongly modulated around the ionized doping atoms near the oxide interface. The variation in the total gate leakage at high gate bias is relatively small for two reasons. First, the excess of electron charge in the inversion channel, and in the accumulated

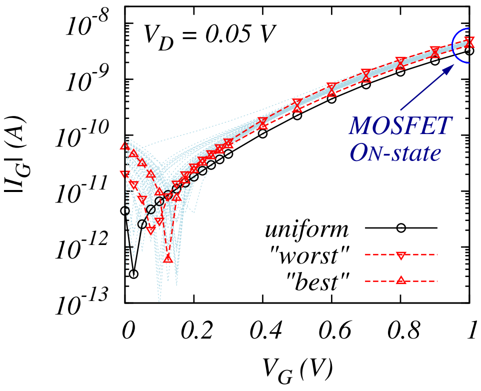


Figure 4.10: $I_G - V_G$ characteristics at low drain voltage $V_D = 50$ mV for a sub-sample of 50 devices from the RDF set. For CMOS logic, the area of interest is at V_G of 1.0 V, corresponding to the MOSFET On-state.

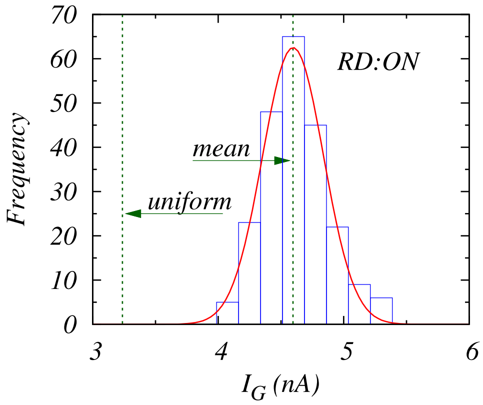


Figure 4.11: Gate current histogram for the ensemble of 230 devices with RDF, under On-state bias conditions. A Gaussian distribution with mean $\langle I_G \rangle = 4.6$ nA and standard deviation $\sigma = 0.2$ nA is fitted.

S/DE overlaps, screens the Coulomb potential of the impurities. Second, self-averaging of the fluctuations in tunnelling current density happens over the entire surface of the gate. The resulting gate leakage dispersion at $V_G = 1.0$ V, for the entire ensemble of 230 devices, is shown in Fig. 4.11. The spread is confined to less than a factor of two of the mean current (4.6 nA), which is increased by 50 % from the leakage of the uniform device. The standard deviation is less than 5 % of the mean ($\sigma = 0.2$ nA).

It is interesting to examine in more detail the reason for the observed gate current variations, in order to understand if they would be augmented, or further reduced by an alternative device architecture. In this respect, we recall a previous study of the subject, which stated that the RDF-induced gate leakage variations are not associated with local electric field fluctuations, but only with the uncertainty in the definition of source and drain overlaps, caused by the random distribution of acceptor atoms in the contact extensions (31). There are a couple of important shortfalls in the methodology of this study however. First, a drift-diffusion simulator has been used without consideration of quantum confinement effects, meaning that a classical electron distribution is obtained, with an unrealistically high peak at the Si/SiO$_2$ interface. Second, discrete doping atoms are accounted for only by the explicit incorporation of the long-range potential of the ionised impurities in the Poisson equation (31; 196). The disregard of the short-range part of the Coulomb potential of the impurities, and the overestimated electron concentration at the top of the substrate render the conclusions of the study questionable.

To obtain a clearer understanding of the phenomena involved, we look at the local fluctuations of the tunnelling current density and its two key variables, according to our model, namely, the normal component of the oxide field and the distribution of electrons in the substrate. These are shown in Fig. 4.12 for the *best* and *worst* devices for $V_G = 1.0$ V. The identical colour maps and displacement scale for the rubber-sheets in both cases facilitate an objective comparison. We draw the attention of the reader to the following important observations. First, the leakage contributions of the source and drain extension overlaps seems equal, but the *best* device has appreciably lower tunnelling current density within the area of the channel, compared to the *worst* device. Second, the normal component of the oxide field is rather uniform, except a few highly localised sharp peaks and valleys, associated with ionised acceptors and donors respectively, at the interface, which cannot be screened by electrons due to quantum confinement effects. A peak in the field should contribute to a peak in the current density. Such a correlation is not observed however, except for a single case (in each of the displayed devices), where the highest leakage density is associated with the highest field magnitude, due

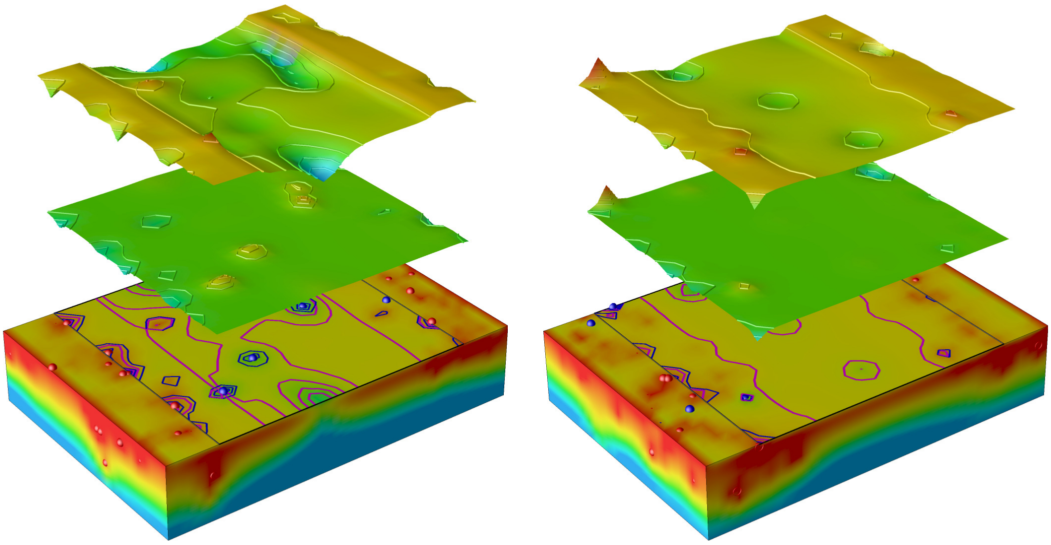


Figure 4.12: Substrate electrons distribution (3D slab), normal component of the oxide field (middle), and gate leakage density (top,*Log*-scale) for the *best* (left) and *worst* (right) devices under ON-state bias. The iso-lines of the leakage current (purple) are clearly correlated with the electron distribution at the Si/SiO$_2$ interface. The iso-lines of the normal electric field (blue) at the interface reflect the very strong localisation of the disturbances in surface potential due to the discrete impurities, shown as small spheres (blue/red - acceptor/donor atoms).

to an acceptor atom within the area of the drain extension, where donor - and by implication, electron - concentration is high. The third observation is that the overall pattern of leakage fluctuations promptly follows the distribution of electron density at the Si/SiO$_2$ interface, which appears to be, overall, higher for the *worst* device. It is worth considering that the mean field of the *uniform*, *best*, and *worst* devices are almost the same - 6.14, 6.24, and 6.18 MV/cm, respectively. However, the corresponding gate current of 3.0, 2.9 and 4.5 nA, obtained from the mean tunnelling current density is clearly different for the *uniform*, *best*, *worst*, and similar in magnitude to the corresponding actual gate current, 3.2, 4.0, and 5.2 nA, respectively.

The above considerations suggest that the variation in tunnelling current from device to device is due to variation in the inversion charge density, and hence to variation in the effective doping concentration, which is the only free variable affecting the inversion charge density at a fixed bias, for the devices being compared. To verify this hypothesis, we show, on Fig. 4.13, the impurity distribution in the first 6 nm of the substrate, corresponding to the depth of the shallow contact extensions, for the *best* and *worst* devices. An obvious feature of the *worst* device is a much smaller number of acceptors under the gate. In particular, there are only 14 of

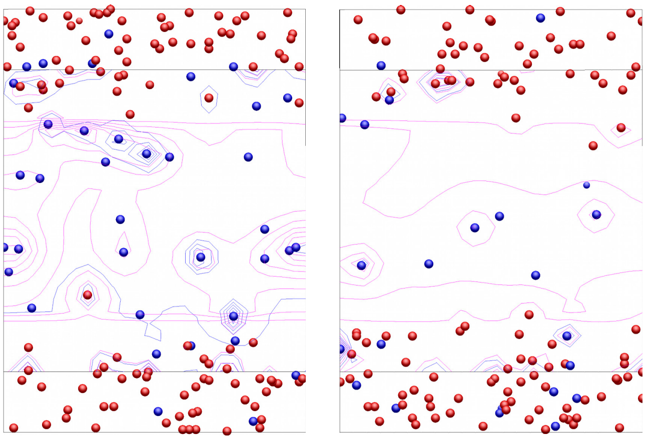


Figure 4.13: Impurity distributions (red/blue spheres representing donor/acceptor atoms) within 6 nm below the Si/SiO$_2$ interface. Source contact is up. Under the gate of the device, there are 29 acceptors in the *best* (left) device, and only 14, in the *worst* (right) device. Blue/magenta iso-lines of the oxide field/gate leakage density, are also shown.

them, while for the *best* device, there are 29. The number of acceptors, related to the considered volume under the gate, yields an equivalent acceptor concentration of 3.7 and 7.7 ($\times 10^{18}$ cm^{-3}), for the *worst* and *best* device, respectively. [1] It is well known that given two MOS structures with identical gate and oxide parameters, the inversion charge density at certain bias is larger for the one with lower substrate doping level, due to the smaller amount of depletion charge and corresponding threshold voltage.[2] This leads to an increase in the tunnelling current, as the substrate doping is decreased, and *vice versa*, the effect being stronger at low gate voltage (72).

We conclude, that for the selected macroscopic parameters of the simulated ensemble, the gate leakage variability arises from the variation in the number of acceptor atoms in the first few nm of the substrate below the gate, and not from the local fluctuations in the leakage current density that are due to ionized impurities near the Si/SiO$_2$ interface. The fluctuations in the leakage current density are strong indeed, with the maximum being nearly a 100 times bigger than the mean value. Similarly to the peaks in the oxide field however, the peaks in tunnelling current are confined to a very small area - less than a hundredth of the gate surface - and as such, cannot dominate the value of the total leakage current. This is clarified with the help of the component density histograms shown in Fig. 4.14, comparing the distribution of leakage current density magnitudes in the *uniform*, *best*, and *worst* devices. For each device, the frequency of the histograms is normalised so that the area under the histogram amounts to unity. Two peaks in the component density are easily distinguished for each device, with the lower component associated with the mean tunnelling from the channel, and the higher

[1] Recall that the nominal (uniform) acceptor doping is 5×10^{18} cm^{-3}.

[2] The depletion charge depends on the entire number of acceptors: 182(135) for best(worst) device.

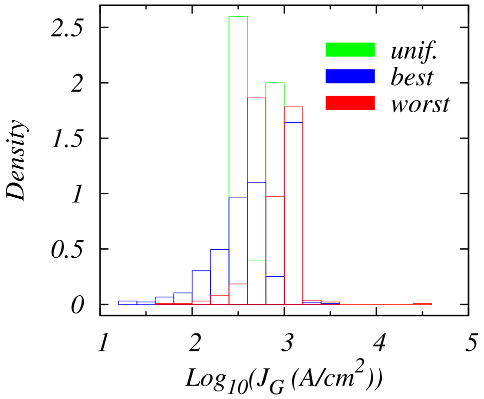


Figure 4.14: Component density histograms of the tunnelling current density in the *uniform*, *best* and *worst* devices with RDF at high V_G and low V_D. The extreme values of tunnelling current density represent a negligible component in the area of the histogram, implying these are associated with a small number of elements of the lateral gate mesh. Their surface is insufficiently large to add appreciable current to the total gate leakage magnitude.

component associated with the mean tunnelling from the gate-overlapped areas of the source and drain extensions. Note that for the best device, a larger number of components of lower value contribute appreciably to the area of the histogram – this is consistent with the bigger number of acceptor dopants in the channel of the *best* device, leading to a wider dispersion of tunnelling current density within the channel area. The components with maximum value for the devices with RDF contribute negligible density to the area of the histogram, as would be implied by a high localisation of the corresponding peaks in tunnelling current.

At this point it is worth re-examining our initial assertion that the slight increase in the mean current for the devices with random dopants, compared to the gate leakage of the uniform device, is due to the exponential sensitivity on the oxide field. In fact, a density histogram of the oxide field (not shown) suggests identical magnitudes of the field contribute to the mean, while it can be clearly seen in Fig. 4.14 that the components with highest density in the distribution of J_G for the uniform device are with lower magnitude than the ones for the devices with RDF. Therefore, the reason for the increase in leakage current, must be an average increase of the electron density at the interface, for the non-uniform devices. What is causing it is the fact that away from the discrete ionised acceptors, which are relatively far apart, the depletion charge is sub-nominal, compared to the one of the uniform device. This is supported by the correlation of higher gate leakage to higher drain current, clearly observed in the scatter plot on Fig. 4.15.

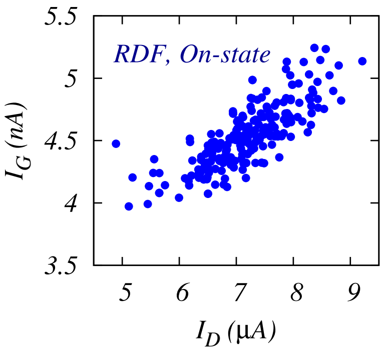


Figure 4.15: $I_G - I_D$ scatter plot.

Finally, we anticipate an increase of RDF-induced gate leakage variability with gate area reduction, as the local peaks in tunnelling density start to represent a more significant fraction of the total gate current. An alternative device geometry, with an underlap of the source and drain contacts (197; 198), will increase the gate leakage variability in the On-state of an n-type bulk MOSFET, since the gate current will be entirely from the inverted channel that is subject to a variable doping concentration, but the magnitude of the gate current will decrease.

4.3.2.2 $I_G - V_D$ characteristics

Figure 4.16 shows (light-blue lines) the $I_G - V_D$ characteristics of the same 50 devices, discussed above, simulated at 0 V gate voltage. The black curve corresponds to the uniform device, while the red curves represent the devices with highest and lowest tunnelling current at $V_D = 1.0$ V. Such a gate bias is important for digital CMOS technology, and corresponds to the Off-state of the transistor. The spread in gate leakage characteristics in this case is very broad, and for the ensemble of 230 devices spans over one order of magnitude, as shown in Fig. 4.17. Note that under the bias conditions of low gate voltage and high drain voltage, the tunnelling flux is composed of electrons from the accumulation charge at the gate interface into the drain overlap region. This is a relatively small area, over which self-averaging of the local fluctuations in local tunnelling current density is less effective. It is interesting to consider if the inter-device

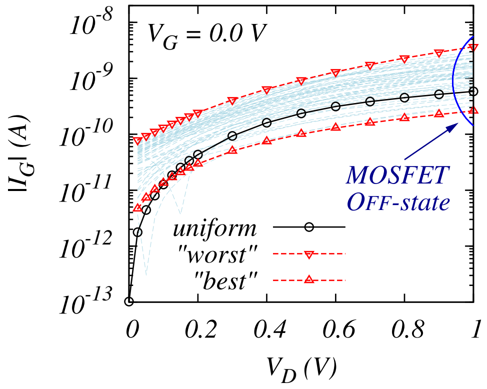


Figure 4.16: $I_G - V_D$ characteristics at low V_G for a sub-sample of 50 devices from the RDF set. For CMOS logic, the area of interest is the highest V_D of 1.0 V, corresponding to the Off-state of an n-channel transistor.

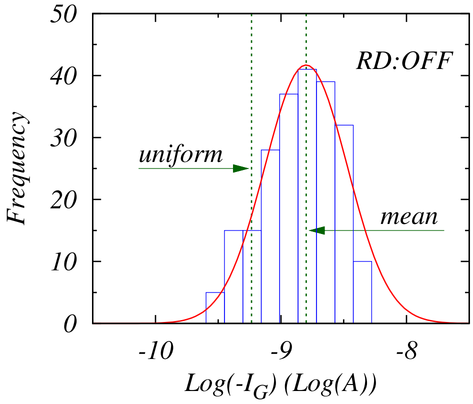


Figure 4.17: Gate current histogram for the ensemble of 230 devices with RDF, under Off-state bias conditions. A Gaussian distribution with a mean of $\langle \log_{10}(I_G) \rangle = -8.8$ and a standard deviation of $\sigma = 0.3$ is fitted.

variability arises from the local fluctuations in leakage density, or by the uncertainty of the *pn*-junction position, due to acceptor dopant fluctuations.

We first look at the leakage current density fluctuations. The peak tunnelling currents are 1.2×10^3 A/cm^2, 1.4×10^3 A/cm^2, and 17.0×10^3 A/cm^2, for the *best*, *uniform*, and *worst* devices respectively, differing by more than an order of magnitude. Within the direct tunnelling model that we have deployed, the tunnelling current density J_G is given by $J_G = QfT$, where Q is the charge available for tunnelling, f is the impact frequency, and T is the tunnelling probability. The charge density in the accumulated gate is relatively insensitive from position to position within a device, and from a device to device, due to the very high, uniform doping level in the poly-Si. The impact frequency slowly varies with the field (refer to Eq. 3.19), but the tunnelling probability is exponentially sensitive to the oxide field. At the given bias conditions the substrate is depleted and the unscreened Coulomb potential of the impurities strongly modulates the electrostatic field. Figure 4.18 shows a density histogram of the normal component of the oxide field F_{ox} in the *uniform*, *best* and *worst* devices. The components with negative value do not contribute to the tunnelling current, because they correspond to the field in the oxide above the channel area of the transistor, which is devoid of electrons. Of importance are the positive components in F_{ox} distribution, causing gate electrons to tunnel into the drain extension. These components are widely dispersed and even the extreme values for each device have appreciably frequency on the histogram. The peak field differs by a factor of more than two, between the *best* (3.7 MV/cm), and *worst* (8.2 MV/cm) devices, and is therefore responsible for the large fluctuations in the leakage density.

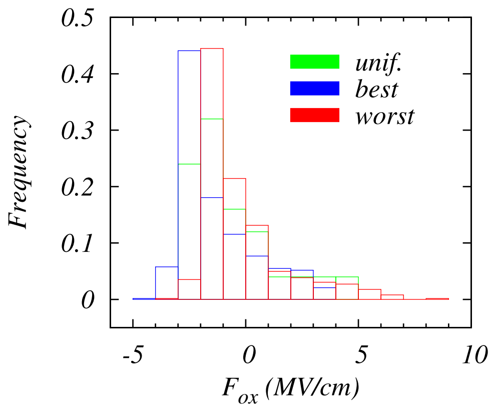


Figure 4.18: Density histogram of the normal component of the oxide field in the *uniform*, *best* and *worst* devices. The components with positive values correspond to the oxide field over the drain overlap region and are widely dispersed. Even the maximum positive values have comparable density to the smallest positive components. The negative values correspond to the oxide field over the channel.

Considering now the area of the drain extension overlap, we compare in Fig. 4.19 the electrostatic potential and impurities in the substrate, and the magnitude of the tunnelling current

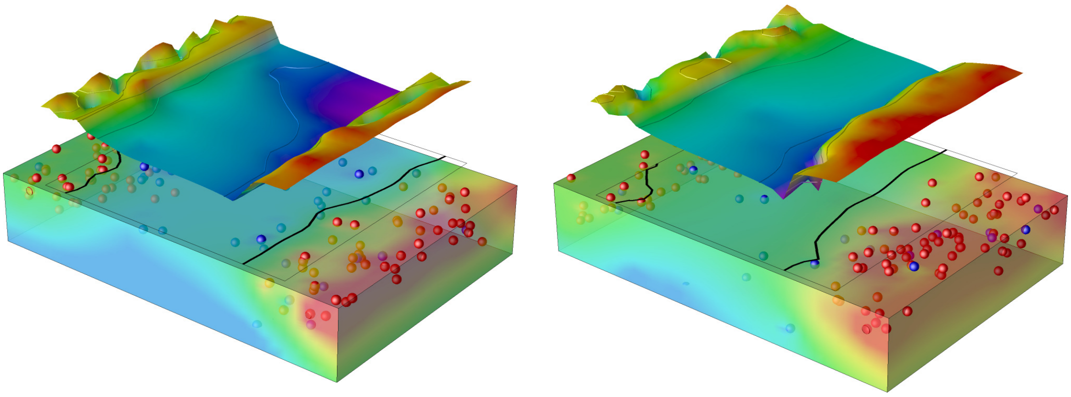


Figure 4.19: Electrostatic potential and discrete impurities in the substrate, and *log* of the gate leakage density for the *best* (left) and *worst* (right) devices. The drain end is closer to sight. Black iso-lines correspond to $F_{ox} = 0$, suggesting drain extension overlap for the *best* device. Local fluctuations in the potential are directly translated in oxide field fluctuations (as the gate could be thought of equipotential), implicating strong fluctuations in the gate leakage density.

density in the *best* and *worst* devices. The overlap area for the *best* device is clearly smaller than the overlap area of the *worst* device, hence correlation clearly exists. The difference in area is in the order of two however, and cannot explain on its own, as previously suggested (31), the large difference in the total gate current magnitude – 0.2 nA, *best* and 4.2 nA, *worst*.

We conclude, that gate leakage variability in the MOSFET Off-state is primarily due to localised fluctuations in the tunnelling current density, associated with the microscopic differences in the number and spacial configuration of doping atoms in the shallow extension region.

4.3.3 Oxide thickness variation

In subsequent simulations we investigate the impact of the Si/SiO$_2$ interface roughness on gate current variability, by allowing for steps of one atomic layer at the Si(001) surface. Considering that the interface between the poly-Si gate and the oxide is modelled as flat, and that Si/SiO$_2$ interface roughness is generated over a nominal oxide thickness of 1 nm, this introduces local variations in the oxide thickness, so that two kinds of regions are formed – with a thickness of either 0.85 or 1.15 nm.

4.3.3.1 $I_G - V_G$ characteristics

The $I_G - V_G$ characteristics of a sub-sample of 50 devices from the OTV set at $V_D = 50$ mV are shown in Fig. 4.20 (left). The $I_G - V_G$ curve of the nominal device with a 1 nm thick,

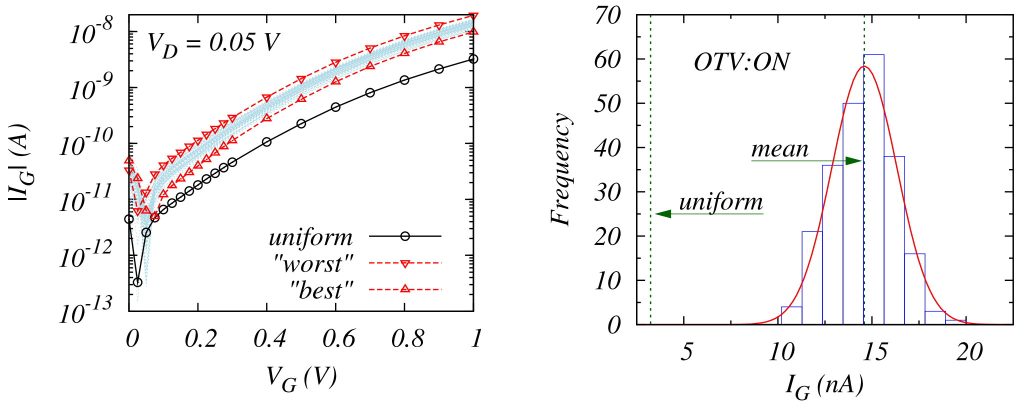


Figure 4.20: $I_G - V_G$ characteristics at low drain voltage $V_D = 50$ mV for a sub-sample of 50 devices from the OTV set (left). Gate current histogram for the simulated ensemble of 230 devices with OTV under the ON-state bias conditions (right). A Gaussian distribution with a mean value of $\langle I_G \rangle = 14.6$ nA and a standard deviation of $\sigma = 2$ nA is fitted.

uniform oxide is shown in black. The curves in red, labelled *worst* and *best*, correspond to the devices with highest and lowest leakage, respectively, at $V_G = 1.0$ V, i.e. in the ON-state of the transistors. There is an appreciable spread of the gate leakage magnitude, and a very clear increase of the mean value, compared to the gate leakage of the uniform device. As can be seen in the right graph of Fig. 4.20, for the full ensemble of 230 devices, the mean gate current is nearly five times higher ($\sim$ 15 nA) than the leakage of the uniform device ($\sim$ 3.2 nA). The increase in the mean gate current is due to the exponential dependence of the tunnelling probability on the oxide thickness and is easily understood with the help of the histogram of the local tunnelling current density shown in Fig. 4.21. On average in the simulations, half of the gate oxide is thinner than the nominal, while the other half is thicker. This leads to the presence of two distinct ranges in the dispersion of the gate tunnelling current density, with similar frequency of occurrence. The mean values of each of the corresponding sub-distributions are separated by nearly two orders of magnitude, with the total current being determined by the higher sub-range. Note that in the calibration of the oxide effective mass against experimental tunnelling data, the oxide roughness was not taken into account, but should be considered in future simulations. We also find a wide dispersion in the oxide field, the normal component of which locally varies between 2 and 7 MV/cm for the simulated devices, and recall that a thinner oxide induces more charge and bears a stronger field, at a fixed bias. Therefore, the variation in the oxide field further accentuates the tunnelling density fluctuations, leading to a larger inter-device gate current variation. For the OTV set of devices, the standard deviation

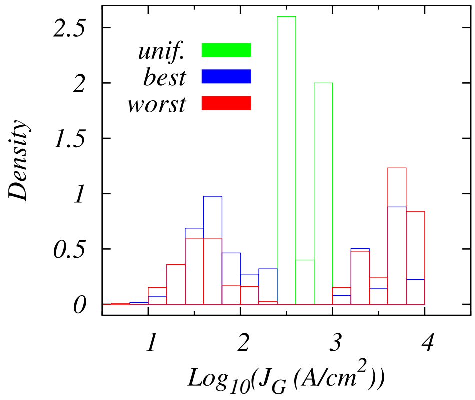


Figure 4.21: Density histogram of the local tunnelling current density in the *uniform*, *best*, and *worst* devices under On-state bias. The devices with oxide roughness exhibit two distinct areas of high contribution, associated with the thinner than nominal, and thicker than nominal local oxide thicknesses.

of the Gaussian fitted to the simulated gate current distribution is 10 times higher (2 nA) than in the case of the RDF set (0.2 nA). For the chosen correlation length of 1.8 nm in generating the interface roughness pattern, the change in oxide thickness happens often enough to yield relatively small regions, compared to the gate area, with a fixed thickness. This allows for self-averaging of the local fluctuations in the gate leakage density. Larger values for the correlation length could be found in literature, however. We anticipate an increase in gate leakage variability if the oxide roughness correlation length becomes comparable to the gate length, i.e. if the gate length is further reduced, or if the correlation length is larger.

4.3.3.2 $I_G - V_D$ characteristics

Figure 4.22 (left) shows (light-blue lines) the $I_G - V_D$ characteristics of the same 50 devices, discussed above, simulated at 0 V gate bias. The black curve corresponds to the uniform device, while the red curves represent the devices with highest and lowest tunnelling current at $V_D = 1.0$ V, i.e. in the Off-state of the transistor. There is a broadening in the dispersion of gate leakage characteristics, compared to the On-state bias, but a similar increase in the mean tunnelling current. The right-hand graph of Fig. 4.22 shows the gate leakage distribution for the entire ensemble of 230 transistors at $V_D = 1$ V. Note that a log scale is used for the current, since the variability in this case exceeds one order of magnitude, with the *min* and *max* leakage magnitudes being 0.3 and 4.4 nA, respectively. The mean gate current is 3.2 nA, and is just over five times larger than the gate current of the uniform device, 0.6 nA. The standard deviation of 0.6-0.8 nA constitutes nearly 15-20 % of the mean. [1] Leakage variability at high drain bias for the OTV devices is entirely due to the different patterns of oxide roughness, and

[1] Figures are indicative only, since the normal distribution, and hence the standard deviation, apply for the *Log* of the current.

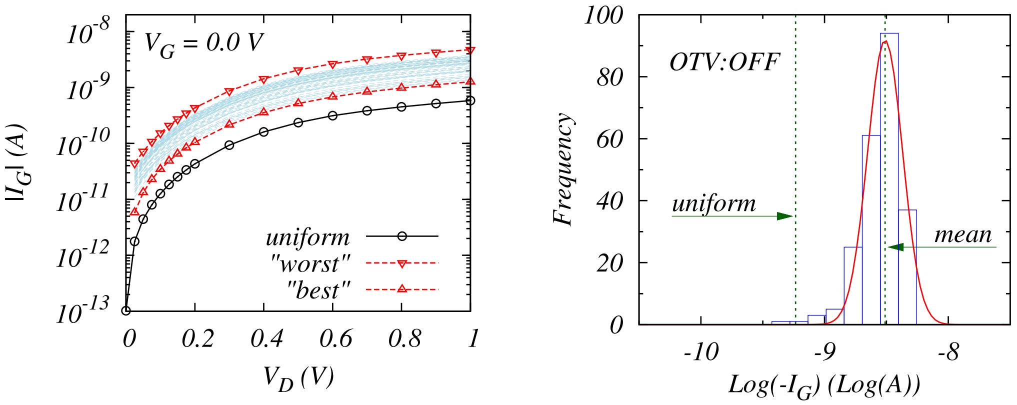


Figure 4.22: $I_G - V_D$ characteristics at zero gate voltage V_G for a sub-sample of 50 devices from the OTV set (left). Gate current histogram for the simulated ensemble of 230 devices with OTV under the Off-state bias conditions (right). A Gaussian distribution with mean value $\langle \log_{10}(I_G) \rangle = -8.5$ and standard deviation $\sigma = 0.1$ is fitted.

hence oxide thickness variation within the area of the drain extension overlap. The area for self-averaging is much smaller, compared to the case of high gate voltage and low drain voltages, and inter-device leakage variability increases. The increase of the mean current, relative to the leakage of the uniform device is due to the same reason as already discussed for the On-state of the transistor. Note that for the device with lowest gate current, nearly the entire drain overlap is under a thicker oxide than the nominal, and this is why the leakage in this case is below that of the uniform device. This is illustrated in Fig. 4.23.

