Alma Mater Studiorum Università di Bologna

Corso di Dottorato in

Tecnologie dell'Informazione

Tesi in cotutela internazionale

Titolo:

Study of silicon-on-insulator multiple-gate MOS structures including band-gap engineering and self-heating effects

Scritta da

Marco BRACCIOLI

Relatore

Prof. Claudio FIEGNA

Presentata e sostenuta pubblicamente il

9 Aprile 2009

Coordinatore

Prof. Claudio FIEGNA

Settore Scientifico-Disciplinare: ING-INF/01 Elettronica

Anni Accademici: 2005/06 - 2006/07 - 2007/08

To my family

Contents

List of Figures

List of Tables

List of Symbols

Constant

$\hbar$	Reduced Planck's constant, 1.05458×10^{-34} Js
m_0	Free electron mass, 9.1095×10^{-31} kg
q	Fundamental electron charge, 1.60218×10^{-19} C
k_B	Boltzmann's constant, 8.617×10^{-5} J/K
ε_{OX}	Oxide dielectric permittivity, 3.45×10^{-13} F/cm
ε_{Si}	Silicon dielectric permittivity, 1.04×10^{-12} F/cm

Symbols

α	Scaling parameter
α_d, α_w	Selective scaling parameters
Γ	Scattering rate, [s^{-1}]
δ	Coefficient of SOI scaling function L_G–t_{Si}
Δ_{fin}	Fin–pitch (i.e. inter–fins distance), [nm]
ΔE	Value of the conduction band offset, [eV]
ΔT_{MAX}	Peak temperature with respect to the room temperature, [K]
λ	Mean free path, [nm]
Λ_S	Phonon mean free path, [nm]
μ	Carrier mobility, [$cm^2V^{-1}s^{-1}$]
ϕ	Electrostatic potential, [V]

ϕ	Work–function, [eV]
ϕ_F	Fermi potential, [V]
ϕ_G	Gate work–function, [eV]
χ	Electron affinity, [eV]
ρ	Charge density, [C/cm^3]
τ	Relaxation time, [s]
c	Lattice heat capacity [JK^{-1}cm^{-3}]
C_{eff}	Effective oxide capacitance per unit area, [F/cm^2]
C_{OX}	Oxide capacitance per unit area, [F/cm^2]
D_n and D_p	Electron and hole diffusion coefficients, [m^{2}s^{-1}]
$\vec{\mathcal{E}}$	Electric field, [V/cm]
E	Carrier kinetic energy, [eV]
E_ν	Carrier kinetic energy of band ν, [eV]
E_C and E_V	Conduction and valence band edge, [eV]
E_F	Fermi energy level, [eV]
E_G	Energy band–gap, [eV]
E_0	Vacuum level, [eV]
$\mathcal{F}$	Distribution function
$\mathcal{F}_{eq}$	Distribution function at equilibrium
g_m	Transconductance $\partial I_{DS}/\partial V_{GS}$, [S]
g_{out}	Output conductance $\partial I_{DS}/\partial V_{DS}$, [S]
G	Electron–hole pair generation rate, [cm^{-3}s^{-1}]
H_{fin}	FinFET's fin height, [nm]
I_{BL}	Ballistic current, [A/μm]

I_{DS}	Drain current per unit width, [A/μm]
I_{OFF}	Sub–threshold current per unit width, [nA/μm]
I_{ON}	MOSFET ON–current (I_{DS} for V_{GS}=V_{DS}=V_{DD}) per unit width, [A/μm]
$\vec{J}$	Current density, [A/cm^2]
$\vec{J}_n$ and $\vec{J}_p$	Electron and hole current density, [A/cm^2]
$\vec{k}$	Wave vector, [m^{-1}]
L_G	MOSFET gate length, [nm]
L_{epi}	Thickness of the epitaxially grown S/D region, [nm]
L_{ext}	Source/Drain Extension Length, [nm]
L_{kT}	Length of the kT–layer, [nm]
L_{spacer}	Distance between gate edge and epitaxially grown S/D regions, [nm]
m^*	Effective mass, units of m_0
n and p	Electron and hole concentration, [cm^{-3}]
N_{CH}	Channel Doping Concentration, [cm^{-3}]
N_D and N_A	Donor and acceptor type doping concentration, [cm^{-3}]
N_{inv}	Inversion charge density, [cm^{-2}]
$N_{S/D}$	Source/Drain Doping Concentration, [cm^{-3}]
$\vec{p}$	Carrier's momentum, [kgm/s]
P_n and P_p	Absolute electron and hole thermoelectric power, [mV/K]
Q_d	Depletion charge, [cm^{-2}]
$\vec{r}$	Real space vector, [m]
r	Back–scattering coefficient
R	Electron–hole pair recombination rate, [cm^{-3}s^{-1}]
R_{TH}	Thermal resistance, [cm^{2}K/W]