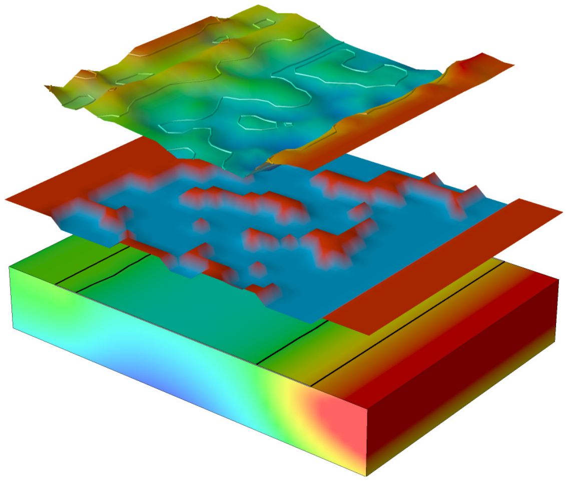


Figure 4.23: Top to bottom: gate leakage density (*Log*-scale); Si/SiO$_2$ roughness pattern (blue dips indicate a region of thicker oxide of 1.15 nm, protruding into the substrate, red islands indicate thinner oxide of 0.85 nm); potential distribution in the substrate. The device with lowest leakage is illustrated. The drain extension (closer to sight) is separated from the gate overlap by a thicker oxide, due to which the leakage of this transistor is lower than the one of the uniform device.

Since the reason for gate leakage variability is essentially the same for both On- and Off-state bias conditions, the increased gate leakage dispersion in the Off-state is a confirmation of

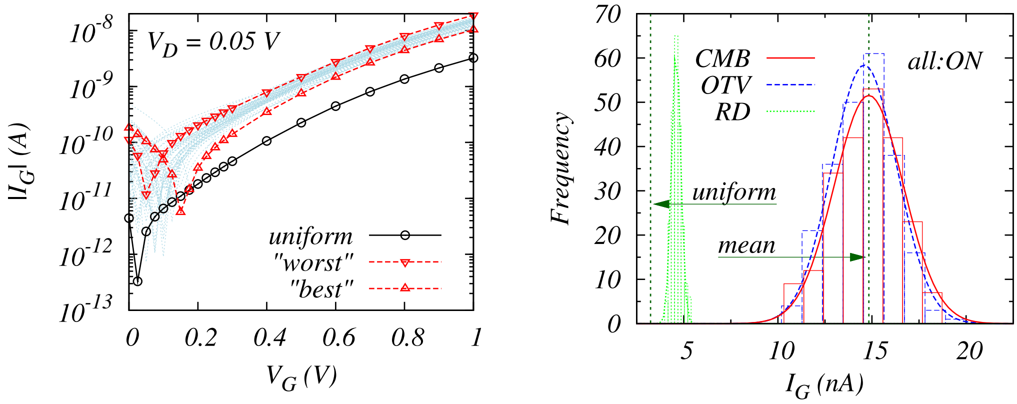


Figure 4.24: $I_G - V_G$ characteristics at low drain voltage of a sub-sample of 50 devices from the CMB set (with simultaneous OTV and RDF), to the left. Gate current histograms for each of the ensembles of 230 devices (RDF - dotted, OTV - dashed, CMB - solid) simulated under ON-state bias, to the right. A Gaussian distribution with mean $\langle I_G \rangle = 14.8$ nA and standard deviation $\sigma = 2$ nA is fitted to the data from the CMB set.

our earlier hypothesis, that a reduction of the gate length at a given correlation length increases leakage variability.

4.3.4 Combined effect of RDF and OTV

In this subsection we report the simulations of the CMB ensemble of devices, featuring both RDF and OTV. We compare the results against the data from the simulations of individual sources of intrinsic parameter fluctuations.

4.3.4.1 $I_G - V_G$ characteristics

The $I_G - V_G$ characteristics of a subset of 50 devices from the COMBINED (CMB) set are shown in Fig. 4.24 (left). As in the earlier discussions, the $I_G - V_G$ curve of the uniform device is shown in black, while the lowest and highest lying curves at $V_G = 1$ V are in red. The spread of characteristics bears the effects of OTV induced variability at high gate voltage, while at low gate voltage, the dispersion is similar to the one induced by RDF. This is expected, since at high gate bias, the excess of electron charge in the substrate screens the bare potential of the ionised impurities, and the RDF induced fluctuations of the tunnelling current density become too localised, compared to the OTV induced fluctuations. At low gate bias and low drain voltage, the screening of the impurities, particularly in the channel of the transistor, is not sufficient, and the RDF induced fluctuations lead to the increase of gate leakage variability. This bias regime is of less importance to the operation of digital CMOS devices, however.

The dispersion of gate current in the On-state of the transistors of each of the simulated ensembles of devices are shown in the right graph of Fig. 4.24. Clearly, for this gate bias and simulation parameters (macroscopic device geometry, and oxide roughness correlation length), gate leakage variability is dominated by the effects of oxide thickness variation, and discrete doping atoms have a negligible impact. This is confirmed also by Fig. 4.25, showing a correlation of the direct tunnelling current density to the features of the oxide interface roughness, for the *worst* device in the CMB set, biased in the On-state. Note that the impurities near the drain (closer to sight) would have lead to sharp local modulations in the leakage density, in the case of a flat interface. In the present case however, the stronger dependence of the direct tunnelling on the barrier thickness dominates.

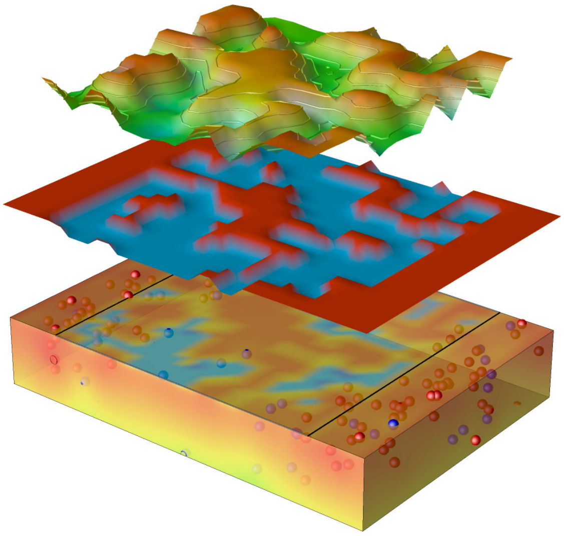


Figure 4.25: Top to bottom: direct tunnelling current density (*Log*10 of the magnitude), Si/SiO$_2$ interface roughness features (blue identifies regions of SiO$_2$ protrusions into the substrate, i.e. thicker oxide; red corresponds to thinner oxide), and electron density distribution in the substrate (to the depth of the shallow source/drain extensions) of the *worst* device from the CMB set. Tunnelling density fluctuations are correlated to the features of the oxide, although electron density is affected by impurities near the interface too.

In the On-state of the transistor, there appears to be some dependence between the RDF and OTV induced gate leakage variability, expressed in a small covariance of 0.16, obtained from $Cov = 0.5(\sigma^2_{CMB} - \sigma^2_{RDF} - \sigma^2_{OTV})$.

4.3.4.2 $I_G - V_D$ characteristics

The $I_G - V_D$ characteristics of a subset of 50 devices from the Combined (CMB) set are shown in the left graph of Fig. 4.26. This graph suggests that both random dopant fluctuations and oxide thickness variation contribute to the simulated gate leakage variability, since the spread is even wider than the one from the simulations of the RDF ensemble, and the mean

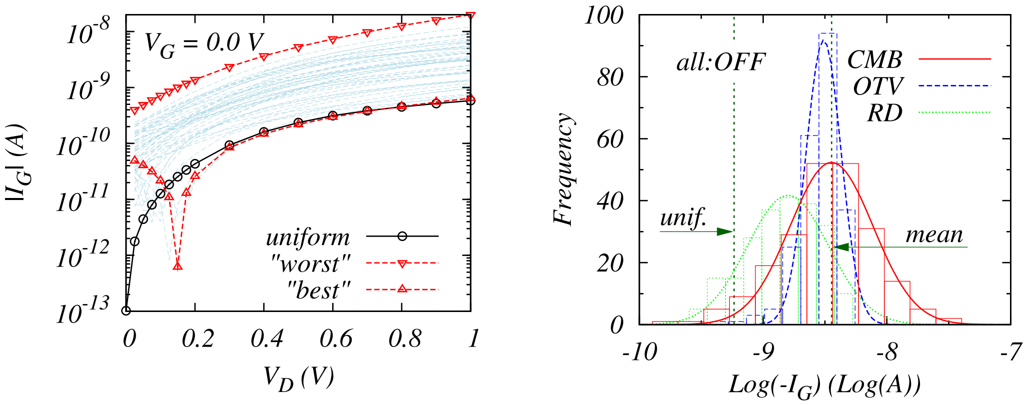


Figure 4.26: $I_G - V_D$ characteristics at 0 gate voltage of a sub-sample of 50 devices from the CMB set (with simultaneous OTV and RDF), to the left. Gate current histograms for each of the ensembles of 230 devices (RDF - dotted, OTV - dashed, CMB - solid) simulated under OFF-state bias, to the right. A Gaussian distribution with mean value $\langle \log_{10}(I_G) \rangle = -8.4$ and standard deviation $\sigma = 0.3$ is fitted to the data from the CMB set.

gate current is significantly increased, as in the simulations of the OTV ensemble of devices. Such an effect is anticipated, because in this case, gate leakage variability is due to the lack of sufficient self-averaging of the fluctuations in direct tunnelling current density that are local to the area of the drain extension overlap. While the density of electrons in the accumulated gate is essentially constant, the change in tunnelling probability (due to the features of the oxide roughness), and the sharp modulation of the oxide field, (due to unscreened impurities in the depleted substrate), are independent phenomena. Their simultaneous manifestation increases the range of fluctuations in the tunnelling current density, and implicates a wider dispersion in the total gate current.

A quantitatively clearer picture is obtained from the histograms of gate tunnelling current in the OFF-state of the transistors in each of the simulated ensembles, shown in the right graph of Fig. 4.26. Note the use of a logarithmic scale, because the spread of gate current exceeds an order of magnitude. The gate leakage distribution of the COMBINED set has the highest mean value of 4 nA, but very close to the mean corresponding to the OTV set, 3.2 nA. For the CMB set, the standard deviation of the logarithm of the gate current is identical to the one for the RDF set. Due to the larger mean value however, this translates to 50-100 % of the mean current (i.e. the interval from $-\sigma$ to $+\sigma$ is equivalent to the range of 2 - 8 nA.

In the OFF-state of the transistor, the OTV and RDF appear to be statistically independent sources of variability, as suggested by the minimal covariance of 0.008 that we establish from the standard deviations of the fitted Gaussian distributions.

4.4 Summary

The study of MOSFET gate leakage variability requires a physically sound 3D device modelling technique and simulations of large ensembles of *macroscopically* identical, but *microscopically* different transistors. We reviewed the three most established device modelling frameworks for performing such analysis – density-gradient (DG), Monte Carlo, and non-equilibrium Green's functions (NEGF) – in view of these requirements. Currently, the NEGF and MC simulations cannot meet the demands of computation efficiency, while in terms of accuracy of device electrostatics, the DG approximation is sufficient, due to its proper accounting of quantum confinement effects. MCs advantages in accounting for hot-carrier effects and velocity overshoot are not critical in ultra-scaled MOSFETs in the regimes relevant for static power dissipation – low drain and high gate voltage, referred to as On-state, and low gate and high drain voltage, referred to as Off-state. Therefore we can dismiss MC as a tool to analyse these particular effects. We cannot profit by the inherent quantum-mechanical treatment of transport in NEGF and its handling of the gate contact and gate current on equal basis to the source and drain, due to the paramount complexity involved in such simulations.

By choosing the DG framework, we benefit from the already established and proven capabilities of the Glasgow *3D Atomistic* simulator for studying device variability. A section of this chapter clarifies the mathematical formulation and implementation aspects of this simulator. In particular, the use of *control volume* discretization of the Poisson and current continuity equations, is essential to allow the modelling of microscopic variations between devices. Physical models of the two principle sources of variability considered in this study - discrete, random dopant fluctuations (RDF), and oxide thickness variation (OTV) due to oxide interface roughness, are clarified to anticipate their effects on direct tunnelling gate leakage in the transistor.

The advancement of the simulator to account for gate leakage is based on the inclusion, at a post-processing stage, of the 1D direct tunnelling model described and used in the previous chapter. Presently, the essential quantities of the model are obtained directly from the electrostatic potential and electron distribution calculated by the 3D DG simulator.

We first simulate a 25 nm square gate MOSFET with a uniform doping profile, flat interfaces, and 1 nm SiO_2 dielectric, and analyse the gate and drain voltage dependence of the gate leakage.

Our results qualitatively agree with previous experimental and modelling works (30; 175; 192; 193; 195), and show that

- maximum gate leakage flows at $V_G = V_{DD}$ and $V_D \sim 0$, due to simultaneous electron tunnelling from all geometrical partitions of the substrate (i.e. source and drain extensions, and the channel);
- gate leakage at $V_G \sim 0$ and $V_D = V_{DD}$ is due to the dominant tunnelling from the gate into the drain extension, and is larger than the sub-threshold (source-to-drain) leakage;
- at high gate and high drain voltage, the source extension and channel provide the dominant tunnelling from the substrate, while the opposite electron flux at the drain extension overlap is minimised.

We further study three different ensembles of 230 devices, macroscopically identical to the 25 nm uniform MOSFET, but microscopically different. Devices differ from set to set in their sources of microscopic features – the RDF set has discrete dopants, random in number and distribution; the OTV set has a rough Si/SiO$_2$ interface introducing random fluctuations of the oxide thickness by one inter-atomic layer of 0.28 nm; the CMB set combines both sources of microscopic differences. The effects of RDF and OTV are relatively independent, and

- OTV increases the mean gate leakage 5 times, over the gate current of the nominal (uniform) device, which is due to the exponential sensitivity of the direct tunnelling to the oxide thickness.
- At $V_G = V_{DD}$ and $V_D \sim 0$, OTV induces appreciable variability with a standard deviation of a few % of the mean, while the RDF adds insignificantly to the spread. This is because the large electron concentration in the substrate at this bias screens the potential of the ionized impurities, while surface roughness affects both oxide field and carrier concentration near the interface. This variability is expected to increase if oxide roughness correlation length becomes comparable to the gate dimensions.
- At $V_G \sim 0$ and $V_D = V_{DD}$, both RDF and OTV contribute to a large spread in the gate current, which for the chosen geometry is nearly two orders of magnitude. The contribution from RDF is stronger since large fluctuations in the local tunnelling density are implicated by the exposed impurities in the depleted substrate. Due to the narrow region (drain-extension overlap) determining the tunnelling magnitude, self-averaging is not effective.

It is evident that gate leakage variability is a very important issue, particularly at high drain, and low gate bias, where gate tunnelling is the major leakage component in the transistor, and variability is large. Further effort should address the self-consistent inclusion of the gate leakage into the drain current calculations, and improve the tunnelling model to account for some multidimensional aspects of the leakage process, currently omitted – lateral field dependence, tunnelling paths through oxide spacers, and effects of width scaling. Naturally, such effort must be put in the perspective of metal/high-κ dielectric gate stacks.

Chapter 5

Oxide Interface Transition

In the review of tunnelling current models (Chapter 3), it was suggested that the oxide interface transition region has an appreciable impact, on the direct tunnelling probability for ultra-thin oxides. This chapter is an extensive study of the non-abrupt change of electronic properties at such an interface and its impact on the quantisation, capacitance and leakage characteristics of a MOS inversion layer. After introducing the issue, as revealed by experiments and first-principles simulations, we detail out the advances made to the 1D Poisson/Schrödinger solver, which were required for this study. Simulations compared with traditional, abrupt interfaces demonstrate an order of magnitude increase in leakage current, enhancement of capacitance, and reduction in subband splitting.

5.1 Introduction

The smallest critical dimension in ultra-scaled MOSFETs is the gate insulator equivalent oxide thickness (EOT), which at the current VLSI technology generation is of the order of 1 nm, and is projected to become as small as 0.5 nm (10). For a pure silicon dioxide or oxynitride, a 1 nm physical thickness is realised in 3 molecular layers (each 0.32 nm thick (199)); for a high-κ dielectric stack, the interfacial oxide is even thinner (11). Despite the renowned quality of the Si/SiO$_2$ interface, a physico-chemical transition happens at this interface over a depth of a few Å, as comprehensively reviewed in Refs. (200; 201). This compositional and structural transition imparts a non-abrupt variation of the electronic and dielectric properties across the interface (186; 199; 202; 203; 204; 205; 206; 207). Such variation is intrinsic to the interface and its impact on the electrical characteristics of devices should become more pronounced with the aggressive scaling of the oxide.

5.1.1 Prior art

So far, only a few device-related modelling studies address the physical consequences of a Si/SiO$_2$ transition layer, the first one dating back to 1977, by Stern (208). Assuming the band-gap varies linearly from that of Si, to that of the oxide, over a distance of 0.5 nm above the substrate, his low temperature (4.2 K) calculations showed about 1 % of the lowest subband inversion charge resides in the dielectric under conditions of high field normal to the interface; depletion charge dependence of this fraction suggested that scattering of charge in the transition layer may be partially responsible for mobility degradation that was normally attributed to surface roughness. The same study, done in the envelope wave function, effective mass approximation, also showed that wave function penetration in the transition layer reduces the split between the lowest lying subbands of the four-fold and two-fold degenerate valleys, by some tens of mV. It is not trivial, however, to extrapolate the low temperature results to room temperature characteristics with a higher level of depletion charge, as found in modern bulk MOSFETs. Later, Maserjian reported a dependence of the Fowler-Nordheim tunnelling current oscillatory component on the width of the transition layer, obtaining best fit for a linear potential barrier transition of 0.25 nm (209). More recently, and adopting similar barrier configuration (0.27 nm wide transition), Yang *et al.* reported an order of magnitude increase in the direct tunnelling leakage, relative to the abrupt band-gap change (210). Watanabe *et al.* devised a self-consistent $C-V$ and J_G-V fitting procedure for the extraction of oxide thickness and oxide tunnelling effective mass (131). They report that best fit to the experimental characteristics is obtained by considering simultaneous transitions of both band-gap and dielectric constant, over a distance of 0.4 nm from the semiconductor. The tunnelling mass in this case, 0.85, is much larger than the one obtained for an abrupt oxide barrier, 0.49, compensating for the enhanced tunnelling due to the gradual conduction band offset. The methodology employed in these two studies (131; 210), does not give detailed information about the subband levels and subband carrier populations.

In contrast to the above findings, de Sousa *et al.* report a decrease in the tunnelling current associated with a gradual transition of the band-gap, compared with an abrupt change (211). This is counter intuitive (particularly for their interface localisation parameter $\alpha = 0.0$), and could be due to either a systematic limitation of their gate leakage model, [1] to an implementation

[1] De Sousa *et al.* use a semiclassical WKB approach for the calculations of tunnelling current (211). It might be that if one considers only the lowest lying subband in the inversion layer, as a result of the widening of the potential well, the effective thinning of the tunnel barrier is overcompensated by the lowering of subband level and impact frequency, associated with the widening of the potential well.

mistake, [1] or both. Another unexpected result from their calculations is the degradation of the total gate capacitance in inversion. [2] We note that the gradual potential barrier at the interface allows stronger penetration of the subband wave functions, in effect bringing the inversion charge centroid closer to the gate electrode and is therefore expected to enhance the gate capacitance.

5.1.2 This work

The contradictory results summarised in the previous sub-section require some clarification. A better understanding of subband quantisation and subband carrier distribution in the presence of interface transition is also necessary, due to the technological importance of the interface – particularly in view of modelling and characterisation of high-κ gate stacks (HKGS) with a sub-nm effective oxide thickness (EOT).

This chapter is an extensive investigation of the impact that interface band-gap and permittivity transition, have on MOS inversion layer characteristics. The next section is a summary of the experimental and simulation findings regarding the electronic and dielectric properties of the Si/SiO$_2$ interface. This is followed by a study of the quantisation, capacitance, and direct tunnelling gate leakage characteristics of a metal/SiO$_2$/p-Si(100) structure. A systematic comparison is drawn between three different models of the interface barrier – the traditional, *abrupt* band-gap transition; a gradual *linear* variation of the band gap across the interface; and a *realistic* conduction and valence band evolution obtained from *ab initio* calculations of the interface. With the new, non-abrupt barrier model, direct tunnelling gate leakage is simulated in devices with sub-nm EOT HKGS, in accord with the ITRS projections.

5.2 The Si/SiO$_2$ interface

5.2.1 Atomic structure

The silicon dioxide used as a gate insulator in MOSFETs is typically a thermally grown amorphous material (212). It is composed of disordered SiO$_4$ tetrahedral structures, in which a Si atom – being at the centre – is bonded to the 4 neighbouring O atoms. Each O atom at a vertex

[1] The published set of equations is incorrect – in particular, the units on each side of Eq. 11 are inconsistent, considering Eq. 13 (211).

[2] De Sousa *et al.* establish capacitance degradation by considering normalised capacitance; however, the normalisation is done by dividing to the oxide capacitance, calculated by excluding the interfacial layer. This is inconsistent with their oxide thickness definition, and with the fact that the interfacial layer in their calculations has the oxide permittivity; hence the normalisation capacitance (i.e. the denominator) of abrupt interface is smaller than the one for non-abrupt interface (211).

is shared between two tetrahedra, as depicted on Fig. 5.1, bringing the average ratio between Si and O atoms to 1:2 – hence, the known stoichiometry of the oxide, expressed as SiO_2 (200). The oxide is grown on the surface of a hydrogen-passivated Si substrate (212). Various models have been proposed to describe the initial oxidation process, in an attempt to predict the possible atomic configurations at the interface (212; 214; 215; 216; 217; 218; 219; 220). However, a more detailed knowledge of the interface has been obtained from spectroscopic measurements at a post-oxidation stage (201), and first principles (*ab initio*) simulations of Si/SiO_2 atomic models (205; 206; 221). While the exact atomic arrangement at the interface remains unclear, and although crystalline SiO_2 forms near the substrate have also been observed (206; 212; 222; 223), two types of transitional layers are commonly distinguished – compositional, and structural – following the review in Ref. (201).

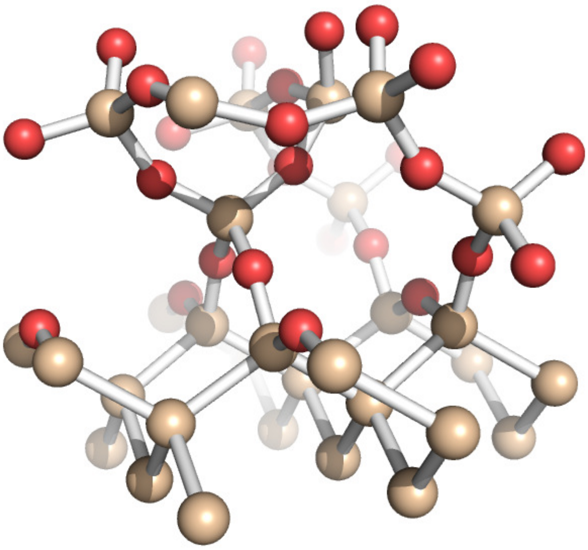


Figure 5.1: Atomic structure of the Si(001)/SiO_2 interface (ball-and-stick model fragment), red (smaller) and beige (bigger) balls represent O and Si atoms respectively. Typical SiO_4 tetrahedral configuration (angles and bond-lengths) are annotated. An ideal (for Si(001) surface) interface is shown, having the minimum possible transition layer – a monolayer of partially oxidised Si atoms with two oxygen bonds (Si^{2+}) – considered to be abrupt (213). The suboxide and the different (relative to bulk) atomic arrangement impart non-abrupt transition in the electronic and dielectric properties at the interface, as summarised later in the text.

5.2.1.1 Compositional transition layer

The *compositional transition layer* (CTL) is characterised by the high content of partially oxidised Si atoms; between 75 and 90 % of the Si atoms have less than four bonds to O atoms, and are commonly denoted as Si^{1+}, Si^{2+} and Si^{3+} (200; 202). This layer, also referred to as the *suboxide* or sub-stoichiometric oxide, links the Si lattice of the substrate to the stoichiometric oxide (the part of the oxide with fully oxidised Si atoms) (201). Although the formation of a suboxide layer over a Si(100) substrate is considered to be energetically costly, its presence is ascribed to the lateral boundaries of energetically equivalent, but structurally different Si-surface orders, and to atomic steps on the Si surface (200; 213; 220).

The suboxide width, reported from recent experiments, varies in the range of 0.2 – 0.5 nm; the variation is attributed to the different quality of the Si surface and oxide growth conditions. In particular, a suboxide depth of 0.16 nm has been inferred from electron-energy-loss spectroscopy (EELS) (186), 0.21 nm was deduced from photoemission (PS) experiments (199), 0.4 – 0.51 nm obtained through angle-resolved Si 2p core level X-ray PS (XPS) (201; 224), and 0.53 nm was fitted to high-resolution Rutherford backscattering spectroscopy (HRBS) data (225). It should be noted that the value of 0.23 nm is the width of the CTL corresponding to an abrupt interface of oxide grown on ideal Si(100) surface; in such case the suboxide consists of one monolayer of Si^{2+} atoms (200; 201; 213). The wider compositional transition is observed together with a distinct localisation of the peak density of the different sub-oxide species (Si$^{1,2,3+}$), as shown in Fig. 5.2 (201; 224).

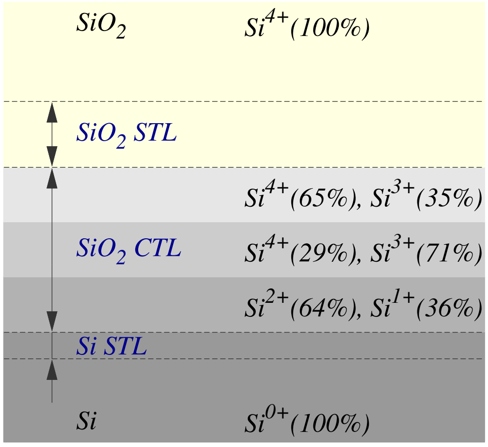

SiO$_2$
Si^{4+}(100%)
SiO$_2$ STL
Si^{4+}(65%), Si^{3+}(35%)
SiO$_2$ CTL
Si^{4+}(29%), Si^{3+}(71%)
Si^{2+}(64%), Si^{1+}(36%)
Si STL
Si
Si^{0+}(100%)

Figure 5.2: Schematic diagram of the Si/SiO$_2$ interface, showing the physical transitional layers that can be identified (201). The structural transition layer (STL) differs from its bulk counterpart by atomic arrangement. The compositional transition layer (CTL) differs from SiO$_2$ by chemical composition, containing sub-oxidised Si atoms ($Si^{1+,2+,3+}$). Sub-layers within the CTL are identified by angle resolved XPS measurements (201; 224). Distribution of suboxide moieties is from (224).

Reports from *ab initio* molecular dynamics simulations typically indicate a 0.4 – 0.6 nm wide suboxide region (206; 216; 226; 227). Abrupt interface models exist in literature (203; 213;

228), but their validity is questioned by ion-scattering experiments and simulation (229). The interface models in Ref. (227) are built to reproduce a range of experimental data for Si/SiO$_2$, emphasising the density ratio between Si$^{1,2,3+}$ atoms, so not surprisingly they yield a similar CTL width to the angle-resolved XPS studies (0.5 nm). However, the 2 nm amorphous-SiO$_2$ on Si(001), reported in Ref. (206) has been obtained from a restraint-free structural optimisation involving 2000 atoms, and features a suboxide of similar width (0.6 nm).

The importance of a relatively wide suboxide cannot be overstated, since the local dielectric response is found to be largely affected by the oxidation state of a Si atom and the particularities of its bonding (201; 230).

5.2.1.2 Structural transition layer

Two structural transition layers (STL) are identified, on each side of the suboxide (201). Their chemical composition is that of the bulk Si (on the side of the substrate), or that of the stoichiometric SiO$_2$ (on the opposite side), as indicated in Fig. 5.2. Their atomic arrangement differs, however, from the corresponding bulk counterparts.

The structural transition in Si appears as a perturbation of the lattice over the topmost two or three monolayers, with atoms being displaced from their regular lattice position by more then 0.09 Å. This result has been obtained from ion backscattering experiments and simulations (229; 231), which probe the Si-side of the interface. While this level of disorder induces strain amongst the Si surface atoms, and affects the oxidation process and generation of defects (as revealed by real time oxidation kinetics observations using XPS (220)), we find no data for its impact on the electronic and dielectric properties of the interface.

The transition on the SiO$_2$ side is studied mostly with XPS, and is believed to extend up to 1 nm into the stoichiometric oxide (201). Experiments suggests that the structural irregularity in this case is expressed in a reduced Si-O-Si bonding angle (135° – 140°) between the SiO$_4$ tetrahedra, compared to the known value of 144° for bulk SiO$_2$ (200; 201). This irregularity is found to affect the valence band offset, which increases within the SiO$_2$ STL by up to 0.2 V (201; 204).

5.2.2 Electronic and dielectric properties

5.2.2.1 Band gap transition

The energy band diagram of the Si/SiO$_2$ interface was established by Williams (232), who determined the band-gap alignment between Si and SiO$_2$, by measuring the energy required to excite an electron from the valence band of Si into the conduction band of SiO$_2$. It was assumed that the change of electronic band gap happens abruptly, as depicted by the solid line in Fig. 5.3. The offset of the SiO$_2$ conduction and valence band edges, relative to those of Si, is often considered sufficiently large to justify an even further simplification – that of an infinite potential barrier at the interface – for the purposes of device modelling, following Ref. (233).

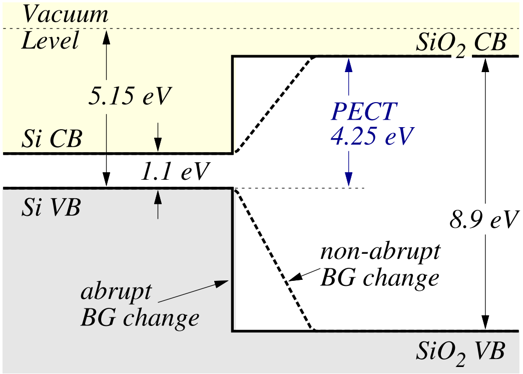


Figure 5.3: Energy band-diagram at the Si/SiO$_2$ interface. Band gap (BG) alignment between Si and SiO$_2$ is established from the photo-emission current threshold (PECT) for Si valence band (VB) electrons excited to the SiO$_2$ conduction band (CB) (232). Recent studies indicate a non-abrupt BG transition at the interface (dashed) (186; 201), rather abrupt (solid).