$\mathcal{S}$	Collision event probability
t	Time variable, [s]
t_{BOX}	Buried Oxide Thickness, [nm]
t_{FF}	Free–flight, [s]
t_{OX}	Gate oxide thickness, [nm]
t_{OX}^{top}	top–gate oxide thickness for FinFET, [nm]
t_{PAS}	Passivation layer thickness, [nm]
t_{Si}	Silicon film thickness, [nm]
T	Lattice temperature, [K]
T_{MAX}	Maximum temperature inside the device, [K]
$\vec{v}$	Carrier velocity, [cm/s]
v_G	Carrier group velocity, [cm/s]
v_{avg}	Carrier average velocity, [cm/s]
v_{inj}	Carrier injection velocity, [cm/s]
v_{sat}	Inversion layer saturation velocity, [cm/s]
v_{th}	Thermal velocity, [cm/s]
V_t	Threshold voltage, [V]
V_{DD}	Supply Voltage, [V]
V_{DS}	Drain–to–Source voltage, [V]
V_{FB}	Flat band voltage, [V]
V_{GS}	Gate–to–Source voltage, [V]
V_t	Threshold voltage, [V]
V_{t0}	Threshold voltage at zero bias, [V]
W or W^{ch}	Device width, [nm]

W_{fin}	FinFET's fin width, [nm]
x	Horizontal axis, parallel to the Si–SiO_2
x_{HJ}	Position of E_C discontinuity, [nm]
x_{inj}	Position of the Virtual Source, [nm]
y	Vertical axis, normal to the Si–SiO_2

Acronyms

3DEG	Three–dimensional Electron Gas
BC	Boundary Condition
BJT	Bipolar Junction Transistor
BOX	Buried Oxide
BR	Ballistic Ratio
CBO	Conduction Band Offset
CVD	Chemical Vapor Deposition
DC	Direct current
DD	Drift–Diffusion
DIBL	Drain Induced Barrier Lowering, [mV/V]
EOT	Equivalent Oxide Thickness ,[nm]
ET	Electro–thermal
FBZ	First Brillouin Zone
FET	Field Effect Transistor
HJ	Heterojunction
IC	Integrated Cirtcuits
IT	Isothermal
ITRS	International Technology Roadmap for Semiconductors

MC	Monte Carlo
MOS	Metal–Oxide–Semiconductor
RHS	Right–Hand Side
RSD	Raised Source–Drain
S, D, G	Source, Drain, Gate
SCE	Short–Channel Effects
SHE	Self–Heating Effects
SOI	Silicon–On–Insulator
TN	Technological Node
UTB	Ultra–Thin–Body
VS	Virtual Source

Abstract

The progresses of electron devices integration have proceeded for more than 40 years following the well–known Moore's law, which states that the transistors density on chip doubles every 24 months [1]. This trend has been possible due to the downsizing of the MOSFET dimensions (*scaling*); however, new issues and new challenges are arising, and the conventional "bulk" architecture is becoming inadequate in order to face them. In order to overcome the limitations related to conventional structures, the researchers community is preparing different solutions, that need to be assessed.
Possible solutions currently under scrutiny are represented by:

- devices incorporating materials with properties different from those of silicon, for the channel and the source/drain regions;
- new architectures as Silicon–On–Insulator (SOI) transistors: the body thickness (t_{Si}) of Ultra-Thin-Body (UTB) SOI devices is a new design parameter, and it permits to keep under control Short–Channel–Effects (SCE) without adopting high doping level in the channel.

Among the solutions proposed in order to overcome the difficulties related to scaling, we can highlight heterojunctions at the channel edge, obtained by adopting for the source/drain regions materials with band–gap different from that of the channel material. This solution allows to increase the injection velocity of the particles travelling from the source into the channel, and therefore increase the performance of the transistor in terms of provided drain current.

The first part of this thesis work addresses the use of heterojunctions in SOI transistors: chapter 3 outlines the basics of the heterojunctions theory and the adoption of such approach in older technologies as the heterojunction–bipolar–transistors; moreover the modifications introduced in the Monte Carlo code in order to simulate conduction band discontinuities are described, and the simulations performed on unidimensional simplified structures in order to validate them as well.
Chapter 4 presents the results obtained from the Monte Carlo simulations performed on double–gate SOI transistors featuring conduction band offsets between the source and drain regions and the channel. In particular, attention has been focused on the drain

current and to internal quantities as inversion charge, potential energy and carrier velocities. Both graded and abrupt discontinuities have been considered.

The scaling of devices dimensions and the adoption of innovative architectures have consequences on the power dissipation as well. In SOI technologies the channel is thermally insulated from the underlying substrate by a SiO_2 buried–oxide layer; this SiO_2 layer features a thermal conductivity that is two orders of magnitude lower than the silicon one, and it impedes the dissipation of the heat generated in the active region. Moreover, the thermal conductivity of thin semiconductor films is much lower than that of silicon bulk, due to phonon confinement and boundary scattering. All these aspects cause severe self–heating effects, that detrimentally impact the carrier mobility and therefore the saturation drive current for high–performance transistors; as a consequence, thermal device design is becoming a fundamental part of integrated circuit engineering.
The second part of this thesis discusses the problem of self–heating in SOI transistors. Chapter 5 describes the causes of heat generation and dissipation in SOI devices, and it provides a brief overview on the methods that have been proposed in order to model these phenomena. In order to understand how this problem impacts the performance of different SOI architectures, three–dimensional electro–thermal simulations have been applied to the analysis of SHE in planar single and double–gate SOI transistors as well as FinFET, featuring the same isothermal electrical characteristics.
In chapter 6 the same simulation approach is extensively employed to study the impact of SHE on the performance of a FinFET representative of the high–performance transistor of the 45 nm technology node. Its effects on the ON–current, the maximum temperatures reached inside the device and the thermal resistance associated to the device itself, as well as the dependence of SHE on the main geometrical parameters have been analyzed. Furthermore, the consequences on self–heating of technological solutions such as raised S/D extensions regions or reduction of fin height are explored as well.

Finally, conclusions are drawn in chapter 7.

Prefazione

I progressi nell'integrazione dei dispositivi elettronici sono proseguiti per più di 40 anni seguendo la legge di Moore, la quale stabilisce che il numero di transistori integrati su chip di silicio raddoppia ogni due anni [1]. Queste tendenza è stata possibile grazie alla riduzione delle dimensioni del transistore MOS (*scaling*). Tuttavia, nuovi problemi e nuove sfide stanno emergendo, e l'architettura convenzionale di tipo "bulk" sta divenendo inadeguata ad affrontarle. Per superare i limiti legati a strutture convenzionali, la comunità dei ricercatori sta preparando diverse alternative, che necessitano di essere messe alla prova.
Possibili soluzioni attualmente allo studio sono rappresentate da:

- dispositivi che includono materiali con proprietà diverse da quelle del silicio per le regioni di source, drain e canale;
- strutture innovative quali, ad esempio, i transistori silicio–su–isolante (*Silicon–On–Insulator*, SOI); lo spessore dello strato di silicio utilizzato per realizzare il dispositivo è un nuovo parametro di progetto, che permette di tenere sotto controllo gli effetti di canale corto (*Short–Channel Effects*, SCE) evitando di drogare pesantemente la regione di canale.

Tra le diverse soluzioni proposte per superare le difficoltà relative allo scaling, è possibile evidenziare le eterogiunzioni realizzate agli estremi del canale, ottenute utilizzando per le regioni di source/drain materiali con un band–gap diverso da quello del materiale utilizzato per il canale. Questa soluzione permette di aumentare la velocità di iniezione delle particelle che si muovono dal source ed entrano nel canale, e di conseguenza di aumentare le prestazioni del transistore in termini di corrente erogata.

La prima parte di questo lavoro di tesi è relativa all'utilizzo di eterogiunzioni in transistori SOI: il capitolo 3 riassume le basi della teoria delle eterogiunzioni e l'utilizzo di una tale soluzione in tecnologie quali, ad esempio, i transistori bipolari ad eterogiunzione (*Heterojunction–Bipolar–Transistor*, HBT). Vengono inoltre descritte le modifiche che sono state introdotte nel codice Monte Carlo per poter simulare discontinuità della banda di conduzione, e le simulazioni effettuate su strutture semplificate unidimensionali per verificare tali modifiche.

Il capitolo 4 presenta i risultati ottenuti da simulazioni Monte Carlo effettuate su transistori SOI a doppio gate caratterizzati da discontinuità della banda di conduzione tra le regioni di source e drain ed il canale. In particolare, l'attenzione è stata focalizzata sulla corrente di drain e su quantità interne come la carica di inversione, l'energia potentiale e le velocità dei portatori. Sono state considerate sia discontinuità brusche che graduali.