The assumption of an abrupt band-gap change at the interface was first revisited by Stern who – motivated by the early stoichiometry-related studies of the Si/SiO$_2$ interface – assumed a linear band-gap transition over 0.5 nm in the oxide, as schematically illustrated by dashed line in Fig. 5.3, and briefly addressed, within the envelope wave-function approximation, the effect of this transition on MOS inversion layer quantisation and mobility (208). He found that for a strongly confining potential in the substrate, there is a significant wave function penetration into the oxide that results in lower subband levels, relative to the abrupt case, and decreased splitting between the lowest lying subbands of the four-fold and two-fold degenerate valleys by some tens of mV. Later, Maserjian considered the effect of the transition layer on the frequency of Fowler-Nordheim (FN) tunnelling current oscillations, and obtained the best fit to experimental data by assuming a 0.26 nm linear band-gap transition (209). It is worth noting that the band-gap transition width in both these cases is of the order of the sub-stoichiometric oxide, discussed already.

While most of the experimental work on determining the stoichiometry and atomic arrangement at the Si/SiO$_2$ interface is based on XPS, it yields information about the atomic core levels (e.g. Si 2p, O 1s), or the valence band, but lacks information about the conduction band. Definitive experimental evidence for the non-abrupt transition of the interface band-gap was published by Muller and co-workers, who used electron energy loss spectroscopy (EELS) to locally resolve the density of unoccupied electronic states by atomic sites and atomic species near the interface (186). In this experiment, they observed the electronic states in the interface transition region to be roughly aligned with the conduction band of Si, rather than that of SiO$_2$, and found a satisfactory large band gap in the oxide only after 0.43 nm. At the same time, the deduced depth of the suboxide was merely 0.16 nm. Subsequent investigations correlate these additional unoccupied states with bulk-Si induced conduction states, and with O p-projected unoccupied densities of states (due to O atoms near the surface, with less than six O second nearest neighbours, cf. Fig. 5.1) (199; 204; 205; 207; 234; 235).

The fundamentally important conclusion from these studies is that regardless of the type of Si-to-SiO$_2$ chemical transition – abrupt, or graded through a suboxide of a finite width – the band gap transition at the interface is not discontinuous, but progressive, over a distance of about 0.5 nm, with a minor change in the first 0.2 – 0.3 nm away from the Si lattice. Moreover, this intrinsic feature of the interface should be relatively independent of the specific atomic structure of the interfacial oxide (crystalline or amorphous).

Further to the above experiments, a number of density functional theory (DFT) simulation studies of the interface electronic properties report a progressive band gap transition over 0.5 – 0.6 nm. These studies consider different SiO$_2$ model structures (crystalline or disordered polymorphs) varying the complexity of the simulated entity (2D hydrogen-terminated or 3D periodic supercells of 30 to 2000 atoms) (203; 205; 206; 236; 237; 238).

Ab initio simulation results are of particular relevance to device modelling, since they allow one to trace the evolution of the conduction and valence band edges versus distance from the Si interface (205; 236; 239). Typically, DFT calculations of the electronic structure underestimate the band gap of bulk Si and SiO$_2$, and have some dependency on the choice of an empirical functional. Recent calculations overcome these issues by the use of the GW-approximation (240; 241; 242). Further to the good agreement with the known experimental band gaps, these works confirm that DFT well describes the interface charge density and dipole moments at the interface.

5.2.2.2 Permittivity variation

Permittivity variation that is local to the interface region has been suggested, based on the compositional and structural transitions at the interface described before (186; 200). Experimentally, an enhanced high-frequency dielectric response has been deduced from a comparison of thin oxide film thicknesses measured by ellipsometry, against the thicknesses obtained from other techniques, which are independent of permittivity or refractive index (e.g. ion scattering and high resolution transmission electron microscopy). A number of authors confirmed an increased refractive index of oxides below 7 nm, and attributed this to the interface transition (243; 244; 245). Detailed investigations, reported in Refs. (230) and (201), revealed that the increase of both the static and high-frequency permittivity is due to the sub-stoichiometric oxide. The authors assign this to the larger polarizability of the electron distribution associated with partially oxidised Si atoms, and agree quantitatively in their estimates, as shown in Fig. 5.4, despite employing different methodologies to calculate the effect. A bulk-like dielectric constant is restored as soon as the Si atoms become fully oxidised, independently of the density of induced gap states (246; 247; 248). This fact leads to the important conclusion that the dielectric transition at the interface happens over a shorter distance, compared to the band-gap transition, since the former is sensitive to the first nearest neighbour (Si-O) atomic arrangement (246), while the latter is also sensitive to the second nearest neighbour (O-O) (186; 234), and exceeds the width of the suboxide, as previously discussed. This is also evident in the band-gap and permittivity profiles versus distance, reported in Ref. (205).

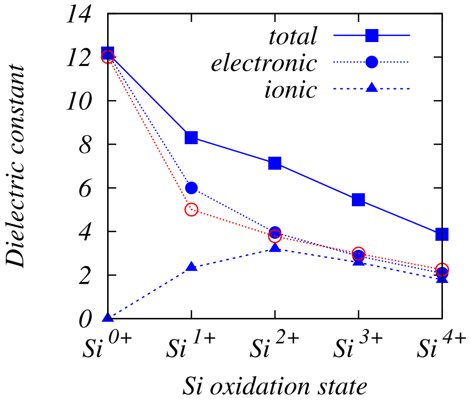


Figure 5.4: Static dielectric constant as a function of Si oxidation state. Total (squares), as well as electronic (disks) and ionic (triangles) contributions, obtained from DFT calculations are shown (205); electronic contribution deduced from XPS (open circles) are in agreement (201). Enhanced permittivity occurs in the suboxide (Si^{n+}, with $n = 1, 2, 3$).

This permittivity enhancement at the interface is of crucial importance to both capacitance (CV) and gate leakage ($J_G V$) characterisation of thin oxides, since an increased oxide capacitance, for a given physical thickness, induces more charge in the inversion or accumulation layer at a particular bias, to which both measurements are sensitive. This issue has already been addressed by Watanabe, who established an improved fitting of oxide thickness and oxide tunnelling effective mass, by accounting for a linear transition of both band gap and dielectric constant, over a distance of 0.4 nm (131).

5.3 Methodology details

5.3.1 Simulation approach

It is important to understand the impact of the non-abrupt electronic and dielectric transition at the semiconductor-oxide interface, on the electrical (leakage and capacitance) characteristics of a MOS structure. For this purpose we use a one-dimensioanal (1D) quantum-mechanical simulator employing the envelope wave-function, effective-mass approximation (EMA) (249). Such a tool is suitable for analysing inversion layer properties (121; 233), and is often used for the characterisation of tunnelling oxides through $C-V$ and J_G-V measurements (245; 250). A more direct approach for leakage modelling might be based on first-principles or tight-binding simulations (132; 133; 251; 252), but these techniques are still not suitable for device simulation over wide bias range – the wide extent of an accumulation or depletion region, containing thousands of atoms, represents too onerous a computational burden, and renders them unsuitable.

The mathematical formulation and implementation details of the simulator, the development of which constitutes part of this work, are given in Appendix A. In summary, the commonly available 1D self-consistent Poisson-Schrödinger solver SCHRED-2.0 (253) has been modified to allow for

- solution of the Schrödinger equation (SE) with spatially varying effective mass (to account for wave-function penetration into the dielectric),
- solution of the Poisson equation (PE) with spatially varying permittivity (to account for non-zero charge density in the dielectric, due to wave-function penetration),
- calculation of the direct tunnelling gate current,
- external definition and spacial dependence of material parameters - band gap, $E_G(x)$, dielectric constant, $\kappa(x)$, and effective mass, $m(x)$,

where x is the direction normal to the Si/SiO$_2$ interface. The implementation delivers wave-function solutions subject to open boundary conditions at the dielectric interfaces, and non-zero charge density in the oxide. A 6-ellipsoidal electron band structure is assumed for the Si conduction band, and a single, parabolic band structure with effective mass $m_{ox} = 0.5m_0$ approximates the SiO$_2$ for the greater part of this study. Inversion layer leakage is calculated as direct tunnelling of electrons in quasi-bound states (QBS), treated quantum mechanically, but non-self consistently – this is a reasonable approximation under the assumption that the inversion charge density is guaranteed by an adjacent source contact (as in a MOSFET), and QBS time evolution needs not be followed (114).

Our modifications of the solver make it suitable for the simulation of gate stacks with high-κ dielectric layers and interfacial oxide. However, we start our investigation with the simplest structure – metal/SiO_2/p-Si(100) – to avoid complications from poly-gate depletion effects, and to understand the effects pertaining to the semiconductor-oxide interface transition, in comparison with an ideal, sharp interface. The effects of interfaces in hafnia-based gate dielectric stacks are presented later in this chapter.

All calculations assume room temperature of 300 K. Exchange and correlation corrections to the electrostatic potential are not considered in this study, although they could be accounted for in the original SCHRED code via the local density approximation (254). Their effect is known to reduce slightly the subbands in the inversion layer and correspondingly increase the carrier density at a given bias by about 5 % (255); that impact, in the scope of the current study, is expected to be the same, regardless of the profile of the interface band-gap and permittivity transition.

5.3.2 Interface barrier model

Hereafter, the *interface barrier model* (or simply, interface model) refers to the position of a *nominal interface*, with respect to the end of the ordered Si-lattice of the substrate, and the specifics of the conduction and valence band profiles around that *nominal interface*. The *nominal interface* constitutes the boundary between the substrate and the gate dielectric, from a device modelling perspective.

In view of the chemical and structural properties of the Si/SiO_2 transition, a definition of a nominal interface is ambiguous. Nevertheless, the topmost fully Si-coordinated atoms of the substrate lie on a (001) plane, and we associate the nominal interface with this plane. This is justifiable if we ignore interface roughness in the form of Si-monolayer steps, and the relatively minor lattice displacements due to residual, oxidation-induced stress [1]. By implication, the entirety of the compositional transition layer is assigned to be part of the oxide, in which case, the band gap and dielectric constant gradually transition within the oxide, or abruptly change at the nominal interface. For the metal/SiO_2/p-Si(100) structure in mind, the two possibilities are schematically illustrated in Fig. 5.5, and represent the *linear* and *abrupt* interface barrier

[1] Cf. 5.2.1.2 Structural transition layer

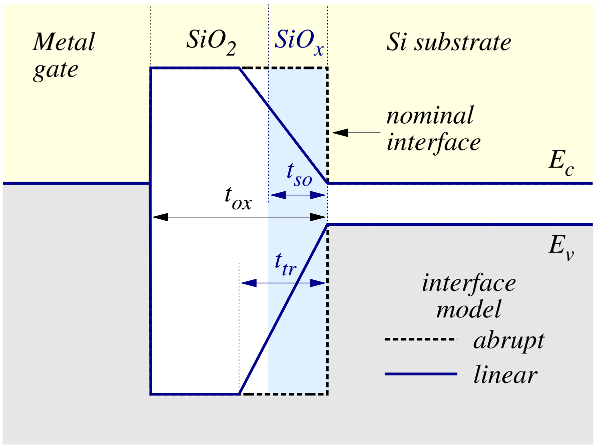


Figure 5.5: Schematic (flat-) band diagram of the *abrupt* and *linear* interface barrier models, showing the position of the *nominal interface* (the same for both models) that defines the oxide thickness (t_{ox}). For the *linear* barrier model, the band-gap transition width (t_{tr}), and the suboxide (light-blue shaded region) width (t_{so}) are also shown. Permittivity transition pertains to the suboxide only.

models, to be compared in the next section. Note that the physical thickness of the oxide is defined as the distance between the nominal interface and the metal/oxide interface; the latter is assumed to be structurally and electrically sharp.

An alternative view is to conceive a plane, parallel to (001), in the middle of the suboxide, as a representative of the nominal interface; such view is considered in Refs. (131; 132).

The last interface barrier model we introduce in this work is obtained through DFT calculations of the interface electronic structure, as described in Appendix B, and is referred to as *realistic*.

5.4 Impact on MOS inversion layer

This section reports a comparison between the inversion layer characteristics simulated with the different interface barrier models described previously. The modelled structure is metal/SiO_2/p-Si(100) with 4.05 eV gate work function, and a uniformly doped substrate to 2×10^{18} cm^{-3}, unless otherwise stated. Oxide thickness, defined as the distance between the nominal and the metal-gate interfaces, is varied in the range of 1 – 3.5 nm. To make a meaningful comparison and estimate the impact of the band-gap transition alone, the permittivity between the two interfaces is at first fixed to that of bulk oxide, preserving an identical EOT for all barrier models. The *abrupt* interface model represents a finite potential barrier, and is associated with an open boundary condition for the wave functions, so that a finite amount of charge penetrates in the oxide. This model is the reference point for the subsequent comparisons. An *infinite* barrier model is also simulated – it reflects a closed boundary condition for the envelope wave functions at the nominal interface, so that charge density at this interface vanishes. Together with the *abrupt* barrier, it represents the traditional device modelling perspective of the Si/SiO$_2$ interface. The *linear* barrier model is regarded as an evolution over the *abrupt* model, with the aim to obtain a more accurate physical picture. It is characterised by a band-gap transition layer, t_{tr}, that is varied in the range of $0.2 - 0.6$ nm in the following simulations. The widest transition layer of 0.6 nm approximately corresponds to the total transition width in the *realistic* interface model. The evolution of the band edges in this model is markedly different however, in the latter case (cf. Appendix B). A comment on the non-linearity of the band-transition profiles is finally presented.

5.4.1 Electrostatics

We first consider the electrostatic effects, reporting the conduction band profile and electron density distribution for three different interface barrier models – *linear*, *abrupt*, and *infinite* – in Fig. 5.6. The band-gap transition region of the *linear* barrier model is $t_{tr} = 0.5$ nm. The non-zero wave density in the oxide for the finite *abrupt* and *linear* barriers corresponds to a finite amount of charge being in the dielectric. Clearly, in Fig. 5.6, electron penetration for the *linear* case is stronger, bringing the inversion charge centroid closest to the gate. Simulations with different oxide thickness, $t_{ox} \in [1.0, 3.5]$ nm, doping level, $N_A \in [1 \times 10^{17}, 3 \times 10^{18}]$ cm^{-3}, and at different gate bias, $V_G \in [0.0, 2.5]$ V, result in the same trend.

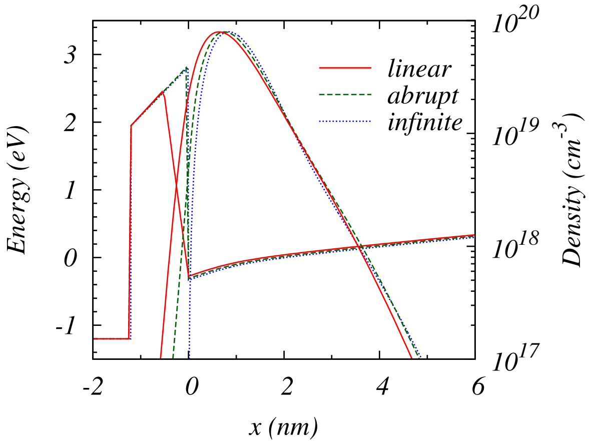


Figure 5.6: Conduction band profile and electron density distribution at 1.2 V gate bias ($t_{ox} = 1.2$ nm, $t_{tr} = 0.5$ nm, $N_A = 2\times10^{18}$ cm^{-3}). Identical magnitude of the peak electron density and band-bending in Si is observed, regardless of the interface barrier (linear, abrupt, or infinite). To the linear barrier corresponds the strongest penetration of electrons however, and charge centroid moves closest to the gate.

It is interesting to know the fraction of the inversion population that penetrates the dielectric, since carriers in the oxide will exhibit different lateral transport properties than electrons in the substrate (208). This fraction reaches a few percent at strong inversion, and is shown in Fig. 5.7 as a function of inversion charge sheet density, N_i, for a few different levels of depletion

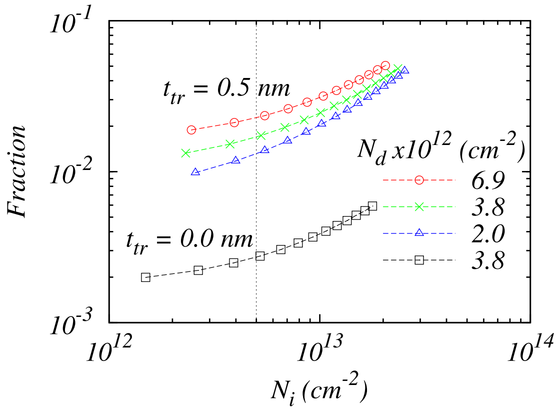


Figure 5.7: Fraction of electrons in the oxide, as a function of the inversion charge sheet density, N_i, with the depletion charge sheet density, N_d, as a parameter. Quoted N_d values correspond to substrate doping level in the range of 5×10^{17} to 3×10^{18} cm^{-3}.

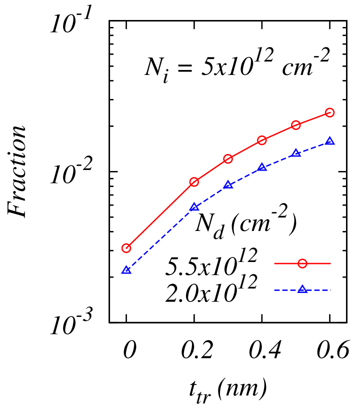


Figure 5.8: Fraction of electrons in the oxide versus band-gap transition width, t_{tr}, at fixed electron sheet density, N_i. Results are independent of the oxide thickness.

charge sheet density, N_d. A ten-fold increase, relative to the *abrupt* barrier, is observed when the gradual band-gap transition happens over 0.5 nm, and is due to the smaller potential barrier within the transition region. There is also a dependence on the depletion charge sheet density. This latter fact is due to the stronger band bending required to obtain a given level of inversion carrier density; as a result, subband levels raise, making the potential barrier effectively lower, hence wave-function penetration stronger (*cf.* discussion on quantisation, 5.4.2). The dependence of the fraction of electrons in the oxide on the width of the band-gap transition region, t_{tr}, is shown in Fig. 5.8, at a fixed inversion charge density, N_i, of 5×10^{12} cm^{-2}. Despite the strong sensitivity on the slope of the interface potential barrier, there is essentially no dependence on the simulated oxide thickness, for $t_{ox} > 1.0$ nm, which is a direct consequence of the sufficiently large conduction band offset of the oxide, leading to a fast wave function decay. A characteristic length λ of a wave function penetration in the oxide is obtained from the slope of its logarithmic derivative at the nominal interface, and for $t_{tr} = 0.6$ nm the value is $\lambda \sim 0.4$ nm – two times larger than that for the abrupt interface, 0.2 nm. The characteristic length for charge penetration is $\lambda/2$, as the charge density is proportional to the square of the wave function modulus.

The results above lead to the conclusion that the non-abrupt interface further enhances the effects of wave function penetration, leading to a relative reduction of the electrical oxide thickness and increased capacitance, with respect to the *abrupt* barrier (256; 257).

Gate capacitance enhancement is readily observed in Fig. 5.9, reporting the normalised capacitance for the three MOS capacitors corresponding to Fig. 5.6. Total capacitance, C_G, is obtained by differentiating the integrated sheet charge density, Q_t with respect to the gate voltage, V_G, as follows: $C_G = dQ_t/dV_G$. The oxide capacitance used for normalisation is identical for the three devices, $C_{ox} = \epsilon_{ox}/t_{ox}$, since permittivity transition is not considered here, and the choice of interface barrier model yields identical oxide thickness, t_{ox}. The relative difference in gate capacitance is bias dependent, as clearly shown in the same figure, 5.9, for the *linear* case, with respect to the abrupt. Given the oxide thickness of 1.2 nm and 0.5 nm band transition width in this case, the strong inversion capacitance is increased by over 5 %.

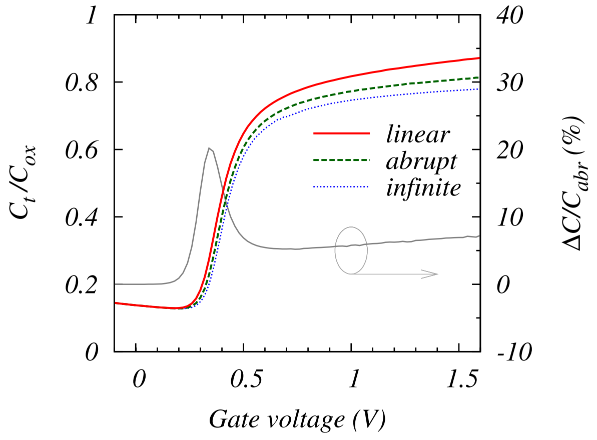


Figure 5.9: Normalised $C-V$ characteristics for the devices from Fig. 5.6. Enhancement of wave-function penetration in the oxide, greatest for the *linear* interface barrier model, moves the inversion charge centroid closer to the dielectric, hence increasing the gate capacitance relative to the case of infinite potential barrier.

The strongest discrepancy however is at lower gate bias, near the threshold voltage. This is consistent with the fact that the inversion charge centroid is closer to the gate electrode, resulting in a lower effective oxide thickness, and hence a lower threshold voltage V_T. V_T reduction is reflected in the left shift of the $C-V$ curve, for linear transition profile, in Fig. 5.9. We found this reduction to be in the order of 20 mV, independent of the simulated oxide thickness in the range of 1.0 − 2.0 nm, for a fixed transition width of 0.5 nm.

5.4.2 Quantisation

At this stage it is important to address the impact of the oxide interface transition on subband energy levels and subband occupancy, on which the properties of the 2D electron gas in the inversion layer strongly depend.

Figure 5.10 compares the ground state wave functions (WF) of the 2-fold (denoted $\Psi_{\Delta 2}$) and 4-fold (denoted $\Psi_{\Delta 4}$) degenerate Si Δ-valleys, for a device simulated with *infinite*, *abrupt*, and *linear* barrier models; the conduction band profiles (linear and abrupt) are also shown for clarity. Due to the lower quantisation mass and higher energy level, the wave functions of the $\Delta 4$ valley are more seriously affected by the type of barrier, and their peak density shifts closer to the interface, as the confinement to the left is relaxed. This corresponds to the increased electron density near the interface, discussed in the previous subsection. Noticeably, for the *linear* barrier model, the peak density is equidistant from the nominal interface for both wave functions; if lateral transport in the inversion layer is considered, essentially all carriers will be subject to interface-related scattering to the same extent.

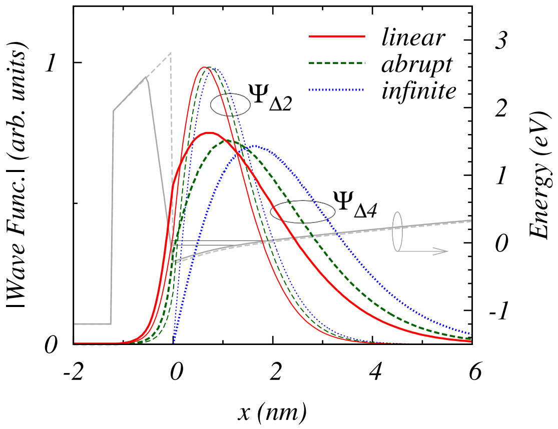


Figure 5.10: Ground state wave functions of Si $\Delta 2$ and $\Delta 4$ valleys, for three different barrier models, as indicated. The conduction band profile for abrupt and linear interfaces is also shown. The gradual conduction band discontinuity impacts mostly $\Psi_{\Delta 4}$, making its peaks, and that of $\Psi_{\Delta 2}$ equidistant from the nominal interface. Horizontal lines indicate the lowest two subband levels for the linear barrier model. V_G is 1.2 V.

The top graphs of Fig. 5.11 report the variation in energy level of the lowest two subbands, denoted $E_0^{\Delta 2}$ and $E_0^{\Delta 4}$, for the three different interface barrier models. An obvious consequence of the weaker confinement at the interface, due to the open-boundary condition for the wave functions, is the lowering of both subbands, which is most pronounced for the *linear* barrier model. In that case, the broadening of the quantum well with energy means the higher subbands are affected more strongly, hence reducing the splitting between the $\Delta 4$ and $\Delta 2$ ground states, which are closest to the Fermi level. A less intuitive result of this process is the significant redistribution of carriers between the two valleys – their occupancy (percentage of the total inversion population) is shown in the lower graphs of Fig. 5.11. While for an infinite barrier the lowest 2-fold degenerate subband contains nearly the entire inversion population (more than 90 %), the finite abrupt barrier reduces this figure by 10 % and increases the population in the lowest 4-fold degenerate subband accordingly. In the case of a 0.5 nm linear interface transition,

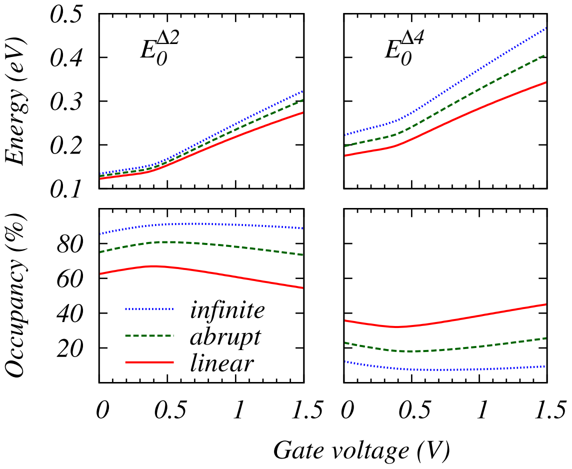


Figure 5.11: Ground state levels (above) and occupancy (below) versus gate voltage for the Si Δ2 and Δ4 valleys. Lowering of the lowest Δ4 subband, $E_0^{\Delta 4}$, is stronger, and reduces the split between the two valleys by ∼ 25 meV, leading to a significant change in their occupancy, more dramatic for higher gate voltage. Device oxide is 1.2 nm; interface transition width is 0.5 nm; substrate doping is 2×10^{18} cm^{-3}.

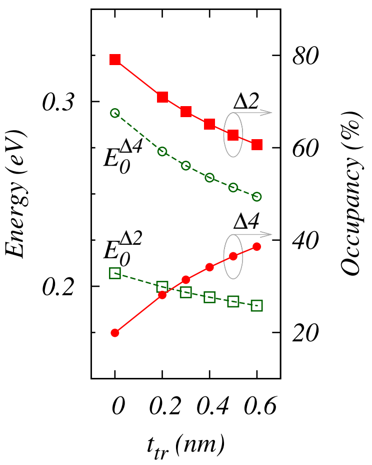


Figure 5.12: Subband energy and occupancy versus barrier transition width, t_{tr}, at a fixed inversion charge, $N_i = 5 \times 10^{12}$ cm^{-2}. At the widest simulated transition, the occupancy of Δ4 (Δ2) valley increases (decreases) by 20 %, independently of oxide thickness.

the effect is even more dramatic – nearly 40 % of the carriers are in the Δ4 valley, and only around 60 % in the Δ2 valley. Again, in consideration of lateral transport in the inversion layer, a big fraction of the electrons contained in the Δ4-valley will respond to an applied electric field with heavier effective mass of $0.315m_0$, rather than with $0.19m_0$ of the electrons for the Δ2 valley (258).

The trends of sub-band levels and occupancy changes with the total width of a *linear* band-gap transition are shown in Fig. 5.12, for a fixed inversion charge density, N_i, of 5×10^{12} cm^{-2}. Given the same level of depletion and inversion charge sheet densities in constructing this graph, the results are independent of the oxide thickness, which is larger than the characteristic length of wave function penetration into the oxide.

5.4.3 Direct tunnelling gate current

Direct tunnelling gate current density dependence on gate voltage is shown in Fig. 5.13, for two different oxides – 1.2 and 1.8 nm thick – and two different interface barrier models – *abrupt* and *linear*. Comparison between the two barrier model suggests a ten-fold increase in gate leakage due to the non-abrupt band-gap transition of 0.5 nm in the *linear* case. The enhancement in tunnelling is consistent over the entire range of simulated gate-voltage, and is brought about by the effective thinning of the tunnelling barrier, as well as the increased electron density in immediate proximity to the barrier at a given gate bias. The results are in agreement with simulations reported in Ref. (210) and fitting to leakage data in Ref. (131).[1] Therefore the choice of interface barrier model, and width of the band-gap transition appears to be of great importance for leakage characterisation of tunnelling oxides.

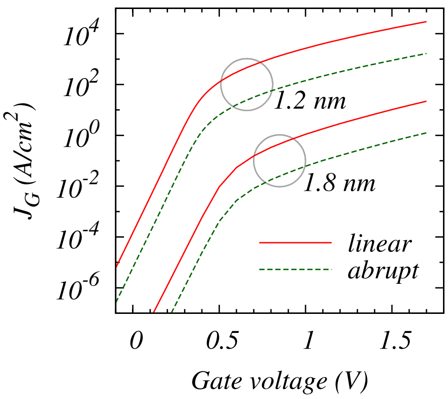


Figure 5.13: Direct tunnelling gate current characteristics for oxide thicknesses of 1.2 and 1.8 nm, with *abrupt* and *linear* interface barrier models. Transition width for the *linear* interface model is 0.5 nm; substrate doping is 2×10^{18} cm^{-3}.

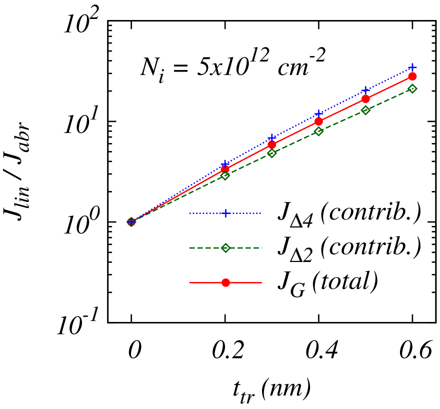


Figure 5.14: Tunnelling current density dependence on the width of the band-gap transition of the linear model – relative difference with respect to the abrupt case is shown; the change in contributions from the Δ2 and Δ4 valleys is also displayed.

The relative increase in tunnelling current density, J_G, with respect to the abrupt interface model, at a fixed inversion charge density, N_i, of 5×10^{-12} cm^{-2}, is shown in Fig. 5.14, as a function of the total width of the linear band-gap transition, t_{tr}. The leakage enhancement at the widest simulated transition of 0.6 nm exceeds one order of magnitude, and since comparison is reported at a constant level of inversion charge, the results are independent of the oxide thickness. The relative increase in the contributions from carriers in the Δ2 and Δ4 Si valleys, denoted $J_{\Delta 2}$ and $J_{\Delta 4}$ respectively, is also displayed in Fig. 5.14. Note that the largest relative increment corresponds to tunnelling of electrons from the four-fold degenerate sub-bands. This is due to the relative increase of the occupancy of this valley, arising from the redistribution

[1] In Ref. (131), there is no direct comparison of leakage with abrupt and non-abrupt interface, but more than a 2-fold increase in oxide effective mass is required to fit leakage with non-abrupt interface to experimental data

of carriers discussed earlier (*cf. Quantisation* 5.4.2), and to the thinner potential barrier, since energy-wise, the Δ4 sub-bands are higher than the Δ2 ones.