La riduzione delle dimensioni dei dispositivi elettronici e l'utilizzo di architetture innovative hanno conseguenze anche sulla generazione di potenza. Nelle tecnologie SOI il canale è isolato termicamente dallo strato di silicio sottostante da uno strato di biossido di silicio; tale strato di ossido presenta una conducibilità termica che è due ordini di grandezza più bassa rispetto a quella del silicio, ed impedisce la dissipazione del calore generato nella regione attiva. Inoltre, la conducibilità termica di strati sottili di silicio è molto inferiore rispetto a quella del silicio di tipo "bulk", a causa del confinamento dei fononi e dei fenomeni di scattering al contorno. Tutti questi aspetti provocano severi effetti di autoriscaldamento, che degradano la mobilità dei portatori e quindi la corrente di saturazione per dispositivi ad elevate prestazioni; di conseguenza, il progetto di dispositivi che tenga conto anche degli effetti termici sta divenendo parte fondamentale dell'ingegneria dei circuiti integrati.
La seconda parte della tesi affronta il problema dell'autoriscaldamento nei transistori SOI. Il capitolo 5 descrive le cause della generazione e dissipazione di calore nei dispositivi SOI, e fornisce una breve panoramica sui metodi che sono stati proposti per modellare tali fenomeni. Per comprendere come questo problema impatta le performance di diverse srchitetture SOI, sono state effettuate simulazioni elettrotermiche in tre dimensioni per l'analisi del SHE in dispositivi planari a singolo e doppio gate, e dispositivi di tipo FinFET, caratterizzati dalle stesse caratteristiche elettriche in condizioni isoterme.
Nel capitolo 6 lo stesso approccio simulativo è largamente utilizzato per studiare l'impatto dell'autoriscaldamento sulle prestazioni di un transistore di tipo FinFET rappresentativo dei transistori per elevate prestazioni del nodo tecnologico a 45 nm. Vengono analizzati i suoi effetti sulla corrente erogata, sulla temperatura massima raggiunta all'interno del dispositivo e la resistenza termica associata al dispositivo stesso, nonchè la dipendenza del SHE dai principali parametri geometrici quali spessore dello strato di ossido, lunghezza delle regioni di accesso e distanza tra "fins" adiacenti. Sono inoltre indagate le conseguenze sull'autoriscaldamento di soluzioni tecnologiche quali estensioni di source e drain elevate o la riduzione dell'altezza del fin di silicio.

Infine, le conclusioni sono presentate nel capitolo 7.

Avant–propos

Le progrès de l'intégration des dispositifs électroniques à l'échelle nanométrique est avancé pour plus de 40 années en suivant la loi de Moore, qui spécifie que la densité des transistors double chaques 24 mois [1]. Cette tendance a été possible en raison de la miniaturisation des dimensions du transistor MOS (nommée *scaling*) ; cependant, nouveaux problèmes et défis surviennent, et le transistor conventionnel de type "bulk" devient insuffisant pour les affronter. Pour surmonter les limitations liées aux architectures conventionelles, la communauté des chercheurs est en train de préparer un grand nombre de différents alternatives, qui doivent être évaluées.
Solutions possibles qui sont actuellememt étudiées sont representées par :

- dispositifs qui includent materiaux avec propriétés différentes de celles du silicium pour réaliser les regions de source, de drain et de la grille ;
- nouvelles architectures comme les transistors silicium–sur–isolant (*Silicon–On–Insulator*, SOI) : l'épaisseur de la couche de silicium (t_{Si}) des transistors à couche trés mince est un nouveau paramètre pour la conception du transistor, qui permit de tenir sous control les effets de canaux courts sans utiliser un dopage trés elevé dans le canal.

Entre les solutions proposées pour surmonter les difficultés liées à la miniaturisation du transistor MOS, on peut souligner les hétérojonctions aux bords du canal, qui sont obtenues en utilisant pour les regions de source et de drain materiaux qui ont un interval de la bande interdite différent que celui qui forme le canal. Cette solution peut augmenter la vitesse d'injection des porteurs qui sortent de la source et ils entrent dans le canal, et par conséquent elle peut améliorer les performances du transistor.