5.4.4 Interface permittivity increase

Permittivity enhancement at the interface is a bias independent phenomenon that is due to the sub-oxidised Si atoms in the compositional transition layer (CTL) (201; 230). Consequently, the most simplistic view is that it imparts a reduction of the effective oxide thickness (EOT) of a MOS capacitor. A dual layer, classical electrostatics model of the oxide is capable of reproducing the EOT obtained from microscopic permittivity calculations in the DFT formalism (246). In the dual layer model, the oxide is composed of bulk SiO_2 with a constant permittivity $\varepsilon_{ox} = 3.9\varepsilon_0$, and an interfacial suboxide with varying dielectric constant. If the oxidation index n of the Si^{n+} moieties is linearly graded over the suboxide, the transition of the dielectric constant is also linear (230). While the most recent experiments reveal a non-linear distribution of suboxide species in the CTL, the linear approximation allows for analytic estimate of the EOT in the presence of a non-abrupt permittivity change. The following relation determines the EOT of the suboxide itself (EOT_{so})

$$EOT_{so} = \varepsilon_{ox} \int_{-t_{so}}^{0} \frac{1}{\varepsilon(z)}\,\mathrm{d}z, \tag{5.1}$$

and with the help of the linear approximation it yields

$$EOT_{so} = t_{so} \frac{\varepsilon_{ox}}{\varepsilon_{si} - \varepsilon_{ox}} \ln \frac{\varepsilon_{si}}{\varepsilon_{ox}}, \tag{5.2}$$

where t_{so} is the suboxide width, ε_{ox} and ε_{si} are the bulk SiO_2 and Si permittivity respectively. Figure 5.15 shows the departure of the electrical from the physical oxide thickness due to a linearly graded suboxide of different width. The EOT is reduced by nearly a half of the suboxide width, when only the Si/SiO_2 interface is considered (e.g. for a device with a metal gate). For a poly-Si gate, the EOT is smaller than the physical thickness almost by t_{so} itself, since the poly-Si/SiO_2 interface exhibits identical properties to the Si/SiO_2 interface (186).

It is interesting to know if the permittivity enhancement might be accounted for in device modelling by assuming the dual layer (DL) model with a fixed permittivity in both layers – compare the results for linear $\varepsilon_{so}(x)$ and for fixed $\varepsilon_{so} = 7.1\varepsilon_0$ in Fig. 5.16. The EOT for the simulated devices is 1.2 nm, $t_{ox} = EOT$ for the abrupt interface (homogeneous, bulk-like oxide), and $t_{ox} = 1.43$ nm for the DL model, since the suboxide is 0.5 nm. Due to the identical EOT, the electrostatic potential and electron density profiles overlap in the inversion layer, to the right of

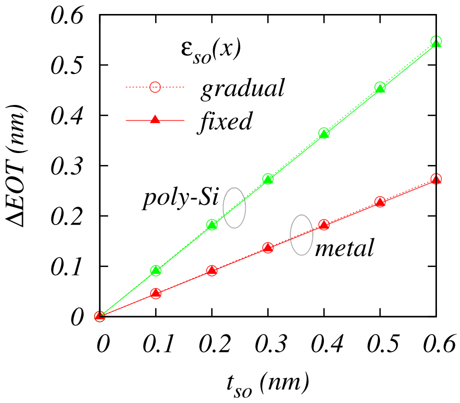


Figure 5.15: EOT reduction due to suboxide permittivity enhancement, obtained from $\Delta EOT = t_{so} - EOT_{so}$. EOT_{so} is obtained from Eq. 5.2 (gradual) or from $t_{so}(\varepsilon_{so}/\varepsilon_{ox})$ (fixed $\varepsilon_{so} = 7.1\varepsilon_0$ (246)). Upper curves are for a poly-Si gate with identical suboxide region on both sides of the oxide; lower curves are for a metal gate having a sharp interface with the oxide.

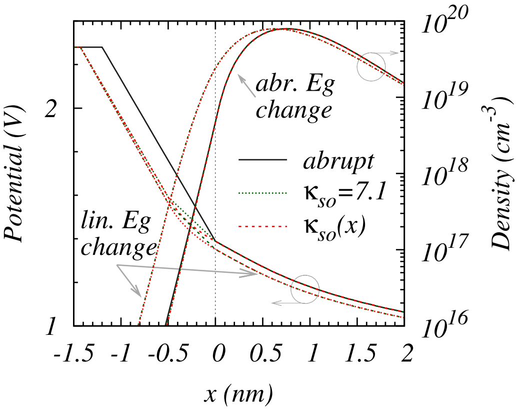


Figure 5.16: Electrostatic potential and electron density distribution for MOS capacitors with $EOT = 1.2$ nm. Three cases are compared – abrupt interface (solid), with homogeneous oxide ($t_{ox} = 1.2$ nm); dual layer (DL) oxide with $t_{ox} = 1.43$ nm, of which the suboxide is 0.5 nm with a fixed permittivity $\varepsilon_{so} = 7.1\varepsilon_0$ (dotted), or a linearly varying $\varepsilon_{so}(x)$ (dashed). Results with the two DL models are identical even if non-abrupt band-gap transition is considered.

the thin vertical line denoting the nominal interface. The oxide field in the bulk-like part of the oxide is essentially that of the abrupt case. The only difference appears in the suboxide region, where the field for the DL with linear $\varepsilon_{so}(x)$ is initially lower than that for the fixed suboxide permittivity. The situation is the same even if interface band-gap is non-abrupt, but linearly varying (over 0.6 nm). Note however, that at a given EOT, the increase of physical thickness due to the interface permittivity enhancement leads to a decrease in the leakage current (not shown here). This is of great relevance for fitting leakage current density to experimental data, where the oxide thickness is obtained from $C - V$ measurements or ellipsometry. For example, at $V_G = 1.2$ V, the device with single layer oxide ($EOT = t_{ox} = 1.2$ nm) and abrupt interface has gate leakage density of 3.15×10^2 A/cm^2, the DL model with abrupt band-gap transition yields 0.19×10^2 A/cm^2, while the same model with band-gap transition of 0.6 nm yields 5.67×10^2 A/cm^2.

5.4.5 Effective mass change

The results presented so far reflect the assumption of a fixed effective mass in the oxide, $m_{ox} = 0.5m_0$, further implying that the effective mass characterising the two- and four-fold degenerate valleys of the Si substrate abruptly changes its value at the nominal interface. In what follows we demonstrate that the magnitude of the observed effects due to non-abrupt transition of the band-gap largely depend on the value of the oxide mass m_{ox}, and that a non-abrupt effective mass profile across the interface reduces the impact of the band-gap transition.

It is useful to first consider the boundary conditions for the wave function (WF) at the interface (*cf.* Appendix A). The matching of the WF derivative at the interface guarantees conservation of particle flux across the interface, and if written independently of the discretization scheme is

$$\frac{1}{m_{ox}}\psi'_{ox}(x)|_{x=0}= \frac{1}{m_{Si}}\psi'_{Si}(x)|_{x=0}\,, \tag{5.3}$$

where ψ_{ox} and ψ_{Si} are the WFs to the left and right of the nominal interface (at $x = 0$), m_{ox} is the effective mass to the left of the interface, and $m_{Si} = 0.92m_0$, for $\psi_{Si} \equiv \psi_{\Delta 2}$, or $m_{Si} = 0.19m_0$, for $\psi_{Si} \equiv \psi_{\Delta 4}$. Since in the current problem the wave functions are solutions of the one-dimensional Schrödinger equation for real energies, they could be real too (126), in which case the slope of the WF at each side of the barrier is given by the corresponding WF derivative. [1] It is convenient to recast Eq. 5.3 into

$$\psi'_{Si}(x)|_{x=0}= \frac{m_{Si}}{m_{ox}}\psi'_{ox}(x)|_{x=0} \tag{5.4}$$

We observed that ψ'_{ox} is relatively insensitive to m_{ox}, compared to the ratio m_{Si}/m_{ox}, which means that an increase (decrease) of m_{ox} translates in a decrease (increase) of ψ'_{Si} – the slope of the wave-function to the right of the interface. More formally, a real solution for ψ_{ox} could be constructed by a linear combination of a growing and a decaying exponents, and considering only the growing exponent (a limiting case for an infinitely thick barrier for negative x), we let

$$\psi_{ox}(x) = Ce^{\eta x}\,, \quad \eta = \sqrt{\frac{2m_{ox}}{\hbar^2}(E_{c,ox} - E_{sb})}\,, \tag{5.5}$$

where $C = \psi_{ox}(x = 0)$ is a constant, $E_{c,ox}$ is the oxide conduction band edge, [2] and E_{sb} is the subband under consideration. Differentiating ψ_{ox} with respect to x, and with the help of

[1] If complex exponents are used for the WFs, Eq. 5.3 could be formulated in terms of $d|\psi_{ox}|/dx$ and $d|\psi_{Si}|/dx$, i.e. the wave function modulus, rather than the WFs themselves, due to the second boundary condition, $\psi_{ox}(x = 0) = \psi_{Si}(x = 0)$.

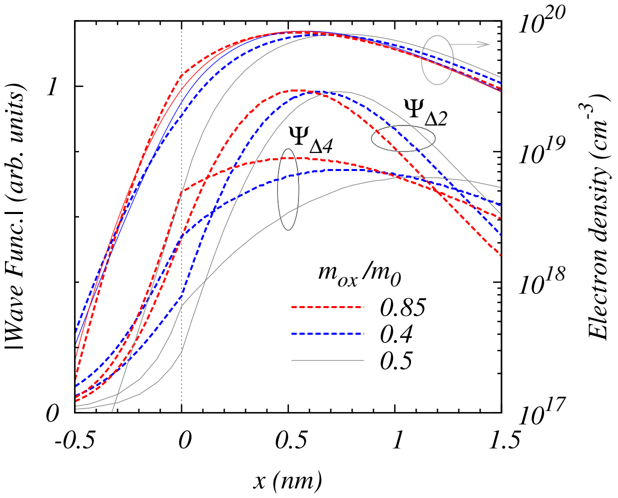
[2] $E_{c,ox}$ is assumed invariable with x, to simplify the argument.

Eq. 5.4, we obtain

$$\psi'_{Si}(x)|_{x=0} = \frac{m_{Si}}{\sqrt{m_{ox}}}\psi_{ox}(x=0)\sqrt{\frac{2}{\hbar^2}(E_{c,ox}-E_{sb})}\,, \tag{5.6}$$

which leads to the conclusion stated above (an increase of m_{ox} decreases $\psi'_{Si}(x)|_{x=0}$, and vice versa), since m_{ox} appears in the denominator, and physically, an increase (decrease) of m_{ox} corresponds to a decrease (increase) of the wave density in the oxide, hence of $\psi_{ox}(x=0)$.

The above argument is supported by Fig. 5.17, showing the modulus of the lowest subband WFs for two values of m_{ox} – $0.4m_0$ (blue dashed line) and $0.85m_0$ (red dashed line). The simulated devices are otherwise identical, with 1.2 nm EOT, 0.5 nm suboxide over which the permittivity varies linearly, and a linear band-gap transition over 0.6 nm. Noticeably, the increase of m_{ox} nearly doubles the effect of the band-gap transition, relative to a device with an abrupt band-gap at the interface and physical oxide thickness t_{ox} = 1.2 nm (shown with grey solid line). There is a shift in the peak of both WFs towards the interface, increasing the



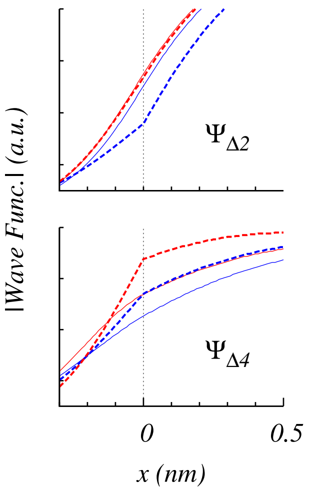


Figure 5.17: Wave function modulus (left axis) and electron density (right axis) profiles for three different values of m_{ox}. Solid grey lines – a device with $EOT = t_{ox}$ = 1.2 nm and abrupt band-gap change at the nominal interface ($x = 0$). Red and blue dashed lines – a device with the same EOT, gradual band-gap transition over 0.6 nm, 0.5 nm suboxide (t_{ox} = 1.43 nm) and effective mass changes abruptly. Red and blue thin, solid lines (electron density only) – the same device but effective mass transitions gradually over the suboxide.

Figure 5.18: Wave function modulus as per Fig. 5.17 (the same line coding). Abrupt mass change (dashed lines), and linear effective mass profile (solid lines), for m_{ox}/m_0 = 0.4(0.85) in blue(red) colour. Upper (lower) graph is for the lowest lying Δ2 (Δ4) subband.

electron density too, as shown on the same graph (right axis). This is explained as follows. A decrease in the WF slope to the right of the barrier, due to the increase of m_{ox}, means an advancement in the phase of the WF at the interface, since the WF in this region is an oscillatory solution for a quasi-bound state (QBS), and is growing in magnitude. The advancement of the phase itself means higher WF density at the interface and entails more electrons populating the corresponding QBS (hence associated with lowering of the QBS level).

It is clear from Fig. 5.17 that the simulated consequences of the non-abrupt band gap transition are markedly dependent on the value of the oxide effective mass due to its discontinuity at the interface. Is this a realistic physical picture? A gradual transition of the effective mass at the interface would bring the ratio m_{Si}/m_{ox} closer to 1, making WF derivative nearly continuous over the interface. Simulation results for such case (assuming effective mass changes over the thickness of the suboxide, 0.5 nm here), are shown in Fig. 5.18 – thin, solid lines (thicker, dashed lines) correspond to gradual (abrupt) effective mass transition profiles. What is interesting to note is that the effect of the gradual effective mass profile on $\psi_{\Delta 2}$ is opposite to the corresponding effect on $\psi_{\Delta 4}$ – compare the blue lines (for $m_{ox} = 0.4m_0$) on the upper and lower graphs. This is expected however, since $m_{Si\Delta 2}$ is higher than m_{ox}, while $m_{Si\Delta 4}$ is lower than m_{ox} for both values of m_{ox}; the gradual transition of the mass profile brings about an average increase and decrease of the m_{ox} associated with $\psi_{\Delta 2}$ and $\psi_{\Delta 4}$, respectively. The results reported in Fig. 5.18 are therefore consistent with the argument in the previous paragraph, and show that gradual mass transition leads to enhancement of $\psi_{\Delta 2}$ penetration, and diminution of $\psi_{\Delta 4}$ penetration. These competing trends between the two WFs are influenced (in magnitude) by the value of m_{ox}, as can be seen from the electron density distribution plotted with thin, solid lines in Fig. 5.17 for the gradual mass profile – both lines fall between the corresponding electron density distributions for discontinuous effective mass transition. This suggests that there is a value of m_{ox}, for which the electron density distribution (hence the impact of band-gap transition on electrostatics and capacitance) is the same, regardless of the type of effective mass transition.

It is interesting to see how the value and the interface transition profile of the effective mass affect the leakage current in the presence of a gradual band-gap change. This is shown in Fig. 5.19 for three different values of m_{ox}/m_0 – 0.4, 0.5, and 0.85. The difference between the two extreme values of m_{ox} translates into a two orders of magnitude difference between the corresponding leakage currents. The leakage for a reference device with traditional, abrupt interface and physical oxide thickness of 1.2 nm is shown as a black, solid line. The rest of

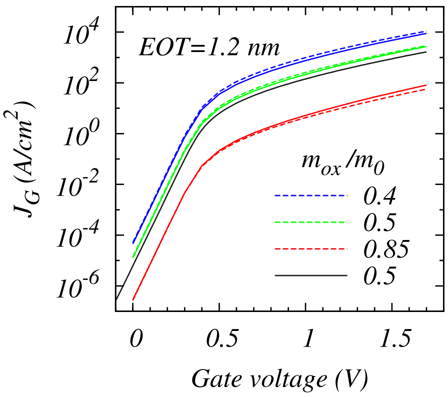


Figure 5.19: Direct tunnelling leakage characteristics with oxide effective mass as a parameter. Solid black line is for the reference device with abrupt interface (band-gap, permittivity, effective mass). The other curves are for a device with 1.2 nm EOT, 0.6(0.5) nm linear band gap(permittivity) transition profile and abrupt (dashed) or linear (thin, solid) effective mass transition profile.

the simulated devices have the same EOT with larger physical thickness of 1.43 nm, due to a linear permittivity transition over the suboxide (0.5 nm), and a linear band-gap transition (over 0.6 nm) at the interface. Despite the difference in the barrier model, the reference and the non-abrupt device with identical m_{ox} have very similar leakage magnitude. Consequently, using m_{ox} as a phenomenological constant in the usual way, one could fit the non-abrupt interface barrier model to experimental leakage data. Note further, that the influence of the gradual effective mass profile at the interface is rather small (compare the thin solid lines with the dashed lines in the same colour), consistent with the small variations in electron distribution for the two cases. The tunnelling barrier is unchanged.

Last on the subject of effective mass value and transition, we look at the subband levels and occupancy, reported in Fig. 5.20, resulting from the simulation of the devices described in the previous paragraph. Consistent with the discussion regarding Fig. 5.17 and Fig. 5.18, the bigger the m_{ox}, the more exaggerated the effect of a gradual band-gap transition, in case of abruptly changing m_{ox} (dashed lines). In the case of a continuous effective mass transition profile (thin, solid lines) and $m_{ox} = 0.4m_0$ (blue), the enhanced penetration of $\psi_{\Delta 2}$ reduces the corresponding subband level much more strongly than that of $\psi_{\Delta 4}$, so that the split between the two subbands actually increases, and no carrier redistribution between the two- and four-fold degenerate subbands is observed. For $m_{ox} = 0.85m_0$, the effect on $\psi_{\Delta 4}$ is sufficiently strong to reduce the Δ-valley splitting and shift about 10% of the carriers from the two-fold to the four-fold degenerate subbands. Thus, the impact of a non-abrupt band-gap transition on inversion layer quantisation levels and subband occupancy is significantly reduced if the effective mass also transitions gradually, rather than being discontinuous. Unfortunately, the

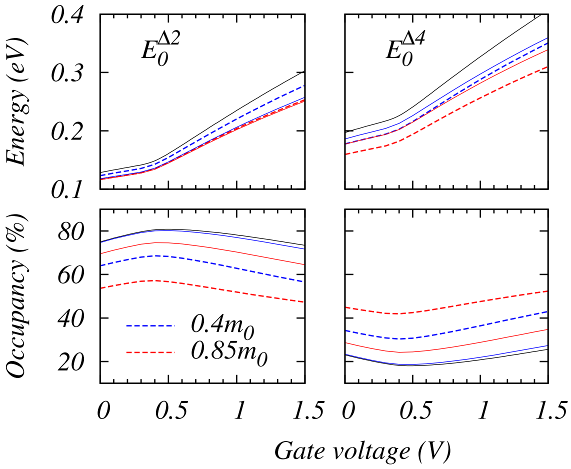


Figure 5.20: Subband levels (top) and occupancy (bottom) for a device with $t_{tr} = 0.6$ nm, $t_{so} = 0.5$ nm, and abrupt (dashed) or gradual (solid) effective mass transition. At the lower value of m_{ox}, the continuous transition nearly completely damps the effect of gradual band-gap transition (compare the black solid line – a device with abrupt interface).

value and spacial profile of the effective mass (being essentially a macroscopic parameter for bulk materials) could not be measured, nor directly calculated around the interface.

5.4.6 Beyond the linear transition approximation

This subsection discusses the effect of non-linearity in the band-gap transition profile. Density functional theory (DFT) simulations of different interface models (both crystalline and disordered structures) show a rather slow initial development of the band-gap, followed by a steeper rise (203; 205), which is in contrast to the constant gradient of the band-edges for a linear transition. A similar trend is observed in the work reported in Appendix B. The resulting conduction and valence band profiles, referred to as the *realistic* barrier model, are compared against a *linear* transition of the band-gap in Fig. 5.21; the total width of the transition is similar, of the order of 0.6 nm, but the actual profile of the band edges is quite different.

Simulation results for identical devices, aside from the two different barrier models, are reported in Fig. 5.22. The energy levels of the two lowest lying subbands (hosting almost the entire electron population in the substrate) are close to the bottom of the quantum well, so that for the corresponding wave functions in the *realistic* case, the quantum well is slightly wider, while the potential barrier is slightly thinner. This induces more inversion charge in proximity to the oxide for the *realistic* case, as can be seen in the electron density distribution, (the inset of Fig. 5.22). Therefore the impact of the non-abrupt band-gap transition is enhanced, due to the non-linear band-edge profiles, and the exact profile of the band evolution, rather than the total transition width, determines the magnitude of the observed effect.

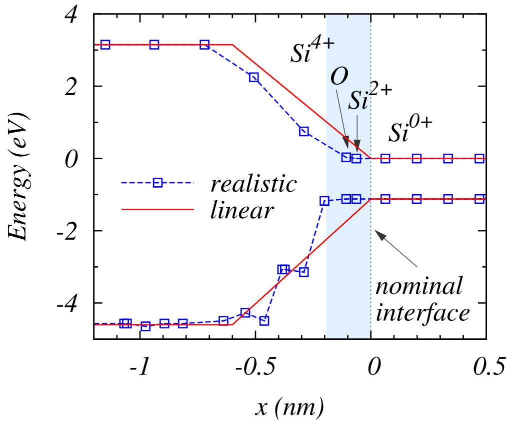


Figure 5.21: Band edge transition profiles of the *realistic* and *linear* interface models. Linear transition width is $t_{tr} = 0.6$ nm. The *realistic* profile is obtained from DFT simulations of an idealised oxide interface with α-quartz SiO_2 structure (Cf. Appendix B), resulting in the narrowest possible suboxide region (light-blue shaded).

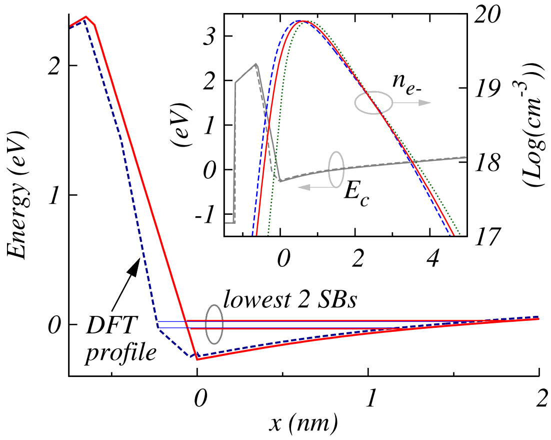


Figure 5.22: Simulated conduction band profile at $V_G = 1.2$ V, for devices with *linear* (solid) and *realistic*, DFT (dashed) interface band-gap profiles. Horizontal lines indicate the lowest two subbands from the Si $\Delta 2$ and $\Delta 4$ valleys. Inset shows the conduction band and electron density at a bigger scale. The density profile for *abrupt* interface is also shown for comparison.

The *realistic* profile is not obtained without ambiguity however. Appendix B describes the method used to obtain the *realistic* profile shown in Fig. 5.21 and discussed in relation to Fig. 5.22. But this is only one of the three slightly different profiles, resulting from the electronic structure calculation of the same atomic model of the Si/SiO_2 interface. The uncertainty in the *realistic* band-gap transition stems partly from the DFT methodology employed for the electronic structure calculation, and partly from the simple translation of the *ab initio* obtained microscopic data, to the macroscopic parameters of band-gap and band-gap offset.

To gauge the extent to which such an uncertainty in the *realistic* profile changes the impact of the transition layer on the device characteristics, we simulated devices with the three different *realistic* profiles, obtained in Appendix B. A comparison of the results is drawn in Fig. 5.23, showing the impact of the transition for different interface band-gap transition profiles, on subband quantisation and occupancy, and on gate tunnelling current. The mismatch in the *realistic* profiles results in a spread of the device characteristics. This spread is small, compared

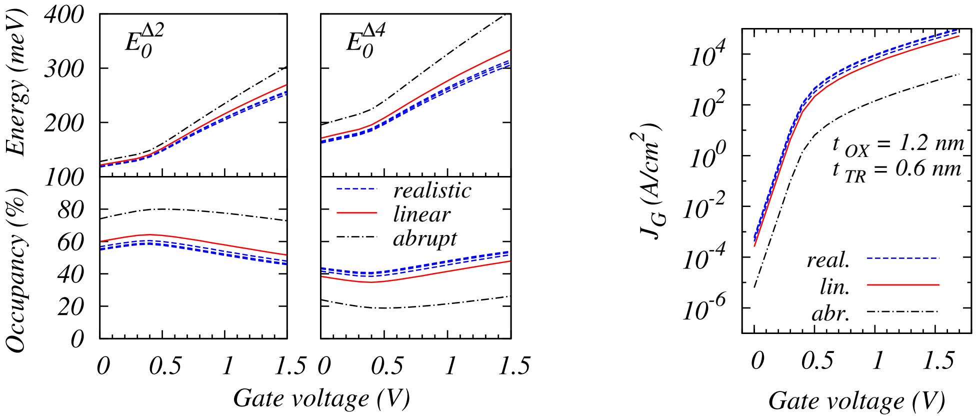


Figure 5.23: Comparison of the impact of *realistic* (blue dash) and *linear* (red line) conduction and valence band profiles on subband quantisation and occupancy (left), and gate leakage (right). Results for abrupt interface band-gap transition are also shown (black dot-dash). Although there is a mismatch between the three *realistic* profiles, their impact on the device characteristics is very similar, and stronger than that of a 0.6 nm *linear* transition.

to the magnitude of the impact that the transition itself implies – compare the set of blue lines (*realistic*), with the black line (*abrupt*), on each graph. Further, it can be noted that the simulations with a *linear* barrier model (red lines in Fig. 5.23), with 0.6 nm transition width in this case, underestimate the impact of the *realistic* transition. We note, that enhancement of inversion gate capacitance (not shown), due to the *realistic* barrier model is ~ 12 %, while the enhancement due to a *linear* barrier model with $t_{TR} = 0.5$ nm is ~ 8 % (239).

In conclusion, any of the simulated *realistic* barrier models exaggerates the magnitude of the impact that we already considered through simulations with the *linear* barrier model. The spread in the characteristics, due to the uncertainty in the *realistic* profile, may be regarded as an uncertainty in the transition width t_{TR}, of a *linear* barrier model.

5.5 Relevance to contemporary and future MOSFETs

The last section revealed the physical consequences of the Si-to-SiO$_2$ transition region on the inversion layer, based on the analysis of a metal/SiO$_2$/p-Si(001) structure. Present and future Si technology heavily relies on a number of performance boosters, e.g. channel strain and alternative gate dielectric stacks, and novel device architectures with ultra-thin Si body with low impurity concentration (259; 260). The relevance of the interface transition related effects to these cases is considered here.

5.5.1 Towards sub-nm gate oxide scaling

It is well known that advanced devices in production nowadays, with EOT in the order of 0.9 - 1.0 nm, use SiON (oxynitride) as gate dielectric to combat oxide reliability issues, at the same time reducing the gate leakage relative to pure SiO_2 of the same EOT (261; 262; 263). The presence of nitrogen at the interface is found to widen the compositional transition layer (247; 264). At the same time, there is evidence that N incorporation at the interface forms an abrupt band gap opening from that of Si, to that of the oxynitride ($\sim$ 2.2 eV or higher) (210). This means the impact of interface transition observed for pure SiO_2 is effectively suppressed for a SiON dielectric. However, to implement an efficient barrier for Boron penetration without incurring undesirable flat-band voltage shift and mobility degradation, the Nitrogen profile is tailored to maximize away from the Si/SiO_2 interface (49).

Oxynitrides are exploited to their scaling limit, and further EOT reduction to 0.5 nm, while keeping the gate leakage below 10^3 A/cm^2, imposes the deployment of high-κ dielectric (HK) materials, replacing SiO_2 and oxynitrides (10). Hafnia-based (HfO_2) dielectric stack is already introduced for the 45 nm technology (11), and together with $HfSiO_x$ (hafnia-silicate) emerges as a potential solution, even compatible with the conventional *gate-first* fabrication process (12; 13; 43; 265). However, the formation of a few atomic layers thick SiO_2 interfacial layer (IL) is unavoidable (due to the high-reactivity between O and Si), and is in fact desirable (to reduce remote scattering mechanisms due to the high-κ layer) (14; 15; 16).

5.5.1.1 Interfacial SiO_2

To realise EOT thickness below 1 nm with HfO_2 as a high-κ material, one must control the physical extent of the IL in the range of 0.7 - 0.3 nm. Justifiably, at such thickness, the IL is regarded as a suboxide (SiO_x), and both of its interfaces – with the hafnia, and with the Si substrate – are important. Due to phase separation at the HfO_2/SiO_2 interface, the compositional and structural transition at this end is effectively abrupt, if physical roughness due to hafnia protrusions into the IL is not considered (266). The local dielectric response is linearly related to the stoichiometry (the oxidation state of Hf and of Si atoms in this case), similarly to the Si/SiO_2 interface, and therefore the change in permittivity is also nearly abrupt (248). Although the transition of the electronic band-gap from that of SiO_2 to that of HfO_2 is not abrupt (267; 268), DFT modelling suggests it is happening over a very short distance, in the

Table 5.1: DIELECTRIC STACK PARAMETERS

EOT= 0.75 nm	SET	
	(a)	(b)
HK type	LLDA c-HfO$_2$	a-HfO$_2$
EOT (t_{HK}), nm	0.52(4.0)	0.4(2.05)
κ_{HK}	30	20
IL type	CCIL HfSiO$_x$	SiO$_2$
EOT (t_{IL}), nm	0.23(0.4)	0.35(0.5)
κ_{IL}	6.7	5.6

Set-(a) uses typical κ values for HfSiO$_x$ IL formed by cycle-by-cycle HfO$_2$ ALD and RTA, and crystalline c-HfO$_2$ formed through layer-by-layer deposition and anneal (13; 265). Set-(b) uses typical κ values for SiO$_2$ IL formed as a by-product during the amorphous a-HfO$_2$ layer growth (12; 43).

order of 0.2 nm (269), consistent with the non-linear dependence on Hf contents (248). Concerning the Si/SiO$_2$ interface transition, similar properties as already discussed for ordinary silicon dioxide dielectric are anticipated and simulated (267).