La première partie de ces mémoires concerne l'utilisation des hétérojonctions pour les transistors silicium–sur–isolant. Le chapitre 3 expose les bases de la théorie des hétérojonctions et l'utilisation de cet approche en vieilles technologies comme celle des transistors bipolaires à hétérojunction (*Heterojunction–Bipola–Transistor*, HBT) ; en plus, on décrit les modifications qui ont été introduites dans le logiciel de simulation Monte Carlo pour simuler discontinuités de la bande de conduction et aussi les simulations qui ont été effectuées sur structures simplifiées pour valider ces modifications.
Le chapitre 4 présente les résultats obtenus avec simulations Monte Carlo qui ont été ef-

fectuées sur transistors silicium–sur–isolant double grille avec discontinuité de la bande de conduction entre les regions de source et de drain et le canal. En particulier, on a focalisé l'attention sur le courant de drain fourni par le transistor et aussi sur quantités internes comme la charge d'inversion, l'énergie potentielle et la vitesse des porteurs. On a consideré soit discontinuités abruptes soit discontinuités graduelles.

La miniaturization des dimensions des dispositifs et l'utilisation d'architectures innovatives ont des conséquences aussi sur la dissipation de la chaleur. Dans la technologie silicium–sur–isolant le canal est thermiquement isolé de la couche de silicium sous–jacente par une couche enterrée d'oxide (SiO_2) ; cette couche enterrée est caractérisée par une conductivité thermique qui est deux ordres de grandeur inférieure que celle du silicium, et elle bloque la dissipation de la chaleur generée dans la region active. De plus, la conductivité thermique des couches trés minces de material semiconducteur est beaucoup inférieure que celle du silicium de type "bulk", à cause du confinement des phonons et des interactions aux limites. Tous ces aspets provoquent sévères effets d' auto–échauffement (*Self–Heating Effects*, SHE), qui affectent nuisiblement la mobilité des porteurs et par conséquent le courant de saturation pour les transistors à hautes perfomances ; comme conséquence, le projet thermique du dispositif devient une partie fondamentale de l'ingénierie des circuits intégrés.
La deuxième partie de ces mémoires de thèse discute le problème de l'auto–échauffement dans les transistors fabriqués en technologie silicium–sur–isolant. Le chapitre 5 décrit les causes de la génération et de la dissipation de la chaleur dans les dispositifs SOI, et aussi elle fournit une brève vue d'ensemble sur les méthodes qui ont été proposées pour les simuler. Pour comprendre comment ce problème affecte les performances des différentes architectures SOI, on a effectué simulations électro–thermiques en trois dimensions pour étudier l'auto–échauffement dans différent architectures SOI (single– et double–grille SOI transistors et aussi FinFET) avec les mêmes caractéristiques isothermiques.
Dans le chapitre 6 la même approche de simulation est largement utilisée pour étudier l'impact du SHE sur les performances d'un FinFET qui représente les transistors du node technologique 45 nanometres, et ses effets sur le courant de saturation, la temperature maximale atteinte dans le dispositif et la résistance thermique. La dépendance de l'auto–échauffement par les principals paramètres géomètriques, comme l'épaisseur de la couche d'oxide enterré, la longueur des region d'accès et la distance entre "fins" de silicium adjacents, est aussi analysée. De plus, les consequences sur le SHE de solutions technologiques comme les extensions de source et drain surélevées ou la réduction de l'hauteur du fin sont pareillement explorées.

Enfin, les conclusions sont tracées dans le chapitre 7.

Riassunto e presentazione generale della tesi

Introduzione

La velocità di integrazione dei circuiti CMOS è stata mantenuta fino ad ora, seguendo la nota legge di Moore secondo cui il numero di transistori integrati su chip raddoppia ogni due anni [1]. Anche se la maggior parte dei ricercatori aveva predetto in passato che questa progressione potrebbe fermarsi ben presto, questa tendenza non ha ancora incontrato dei problemi insormontabili. I dispositivi MOSFET moderni sono caratterizzati da lunghezze di gate inferiori a 100 nanometri, e di conseguenza possiamo dire che si tratta di un regime decananometrico.

La miniaturizzazione dei MOSFET non può essere realizzata tramite una semplice riduzione delle dimensioni geometriche del dispositivo, poichè questo provocherebbe problemi significativi in termini di realizzabilità a causa dell'aumento dei campi elettrici interni al dispositivo stesso. Infatti, una scalatura efficace può essere ottenuta grazi ad un approccio più elaborato. Questo approccio è chiamato, in inglese, *scaling*. Il processo di scaling riguarda tutti i parametri fisici del dispositivo, non solo geometrici come lunghezze, larghezze e spessori, ma anche le concentrazioni dei profili di drogaggio e le tensioni di alimentazione. Lo scaling ha come obiettivo quello di aumentare il grado di integrazione dei transistori e di ottenere prestazioni migliori rispetto alle generazioni tecnologiche precedenti, salvaguardandone in ogni caso la realizzabilità.
Diverse tecniche di scaling sono state proposte:

- scaling a campo elettrico costante (*constant field scaling*), proposto da Dennard nel 1975 [4];
- scaling a tensione costante (*constant voltage scaling*);
- scaling generalizzato (*generalized scaling*) in cui la riduzione delle dimensioni geometriche del dispositivo e della tensione di alimentazione sono trattati separatamente;

- scaling selettivo (*selective scaling*), in cui la riduzione della lunghezza di gate e la riduzione delle dimensioni delle interconnessioni elettriche sono trattate separatamente.