Taking the above considerations into account, we construct two sets of MOS devices with a high-κ gate stack of 0.75 nm EOT and different parameters of the dielectrics (permittivity and thickness), referred to as set-(a) and set-(b), as shown in Table 5.1. Common parameters for the two stacks, with their typical values for hafnia, are the band gap, $E_g = 5.5$ eV, and the conduction band offset (with respect to bulk-Si CB edge), $\Delta E_c = 1.6$ eV, since incorporation of Si in HfO$_2$ does not change significantly the electronic structure (15; 43; 270). Similarly, the HK layer effective mass m^*_{HK} is assumed to be $0.2m_0$ in all cases (43; 265; 271; 272), although values in the range $0.1m_0 - 0.4m_0$ are found in literature (273; 274; 275; 276). The corresponding parameters of the IL have the values for bulk-SiO$_2$: $E_g = 8.9$ eV, $\Delta E_c = 3.15$ eV, and $m^*_{IL} = 0.5m_0$. Single, parabolic band structure describes both the HK and the IL layer. Within each set, devices differ by the band-gap transition profile – *abrupt* and *realistic*. In the latter case, the Si/IL transition is obtained from the band-edge profile obtained *ab initio* for the Si/SiO$_2$ interface; the IL/HK band gap variation is assumed to be linear, over a distance of 0.2 nm. Permittivity changes abruptly at the physical boundaries of the IL, but permittivity enhancement is accounted for in the dielectric constant used, as quoted in Table 5.1. Substrate impurity concentration is 7.3×10^{18} cm^{-3}, in line with the ITRS projection for a bulk-MOSFET with a metal gate and high-κ dielectric stack of 0.75 nm EOT.

Figure 5.24 shows a comparison of the conduction band profile and electron density distribution, between the realistic cases of set-(a) (solid, red line), and set-(b) (dashed, green line). The abrupt case of set-(a) is also shown (solid, grey line). The curves for the realistic case of the two sets overlap, and are clearly shifted from the abrupt case (shown only for set-(a)). The impact of the Si/SiO$_2$ transition is very similar to what was earlier observed for a pure SiO$_2$ insulator, and the IL-HK transition appears to have a negligible impact. Note that electrons populating the lowest three subbands (denoted with red, horizontal lines on Fig. 5.24, and containing more than 99% of the inversion charge) experience the entire physical thickness of the gate stack. Consequently, it is the first few Å of the Si/SiO$_2$ transition that are responsible for the observed effects, just like in pure SiO$_2$ dielectric. This is expected also in consideration of the characteristic length of the electron density decay in the dielectric (established earlier, Cf. Section 5.4.1), which is in the order of 0.2 nm, i.e. half the thickness of the interfacial layer. The characteristic length increases in the HK region however, which is due to the simultaneous reduction of the conduction band offset and effective mass. Figure 5.24 confirms, that for the simulated devices, the properties of the inversion charge distribution do not depend strongly on these two parameters of the HK dielectric. Consider the simulated hypothetical dielectric stack, differing from the *realistic* case of set-(a) only in that ΔE_c, E_g, and m^*_{HK} have the same values as for SiO$_2$ – the resulting density distribution (dotted, light-blue line) is indistinguishable from the other two curves, corresponding to the *realistic* profile.

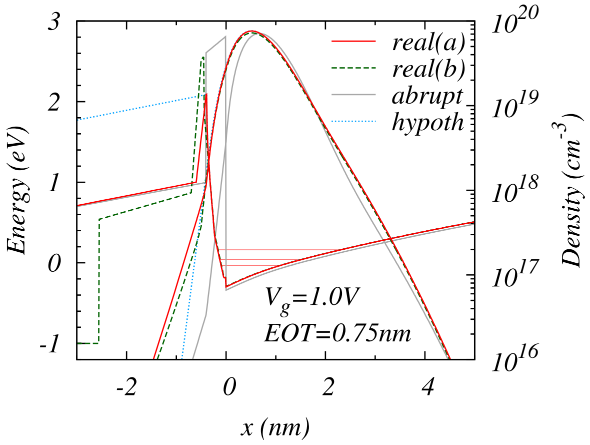


Figure 5.24: Conduction band and electron density profiles in an MOS device with high-κ gate stack, 0.75 nm EOT, simulated with different gate stack parameters, described in the text. $Real(a)$ and $Real(b)$ differ in permittivity and physical thickness of the oxides, but have identical band gap transition. For the non-abrupt cases, electron distributions in the channel overlap, and are offset from the corresponding distribution for abrupt interface.

Figure 5.25 shows the normalised $C-V$ characteristics for devices that differ only in their EOT (t_{HK} respectively; κ_{HK} and the IL parameters being the same as for set-(a) in Table 5.1). The difference in capacitance for lower EOT is more significant (nearly 20 %, for 0.5 nm EOT), and we find a negligible dependence on substrate impurity (varied in our simulations from 2 to 9×10^{18} cm^{-3}). As shown in the inset of the same figure, at strong inversion, an identical absolute capacitance is obtained for a 0.65 nm EOT device, modelled with an *abrupt* interface, and a 0.75 nm device, modelled with a *realistic* interface barrier. It is important to note that this result is relatively independent of the effective mass used in the interfacial layer, because it is due mostly to the shift of carriers closer to the nominal interface, i.e. the profile of the quantum well is of prime importance. Therefore, accounting for the non-abrupt band-gap transition is very important for the accurate characterisation and predictive modelling of ultra-thin EOT devices with HK gate stack.

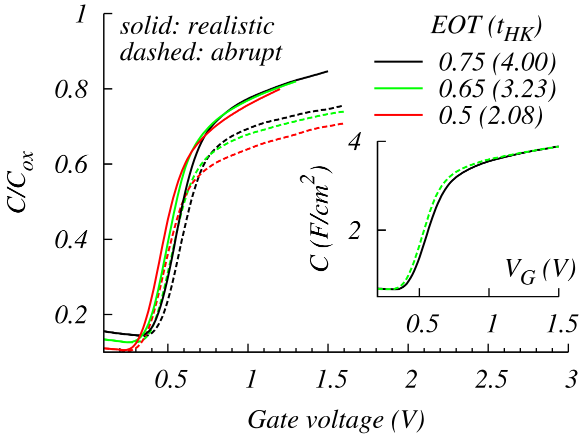


Figure 5.25: Normalised $C - V$ characteristics (EOT(t_{HK}) is a parameter; the same IL and κ_{HK} of set-(a) are used). The *realistic* band gap transition (solid) brings more charge near the interface, resulting in gate capacitance increase over the *abrupt* case (dashed). A *realistic* band profile at thicker (0.75 nm) EOT effects the same capacitance as a thinner (0.65 nm) EOT with an *abrupt* interface (inset).

Concluding this subsection, we state that the impact of progressive band-gap interface transition on subband levels and occupancy in devices with HK gate stack, is very similar to that in devices with pure SiO$_2$ gate insulator, decreasing the lowest subband splitting by more than 70 meV and populating an equal amount of electrons in the two- and four-fold degenerate Si Δ-valleys. About 5 % of the inversion charge resides in the oxide, at inversion sheet charge density of 10^{13} cm^{-2} and high impurity concentration of 7.3×10^{18} cm^{-3} (108).

5.5.1.2 Gate leakage

Here we consider the influence of the band-gap transition at interfaces, on the gate leakage – the main reason for the introduction of high-κ dielectric stacks. Figure 5.26 shows the gate leakage characteristics of the devices from set-(a) (blue lines), and set-(b) (red lines), described earlier, with 0.75 nm EOT. Another device is also simulated, with a SiO$_2$ IL of 0.5 nm, and a 2.26 nm thick HK layer of permittivity $16\epsilon_0$ (typical for hafnia silicates (43)), resulting in a 0.9 nm EOT of the gate stack. The effect of a non-abrupt band-gap transition at the interface is to increase the direct tunnelling current density by over a factor of 10, similarly to the case of a single layer SiO$_2$ insulator. Due to this increase, the gate current for the simulated devices with a thicker IL of 0.5 nm exceeds the ITRS gate leakage density limit at 1.0 V. Note however, that the effective mass of the IL is not calibrated against experimental data – this is beyond the scope of the current work, but is essential for the projections of gate leakage in high-κ gate stacks according to the ITRS.

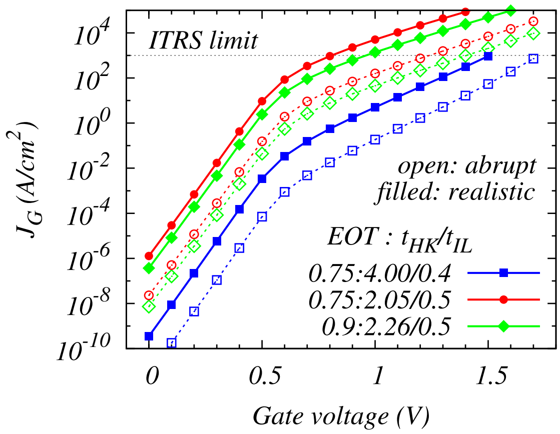


Figure 5.26: Direct tunnelling gate leakage in three devices, simulated with *abrupt* (dashed) and *realistic* (solid) band-gap transition. The progressive band-gap transition raises the leakage current over 10 times, as for pure SiO_2 dielectric. The doubling of the HK layer thickness lowers gate tunnelling by over 3 orders of magnitude. The transparency of the IL is reflected in the steeper curves for $t_{IL} = 0.4$ nm.

Comparing the curves for 0.75 nm EOT and different stack composition, we observe a difference in the slope at high gate bias – the steeper curves correspond to the gate stack with the thicker HK layer. We verified that this is the trend for t_{HK} in the range of 2 – 5 nm, for gate stacks with the same IL of 0.4 nm. It is indicative of the transparency of so thin an interfacial layer, effectively putting the inversion charge in contact with the HK layer that has a lower conduction band offset, and hence, lower tunnelling barrier. As a result, the suppression of direct tunnelling at high gate voltage is much weaker (about a 100 times leakage reduction for two times increase in t_{HK}, at $V_g = 1.5$ V) than it is for lower gate bias (over a 1000 times leakage reduction for the same increase in t_{HK}, at $V_g = 0.5V$).

Figure 5.27 shows the dependence of the gate leakage on inversion sheet charge density, for different levels of impurity concentration. The gate leakage increases by a factor of 100, with the increase of doping concentration from 0.5×10^{18} cm^{-3} to 7.3×10^{18} cm^{-3}. This is consistent with the stronger band-bending needed to induce the same amount of inversion charge at a higher level of depletion charge, hence raising the subbands in the inversion layer closer to the top of the tunnelling barrier. It is also shown that the impact of the non-abrupt band-gap transition is of the same magnitude, regardless of the impurity concentration, for the relevant range of inversion charge sheet density.

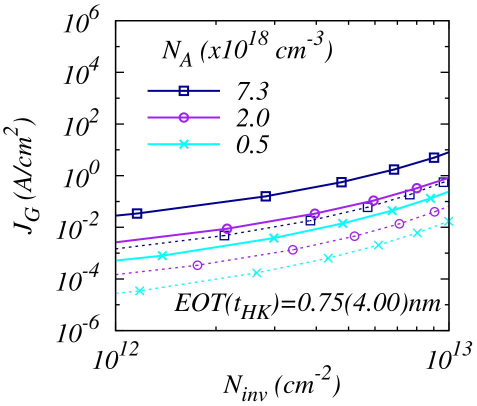


Figure 5.27: Gate leakage dependence on inversion sheet-charge, N_i, with impurity concentration, N_A as a parameter. N_A values are dictated by the ITRS for bulk-MOSFET scaling (10). The increase of impurity concentration increases leakage current too. The progressive band gap transition has the same impact, regardless of N_A in the given range of N_i.

5.5.2 Channel strain

Of interest to a p-Si(001) substrate is tensile strain, which improves drive current in n-channel MOSFETs (277; 278). We model the effect of strain by accounting for the splitting of the conduction band minima that lowers the conduction band edge of the 2-fold degenerate Si Δ-valley with respect to the band edge of the 4-fold degenerate valley (which experiences only a minor shift from its ordinary position in relaxed Si), and by an adjustment of the corresponding valley effective masses (279; 280). The valley splitting enlarges the band offset at the interface for carriers in the Δ2 valley and also increases the difference between the lowest two subbands in the quantised inversion layer (281). Tentative simulations with a 0.5 nm interface band-gap transition width result in a subband splitting of about 300 mV for the 2 % strained-Si, against about 100 mV for the relaxed substrate. In effect, less than 1 % of the inversion carriers populate the Δ4 valley in the case of strain, and the electron distribution is characterised by the density profile of $\psi_0^{\Delta 2}$. This WF is less affected by the non-abrupt interface band-gap change, due to the increased potential barrier in the case of strain. Moreover, the band-gap transition width is reduced with the application of tensile strain (238). Therefore, tensile strain reduces the impact of the interface transition on the inversion layer characteristics.

5.5.3 Ultra-thin body and low substrate doping

Low-doped, ultra-thin body devices are promoted for their superior electrostatic integrity, and less pronounced short-channel effects (260). We previously showed that the magnitude of the impact from Si/SiO$_2$ interface transition depends on the depletion charge sheet density, which in low-doped devices is smaller. However, the ultra-thin body introduces additional confinement of carriers near the interface, so that wave functions in the channel have similar properties to the case of confinement due to very high impurity concentration (258; 282). Therefore the relevance of the Si/SiO$_2$ interface is the same as in a bulk device. We simulated bulk devices with low doping concentration of 2×10^{16} cm^{-3} – such MOS structures are often used in mobility and leakage characterisation experiments. The impact of the interface transition in this case is of similar magnitude to the one observed in simulations of highly doped devices (283). This is due to the following fact. The low confinement from the ionised impurities significantly reduces the subbands splitting in the inversion layer – for a 1.1 nm EOT MOS device, all of the first four subbands contain appreciable amount of carriers. The upper bands, seeing wider opening of the well due to the non-abrupt band-gap transition, appreciably modulate the electron distribution and hence capacitance and leakage characteristics. Consequently, consideration of the non-abrupt transition of interface electronic and dielectric properties remains important for novel device architectures too.

5.6 Summary

The change of atomic structure at the oxide interface implicates the existence of a transition region, in which the chemical composition and structural arrangement are different from the corresponding ones in either bulk Si or SiO$_2$. Detailed experimental and *ab initio* theoretical investigations reveal that this transition region imparts a gradual change in the electronic and dielectric properties over a distance of 0.2 - 0.6 nm away from the top-most atomic plane of the Si substrate. The extent of this transition is comparable to the gate insulator thickness in modern MOSFETs and will have an appreciable impact on the electrical properties of the inversion layer. Our literature review showed however, a lack of comprehensive understanding of this impact, and the existence of contradictory opinions regarding the qualitative effects.

We developed a self-consistent 1D Poisson-Schrödingersolver that allows the simulation of MOS inversion layer characteristics with the account of non-abrupt band gap, dielectric constant, and effective mass transition at the oxide interface.

Our simulations of devices with oxides thinner than 3.5 nm, and a linear interface band-gap variation over a distance of up to 0.6 nm, show that compared to an abrupt interface, the progressive band-gap transition

- incurs a relatively small change in the electrostatic potential and the electron sheet density, but shifts the inversion charge centroid significantly closer to the interface – reducing the average inversion layer thickness by 8 - 10 % for the widest simulated transition

- increases the fraction of electrons in the oxide barrier, by a factor of 10 (up to 5 % of the carriers reside in the oxide, at an inversion sheet density of 1×10^{13} cm^{-2}), at high impurity concentration

- increases the inversion gate capacitance by nearly 10 %, consistent with the reduction of the inversion layer thickness, with even greater impact at lower gate bias, and correspondingly lowers the threshold voltage of the structure by 20 - 25 mV

- lowers the subband levels and subband splitting in the inversion quantum well by 50 meV, leading to a very significant change in the occupancy of the Si two- and four-fold degenerate Δ-valleys – nearly 20 % more carriers accumulate in the $\Delta4$ subbands, at the expense of the depleted $\Delta2$ subbands

- increases the gate leakage more than 10 times (at a fixed oxide effective mass, as in the case of an abrupt interface), consistent with the effective thinning of the potential barrier due to the progressive band-gap change, and the increase of electron density in the proximity of the barrier

We established a link between first-principles simulations of the interface, and device simulations, by incorporating conduction and valence band-edge profiles obtained from *ab initio* (density functional theory) to calculations of the electronic structure, to investigate the effects of non-linear variation of the interface band-gap. The magnitude of the impact depends mostly on the particularities of the band-edge evolution across the interface. The more realistic, DFT band-gap profile enhances the observed effects to the magnitude typical of a 0.6 - 0.7 nm wide linear transition.

For the first time we presented a detailed analysis of the influence of the oxide effective mass and its non-abrupt transition. Increment of this mass from the typical value of $0.5m_0$ enhances the impact of non-abrupt band-gap transition, and vice versa, with the simulated direct tunnelling current being an exception (i.e. reduced). This is a direct consequence of the boundary conditions imposed on the electron envelope wave functions, to guarantee particle flux conservation, and on the fact that carriers in the $\Delta4$-valley have a smaller quantisation mass and are affected more strongly by the wider band-gap transition and by a larger effective mass discontinuity. The simulated impact of progressive band-gap transition is greatly suppressed however, when the effective mass is linearly varied within the transition layer.

We find that permittivity transition is directly reflected in a reduction of the equivalent oxide thickness, and could be reliably accounted for in device simulations by adopting a dual layer model of the oxide. In such a model, the transition region must be assigned a higher dielectric constant, of ~ 7, while the rest of the oxide must be attributed a bulk-SiO_2 permittivity.

Our results not only agree with the previous analysis done by Stern (208), Yang *et al.*(210), and Watanabe *et al.*(131) (which bare good reference to experimental data), but also provide a much more comprehensive understanding of the physical consequences of the non-abrupt interface transition.

The results obtained for SiO_2 dielectric have immediate relevance to contemporary and future devices with high-κ dielectric stacks, because of the presence of an interfacial SiO_2 at the interface with the Si substrate. The Si/SiO_2 interface in such devices has a dominant role, and the impact of its non-abrupt transition on the inversion layer is of the same, or greater magnitude, compared to the case with a pure SiO_2 gate insulator.

Future research on the subject should focus on establishing the correct picture for the effective mass transition at the interface. A possible approach is through self-consistent fitting of simulations results to experimentally obtained gate leakage and capacitance data, where the oxide is well characterised from independent (possibly spectroscopic) experiments determining the bulk-like band gap and physical thickness. An alternative could lie in using first principles simulations. The importance of such work could not be overestimated, since high-κ gate stack characterisation currently relies on interface models that require *ad hoc* adjustments of the parameters associated with the interfacial layer, ignoring effects of its non-abrupt transition.

Chapter 6

Conclusions

Two goals that advance the modelling and understanding of gate leakage in nano-scaled CMOS transistors were defined at the onset of this research. First, to establish a 3D simulation framework for the study of gate leakage variability in order to study the factors that affect gate leakage variability, and quantify it, for a sub-30 nm gate length, n-channel bulk-MOSFET. Second, to investigate the impact of a gradual transition of the electronic properties at the Si/SiO$_2$ interface, on the characteristics of a MOS inversion layer. Below is a summary of the principal finding of the work, followed by a discussion, regarding future directions for research.

6.1 Summary of results and implications

6.1.1 Gate leakage modelling: The pragmatic approach

Gate leakage in contemporary CMOS transistors is dominated by direct tunnelling through the oxide, and is more severe in n-channel devices, where it is due to electrons tunnelling from quasi-bound states of an inverted or accumulated surface layer. Direct tunnelling gate current depends mostly on the interface-normal component of the electric field, even in scaled MOSFETs, which justifies the wide use of 1D direct tunnelling models for the estimation of gate leakage. Throughout our survey of tunnelling models, we selected one numerical, quantum-mechanical model, based on the approach in Ref. (126), and one analytical, based on an improved WKB expression for the tunnelling probability (1). The former is more accurate, and convenient to implement in the 1D Poisson-Schrödinger solver, used for the study of the Si/SiO$_2$ transition layer, but unsuitable for incorporation in a 3D device simulator. The latter is very efficient, but sufficiently accurate, and easy to incorporate in any device simulator, and we choose it for the study of gate leakage variability. We incorporated this model in a 1D Poisson-density-gradient solver, and improved the way tunnelling charge is evaluated, by assuming the self-consistently obtained, electron sheet charge. Hence we demonstrated a very good agreement with a range of experimental data, using a single value of the oxide effective mass ($m_{ox} = 0.67m_0$), for tunnelling from either accumulation, or inversion. The 1D simulator served as a prototype of the 3D device simulator.

6.1.2 Gate leakage variability: Not a myth!

The study of MOSFET gate leakage variability requires a physically sound 3D device modelling technique and simulations of large ensembles of *macroscopically* identical, but *microscopically* different transistors. The three most established approaches applicable to the simulation of advanced CMOS transistors are drift-diffusion (DD), often complemented by the density-gradient (DG) quantum corrections, Monte Carlo (MC), and non-equilibrium Green's functions (NEGF). Currently, the NEGF and MC simulations are used mainly for gaining insight in the quantum transport phenomena and their impact on variability, but are computationally too costly for the simulation of large statistical ensembles. MC is suited for non-equilibrium transport phenomena but is also computationally expensive. Only the DD/DG method have the required efficiency, and accurately model device electrostatics with the account of quantum confinement effects in terms of device electrostatics, due to its proper accounting of quantum confinement effects.

By choosing the DD/DG framework, we benefit from the already established and proven capabilities of the *Glasgow 3D Atomistic* simulator for studying device variability. The simulator models the most essential sources of intrinsic parameter fluctuations, including random dopant fluctuations (RDF), and microscopic oxide thickness variation (OTV), which are essential for the present study. The advancement of the simulator to account for gate leakage is based on the inclusion, at a post-processing stage, of the 1D, analytical direct tunnelling model, as stated in the preceding sub-section. Presently, the essential quantities of the model are obtained directly from the electrostatic potential and electron distribution calculated by the 3D DG simulator.

We first simulate a 25 nm square gate MOSFET with a uniform, continuous doping profile throughout, flat interfaces, and 1 nm SiO_2 dielectric, and analyse the gate and drain voltage dependence of the gate leakage, for a uniform, continuously doped transistor. Our results qualitatively agree with previously reported experimental and modelling results (30; 175; 192; 193; 195) and show that – *i)* the gate current is maximum at high gate voltage, and low drain voltage, corresponding to the On-state of the n-channel transistor in a CMOS inverter circuit, and is due to electrons tunnelling from the substrate (including the overlapped source/drain extension and channel areas) to the gate; and *ii)* at low gate voltage and high drain voltage (the Off-state of the n-channel transistor), the magnitude of the gate current exceeds that of the drain sub-threshold current, and is due to electrons tunnelling from the gate, to the drain extension.

We further study three different ensembles of 230, 25 nm gate length, uniform MOSFETs, macroscopically identical to the, but microscopically different. Devices differ from set to set in terms of random discrete dopants distribution (RDF), atomic scale interface roughness and corresponding oxide thickness variation (OTV), or the combination of both. The effects of RDF and OTV summarised below are relatively independent

- *OTV increases the mean gate leakage 5 times*, over the gate current of the nominal (uniform) device, which is due to the exponential sensitivity of the direct tunnelling to the oxide thickness.

- At $V_G = V_{DD}$ and $V_D \sim 0$, *OTV induces appreciable variability with a standard deviation of a few % of the mean*, while the RDF adds insignificantly to the spread. This is because the large electron concentration in the substrate at this bias screens the potential of the ionized impurities, while surface roughness affects both oxide field and carrier concentration near the interface. *This variability is expected to increase if oxide roughness correlation length becomes comparable to the gate dimensions.*

- At $V_G \sim 0$ and $V_D = V_{DD}$, *both RDF and OTV contribute to a large spread in the gate current*, which for the chosen geometry is nearly *two orders of magnitude.* The contribution from RDF is stronger since large fluctuations in the local tunnelling density are implicated by the exposed impurities in the depleted substrate. Due to the narrow region of the drain-extension overlap, determining the tunnelling magnitude, self-averaging is not effective.

It is evident that gate leakage variability is a very important issue, particularly at high drain, and low gate bias, where gate tunnelling is the major leakage component in the transistor, and variability is very large. Both RDF- and OTV-induced variability will increase with the reduction of gate length and the further scaling of the oxide thickness. Since gate leakage in the Off-state of the transistor exceeds the sub-threshold current from the source to drain, accounting for it and its variability is required, for the accurate estimation of leakage power in ultra-scaled devices.

6.1.3 Si/SiO$_2$ interface transition: Can it be ignored?

The change of atomic structure at the oxide interface implies the existence of a transition region, in which the chemical composition and structural arrangement are different from the corresponding ones in bulk Si or SiO$_2$. Detailed experimental and *ab initio* theoretical investigations reveal that this transition region imparts a gradual change in the electronic and dielectric properties over a distance of 0.2 - 0.6 nm away from the top-most atomic plane of the Si substrate. The extent of this transition is comparable to the gate insulator thickness in modern MOSFETs, or to the thickness of the interfacial layer in high-κ gate stacks, and has an appreciable impact on the electrical properties of the inversion layer.

We developed a self-consistent 1D Poisson-Schrödinger solver that allows the simulation of MOS inversion layer characteristics taking into account the non-abrupt band gap, dielectric constant, and effective mass transition at the oxide interface.

Our simulations of devices with oxides thinner than 3.5 nm, and a *linear* interface band-gap variation over a distance of up to 0.6 nm, show that compared to an abrupt interface, the progressive band-gap transition

- reduces the average inversion layer thickness by 8 - 10 %;
- increases the fraction of electrons in the oxide barrier, by a factor of 10 (up to 5 % of the carriers reside in the oxide, at an inversion sheet density of $1{\times}10^{13}$ cm^{-2});
- increases the inversion gate capacitance by nearly 10 %, and correspondingly lowers the threshold voltage of the structure by 20 - 25 mV;
- lowers the subband levels and subband splitting in the inversion quantum well by 50 meV, leading to a very significant change in the subband occupancy – 20 % more carriers accumulate in the $\Delta 4$ subbands, at the expense of the depleted $\Delta 2$ subbands;
- increases the gate leakage more than 10 times, consistent with the effective thinning of the potential barrier due to the progressive band-gap change.

Further, we established a link between first-principles simulations of the interface, and device simulations, by incorporating conduction and valence band-edge profiles obtained from *ab* *initio* (density functional theory) calculations of the electronic structure, in our simulations. Hence we investigate the effects of a *realistic* variation of the interface band-gap and show that the magnitude of the impact in this case is enhanced, relative to a *linear* interface band-gap transition of similar width.

While the impact of the Si/SiO$_2$ transition appears so large, it is worth asking – is it real? In this regard, we are the first to present a detailed analysis of the influence of the oxide effective mass and its non-abrupt transition. The increase of this mass from the commonly used value of $0.5m_0$, enhances the impact of non-abrupt band-gap transition on the confinement effects, and reduces its impact on the direct tunnelling current. The reduction of the oxide tunnelling mass below $0.5m_0$ has the opposite effect. This is a direct consequence of the boundary conditions imposed on the electron envelope wave functions, to guarantee particle flux conservation. Moreover, the simulated impact of progressive band-gap transition is greatly suppressed, if a gradual transition of the oxide effective mass is adopted.

We find that permittivity transition is directly reflected in a reduction of the equivalent oxide thickness, and could be reliably accounted for in device simulations by adopting a dual layer model of the oxide. In such a model, the transition region must be assigned a higher dielectric constant, of ~ 7, while the rest of the oxide must be attributed a bulk-SiO$_2$ permittivity.

At this stage it is worth asking - can we ignore the Si/SiO$_2$ interface transition, and absorb its impact in the phenomenological parameters (oxide thickness, oxide effective mass, conduction band discontinuity), describing the oxide potential barrier. As far as gate leakage and capacitance characterisation is concerned, this is the actual state of affairs, where the interface is assumed to be abrupt, and oxide thickness and tunnelling mass are fitted to experiments. The answer may not be so simple for high-κ dielectric stacks, however. In addition, it is not possible to account for the impact on quantisation, which is relevant to the transport properties of carriers in the inversion layer.

6.2 Future work

At the time of writing, variability and leakage power have become the most important limitations for the continuation of device scaling, having far reaching implications on all aspects of semiconductor technology and design. At the same time, the projected life of bulk MOSFET device architecture for digital logic application is extended to the year 2015. In order to understand, control and tolerate leakage power variability, it is imperative to accurately evaluate its distribution and standard deviation in scaled devices, with a realistic doping profile, e.g. obtained from process simulations, and with high-κ gate dielectric stacks. While the simulation framework developed in this work, for studying gate leakage variability, is readily applicable to the simulation of realistic bulk-MOSFETs, the following advancements are necessary – *i)* accounting for the gate tunnelling current in the self-consistent calculation of the drain current, so that gate leakage, sub-threshold leakage, and possibly, band-to-band junction to body tunnelling, could be evaluated; *ii)* extending the gate tunnelling model to apply for stacked dielectric layers and metal gate.

The deployment of high-κ (e.g. hafnia based) dielectric stacks entails additional complexity, associated with the lack of well defined phenomenological parameters, relevant to tunnelling, e.g. effective mass, and conduction/valence band offsets from silicon. Hafnia based gate stacks also exhibit acute sensitivity to processing conditions, which affect their morphology and permittivity. The presence of the sub-oxide interfacial layer makes their characterisation even more challenging and typically relies on *ad hoc* adjustments of the parameters associated with the interfacial layer. In this respect, consideration of the Si/SiO$_2$ interface transition may provide room for improvement – if we better understand the interfacial layer, we will be able to infer on the parameters of the high-κ dielectric layer with greater certainty. However, future research on the subject should focus on establishing the correct picture for the effective mass transition at the interface. A possible approach is through self-consistent fitting of simulations results to experimentally obtained gate leakage and capacitance data. An alternative could lie in using first principles simulations.

Appendix A

Self-consistent Poisson-Schrödinger solver with QM tunnelling

A.1 Mathematical Formulation

Here we describe an implementation for the numerical, self-consistent solution of the one dimensional (1D) Poisson (PE) and Schrödinger (SE) equations, using a modified version of Schred 2.0 solver (253), which realises the self-consistent field calculation scheme described in Ref. (249). The modifications are introduced to allow

- solution of the PE with spatially varying permittivity,
- solution of the SE with spatially varying effective mass
- direct tunnelling gate current computation
- external definition of the spacial dependence of material parameters - band gap, $E_G(x)$, dielectric constant, $\kappa(x)$, and effective mass, $m^*(x)$, with x being the direction normal to the Si/SiO$_2$ interface.