Tuttavia con la miniaturizzazione delle dimensione dei transistori sopravvengono numerose nuove problematiche:

- effetti di canale corto (*Short–Channel Effects*, SCE)
- riduzione della barriera di energia potenziale presente tra source e canale a causa della tensione di drain (*Drain Induced Barrier Lowering*, DIBL);
- elevate correnti di perdita sottosoglia (*high off–state leakage current*);
- elevate correnti di gate a causa dell'effetto tunnel dei portatori attraverso l'ossido (*high gate leakage current*);
- elevate correnti di giunzione;
- elevate resistenze parassite;
- riduzione della resistenza di uscita per applicazioni analogiche;
- riduzione della transconduttanza g_m;
- aumento delle capacità legate alle interconnessioni;
- aumento della generazione di calore nella regione attiva del dispositivo;
- problemi di variabilità legata ai processi tecnologici.

Per superare i problemi che emergono nel momento in cui la riduzione dei dispositivi entra nella scala decananometrica, diverse soluzioni sono in fase di studio, tra le quali ricordiamo:

- architetture innovative come i transistori silicio–su–isolante (*Silicon–On–Insulator*, SOI), nei quali lo spessore dello strato di silicio utilizzato per la realizzazione del dispositivo (t_{Si}) è un nuovo parametro per la progettazione del transistore, che permette di tenere sotto controllo gli effetti di canale corto evitando di drogare pesantemente la regione di canale;
- dispositivi che includono materiali con proprietà diverse da quelle del silicio per realizzare le regioni di source, drain o canale;
- miglioramento della mobilità dei portatori, ottenuta tramite stiramento o compressione del silicio (*strain*), o grazie allo sfruttamento della dipendenza della mobilità dall'orientazione della superficie del cristallo;

	Affinità elettronica χ	[A]	[B]	[C]
Source	χ_1	Relaxed–SiGe	Relaxed–Si	Relaxed–$Si_{1-x}C_x$
Canale	χ_2	Strained–Si	Strained–Si	Si

Tabella 1: Tre possibili candidati per i materiali di source e canale per realizzare transistori MOS con eterogiunzioni tra source/drain e canale. Queste soluzioni soddisfano la condizione che $\chi_2 > \chi_1$ e dunque il canale ha un band–gap inferiore rispetto al source.

- ingegnerizzazione dell'ossido di gate, ottenuta grazie a materiali dielettrici con permettività più elevata rispetto a quella del biossido di silicio (*high–k oxides*), per limitare le correnti di perdita attraverso l'ossido di gate;
- nuove tecnologie per i contatti finalizzate a ridurre le resistenze parassite.

Inoltre, dal momento che il transistore di tipo "bulk" diventerà obsoleto per i prossimi nodi tecnologici, nuove architetture sono attualmente in fase di studio, come i dispositivi silicio–su–isolante (*Silicon–On–Insulator*, SOI), che permettono di controllare gli effetti di canale corto pur utilizzando una bassa concentrazione di atomi droganti nel canale. Questo è possibile riducendo lo spessore del sottile strato di silicio (t_{Si}), che è un nuovo parametro di progetto del transistore. Tra i vari dispositivi SOI, è possibile riconoscere diverse famiglie, quali ad esempio quella dei dispositivi planari a singolo o doppio gate, i FinFET o i nanotubi.

Simulazione Monte Carlo di eterogiunzioni

Eterogiunzioni e modifiche al programma di simulazione

Per aumentare la velocità di iniezione dei portatori che entrano nel canale, e di conseguenza ottenere una maggiore corrente di drain, di recente è stato proposto l'utilizzo di materiali diversi per realizzare le regioni di source e drain e quella di canale. In questo caso, se si utilizzano materiali con diversa affinità elettronica, viene creata una eterogiunzione [33] tra le region di accesso ed il canale. La discontinuità della banda di conduzione può permettere di aumentare l'energia cinetica dei portatori che entrano nel canale, e dunque aumentarne la velocità. La tabella 1 presenta tre possibili soluzioni per ottenere questa discontinuità.