The PE in this case is:

$$\frac{d}{dx}\left(\kappa(x)\frac{d}{dx}\varphi(x)\right) = -\frac{1}{\varepsilon_0}\rho(x)\,, \tag{A.1}$$

and its solution delivers the electrostatic potential, relative to the equilibrium Fermi level of the substrate. Equation A.1 is subject to Dirichlet boundary conditions with the potential values at the two ends of the device determined by the applied gate voltage and the work-function difference between the substrate and the gate. The charge density distribution, $\rho(x)$, is derived from the envelope wave-function in the effective mass approximation, as in the original reference (253), except that the Schrödinger equation for electrons is (due to the variable effective mass, (103)):

$$-\frac{\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m^*(x)}\frac{d}{dx}\psi_{\nu,\iota}(x)\right)+U(x)\psi_{\nu,\iota}(x)=E_{\nu,\iota}\psi_{\nu,\iota}(x)\,. \tag{A.2}$$

The indexes ν and ι correspond to a given valley and sub-band of the 6-ellipsoidal structure of the Si conduction band (233), $E_{\nu,\iota}$ is the corresponding subband energy, and $\psi_{\nu,\iota}(x)$ is the wave-function solution for the given potential energy profile, $U(x)$. The potential energy is linked to the electrostatic potential $\varphi(x)$ through

$$U(x)=-q\varphi(x)+E_C^{bulk}+\Delta E_C(x)\,, \tag{A.3}$$

where E_C^{bulk} is the conduction band offset from the equilibrium Fermi lever of the semiconductor, $\Delta E_C(x)$ is the position dependent variation of the conduction band (due to material variation) with respect to E_C^{bulk}.

The SE (A.2) is solved for each subband energy, $E_{\nu,\iota}$, which is in fact a quasi-bound state (QBS), due to the existing weak coupling of the inversion layer to the gate through the ultra-thin dielectric. The solution is carried out within the transfer-matrix (TM) formalism similarly to (104). The potential profile $U(x)$ is approximated by N connected intervals of constant potential energy, $U(x)=U_n$, for n from 0 to $N-1$, and $x_n \leq x < x_{n+1}$. For a given sub-band energy, $E_{\nu,\iota}$, the wave-function solution of the Schrödinger equation (A.2) in element n is:

$$\psi_n(x)=A_n e^{ik_n x}+B_n e^{-ik_n x}\,,\quad k_n=\sqrt{\frac{2m_n^*}{\hbar^2}(E_{\nu,\iota}-U_n)}\,. \tag{A.4}$$

Parabolic $E(k)$ dispersion is assumed throughout. The amplitudes A_n and B_n are determined through the additional boundary conditions imposed by the continuity of the wave-function and the conservation of probability (density) current at each interval boundary (103):

$$\psi_{n-1}(x_n)=\psi_n(x_n)\,,\quad \frac{1}{m_{n-1}}\frac{d}{dx}\psi_{n-1}(x_n)=\frac{1}{m_n}\frac{d}{dx}\psi_n(x_n)\,. \tag{A.5}$$

From equations (A.4) and (A.5) the coefficients of the TM for the barrier between the element $n-1$ and n are derived:

$$\mathbf{T}_n=\frac{1}{2}\begin{pmatrix}(1+S_n)e^{-i(k_n-k_{n-1})\,x_n} & (1-S_n)e^{-i(k_n+k_{n-1})\,x_n}\\ (1-S_n)e^{+i(k_n+k_{n-1})\,x_n} & (1+S_n)e^{+i(k_n-k_{n-1})\,x_n}\end{pmatrix},\; S_n=\frac{m_n}{m_{n-1}}\frac{k_{n-1}}{k_n}\,. \tag{A.6}$$

The transfer-matrix for the whole system (gate-dielectric-substrate) relates the amplitude of the envelope wave function in the gate region to that in the substrate region according to

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = \mathbf{T} \begin{pmatrix} A_0 \\ B_0 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} = \prod_{n=1}^{N} \mathbf{T}_n \,. \tag{A.7}$$

The left-most and right-most intervals, to which correspond solutions ψ_0 and ψ_N, could be considered semi-infinite, without loss of generality. Since to quasi-bound states (QBS) correspond wave-functions with evanescent density outside the inversion potential well, it follows that the amplitudes of the incoming waves from left and right of the system, A_0 and B_N, must both be 0. From the matrix equation (A.7), and from the symmetry relation $T_{21}^* = T_{22}$ (125), the above condition translates to

$$T_{22}(E_{\nu,\iota}) = 0. \tag{A.8}$$

This equation, (A.8), is used as a criterion for the determination of the QBS energy levels as in (105). Once a subband energy is determined, the corresponding wave-function at each interval boundary, x_n, is obtained from equation (A.4), knowing that

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = (\prod_{i=1}^{n} \mathbf{T}_i) \begin{pmatrix} A_0 \\ B_0 \end{pmatrix} , \tag{A.9}$$

and assuming a plane outgoing wave in the left-most semi-infinite interval, that is $A_n = 0$ and $B_n = 1$.

From the computed wave-function for each subband, the QBS lifetime is obtained following the approach in Ref. (126) (see section 3.2.3.4 for details) as:

$$\tau = \frac{m_{n-1}^*}{\hbar k_{n-1}} \frac{1}{|\psi_n(x_n)|^2} \int_{x_n}^{x_N} |\psi(x)|^2 dx \,. \tag{A.10}$$

The index n here corresponds to the boundary delineating the gate from the dielectric; $\psi_n(x_n)$ is the wave-function at the boundary, while m_{n-1}^*, and k_{n-1}, are the effective mass, and k vector, characteristic for the interval immediately to the left of this boundary, i.e. in the gate.

Finally, the tunnelling current at given bias of the structure is obtained by summing up the contributions from each sub-band (ι) in each valley (ν)

$$J_G = -q \sum_{\nu,\iota} \frac{n_{\nu,\iota}}{\tau_{\nu,\iota}} , \tag{A.11}$$

where $n_{\nu,\iota}$ is the two-dimensional (2D) carrier density (233), and q is the elementary charge.

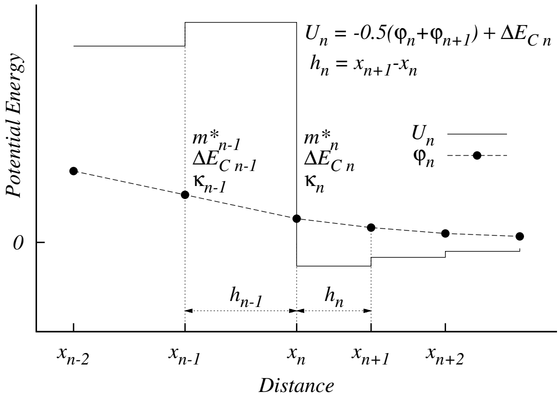


Figure A.1: Segmentation of the device for the discretization of the PE and SE. Each element of width h_n between two neighbouring nodes n and $n{+}1$ is a homogeneous media, with constant conduction band offset, ΔE_{Cn}, dielectric constant, κ_n, and effective mass, m^*_n. The electrostatic potential φ_n is a nodal quantity. It is the solution of the PE and is a smooth continuous function. The potential energy U_n for the solution of the SE is constant within a segment, and discontinuous, due to the discontinuities of the conduction band offset.

A.2 Discretization Scheme

The numerical self-consistent solution of the PE (A.1) and SE (A.2) requires their discretization and their coupling through equation (A.3) for the potential energy, and through the charge density relation to the envelope function density (233). For our specific study, the spatial variation of the material parameters introduce additional boundary conditions through the applicable conservation laws.

The PE is discretized using a finite difference (FD) scheme on an irregular grid, which with reference to Fig. A.1 transforms equation (A.1) into:

$$\frac{-2}{h_n + h_{n-1}}\left(\kappa_n \frac{\varphi_{n+1} - \varphi_n}{h_n} - \kappa_{n-1}\frac{\varphi_n - \varphi_{n-1}}{h_{n-1}}\right) = \frac{1}{\varepsilon_0}\rho_n \tag{A.12}$$

This discretization scheme preserves the continuity of the electric displacement vector across the boundary of two segments with a different dielectric constant, and allows us to model the penetration of electronic charge within the transitional layer of the gate oxide. When applied to all N nodes of the device grid, equation (A.12) yields a linear system with a tri-diagonal coefficient matrix $\mathbf{M}$, with elements (for $1 < n < N$)

$$M_{n\,n-1} = \frac{-2\kappa_{n-1}}{h_{n-1}(h_n + h_{n-1})} \qquad M_{n\,n} = \frac{2(\kappa_n h_{n-1} + \kappa_{n-1} h_n)}{h_{n-1}h_n(h_n + h_{n-1})} \qquad M_{n\,n+1} = \frac{-2\kappa_n}{h_n(h_n + h_{n-1})} \tag{A.13}$$

The non-linear dependence between charge and electrostatic potential is handled in the solver following a relaxation scheme through the iteration between solving PE and SE (249).

Using the transfer-matrix (TM) method described in the previous section, the solution of the SE (A.2) is piece-wise analytical, with constant potential energy within each element of the discretized device, and delivers the wave-function density at each grid node. As is shown on Fig. A.1, the discretization of the potential energy transforms equation A.3 into

$$U_n = -q\frac{(\varphi_n + \varphi_{n+1})}{2} + \Delta E_{C\,n}. \tag{A.14}$$

The average of the electrostatic potential at two adjacent nodes is taken to approximate a constant electrostatic potential within the segment delimited by the corresponding nodes. This introduces small error, since the electrostatic potential is a smooth and continuous function, while the parameters of the medium are constant within an element, but discontinuous at a node.

Appendix B

The Si/SiO$_2$ interface - from atomic structure to device simulations

Here we present the details of the link between *ab-initio* calculations of the Si/SiO$_2$ interface electronic structure, and device simulations. This is relevant to the study of the impact of Si/SiO$_2$ interface transition layer on the electrical characteristics of an MOS structure, presented in Chapter 5. More specifically, the interface band-gap profiles obtain here, are used in the comparison of the effects due to *linear* and *realistic* band transitions in 5.4.6.

B.1 Introduction

Our starting point is a 3D-periodic SiO$_2$/Si/SiO$_2$ super-cell, shown in Fig. B.1. The details of the structural model (α-quartz is assumed for the SiO$_2$) and its optimisation are thoroughly described in Ref. (239). Here, we are concerned with the translation of the *ab initio* calculated electronic structure, to the macroscopic parameters of band gap and band offsets, needed for device simulation.

Electronic structure is calculated in the density functional theory (DFT), with gradient-corrected density functionals, using Gaussian-type basis set.[1] The total density of states (DOS) is obtained using the values of one-electron $E(k)$ dispersion, calculated at a pre-defined set of k-points of the reciprocal space. The corresponding one-electron states are represented as linear

[1] All DFT data is courtesy to P. Sushko, University College of London. Simulations are performed with the Crystal 2003 computer code (284).

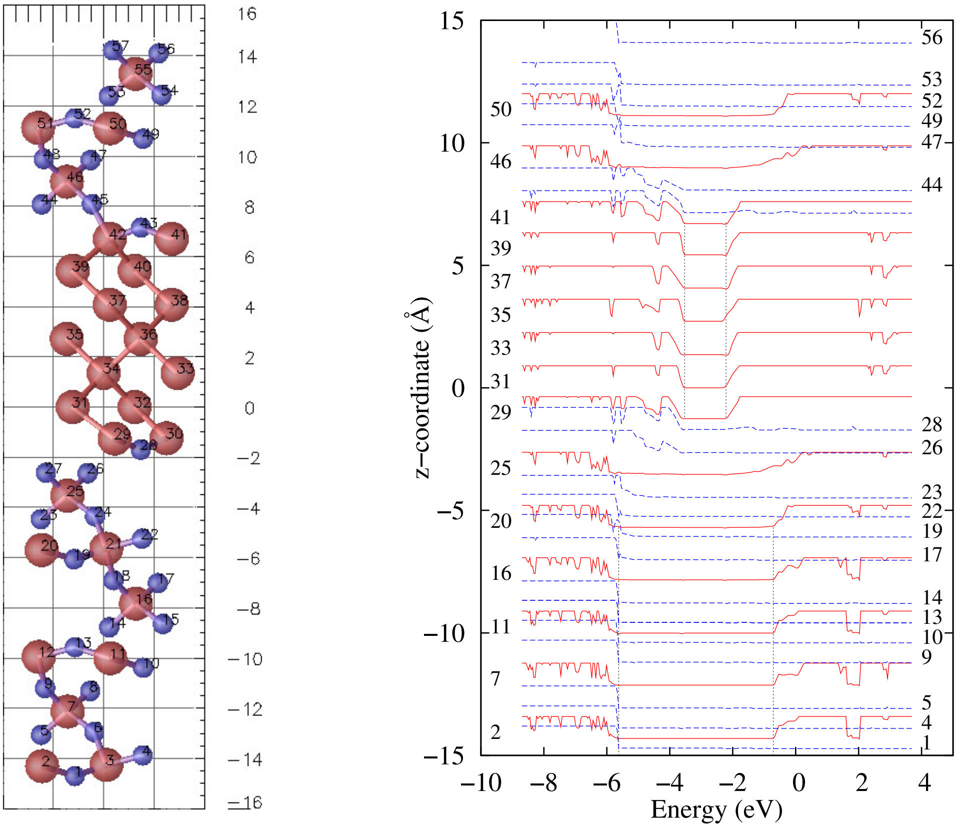


Figure B.1: 2D atomic structure of the simulated unit-cell (ball-and-stick representation) (left), and on atom-projected density of states (pDOS) (right). Interface atomic model is of α-quartz (203). Bigger (red) balls are Si atoms, smaller (blue) are O. Projected DOS are plotted in arbitrary units, and offset to the z-coordinate (vertical axis) of the corresponding atom, as labelled to the side. Vertical lines are a visual guide, suggesting band edges.

combination of atomic orbitals. This allows the projection of the total DOS on such an atomic basis, and therefore, the calculation of the atom-projected DOS (pDOS). An example of such a calculation is presented on Fig. B.1 (right).

Results for pDOS, obtained in this way, are not unique, but depend on the chosen density functional and the choice of pre-defined k-points. It is important to establish the extent to which such dependence is propagated to the simulated device characteristics, when the DFT band-profile and *realistic* transition at the interface is taken into account.

For the unit-cell structure shown in Fig. B.1, we have three sets of pDOS data, resulting from simulation with two different DFT functionals (PBE and B3LYP)), and two different number of pre-defined k- points (for the B3LYP functional).[1] Hereafter, the data sets are referred to as SET1 - PBE (small number of k-points), SET2 - B3LYP-1 (small number of k-points), SET3 - B3LYP-2 (bigger number of k-points). In the rest of this appendix, we elaborate on the procedure to extract conduction and valence band profiles from the pDOS data, and compare the outcome from the three data sets.

B.2 Band-edge extraction

Figure B.2 shows the conduction and valence band-edge profiles extracted from the PBE and B3LYP-2 pDOS data. The band-edges are determined from the highest occupied and lowest unoccupied states associated with each atom. For each data set, we used three different criteria of the minimum pDOS that determines the band-edge - 0.01, 0.05, and 0.1 (in arbitrary units). For 0.05 $< min(pDOS) <$ 0.1 the results for each data set nearly overlap. Below 0.05, the profiles exhibit a 'hump' within the transition layer. This 'hump' is associated with the first O atom in the fully stoichiometric oxide (i.e. atom 24 on Fig. B.1), whose density of states in the energy range close to the Si conduction band edge (CBE) is bigger than the extraction criterion of 0.01. Similar 'hump' in the band-edge is also present in the profiles in Ref. (205), but not in Ref. (203).

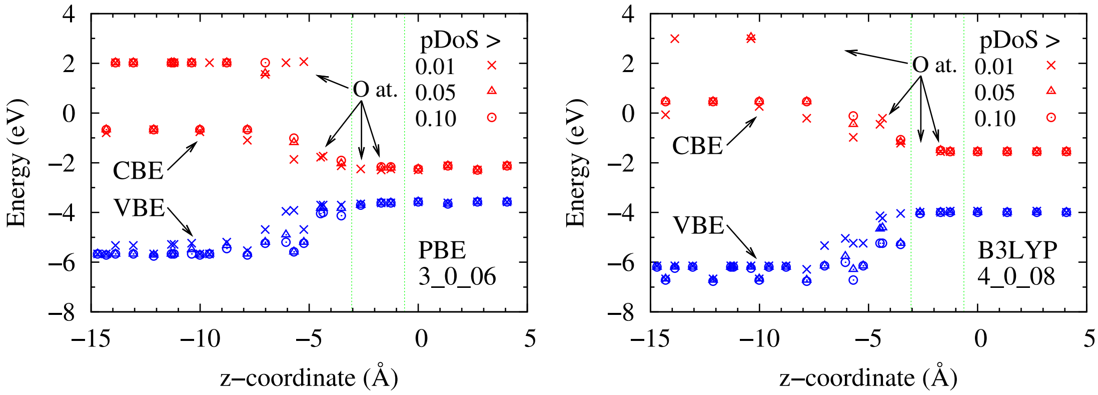


Figure B.2: Conduction and valence band edge profiles for PBE and B3LYP-2 sets, resulting (in each case) from different band-edge determination criteria, $min(pDOS)$, indicated by different symbols. The unoccupied states contributed by oxygen atoms are indicated by arrows. The results obtained for a range of criteria, $0.05 < pDOS < 0.10$, nearly overlap.

[1] The same basis set is used in all cases for atom representation (66-21G for Si, and 8-51G for O).

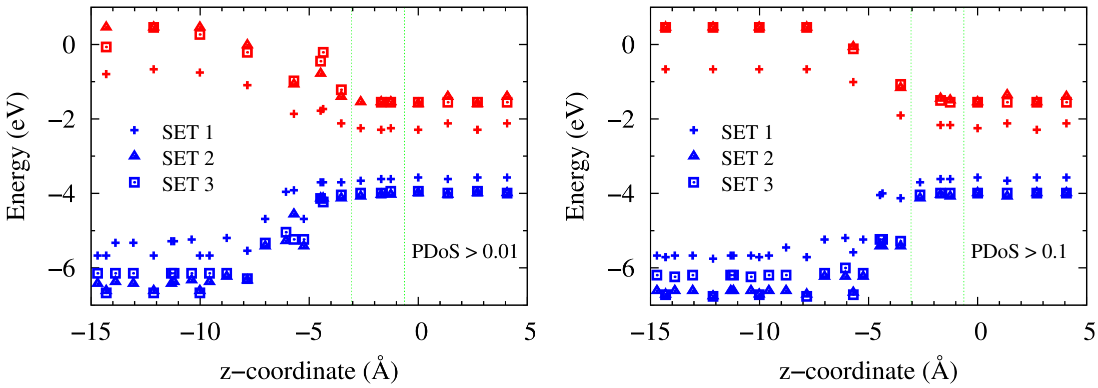


Figure B.3: Comparison between the three data sets at a fixed band-edge determination criterion, $min(PDoS) > 0.01$ (left), and $min(PDoS) > 0.1$ (right). While the profile of the CBE seems to be dependent on the functional only (the oxide CBE for SET 2 and SET 3 nearly overlap), the VBE in the SiO_2 depends strongly on the number of k-points too (there is an obvious discrepancy between the blue symbols in the region of the oxide).

For all band-edge determination criteria, oxygen atoms deeper in the stoichiometric oxide contribute states only for energies much higher than the edge determined by the silicon atoms. However, oxygen atoms within the sub-stoichiometric oxide (delimited by the dotted, green, vertical line on Fig. B.2) contribute states that lay in the same range as the ones contributed from Si atoms. Analogous effect is readily observed for the valence-band edge, where the band-edge determined from the highest occupied states due to O atoms are about 0.5 V above the band-edge determined from Si atoms.

We find similar qualitative picture for all three data sets, at a given band-edge determination criterion. It is noticeable however, that the set with bigger number of k-points for the DOS projection (B3LYP-2) gives more stable (i.e. identical, amongst atoms in the bulk-like region) band-edges in Si, and VBE in SiO_2. Contrary, the SiO_2 CBE of the PBE set is most stable. This trend is independent of the extraction criteria.

Figure B.3 shows a comparison between the extracted band-edge profiles from the three data sets for two values of the band-edge determination criterion – 0.01 (left) and 0.1 (right). The figure highlights a quantitative discrepancy, arising from the use of different functionals. It is clear that both PBE and B3LYP underestimate the bulk SiO_2 band-gap (∼8.9 eV), and overestimate the bulk Si band-gap (1.12 eV). The former effect is attributed to the DFT approach, while the latter is partly due to the very small thickness of the ordered Si (5 Å).

Figure B.3 also shows that regardless of the band-edge determination criterion and the number of k-points, the SiO$_2$ conduction band profiles for a given functional (i.e. B3LYP) nearly overlap. However the different number k-points results in a noticeable disagreement in the SiO$_2$ valence band profile.

Since there is an obvious difference in the profiles, even for an identical band-edge determination procedure, at this stage it is not possible to objectively choose any of them for subsequent device simulation. However, the band-profiles must be scaled, in order to bring their bulk-like band-gaps to the nominal experimental values for Si and SiO$_2$. It is after this stage, that agreement between the profiles is needed, in order to assert quantitative significance of the device simulation results.

B.3 Band-gap scaling

We note that for device simulation, not only the band-gap E_g itself, but also the conduction and valence band offsets, ΔE_c, and ΔE_v, are important. Further, we are interested in comparing the inversion layer quantisation and tunnelling characteristics of a metal/SiO$_2$/p-Si(100) device with and without band-gap transition at the oxide interface. The penetration of electrons in the oxide becomes the main impact factor, and it depends on the potential barrier for electrons at the interface. Therefore, we choose ΔE_c to be the leading parameter for scaling of the oxide region. The accuracy of the oxide band-gap is relaxed in this case, leading to an overestimation of the valence band offset. This is however irrelevant in our study, since calculation of hole density assumes an infinite potential barrier at the interface. For the region of the Si substrate however, the band-gap E_g is the leading parameter for scaling.

The scaled conduction and valence band-edge profiles, $E_c(z)$ and $E_v(z)$ respectively, are obtained from:

$$E_c(z) = E_0 + \alpha(z)\Delta E_c^{DFT}(z) \tag{B.1}$$

$$E_v(z) = E_c(z) - \alpha(z)E_g^{DFT}(z), \tag{B.2}$$

where E_0 is our energy reference, taken as the CBE of atom 35 in Si (refer to Fig. B.1, which has the same value regardless of the band-edge determination criterion. The scaling coefficient α has different value in SiO$_2$, and in Si, in accord with the argument in the preceding paragraph:

$$\alpha(SiO_2) = \Delta E_c^{nom}/\Delta E_c^{DFT}(atom\, 11) \tag{B.3}$$

$$\alpha(Si) = E_g^{nom}/E_g^{DFT}(atom\, 35), \tag{B.4}$$

Table B.1: Band-gap, conduction band offset, and scaling coefficients.

	SET 1		SET 2		SET 3	
	DFT	Scaled	DFT	Scaled	DFT	Scaled
$E_g(Si)$, eV	1.28	1.12	2.44	1.12	2.44	1.12
ΔE_c, eV	1.62	3.15	2.01	3.15	2.01	3.15
$E_g(SiO_2)$, eV	5.0	9.7	7.08	11.1	6.65	10.42
$\alpha(Si)$	0.87		0.46		0.46	
$\alpha(SiO_2)$	1.94		1.57		1.57	

where $\Delta E_c^{nom} = 3.15$ eV, and $E_g^{nom} = 1.12$ eV. In the sub-stoichiometric oxide, α is assumed to change linearly from the value in Si, to that in SiO$_2$:

$$\alpha(SiO_x) = ((\alpha(SiO_2) - \alpha(Si))/t_{SiO_x})z + \alpha(Si) \tag{B.5}$$

where z is the distance from the interface, and t_{SiO_x} is the thickness of the sub-stoichiometric oxide (~ 2 Å for the α-quartz SiO$_2$ structure adopted in this work).

Table B.1 summarises the important values related to the above equations, for the three data sets, based on the highest band-edge determination criterion of 0.10. The corresponding band profiles are shown in Fig. B.4 (left) and a good agreement could be observed. However, the relative difference between the three profiles in the interface transition region is a few times the thermal potential (≈ 26 meV), as shown in Fig B.4 (right).

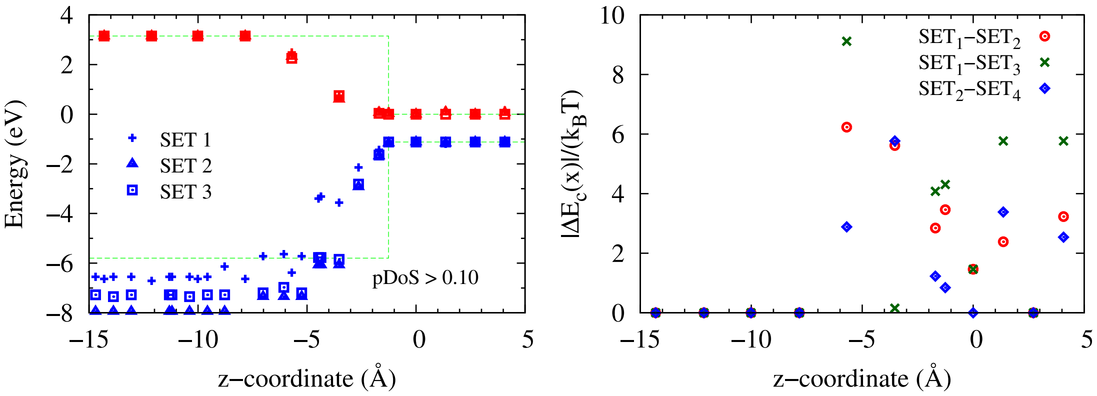


Figure B.4: Scaled band-profiles corresponding to the three data sets, extracted at the same band-edge determination criterion, $min(PDoS) > 0.1$ (left). Dashed (green) line indicates idealised abrupt band-transition assuming $E_g^{si} = 1.12$ eV, $E_g^{ox} = 8.95$ eV, $dEc = 3.15$ eV. Relative error between each two profiles, in multiples of k_BT (right) shows appreciable disagreement between the three conduction band profiles.

B.4 Conclusions

The translation of electronic structure obtained from *ab-initio* DFT simulations, into macroscopic band-gap and band-alignment information, for device simulation, is not unique. The ambiguity stems partly from the first principles simulation routine itself – choice of a SiO_2 structural model, density functional, k-space grid for DOS projection, and probably the choice of atomic basis set. Additional element of uncertainty is implicated by the simplicity of the procedure for extracting the band-edge. Having in mind the sensitivity of the sub-band quantisation on the shape of the inversion quantum well, one could expect the band-profiles compared so far to impact the device simulation result to a slightly different degree.