Un approccio simile era già stato adottato in passato nella tecnologia bipolare: i transistori bipolari ad eterogiunzione (*heterojunction bipolar transistor*, HBT) sono caratterizzati da un emettitore con band–gap maggiore rispetto alla base. Questa soluzione permette di avere una corrente di deriva tra emettitore e base, e anche di drogare pesantemente la base, e queste soluzioni permettono di migliorare le prestazioni rispetto al

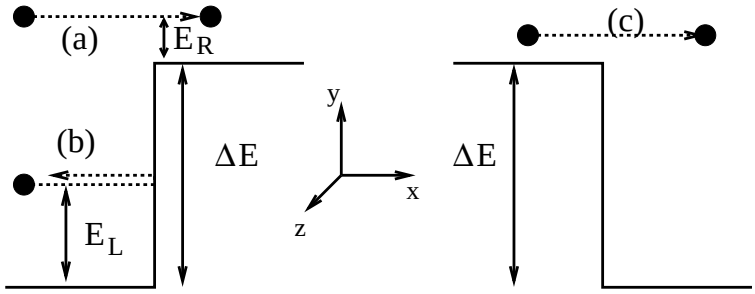

Figura 1: Casi possibili di un elettrone che si muove lungo la direzione x ed incontra una discontinuità brusca della banda di conduzione.

transistore bipolare convenzionale [36].
L'obiettivo della prima parte del manoscritto è studiare con il metodo Monte Carlo gli effetti che le eterogiunzioni hanno sul movimento dei portatori in transistori doppio gate a canale n. Prima di tutto, il programma di simulazione è stato modificato per poter simulare strutture di questo tipo. Sono state introdotte nuove interfacce, per avere la possibilità di definire la posizione ed il valore della discontinuità della banda di conduzione. Nel caso di discontinuità graduali, esse vengono trattate come un campo elettrico aggiuntivo al campo determinato dalla legge di Poisson, seguendo un approccio già utilizzato per gli HBT [52].
Il caso di eterogiunzioni brusche è più complicato; per trattarlo, diversi modelli sono stati utilizzati in funzione dell'energia iniziale e finale del portatore che attraversa la discontinuità:

- energie inferiori a 75 meV: il portatore non ha una energia particolarmente elevata, dunque il modello parabolico fornisce una approssimazione sufficiente per la struttura a bande del silicio;
- energie tra 75 e 500 meV: in questo caso, il modello parabolico non può essere utilizzato e dunque la ricerca dello stato finale è effettuata sull'intera struttura a bande (*full–band approach*);
- energie maggiori di 500 meV: la ricerca dello stato finale è randomizzata, con il solo vincolo di mantenere la stessa velocità di gruppo dell'elettrone che attraversa l'eterogiunzione.

Se lo stato finale esiste, l'energia del portatore è aumentata o diminuita di ΔE e la particella attraversa la discontinuità, viceversa essa è riflessa. La figura 1 presenta i tre casi che si possono avere quando un elettrone incontra una discontinuità della banda di conduzione. Conviene sottolineare che il programma di simulazione non considera l'effetto tunnel dei portatori attraverso la barriera di energia potenziale che si crea in presenza di una eterogiunzione.

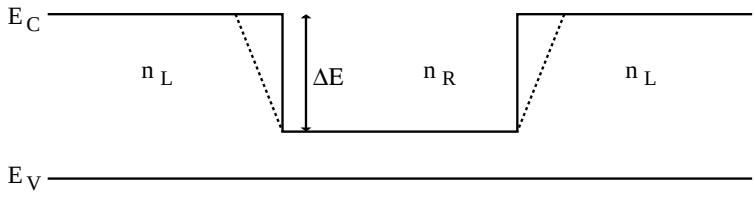

Figura 2: Banda di conduzione (E_C) e di valenza (E_V) per la struttura semplificata che é stata utilizzata per testare il codice. n_L rappresenta la concentrazione dei portatori nella regione con band–gap maggiore, mentre n_R rappresenta la concentrazione di carica nella regione a piccolo band–gap. La linea tratteggiata rappresenta la discontinuità graduale.