Bibliography

[1] L. F. Register, E. Rosenbaum, and K. Yang, "Analytic model for direct tunneling current in polycrystalline silicon-gate metal-oxide-semiconductor devices," *Applied Physics Letters*, vol. 74, pp. 457 – 459, 1999. iv, 28, 29, 33, 34, 35, 38, 40, 41, 42, 116

[2] W. Haensch, E. Nowak, R. Dennard, P. Solomon, A. Bryant, O. Dokumaci, A. Kumar, X. Wang, J. Johnson, and M. Fischetti, "Silicon CMOS devices beyond scaling," *IBM Journal Research & Development*, vol. 50, p. 339, 2006. 1, 2, 8, 9

[3] T. B. Hook, "Technology elements and chip design for low power applications," *Technical Digest - International Electron Device Meeting (IEDM)*, p. 1, 2006. 2

[4] Y.-F. Tsai, A. Ankadi, N. Vijaykrishnan, M. Irwin, and T. Theocharides, "ChipPower, and architecture-level leakage simulator," *Proc. IEEE Int. System-On-Chip Conference*, p. 395, 2004. 2, 3

[5] T. Tuan and B. Lai, "Leakage power analysis of a 90 nm FPGA," *Proc. IEEE Int. Conf. on Custom Integrated Circuits*, p. 57, 2003. 2

[6] N. Yang, W. K. Henson, and J. J. Wortman, "Comparative study of gate direct tunneling and drain leakage currents in n-MOSFET's with sub-2-nm gate oxides," *IEEE Transactions on Electron Devices*, vol. 47, pp. 1636 – 1644, 2000. 2, 41

[7] S. Crowder, "Low power CMOS process technology," *Tutorial, IEEE International Electron Device Meeting (IEDM)*, 2005. 2, 12

[8] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. Das, W. Haensch, E. Nowak, and D. Sylvester, "Ultralow-oltage, mimimum-energy CMOS," *IBM Journal Research & Development*, vol. 50, p. 469, 2006. 2

[9] A. Agarwal, S. Mukhopadahyay, C. Kim, A. Raychowdhury, and K. Roy, "Leakage power analysis and reduction: models, estimation and tools," *IEE Proc.-Comp. Digit. Tech.*, vol. 152, p. 353, 2005. 2

[10] "International technology roadmap for semiconductors (ITRS) 2007, www.itrs.net." 2, 8, 13, 80, 107, 112

[11] K. *et al.* Mistry, "A 45nm logic technology with high-k+metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100% Pb-free packaging," *Technical Digest, IEEE International Electron Devices Metting*, pp. 247–250, 2007. 2, 80, 107

[12] P. *et. al.*. Sivasubramani, "Aggressively scaled high-κ gate dielectric with excellent performance and high temperature stability for 32 nm and beyond," *International Electron Devices Metting, Technical Digest*, p. 543, 2007. 2, 107, 108

[13] M. Takahashi, A. Ogawa, A. Hirano, Y. Kamimuta, Y. Watanabe, K. Iwamoto, S. Migita, N. Yasuda, H. Ota, T. Nabatame, and A. Toriumi, "Gate-first processed FUSI/HfO_2/$HfSiO_x$/Si MOSFETS with EOT=0.5 nm," *International Electron Devices Metting, Technical Digest*, p. 523, 2007. 2, 107, 108

[14] E. Gusev, V. Narayanan, and M. Frank, "Advanced high-κ dielectric stacks with polySi and metal gates," *IBM Journal Research & Development*, vol. 50, p. 387, 2006. 2, 12, 107

[15] J. Robertson, "High dielectric constant gate oxides for metal oxide Si transistors," *Reports on Progress in Physics*, vol. 69, p. 327, 2006. 2, 107, 108

[16] H. Ota, A. Hirano, Y. Watanabe, N. Yasuda, K. Iwamoto, K. Akiyama, K. Okada, S. Migita, T. Nabatame, and A. Toriumi, "Intrinsic origin of electron mobility reduction in high-k MOSFETs – from remote phonot to bottom interface dipole scattering," *International Electron Devices Metting, Technical Digest*, p. 65, 2007. 2, 107

[17] Y. Kamimuta, K. Iwamoto, Y. Nunoshige, A. Hirano, W. Mizubayashi, Y. Watanabe, S. Migita, A. Ogawa, H. Ota, T. Nabatame, and A. Toriumi, "Comprehensive study of V_{FB} shift in high-k CMOS – dipole formation, fermi-level pinning and oxygen vacancy effect," *International Electron Devices Metting, Technical Digest*, p. 341, 2007. 2

[18] K. Bernstein, D. Frank, A. Gattiker, W. Haensch, B. Ji, S. Nassif, E. Nowak, D. Pearson, and N. Rohrer, “High-performance CMOS variability in the 65-nm regime and beyond,” *IBM Journal Research & Development*, vol. 50, p. 433, 2006. 2, 3, 14, 15

[19] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli, “Variation in transistor performance and leakage in nanometer-scale technologies,” *IEEE Transactions on Electron Devices*, vol. 55, p. 131, 2008. 2, 14, 15

[20] H. Tuinhout, “Impact of parametric mismatch and fluctuations on performance and yield of deep-submicron CMOS technologies,” *ESSDERC 2002 - Proceedings of the 28th European Solid-State Device Research Conference*, vol. Florence, Italy, pp. 95–101, 2002. 2

[21] B. Cheng, S. Roy, G. Roy, F. Adamu-Lema, and A. Asenov, “Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells,” *Solid-State Electronics*, vol. 49, pp. 740 – 746, 2005. 2

[22] T. Mizuno, J. Okamura, and A. Toriumi, “Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 41, p. 2216, 1994. 3

[23] P. Stolk, F. Widdershoven, and D. Klaassen, “Modeling statistical dopant fluctuations in MOS transistors,” *IEEE Transactions on Electron Devices*, vol. 45, p. 1960, 1998. 3, 43

[24] P. Andrei and I. Mayergoyz, “Quantum mechanical effects on random oxide thickness and random dopant induced fluctuations in ultrasmall semiconductor devices,” *Journal of Applied Physics*, vol. 94, p. 7163, 2003. 3

[25] C. Diaz, H.-J. Tao, Y.-C. Ku, A. Yen, and K. Young, “An experimentally validated analytical model for gate line edge roughness effects on technology scaling,” *IEEE Electron Device Letters*, vol. 22, p. 287, 2001. 3

[26] A. Asenov, S. Kaya, and J. H. Davies, “Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations,” *IEEE Transactions on Electron Devices*, vol. 49, pp. 112 – 119, 2002. 3, 49, 54

[27] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1837 – 1852, 2003. 3, 15, 17, 45, 49

[28] M. Hane, T. Ikezawa, and T. Ezaki, "Atomistic 3D process/device simulation considering gate line-edge roughness and poly-si random crystal orientation effects," *International Electron Devices Metting, Technical Digest*, p. 9.5.1, 2003. 3

[29] G. Roy, A. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 53, p. 3063, 2006. 3, 14, 15, 45, 62

[30] Cassan.E, S. Galdin, P. Dollfus, and P. Hesto, "Study of direct tunneling through ultrathin gate oxide of field effect transistors using Monte Carlo simulation," *Journal of Applied Physics*, vol. 86, p. 3804, 1999. 3, 16, 26, 47, 61, 78, 117

[31] S. Toriyama, K. Matsuzawa, and N. Sano, "Gate tunneling current fluctuations associated with random dopant effects," *2005 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 23 – 6, 2005. 3, 16, 45, 50, 64, 70

[32] S. Tyaginov, M. Vexler, A. Shulekin, and I. Grekhov, "Statistical analysis of tunnel currents in scaled MOS structures with a non-uniform oxide thickness distribution," *Solid-State Electronics*, vol. 49, pp. 1192 – 1197, 2005. 3, 16

[33] L.-F. Mao, Y. Yang, J.-L. Wei, H. Zhang, M.-Z. Xu, and C.-H. Tan, "Effect of SiO¡sub¿2¡/sub¿/Si interface roughness on gate current," *Microelectronics Reliability*, vol. 41, pp. 1903 – 1907, 2001. 3

[34] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical estimation of leakage current considering inter- and intra-die process variation," *Proc. Int. Symp. on Low Power Electronics Design (ISLPED)*, p. 84, 2003. 3, 19

[35] D. Frank, W. Haensch, G. Shahidi, and O. Dokumaci, "Optimizing CMOS technology for maximum performance," *IBM Journal Research & Development*, vol. 50, p. 419, 2006. 3, 9, 15

[36] R. Dennard, F. Gaensslen, H. Yu, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, p. 256, 1974. 6

[37] Y. Taur and T. Ning, *Fundamentals of modern VLSI design.* Cambridge University Press, 1998. 6, 7, 8, 9, 10, 11, 20, 21, 26, 45, 53

[38] D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H.-S. Wong, "Device scaling limits of Si MOSFETs and their application dependencies," *Proceedings IEEE*, vol. 89, p. 259, 2001. 7, 8

[39] G. Baccarani, M. Wordeman, and R. Dennard, "Generalised scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Transactions on Electron Devices*, vol. 31, p. 452, 1984. 8

[40] D. Frank, "Power constrained CMOS scaling limits," *IBM Journal Research & Development*, vol. 46, p. 235, 2002. 9, 11

[41] K. Roy, S. Mukhopadhayay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proceedings of the IEEE*, vol. 91, p. 305, 2003. 9, 11

[42] H. Iwai, "Technology scaling and roadmap," *International Electron Devices Metting, Short Course – 22 nm CMOS Technology*, 2008. 10, 13

[43] A. Toriumi, K. Kita, K. Tomida, Y. Zhao, J. Widiez, T. Nabatame, H. Ota, and M. Hirose, "Materials science-based device performance engineering for metal gate high-k CMOS," *International Electron Devices Metting, Technical Digest*, p. 53, 2007. 12, 107, 108, 110

[44] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub 50 nm MOSFETs: A statistical 3D 'atomistic' simulation study," *Nanotechnology*, vol. 10, pp. 153 – 158, 1999. 15

[45] S. Borhar, T. Kamik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," *Proc. 40th Desigh Automation Conf. (DAC 03)*, p. 338, 2003. 15

[46] R. Rao, A. Devgan, D. Blaauw, and D. Sylvester, "Analytical yield prediction considering leakage/performance correlation," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, p. 1685, 2006. 15, 19

[47] S. Goodnick, D. Ferry, and C. Wilmsen, "Surface roughness at the Si(100)-SiO_2 interface," *Physical Review B*, vol. 32, p. 8171, 1985. 16, 54, 55

[48] T. Yoshinobu, A. Iwamoto, K. Sudoh, and H. Iwasaki, "Scaling of Si-SiO2 interface roughness," *Journal of Vacuum Science and Technology*, vol. 13, p. 1630, 1995. 16, 54, 55

[49] D. Buchanan, "Scaling the gate dielectric: materials, integration and reliability," *IBM Journal Research & Development*, vol. 43, p. 245, 1999. 16, 107

[50] M. Gotoh, K. Sudoh, H. Itoh, and K. Kawamoto, "Analysis of SiO_2/Si(001) interface roughness for thin gate oxides by scanning tunneling microscopy," *Applied Physics Letters*, vol. 81, p. 430, 2002. 16

[51] D.-Y. Ting, "Tunnelling characteristics of nonuniform ultrathin oxides," *Applied Physics Letters*, vol. 73, p. 2769, 1998. 16

[52] B.-C. Hsu, K.-F. Chen, C.-C. Lai, S. Lee, and C. Liu, "Oxide roughness effects on tunnelling current of MOS diodes," *IEEE Transactions on Electron Devices*, vol. 49, p. 2204, 2002. 16

[53] A. Asenov, M. Jaraiz, S. Roay, G. Roy, F. Adamu-Lema, A. Brown, and V. Moroz, "Integrated atomistic process and device simulation of decananometer MOSFETs," *Proc. SISPAD 2002*, pp. 87–90, 2002. 17

[54] H. Watanabe, "Statistics of grain boundaries in polysilicon," *IEEE Transactions on Electron Devices*, vol. 54, p. 38, 2007. 19

[55] M. W. Koh, M, K. Iwamoto, H. Murakami, T. Ono, M. Tsuno, T. Mihara, K. Shibahara, S. Miyazaki, and M. Hirose, "Limit of gate oxide thickness scaling in MOSFETs due to apparent threshold voltage fluctuation induced by tunnel leakage current," *IEEE Transactions on Electron Devices*, vol. 48, p. 259, 2001. 19, 35

[56] C. Duke, *Tunneling in Solids.* Academic Press, NY, 1969. 20, 22, 26, 27

[57] C. Chang, M.-S. Liang, C. Hu, and R. Brodersen, "Carrier tunneling related phenomena in thin oxide MOSFET's," *International Electron Devices Metting, Technical Digest*, p. 194, 1983. 20, 22

[58] Y. Shi, T. Ma, S. Prasad, and S. Dhanda, "Polarity dependent gate tunneling currents in dual-gate CMOSFETs," *IEEE Transactions on Electron Devices*, vol. 45, p. 2355, 1998. 20

[59] A. Ghetti, C.-T. Liu, M. Mastrapasqua, and E. Sangiorgi, "Characterization of tunneling current in ultra-thin gate oxide," *Solid State Electronics*, vol. 44, p. 1523, 2000. 20, 21, 35

[60] W. Lee and C. Hu, "Modeling CMOS tunneling currents throught ultrathin gate oxide due to conduction and valence band electron and hole tunneling," *IEEE Transactions on Electron Devices*, vol. 48, p. 1366, 2001. 20

[61] M. Lenzlinger and E. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *Journal of Applied Physics*, vol. 40, p. 278, 1969. 21

[62] G. Krieger and R. Swanson, "Fowler-Nordheim electron tunneling in thin Si-SiO_2-Al structures," *Journal of Applied Physics*, vol. 52, p. 5710, 1981. 21

[63] Z. Weinberg, "On tunneling in metal-oxide-silicon structures," *Journal of Applied Physics*, vol. 53, p. 5052, 1982. 21

[64] J. Maserjian, "Tunnelling in thin MOS structures," *Journal of Vacuum Science and Technology*, vol. 11, p. 996, 1974. 21, 22, 25

[65] S. Nagano, M. Tsukiji, K. Ando, E. Hasegawa, and A. Ishitani, "Mechanism of leakage current through the nanoscale SiO_2 layer," *Journal of Applied Physics*, vol. 75, p. 3530, 1994. 21, 22

[66] W.-K. Shih, E. Wang, S. Jallepalli, F. Leon, C. Maziar, and A. Taschjr, "Modeling gate leakage current in nmos structures due to tunneling through an ultra-thin oxide," *Solid State Electronics*, vol. 42, p. 997, 1998. 21, 30

[67] T. Sorsch, W. Timp, F. Baumann, K. Bogart, T. Boone, V. Donnelly, M. Green, K. Evans, C. Kim, S. Moccio, J. Rosamilia, J. Sapjeta, P. Silverman, B. Weir,

and G. Timp, “Ultra-thin, 1.0-3.0nm, gate oxides for high performance sub-100nm technology,” *Proc. Intl. VLSI Symposium*, p. 222, 1998. 22

[68] Fukuda.M, W. Mizubayashi, A. Kohno, S. Miyazaki, and M. Hirose, “Analysis of tunnel current through ultrathin gate oxides,” *Japanese Journal of Applied Physics*, vol. 37, p. L1534, 1998. 22

[69] B. Brar, G. Wilk, and A. Seabaugh, “Direct extraction of the electon tunneling effective mass in ultrathin SiO_2,” *Journal of Applied Physics*, vol. 69, p. 2728, 1996. 22

[70] P. Lundgren and M. Andersson, “Temperature dependence confirmation of tunneling through 2-6 nm silicon dioxide,” *Solid State Electronics*, vol. 36, p. 1143, 1996. 22

[71] A. Yassin and R. Hijab, “Temperature dependence of gate current in ultra thin SiO_2 in direct-tunneling regime,” *Proc. IEEE Intl. Integrated Reliability Workshop*, vol. Stanford Sierra Camp (CA), USA, p. 56, Oct. 13-16, 1997. 22

[72] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, “Modeling study of ultrathin gate oxides using direct tunneling current and capacitance-voltage measurements in MOS devices,” *IEEE Transactions on Electron Devices*, vol. 46, pp. 1464 – 1471, 1999. 22, 29, 33, 35, 39, 40, 41, 66

[73] T. Ando, A. Fowler, and F. Stern, “Electronic properties of two-dimensional systems,” *Reviews of Modern Physics*, vol. 54, pp. 437 – 672, 1982. 23

[74] B. Majkusiak, “Gate tunnel current in a MOS transistor,” *IEEE Transactions on Electron Devices*, vol. 37, p. 1087, 1990. 24

[75] A. Shanware, J. Shiely, H. Massoud, E. Vogel, K. Henson, A. Srivastava, C. Osburn, J. Hauser, and J. Wortman, “Extraction of the gate oxide thickness of n- and p-channel MOSFETs below 20åfrom the substrate current resulting from valence-band electron tunnelling,” *International Electron Devices Metting, Technical Digest*, p. 815, 1999. 24

[76] A. Shanware, H. Massoud, E. Vogel, K. Henson, J. Hauser, and J. Wortman, “Modeling the trends in valence-band electron tunneling in NMOSFETs with ultrathin SiO_2 and SiO_2/Ta_2O_5 dielectris with oxide scaling,” *Microelectronic Engineering*, vol. 48, p. 295, 1999. 24

[77] A. Chou, K. Lai, K. Kumar, P. Chowdhury, and J. Lee, "Modeling of stress-induced leakage current in ultrathin oxides with the trap-assisted tunnelling mechanism," *Applied Physics Letters*, vol. 70, p. 3407, 1997. 25

[78] A. Gehring and S. Selberherr, "Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices," *IEEE Transactions on Device and Material Reliability*, vol. 4, p. 306, 2004. 25, 26, 27

[79] S.-i. Takagi, N. Yasuda, and A. Toriumi, "A new IV model for stress-induced leakage current including inelastic tunneling," *IEEE Transactions on Electron Devices*, vol. 46, p. 348, 1999. 25

[80] T.-K. Kang, M.-J. Chen, C.-H. Liu, Y. Chang, and S.-K. Fan, "Numerical confirmation of inelastic trap-assisted tunneling as SILC mechanism," *IEEE Transactions on Electron Devices*, vol. 48, p. 2317, 2001. 25

[81] D. Ielmini, SpinellinA, M. Rigamonti, and A. Lacaita, "Modeling of SILC based on electron and hole tunneling - Part I Transient effects," *IEEE Transactions on Electron Devices*, vol. 47, p. 1258, 1997. 25

[82] F. Jimenez-Molinos, A. Palma, F. Cagmiz, J. Banqueri, and J. Lopez-Villanueva, "Physical model for trap-assisted inelastic tunneling in MOS structures," *Journal of Applied Physics*, vol. 90, p. 3396, 2001. 25

[83] M. Chang, J. Zhang, and W. Zhang, "Assessment of capture cross section and effective density of electron traps generated in silicon dioxides," *IEEE Transactions on Electron Devices*, vol. 53, p. 1374, 2006. 25

[84] A. Ghetti, M. Alam, J. Bude, D. Monroe, E. Sangiorgi, and H. Vaidya, "Stress induced leakage current analysis via quantum yield experiments," *IEEE Transactions on Electron Devices*, vol. 47, p. 1341, 2000. 25

[85] W. Chang, M. Houng, and Y. Wang, "Electrical properties and modeling of ultrathin impurity-doped silicon dioxides," *Journal of Applied Physics*, vol. 90, p. 5171, 2001. 25

[86] J. Stathis, "Reliaility limits for the gate insulator in CMOS technology," *IBM Journal Research & Development*, vol. 46, p. 265, 2002. 25

[87] W. Chang, M. Houng, and Y. Wang, “Simulation of SILC is silicon dioxides a modified TAT model considering Gaussian-distributed traps and electron energy loss,” *Journal of Applied Physics*, vol. 89, p. 6285, 2001. 25

[88] A. Melik-Martirosian and T. Ma, “Lateral profiling of interface traps and oxide charge in MOSFET devices – charge pumping versus DCIV,” *IEEE Transactions on Electron Devices*, vol. 48, p. 2303, 2001. 25

[89] F. Irrera and G. Puzzilli, “Degradation of ultra-thin oxides,” *IEEE Transactions on Device and Materials Reliability*, vol. 4, p. 530, 2004. 25

[90] D. Fleetwood, P. Winokur, L. Riewe, and R. Reber, “Bulk oxide traps and border traps in metal-oxide-semiconductor capacitors,” *Journal of Applied Physics*, vol. 84, p. 6141, 1998. 25

[91] J. Wu, L. Register, and E. Rosenbaum, “Trap-assisted tunneling current through ultra-thin oxide,” *Annual Proceedings - Reliability Physics (Symposium)*, pp. 389 – 395, 1999. 25, 29, 34, 35, 38, 41

[92] M. Houssa, M. Tuominen, M. Naili, V. Afanas’ev, A. Stesmans, S. Haukka, and M. Heyns, “Trap-assisted tunneling in high permittivity gate dielectric stacks,” *Journal of Applied Physics*, vol. 87, p. 8615, 2000. 25

[93] G. Wilk, R. Wallace, and J. Anthony, “High-κ gate dielectrics: current status and materials properties considerations,” *Journal of Applied Physics*, vol. 89, p. 5243, 2001. 25

[94] Y. Hou, M. Li, Y. Jin, and W. Lai, “Direct tunneling hole currents through ultrathin gate oxides in metal-oxide-semiconductor devices,” *Journal of Applied Physics*, vol. 91, p. 258, 2002. 26, 29

[95] J. Cai and C.-T. Sah, “Gate tunneling currents in ultrathin oxide metal-oxide-silicon transistors,” *Journal of Applied Physics*, vol. 89, pp. 2272 – 2285, 2001. 26, 29

[96] W. A. Harrison, “Tunnelling from an independent-particle point of view,” *Physical Review*, vol. 123, p. 85, 1961. 26, 29

[97] F. Rana, S. Tiwari, and D. Buchanan, "Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides," *Applied Physics Letters*, vol. 69, pp. 1104 –, 1996. 26, 33

[98] J. Albert Thomas Fromhold, *Quantum Mechanics for Applied Physics and Engineering*. Dover Publications Inc., NY, 1991. 27

[99] A. Gehring, T. Grasser, H. Kosina, and S. Selberherr, "Simulation of hot-electron oxide tunneling current based on a non-Maxwellian electron energy distribution function," *Journal of Applied Physics*, vol. 92, p. 6019, 2002. 27, 47

[100] D. Bohm, *Quantum Theory*. Prentice Hall, Inc. NY, 1952. 27, 28, 29

[101] K. Gundlach, "Zur berechnung des tunnelstroms durch eine trapezformige potentialstufe," *Solid State Electronics*, vol. 9, p. 947, 1966. 27

[102] A. Gehring, "Simulation of tunnelling in semiconductor devices," Ph.D. dissertation, TU Wienna, 2003. 27

[103] J. H. Davies, *The Physics of Low-Dimensional Semiconductors - an Introduction*. Cambridge University Press, NY, 1998. 27, 29, 32, 123

[104] Y. Ando and T. Itoh, "Calculation of transmission tunnelling current accross arbitrary potential barriers," *Journal of Applied Physics*, vol. 61, p. 1497, 1987. 27, 123

[105] E. Cassan, "On the reduction of direct tunneling leakage through ultrathin gate oxides by a one-dimensional Schrodinger-Poisson solver," *Journal of Applied Physics*, vol. 87, p. 7931, 2000. 27, 31, 124

[106] O. Simonetti, T. Maurel, and M. Jourdain, "Characterization of ultrathin MOS stuctures using coupled current and capacitance-voltage models based on quantum calculation," *Journal of Applied Physics*, vol. 92, p. 4449, 2002. 27, 33, 35

[107] N. Barin, C. Fiegna, and E. Sangiorgi, "Analysis of strained silicon on insulator double gate MOS structures," *2004 Eur. Solid-State Device Research Conf.*, p. 169, 2004. 27, 31

[108] S. Markov, S. Roy, C. Fiegna, E. Sangiorgi, and A. Asenov, "On the sub-nm EOT scaling of high-k gate stacks," *Proc. Ultimate Limits of Integration in Silicon 2008, Udine, Italy*, pp. 99 – 103, 2008. 27, 110

[109] W. Frensley, "Numerical evaluation of resonant states," *Superlattices and Microsturctures*, vol. 11, p. 347, 1992. 28, 30

[110] F. Li, S. Mudanai, Y.-Y. Fan, L. Register, and S. Banerjee, "Physically based quantum-mechanical compact model of MOS devices substrate-injected tunneling current through ultrathin (EOT$\tilde{1}$ nm sio$_2$ and high-κ gate stacks)," *IEEE Transactions on Electron Devices*, vol. 53, p. 1096, 2006. 28, 29, 35

[111] J. Bardeen, "Tunnelling from a many-particle point of view," *Physical Review Letters*, vol. 6, p. 57, 1961. 29

[112] Khairurrijal, W. Mizubayashi, S. Miyazaki, and M. Hirose, "Analytic model of direct tunnel current through ultrathin gate oxides," *Journal of Applied Physics*, vol. 87, pp. 3000 – 3005, 2000. 29, 35, 41

[113] C. Cohen-Tannoudji, B. Diu, and F. Laloe, *Quantum Mechanics*. Hermann, 1977. 29, 30

[114] W. Magnus and W. Schoenmaker, "Full quantum mechanical model for the charge distribution and the leakage currents in ultrathin metal-insulator-semiconductor capacitors," *Journal of Applied Physics*, vol. 88, p. 5833, 2000. 30, 31, 90

[115] R. Clerc, A. Spinelli, G. Ghibaudo, and G. Pananakakis, "Theory of direct tunneling current in metal-oxide-semiconductor structures," *Journal of Applied Physics*, vol. 91, pp. 1400 –, 2002. 30, 31

[116] S. Mudanai, Y.-Y. Fan, Q. Ouyang, A. Tasch, and S. Banerjee, "Modeling of direct tunneling current through gate dielectric stacks," *IEEE Transactions on Electron Devices*, vol. 47, p. 1851, 2000. 30

[117] J. Sun, W. Wang, T. Toyabe, N. Gu, and P. Mazumder, "Modelign of gate current and capacitance in nanoscale-MOS structures," *IEEE Transactions on Electron Devices*, vol. 53, p. 2950, 2006. 30

[118] M. Karner, A. Gehring, H. Kosina, and S. Selberherr, "Efficient calculation of quasi-bound state tunneling in CMOS devices," *2005 International Conference on Simulation of Semiconductor Processes and Devices*, pp. 2 – 5, 2005. 31

[119] A. Khondker, M. Hkan, and A. Anwar, “Transmission line analogy of resonance tunneling phenomena – the generalised impedance concept,” *Journal of Applied Physics*, vol. 63, p. 5191, 1988. 31

[120] S.-H. Lo, D. Buchanan, and Y. Taur, “Modeling and characterization of quantization, polysilicon depletion, and direct tunneling effects in MOSFETs with ultrathin oxides,” *IBM Journal of Research and Development*, vol. 43, pp. 327 – 337, 1999. 31, 41

[121] S.-H. Lo, D. Buchanan, Y. Taur, and W. Wang, “Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET’s,” *IEEE Electron Device Letters*, vol. 18, pp. 209 – 211, 1997. 31, 41, 89

[122] A. Rahman and A. Haque, “A study into the broadening of the quantized inversion layer states in deep submicron MOSFETs,” *Solid State Electronics*, vol. 45, p. 755, 2001. 31

[123] K. Alam, S. Zaman, M. Chowdhury, M. Khan, and A. Haque, “Effects of inelastic scattering on direct tunneling gate leakage current in deep submicron MOS transistors,” *Journal of Applied Physics*, vol. 92, p. 937, 2002. 31

[124] J. Price, Peter, “Resonant tunneling via an accumulation layer,” *Physical Review B*, vol. 45, p. 9042, 1992. 31

[125] G. Gildenblat, B. Gelmont, and S. Vatannia, “Resonant behavior, symmetry, and singularity of the transfer matrix in asymmetric tunnelling structures,” *Journal of Applied Physics*, vol. 77, p. 6327, 1995. 31, 124

[126] J. Price, Peter, “Electron tunneling from channel to gate,” *Applied Physics Letters*, vol. 83, p. 2080, 2003. 31, 32, 42, 100, 116, 124

[127] ——, “Attempt frequency in tunneling,” *American Journal of Physics*, vol. 66, p. 1119, 1998. 33

[128] E. Nakhmedov, K. Wieczorek, H. Burghardt, and C. Radehaus, “Quantum-mechanical study of the direct tunneling current in MOS structures,” *Journal of Applied Physics*, vol. 98, p. 024506, 2005. 33

[129] H. Kroemer, *Quantum Mechanics for engineering, materials science, and applied physics*. Prentice Hall, 1994. 34

[130] J. Maserjian and G. Petersson, “Tunnelling through thin MOS structures: Dependence on energy (E-k),” *Applied Physics Letters*, vol. 25, p. 50, 1974. 34, 39

[131] H. Watanabe, D. Matsushita, and K. Muraoka, “Determination of tunnel mass and physical thickness of gate oxide including poly-Si/SiO$_2$ and Si/SiO$_2$ interfacial transition layers,” *IEEE Transactions on Electron Devices*, vol. 53, p. 1323, 2006. 35, 81, 88, 91, 97, 115

[132] M. Stadele, B. Tuttle, and K. Hess, “Tunnelling through ultrathin SiO$_2$ gate oxides from microsopic models,” *Journal of Applied Physics*, vol. 89, p. 348, 2001. 35, 36, 48, 89, 91

[133] X.-G. Zhang, Z.-Y. Lu, and S. Pantelides, “First-principles theory of tunnelling currents in metal-oxide-semiconductor structures,” *Applied Physics Letters*, vol. 89, p. 032112, 2006. 35, 89

[134] G. Krieger and R. Swanson, “Electron tunneling in Si-SiO$_2$-Al structures – a comparison between ¡100¿ oriented and ¡111¿ oriented Si,” *Applied Physics Letters*, vol. 39, p. 818, 1981. 36

[135] Z. Weinberg, “Effects of silicon orientation and hydrogen anneal on tunnelling from Si into SiO$_2$,” *Journal of Applied Physics*, vol. 54, p. 2517, 1983. 36

[136] M. Fischetti, S. Laux, and E. Crabbe, “Understanding hot-electron transport in silicon devices - is there a shortcut?” *Journal of Applied Physics*, vol. 78, p. 1058, 1995. 36, 37, 47

[137] A. Schenk and G. Heiser, “Modelling and simulation of tunnelling through ultra-thin gate dielectrics,” *Journal of Applied Physics*, vol. 81, p. 7900, 1997. 37

[138] M. Stadele, F. Sacconi, A. Di Carlo, and P. Lugli, “Enhancement of the effective tunnel mass in ultrathin silicon dioxide layers,” *Journal of Applied Physics*, vol. 93, p. 2681, 2003. 37

[139] E. Goldman, N. Kakharskaya, and A. Zhdan, “The effect of imaging forces in ultra thin gate insulator on the tunnelling current and its oscillations at the region of transition from the direct tunnelling to the Fowler-Nordheim tunnelling,” *Solid State Electronics*, vol. 48, p. 831, 2004. 37

[140] M. Ancona and H. Tiersten, “Macroscopic physics of the silicon inversion layer,” *Physical Review B*, vol. 35, p. 7959, 1987. 37, 45

[141] M. Ancona and G. Iafrate, “Quantum correction to the equation of state of an electron gas in a semiconductor,” *Physical Review B*, vol. 39, p. 9536, 1989. 37, 45

[142] A. Brown, “The Glasgow implementation of density-gradient in drift-diffusion simulations.” 38

[143] R. Keyes, “Effect of randomness in the distribution of impurity ions on FET threshold in integrated electronics,” *IEEE J. Solid-State Circuits*, vol. 8, p. 245, 1975. 43

[144] K. Takeuchi, T. Tatsumi, and A. Furukawa, “Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuations,” *International Electron Devices Metting, Technical Digest*, p. 841, 1996. 43

[145] Y. Taur, D. Buchana, W. Chen, D. Frank, K. Ismail, S. Lo, R. Sai-Hakasz, H. Viswanathan, S. Wind, and H. Wong, “CMOS scaling into the nanometer regime,” *Proc. IEEE*, vol. 85, p. 486, 1997. 43

[146] C. Millar, Reid.D, G. Roy, S. Roy, and A. Asenov, “Accurate statistical description of random dopant-induced threshold voltage variability,” *IEEE Electron Device Letters*, vol. 29, p. 946, 2008. 44

[147] W. Shockley, *Electrons and holes in semiconductors.* D.Van Nostrand Company,Inc, 1966. 44

[148] S. Selberherr, *Analysis and simulation of semiconductor devices.* Springer-Verlag Wien NY, 1984. 44, 51, 52

[149] C. Snowden, *Semiconductor Device Modelling*, 1998. 44, 52

[150] D. Scharfetter and H. Gummel, “Large-signal analysis of a silicon read diode oscillator,” *IEEE Transactions on Electron Devices*, vol. 16, p. 64, 1969. 44, 52

[151] *Synopsis Inc.,* Taurus User manual Vol. 1*, Synopsis 2001.* 44, 45

[152] A. Asenov, A. R. Brown, J. H. Davies, and S. Saini, “Hierarchical approach to ‘atomistic’ 3D MOSFET simulation,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, pp. 1558 – 1565, 1999. 44

[153] D. Caughey and R. Tomas, "Carrier mobilities in silicon empirically related to doping and field," *Proc. IEEE*, p. 2192, 1967. 44

[154] D. Ferry and J.-R. Zhou, "Form of the quantum potential for use in hydrodynamic equations for semiconductor device modelling," *Physical Review B*, vol. 48, p. 7944, 1993. 45

[155] S. Chou, D. Antoniadis, and H. Smith, "Observation of electron velocity overshoot in sub-100-nm-channel MOSFETs in silicon," *IEEE Electron Device Letters*, vol. 6, p. 665, 1985. 45

[156] M. Lundstrom, "Elementary scattering theory of the Si MOSFET," *IEEE Electron Device Letters*, vol. 18, p. 361, 1997. 45

[157] F. *et. al.*. Assad, "Performance limits of silicon MOSFETs," *International Electron Devices Metting, Technical Digest*, p. 547, 1999. 45

[158] A. Lochtefeld and D. Antoniadis, "On experimental determination of carrier velocity in deeply scaled NMOS: how close to the thermal limit," *IEEE Electron Device Letters*, vol. 22, p. 95, 2001. 45

[159] G. Baccarani, E. Gnani, A. Gnudi, S. Reggiani, and M. Rudan, "Theoretical foundations of the quantum drift-diffusion and density-gradient models," *Solid State Electronics*, vol. 52, p. 526, 2008. 45

[160] C. Alexander, "Comparison of density gradient, monte carlo, and NEGF simulations of a 22 nm gate length mosfet." 45

[161] Y. Ashizawa, "Thin body effects to suppress random dopant fluctuations in nano-scaled MOSFETs," *Proc. Intl. Conf. On Simulation of Semiconductor Processes and Devices (SISPAD)*, p. 93, 2007. 45

[162] A. Brown, G. Roy, and A. Asenov, "Poly-Si-gate-related variability in decananometer MOSFETs with conventional architecture," *IEEE Transactions on Electron Devices*. 45

[163] C. Jacoboni and P. Lugli, *The monte carlo method for semiconductor device simulation*, ser. Computational Microelectronics, S. Selberherr, Ed. Wien-New York: Springer-Verlag, 1989. 45

[164] H. Kosina, M. Nedjalkov, and S. Selberherr, "Theory of the Monte Carlo method for semiconductor device simulation," *IEEE Transactions on Electron Devices*, vol. 47, p. 1898, 2000. 45

[165] N. Ashcroft and N. Mermin, *Solid State Physics*. Brooks/Cole, 1976. 46

[166] J. Zimmermann, P. Lugli, and D. Ferry, "On the physics and modeling of small semiconductor devices-IV. generalized, retarded transport in ensemble Monte Carlo techniques," *Solid State Electronics*, vol. 26, p. 233, 1983. 46

[167] K. Tomizawa, *Numerical simulation of submicron semiconductor devices*. Arthech House, 1993. 46

[168] C. Alexander, G. Roy, and A. Asenov, "Random impurity scattering induced variability in conventional nano-scaled MOSFETs: ab initio impurity scattering Monte Carlo simulation study," *International Electron Devices Metting, Technical Digest*, pp. 1–4, 2006. 46

[169] J. Tang and K. Hess, "Theory of hot electron emission from silicon into silicon dioxide," *Journal of Applied Physics*, vol. 54, p. 5145, 1983. 47

[170] C. Huang, T. Wang, C. Chen, M. Chang, and J. Fu, "Modeling hot-electron gate current in Si MOSFETs using a coupled Drift-Diffusion and Monte Carlo method," *IEEE Transactions on Electron Devices*, vol. 39, p. 2562, 1992. 47

[171] A. Duncan, U. Ravaioli, and J. Jakumeit, "Full-band Monte Carlo investigation of hot carrier trends in the scaling of MOSFETs," *IEEE Transactions on Electron Devices*, vol. 45, p. 867, 1998. 47

[172] M. Temple, D. Dyke, and P. Childs, "Hot-electron injection in stacked-gate MOSFET," *Journal of Applied Physics*, vol. 97, p. 104501, 2005. 47

[173] S. Datta, "Nanoscale device modelling: the Green's function method," *Superlattices and Microsturctures*, vol. 28, p. 253, 2000. 47, 48

[174] ——, *Quantum transport: atom to transistor*. Cambridge University Press, NY, 2005. 47

[175] A. Svizhenko, M. Anantram, T. Govindan, and B. Biegel, "Two-dimensional quantum mechanical modeling of nanotransistors," *Journal of Applied Physics*, vol. 91, p. 2343, 2002. 47, 48, 60, 61, 78, 117

[176] S. Datta, "A simple kinetic equation for steady-state quantum transport," *Journal of Physics: Condensed Matter*, vol. 2, p. 8023, 2990. 47

[177] R. Lake and S. Datta, "Nonequilibrium Green's-function method applied to double-barrier resonant-tunneling diodes," *Physical Review B*, vol. 45, p. 6670, 1992. 47

[178] A. Martinez, M. Bescond, J. Barker, A. Svizhenko, M. Anantram, C. Millar, and A. Asenov, "A self-consistent full 3D real-space NEGF simulator for studying nonperturbative effects in nano-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 53, p. 2213, 2007. 48

[179] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the $sp^4d^5s^*$ tight-binding formalism: from boundary conditions to strain calculations," *Physical Review B*, vol. 74, p. 205323, 2006. 48

[180] B. Cheng, A. Martinez, and A. Asenov, "Gate leakage variability due to oxide roughness studied with 2D-NEGF, journal = Proc. Intl. Conf. on Semiconductor Systems, Devices and Materials (SSDM), year = 2005, volume = , pages = , otherinfo = ." 48

[181] V. Do and P. Dolfus, "Oscillation of gate leakage current in double-gate metal-oxide-semiconductor field-effect transistors," *Journal of Applied Physics*, vol. 101, p. 073709, 2007. 48

[182] A. Asenov, G. Roy, C. Alexander, A. Brown, J. Watling, and S. Roy, "Quantum mechanical and transport aspects of resolving discrete charges in nano-CMOS device simulation," *2004 4th IEEE Conference on Nanotechnology*, pp. 334 – 336, 2004. 49

[183] G. Roy, "Simulation of intrinsic parameter fluctuations in nano-CMOS devices," Ph.D. dissertation, University of Glasgow, EEE Dept., 2005. 49, 53

[184] M. Ancona, D. Yergeau, Z. Yu, and B. Biegel, "On Ohmic boundary conditions for density-gradient theory," *Journal of Computational Electronics*, vol. 1, p. 103, 2002. 51

[185] D. Frank, Y. Taur, M. Ieong, and H.-S. Wong, "Monte Carlo modelling of threshold variation due to dopant fluctuations," *Symposium on VLSI Technology, Digest of Technical Papers*, p. 169, 1999. 53

[186] D. Muller, T. Sorch, S. Moccio, F. Baumann, K. Evans-Lutterodt, and G. Timp, "The electronic structure at the atomic scale of ultrathin gate oxides," *Nature*, vol. 399, p. 758, 1999. 54, 80, 84, 86, 87, 88, 98

[187] M. Niwa, T. Kouzaki, K. Okada, M. Udagawa, and R. Sinclair, "Atomic-order planarization of ultrathin SiO2-Si(001) interfaces," *Japanese Journal of Applied Physics*, vol. 33, p. 388, 1995. 54, 55

[188] W. Anderson, D. Lombardi, R. Wheeler, and T. Ma, "Determination of Si-SiO2 interface roughness using weak localisation," *IEEE Electron Device Letters*, vol. 14, p. 351, 1993. 54, 55

[189] Y. Chen, R. Myricks, M. Decker, J. Liu, and G. Higashi, "The origination and optimization of Si-SiO2 interface roughness and its effects on CMOS perfomrance," *IEEE Electron Device Letters*, vol. 24, p. 295, 2003. 54

[190] A. Pirovano, A. Lacaita, G. Ghidini, and G. Tallarida, "On the correlation between surface roughness and inversion layer mobility in Si-MOSFETs," *IEEE Electron Device Letters*, vol. 21, p. 34, 2000. 54, 55

[191] S. Yamakawa, H. Ueno, K. Taniguchi, C. Hamaguchi, K. Miyatsuji, K. Masaki, and U. Ravaioli, "Study of interface roughness dependence of electron mobility in Si inversion layers using the Monte Carlo method," *Journal of Applied Physics*, vol. 79, p. 911, 1996. 55

[192] K. Yang, H. Huang, M. Chen, Y. Lin, M. Yu, Jang.S.M, D. Yu, and M. Liang, "Characterization and modelling of edge direct tunnelling leakage in ultrathin gate oxide MOSFETs," *IEEE Transactions on Electron Devices*, vol. 48, p. 1159, 2001. 60, 61, 78, 117

[193] S.-i. Takagi, M. Tkayanagi, and A. Toriumi, "Experimental examination of physical model for direct tunnelling current in unstressed/stressed ultrathin gate oxides," *International Electron Devices Metting, Technical Digest*, p. 461, 1999. 61, 78, 117

[194] M. Luisier and A. Schenk, “Two-dimensional tunnelling effects on the leakage current of MOSFETs with single dielectric and high-k gate stacks,” *IEEE Transactions on Electron Devices*, vol. 55, p. 1494, 2008. 62

[195] H. Watanabe, K. Matsuzawa, and S.-i. Takagi, “Scaling effects on gate leakage current,” *IEEE Transactions on Electron Devices*, vol. 50, p. 1779, 2003. 62, 78, 117

[196] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, “Role of long-range and short-range Coulomb potentials in threshold characteristics under discrete dopants in sub-0.1 μm Si-MOSFETs,” *International Electron Devices Metting, Technical Digest*, p. 275, 2000. 64

[197] Y. Momiyama, K. Okabe, H. Nakao, M. Kase, M. Kojima, and T. Sugii, “Lateral extension engineering using nitrogen implantation (N-tub) for high-performance 40-nm pMOSFETs,” *International Electron Devices Metting, Technical Digest*, p. 247, 2002. 68

[198] H. Fukutome, T. Saiki, M. Hori, T. Tanaka, R. Nakamura, and H. Arimoto, “Characterization of plasma nitridation impact on lateral extension profile in 50 nm N-MOSFET by scanning tunneling microscopy,” *Japanese Journal of Applied Physics*, vol. 43, p. 1729, 2004. 68

[199] T. Hattori, K. Takahashi, M. B. Seman, H. Nohira, K. Hirose, N. Kamakura, Y. Takata, S. Shin, and K. Kobayashi, “Chemical and electronic structure of SiO_2/Si interfacial transition layer,” *Applied Surface Science*, vol. 212-213, pp. 547–555, 2003. 80, 84, 87

[200] F. J. Grunthaner and P. Grunthaner, “Chemical and electronic structure of the SiO2/Si interface,” *Material Science Report*, vol. 1, p. 65, 1986. 80, 83, 84, 85, 87

[201] K. Hhirose, H. Nohira, K. Azuma, and T. Hattori, “Photoelectron spectroscopy studies of SiO_2/Si interfaces,” *Progress in Surface Science*, vol. 82, pp. 3–54, 2007. 80, 83, 84, 85, 86, 88, 98

[202] F. Himpsel, F. McFeely, A. Taleb-Ibrahimi, J. Yarnoff, and G. Hollinger, “Microscopic structure of the SiO2/Si interface,” *Physical Review B*, vol. 38, p. 6084, 1988. 80, 84

[203] T. Yamasaki, C. Kaneta, T. Uchiyama, T. Uda, and K. Terakura, “Geometric and electronic structures of SiO_2/Si(001) interfaces,” *Physical Review B*, vol. 63, p. 115314, 2001. 80, 84, 87, 104, 128, 129

[204] K. Takahashi, M. B. Seman, K. Hirose, and T. Hattori, “Penetration of electronic states from silicon substrate into silicon oxide,” *Applied Surface Science*, vol. 190, pp. 56–59, 2002. 80, 85, 87

[205] F. Giustino and A. Pasquarello, “Electronic and dielectric properties of a suboxide interlayer at the silicon-oxide interface in MOS devices,” *Surface Science*, vol. 586, pp. 183–191, 2005. 80, 83, 87, 88, 104, 129

[206] D. Fischer, A. Curioni, S. Billeter, and W. Andreoni, “The structure of the SiO2/Si(100) interface from a restraint-free search using computer simulations,” *Applied Physics Letters*, vol. 88, p. 012101, 2006. 80, 83, 84, 85, 87

[207] K. Xue, H. Ho, and J. Xu, “Local study of thickness-dependent electronic properties of ultrathin silicon oxide near SiO_2/Si interface,” *Journal of Physics D: Applied Physics*, vol. 49, pp. 2886–2893, 2007. 80, 87

[208] F. Stern, “Effect of a thin transition layer at a Si-SiO2 interface on electron mobility and energy levels,” *Solid State Communications*, vol. 21, pp. 163 – 6, 1977. 81, 86, 92, 115

[209] J. Maserjian and N. Zamani, “Behaviour of the Si/SiO2 interfrace observed by Fowler-Nordheim tunnelling,” *Journal of Applied Physics*, vol. 53, p. 559, 1982. 81, 86

[210] H. Yang, H. Niimi, J. Keister, G. Lucovsky, and J. Rowe, “The effects of interfacial sub-oxide transition regions and monolayer level nitridation on tunnelling currents in silicon devices,” *IEEE Electron Device Letters*, vol. 21, p. 76, 2000. 81, 97, 107, 115

[211] J. d. de Sousa, P. Leite, E. de Oliveira, and G. Farias, “Consequences of nonstochiometric SiOx interfacial layers on the electrical characteriszation of metal-oxide-semiconductor devices,” *Journal of Applied Physics*, vol. 101, p. 034509, 2007. 81, 82

[212] F. Rochet, S. Rigo, M. Froment, C. d’Anterroches, C. Maillot, H. Roulet, and G. Durour, “The thermal oxidation of silicon in the special case of the growth of very thin films,” *Advances in Physics*, vol. 35, pp. 237–274, 1986. 82, 83

[213] R. Buczko, S. J. Pennycook, and S. T. Pantelides, “Bonding arrangements at the Si-SiO_2 and SiC-SiO_2 interfaces and a possible origin of their contrasting properties,” *Physical Review Letters*, vol. 84, p. 943, 2000. 83, 84

[214] H. Watanabe, K. Kato, T. Uda, K. Fujita, M. Ichikawa, T. Kawamura, and K. Terakura, "Kinetics of initial layer-by-layer oxidation of Si(001) surfaces," *Physical Review Letters*, vol. 80, p. 345, 1998. 83

[215] A. Pasquarello, M. S. Hytertsen, and R. Car, "Interface structure between silicon and its oxide by first principles molecular dynamics," *Nature*, vol. 396, p. 59, 1998. 83

[216] A. A. Demkov and O. F. Sankey, "Growth study and theoretical investigation of the ultrathin oxide SiO_2-Si heterojunction," *Physical Review Letters*, vol. 83, p. 2038, 1999. 83, 84

[217] L. Tsetseris and S. T. Pantelides, "Oxygen migration, agglomeration and trapping: key factros for the morphology of the Si-SiO_2 interface," *Physical Review Letters*, vol. 97, p. 116101, 2006. 83

[218] J.-N. Yeo, G. Jee, B. Yu, H. Kim, C.-H. Chung, H. Yeom, I.-W. Lyo, K. Kong, Y. Miyamoto, O. Sugino, and T. Ohno, "Ab initio study of the oxidation on vicinal Si(001) surfaces: the step-selective oxidation," *Physical Review B*, vol. 76, p. 115317, 2007. 83

[219] T. Yamasaki, K. Kato, and T. Uda, "Oxidation of the Si(001) surface: Lateral growth and formation of P_{b0} centers," *Physical Review Letters*, vol. 91, pp. 146 102–1, 2003. 83

[220] S. Ogawa, A. Yoshigoe, S. Ishidzuka, Y. Teraoka, and Y. Takakuwa, "Si(001) surface layer-by-layer oxidation studied by real-time photoelectron spectroscopy using synchrotron radiation," *Japanese Journal of Applied Physics*, vol. 46, pp. 3244–3254, 2007. 83, 84, 85

[221] A. Bongiorno, A. Pasquarello, M. S. Hybertsen, and L. C. Feldman, "Transition structure at the Si(100)-SiO_2 interface," *Physical Review Letters*, vol. 90, p. 186101, 2003. 83

[222] N. Ikarashi, K. Watanabe, and Y. Miyamoto, "High-resolution transmission electron microscopy of an atomic structure at a Si(001) oxidation front," *Physical Review B*, vol. 62, p. 15989, 2000. 83

[223] P. Donnadieu, E. Blanquet, N. Jakse, and P. Mur, "Detection of subnanometric layer at the Si/SiO_2 interface and related strain measurements," *Applied Physics Letters*, vol. 85, p. 5574, 2001. 83

[224] J. Oh, H. Yeom, Y. Hagimoto, K. Ono, M. Oshima, N. Hirashita, M. Nywa, A. Toriumi, and A. Kakizaki, "Chemical structure of the ultrathin SiO2/Si(100) interface: an angle-resolved Si 2p photoemission study," *Physical Review B*, vol. 63, p. 205310, 2001. 84

[225] K. Kimura and K. Nakajima, "Compositional transition layer in SiO2/Si interface observed by high-resolution rbs," *Applied Surface Science*, vol. 216, pp. 282–286, 2003. 84

[226] A. Pasquarello, M. S. Hytertsen, and R. Car, "Structurally relaxed models of the Si(001)-SiO$_2$ interface," *Applied Physics Letters*, vol. 68, p. 625, 1996. 84

[227] F. Giustino, A. Bongiorno, and A. Pasquarello, "Atomistic models of the Si(100)-SiO$_2$ interface: structural,electronic and dielectric properties," *Journal of Physics: Condensed Matter*, vol. 17, pp. S2065–S2074, 2005. 84, 85

[228] S. T. Pantelides, "Some properties of the oxides of the tetrahedral semiconductors and the oxide-semiconductor interfaces," *Journal of Vacuum Science and Technology*, vol. 14, p. 965, 1977. 85

[229] A. Bongiorno, A. Pasquarello, M. S. Hybertsen, and L. F. Feldman, "Ion scattering simulations of the Si(100)-SiO$_2$ interface," *Physical Review B*, vol. 74, p. 075316, 2006. 85

[230] F. Giustino and A. Pasquarello, "Theory of atomic-scale dielectric permittivity at insulator interfaces," *Physical Review B*, vol. 71, p. 144104, 2005. 85, 88, 98

[231] R. Haight and L. C. Feldman, "Atomic structure at the (111)Si-SiO$_2$ interface," *Journal of Applied Physics*, vol. 53, p. 4884, 1982. 85

[232] R. Williams, "Photoemission of electrons from silicon into silicon dioxide," *Physical Review A*, vol. 140, p. A569, 1965. 85, 86

[233] F. Stern, "Self-consistent results for n-type Si inversion layers," *Physical Review B (Solid State)*, vol. 5, pp. 4891 – 9, 1972. 86, 89, 123, 124, 125

[234] J. Neaton, D. Muller, and N. Ashcroft, "Electronic properties of the Si/SiO2 interface from first principles," *Physical Review Letters*, vol. 85, p. 1298, 2000. 87, 88

[235] H. Ikeda, N. Kurumado, K. Ohmori, M. Sakashita, A. Sakai, S. Zaima, and Y. Yasuda, "Local electrical characteristics of ultra-thin SiO2 films fromed on Si(001) surfaces," *Surface Science*, vol. 493, pp. 653–658, 2001. 87

[236] C. Kaneta, T. Yamasaki, T. Uchiyama, T. Uda, and K. Terakura, "Structure and electronic preperty of Si(100)/SiO2 interface," *Microelectronic Engineering*, vol. 48, pp. 17–120, 1999. 87

[237] M. Watarai, J. Nakamura, and A. Natori, "Band discontinuity at ultrathin SiO_2/Si(001) interfaces," *Physical Review B*, vol. 69, p. 035312, 2004. 87

[238] X.-Y. Liu, D. Jovanovic, and R. Stumpf, "First-principles study of Si-SiO_2 interface and the impact on mobility," *Applied Physics Letters*, vol. 86, p. 082104, 2005. 87, 112

[239] S. Markov, P. Sushko, S. Roy, C. Fiegna, E. Sangiorgi, A. Shluger, and A. Asenov, "Si-SiO_2 interface band-gap transition – effects on MOS inversion layer, journal = Physica Status Solidi: A, year = 2008, volume = 205, pages = 1290, otherinfo = ." 87, 106, 127

[240] L. Hedin, "New method for calculating the one-particle Green's function with application to the electron-gas problem," *Physical Review*, vol. 139, p. A796, 1965. 87

[241] M. Shishkin and G. Kresse, "Self-consistent GW calculations for semiconductors and insulators," *Physical Review B*, vol. 75, p. 235102, 2007. 87

[242] R. Shaltaf, G.-M. Gignanese, X. Gonze, F. Giustino, and A. Pasquarello, "Band offsets at the Si/SiO_2 interface from many-body perturbation theory," *Physical Review Letters*, vol. 100, p. 186401, 2008. 87

[243] H. Cho, Y. Lee, I. Lee, D. Moon, H. Lee, B. Kim, H. Kim, S. Kim, and J. Cho, Y, "Comparison of the EOT determined by ellipsometry with the result by medium energy ion scatering spectroscopy and HRTEM," *Journal of Vacuum Science and Technology B*, vol. 19, p. 1144, 2001. 88

[244] H. Chang, H. Yang, H. Hwang, H. Cho, H. Lee, and D. Moon, "Measurement of the physical and electrical thickness of ultrathin gate oxides," *Journal of Vacuum Science and Technology B*, vol. 20, p. 1836, 2002. 88

[245] J. Ehrstein, C. Richter, D. Chandler-Horowitz, E. Vogel, C. Young, S. Shah, D. Maher, B. Foran, P. Hung, and A. Diebold, "A comparison of thickness values for very thin SiO_2 films by using ellipsometric, capacitance-voltage, and HRTEM measurements," *Journal of The Electrochemical Society*, vol. 153, p. F12, 2006. 88, 89

[246] F. Giustino, P. Umari, and A. Pasquarello, "Dielectric discontinuity at interfaces in the atomic-scale limit: permittivity of ultrathin oxide films on silicon," *Physical Review Letters*, vol. 91, p. 267601, 2003. 88, 98, 99

[247] D. Fischer, A. Curioni, S. Billeter, and W. Andreoni, "Effects of nitridation on the characteristics of silicon dioxide: dielectric and structural properties from *ab initio* calculations," *Physical Review Letters*, vol. 92, p. 236405, 2004. 88, 107

[248] P. Broqvist and A. Pasquarello, "Band gaps and dielectric constants of amorphous hafnium silicates: A first-principles investigation," *Applied Physics Letters*, vol. 90, p. 082907, 2007. 88, 107, 108

[249] F. Stern, "Iteration methods for calculating self-consistent fields in semiconductor inversion layers," *Journal of Computational Physics*, vol. 6, pp. 56 – 67, 1970. 89, 122, 126

[250] B. Sell, D. Schumann, and W. Krautschneider, "Fast interface characterization of tunnel oxide MOS structures," *IEEE Transactions on Nanotechnology*, vol. 1, p. 110, 2002. 89

[251] A. A. Demkov, S. Zhang, and D. A. Drabold, "Towards a first-principles simulation and current-voltage characteristic of atomistic metal-oxide-semiconductor structures," *Physical Review B*, vol. 64, p. 125306, 2001. 89

[252] F. Sacconi, A. Di Carlo, P. Lugli, M. Stadele, and J.-M. Jancu, "Full band approach to tunneling in MOS structures," *IEEE Transactions on Electron Devices*, vol. 51, p. 741, 2004. 89

[253] D. Vasileska and Z. Ren, "SCHRED-2.0 user manual," 2001. [Online]. Available: http://nanohub.org/resources/221 89, 122, 123

[254] W. Kohn and L. Sham, "Self-consistent equations including exchange and correlation effects," *Physical Review*, vol. 140, p. A1133, 1965. 90

[255] D. Vasileska, D. Schroder, and D. Ferry, “Scaled silicon MOSFETs: Degradation of the total gate capacitance,” *IEEE Transactions on Electron Devices*, vol. 44, p. 584, 1997. 90

[256] S. Mudanai, L. Register, A. Tasch, and S. Banerjee, “Understanding the effects of wave function penetration on the inversion layer capacitance of NMOSFETs,” *IEEE Electron Device Letters*, vol. 22, p. 145, 2001. 94

[257] A. Haque and M. Kauser, “A comparison of wave-function penetration effects on gate capacitance in deep submicron n- and p-MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 49, p. 1580, 2002. 94

[258] S.-i. Takagi, J. Koga, and A. Toriumi, “Subband structure engineering for performance enhancement of Si MOSFETs,” *International Electron Devices Metting, Technical Digest*, p. 219, 1997. 96, 113

[259] “International technology roadmap for semiconductors (ITRS) 2005, www.itrs.net.” 106

[260] M. Ieong, B. Doris, J. Kedzierski, K. Rim, and Yang.M, “Silicon device scaling to the sub-10-nm regime,” *Science*, vol. 306, p. 2057, 2004. 106, 112

[261] X. Guo and T. Ma, “Tunneling leakage current in oxynitride: dependence on oxygen/nitrogen content,” *IEEE Electron Device Letters*, vol. 19, p. 207, 1998. 107

[262] Y.-C. Yeo, T.-J. King, and C. Hu, “MOSFET gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations,” *IEEE Transactions on Electron Devices*, vol. 50, p. 1027, 2003. 107

[263] P. Kraus, K. Ahmed, C. Olsen, and F. Nouri, “Model to predict gate tunneling current of plasma oxynitrides,” *IEEE Transactions on Electron Devices*, vol. 52, p. 1141, 2005. 107

[264] M. Green, E. Gusev, R. Degraeve, and E. Garfunkel, “Ultrathin SiO_2 and Si-O-N gate dielectric layers for silicon microelectronics,” *Journal of Applied Physics*, vol. 90, p. 2067, 2001. 107

[265] A. Ogawa, K. Iwamoto, H. Ota, Y. Morita, M. Ikeda, T. Nabatame, and A. Toriumi, “0.6 nm-EOT high-k gate stacks with $HfSiO_x$ interfacial layer grown by solid-phase reaction between HfO_2 and Si substrate,” *Microelectronic Engineering*, vol. 84, p. 1861, 2007. 107, 108

[266] M. P. Agustin, G. Bersuker, B. Foran, L. Boatner, and S. Stemmer, "Scanning TEM investigation of interfacial layers in HfO_2 gate stacks," *Journal of Applied Physics*, vol. 100, p. 024103, 2006. 107

[267] J. Gavartin and A. Shluger, "Modelling of $HfO_2/SiO_2/Si$ interface," *Microelectronic Engineering*, vol. 84, p. 2412, 2007. 107, 108

[268] M. Hakala, A. Foster, J. Gavartin, P. Havu, M. Puska, and R. Nieminen, "Interfacial oxide growth at silicon/high-k oxide interfaces:first principles modelling of the Si-HfO_2 interface," *Journal of Applied Physics*, vol. 100, p. 043708, 2006. 107

[269] O. Sharia, A. Demkov, G. Bersuker, and B. Lee, "Theoretical study of the insulator/insulator interface: band alignment at the SiO_2/HfO_2 junction," *Physical Review B*, vol. 75, p. 035306, 2007. 108

[270] H. Jin, S. Oh, and C. M.-H. Kang, H.J, "Band gap and band offsets for ultrathin $(HfO_2)_x(SiO_2)_{1-x}$," *Applied Physics Letters*, vol. 89, p. 122901, 2006. 108

[271] H. Wu, Y. Zhao, and M. White, "Quantum mechanical modelling of MOSFET gate leakage for high-k gate dielectrics," *Solid State Electronics*, vol. 50, p. 1164, 2006. 108

[272] Y. Hou, M. Li, H. Yu, and D.-L. Kwong, "Modeling of tunneling currents through HfO_2 and $(HfO_2)_x(Al_2O_3)_{1-x}$ gate stacks," *IEEE Electron Device Letters*, vol. 24, p. 96, 2003. 108

[273] C. Hinkle, C. Fulton, R. Nemanich, and G. Lucovsky, "A novel approach for determining the effective tunneling mass of electrons in HfO_2 and other high-k alternative gate dielectrics for advanced CMOS devices," *Microelectronic Engineering*, vol. 72, p. 257, 2004. 108

[274] W. Zhu, T. Ma, T. Tamagawa, J. Kim, and Y. Di, "Current transport in metal/hafnium oxide/silicon structure," *IEEE Electron Device Letters*, vol. 23, p. 97, 2002. 108

[275] A. Campera, G. Iannaccone, and F. Crupi, "Modeling of tunnelling currents in Hf-based gate stacks as a function of temperature and extraction of material parameters," *IEEE Transactions on Electron Devices*, vol. 54, p. 83, 2007. 108

[276] F.-C. Chiu, "Interface characterization and carrier transportation in metal/HfO_2/silicon structure," *Journal of Applied Physics*, vol. 100, p. 114102, 2006. 108

[277] J. Welser, J. Hoyt, and J. Gibbons, "Electron mobility enhancement in strained-Si n-type MOSFETs," *IEEE Electron Device Letters*, vol. 15, p. 100, 1994. 112

[278] D. Antoniadis, I. Aberg, C. Chleirigh, O. Nayfeh, A. Khakifirooz, and J. Hoyt, "Continuous MOSFET performence increase with device scaling - the role of strain and channel material innovations," *IBM Journal Research & Development*, vol. 50, p. 363, 2006. 112

[279] M. Rieger and P. Vogl, "Electronic-band parameters in strained SiGe substrates," *Physical Review B*, vol. 48, p. 14276, 1993. 112

[280] C. Maiti, L. Bera, and S. Chattopadhyay, "Strained-Si heterostructure field effect transistors," *Semiconductor Science and Technology*, vol. 13, pp. 1225–1246, 1998. 112

[281] D. Rideau, M. Feraille, L. Ciampolini, M. Minondo, C. Tavernier, and H. Jaouen, "Strained Si, Ge and $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ alloys modeled with a first-principles-optimized full-zone $k\dot{p}$ method," *Physical Review B*, vol. 74, p. 195208, 2006. 112

[282] N. Barin, M. Braccioli, C. Fiegna, and E. Sangiorgi, "Scaling the high-performance DG SOI MOSFET down to the 32 nm technology node with SiO_2-based gate stacks," *International Electron Devices Metting, Technical Digest*, p. 609, 2005. 113

[283] S. Markov, A. Schenk, B. Majkusiak, C. Compagnoni, and C. Fiegna, "Comparison of gate current in high-k stacks including calculations accounting for gradual transition regions at the interfaces," EC-IST Technical Report, PULLNANO-IST-026828, Tech. Rep., 2008. 113

[284] V. Saunders, R. Dovesi, C. Roetti, R. Orlando, C. Zichovich-Wilson, N. Harrison, K. Doll, B. Cavalleri, I. Bush, P. D'arco, and M. Llunell, CRYSTAL-*2003 User manual Vol. 1*, University of Torino, Torino, 2003. 127