Validazione del modello

Per verificare le modifiche che sono state introdotte nel codice di simulazione, é stata simulata una struttura semplificata unidimensionale, drogata uniformemente, con due discontinuità di ampiezza ΔE simmetriche della banda di conduzione. La figura 2 presenta la banda di conduzione e di valenza per la struttura utilizzata. Per verificare le modifiche apportate, sono stati seguiti due approcci:

- simulazioni non autoconsistenti, senza campo elettrico applicato. n_L e n_R verificano l'equazione n_R/n_L=$\exp(\Delta E/k_B T)$, come previsto dalla teoria. La figura 3 presenta n_R/n_L per ΔE tra 50 e 200 meV;

- simulazioni autoconsistenti, senza campo elettrico applicato. Una regione di svuotamento si forma vicino all'eterogiunzione ed origina una caduta di potenziale (potentiale di *built–in*) pari a ΔV che compensa ΔE. La figura 4 presenta ΔV per ΔE tra 50 e 200 meV.

Simulazioni Monte Carlo di transistori a doppio gate con eterogiunzioni tra source/drain e canale

Simulazioni Monte Carlo, ottenute con il programma di simulazione modificato, descritto nella sezione precedente, sono state effettuate per studiare il meccanismo di trasporto dei portatori di carica in transistori SOI a doppio gate, caratterizzati da una lunghezza di gate di 34 nm, uno spessore dell'ossido di gate di 1 nm, uno spessore dello strato di silicio di 10 nm, ed un drogaggio di canale molto basso. La funzione lavoro dell'elettrodo di gate è stata modificata per avere una corrente sottosoglia pari a 100 nA/μm, e questa operazione é stata ripetuta per tutti i dispositivi considerati. La figura 5 presenta un disegno dei dispositivi utilizzati nell'analisi. Per ridurre il carico computazionale, le correzioni quantistiche non sono state incluse nelle simulazioni.

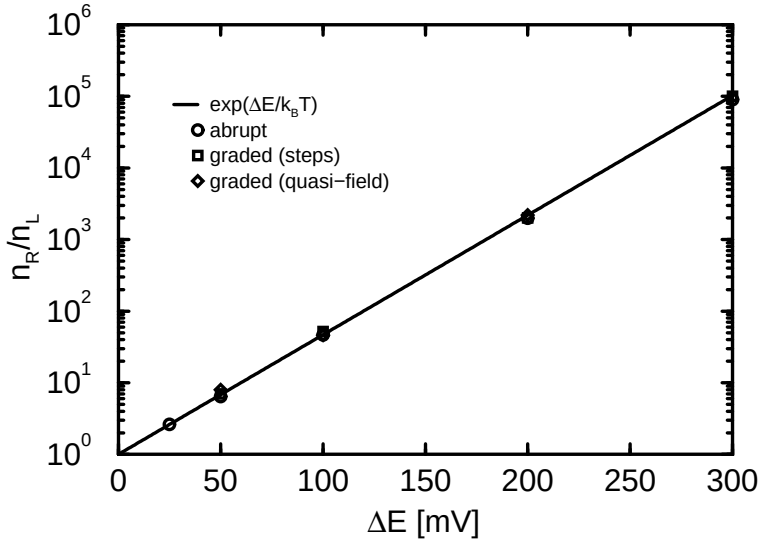

Figura 3: Rapporto tra la concentrazione di elettroni a destra e a sinistra della discontinuitá, per differenti ΔE. La teoria e' rispettata.

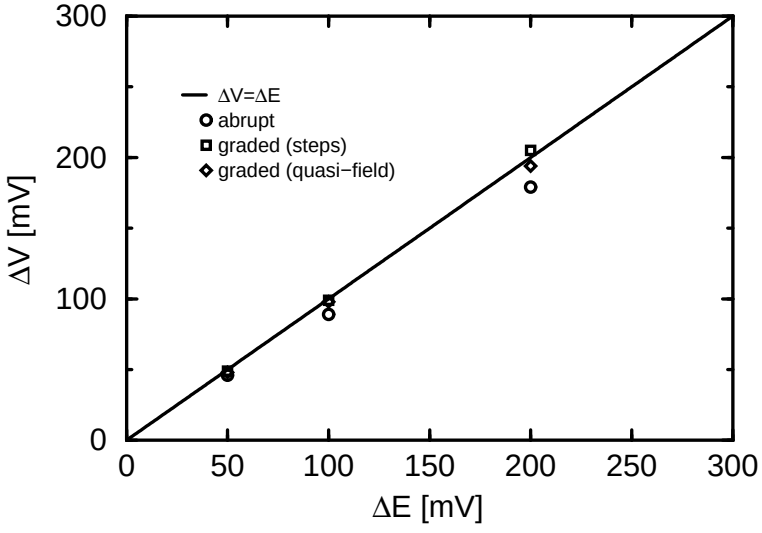

Figura 4: Confronto tra il potenziale di built–in ΔV et ΔE. La teoria é rispettata per differenti valori di ΔE

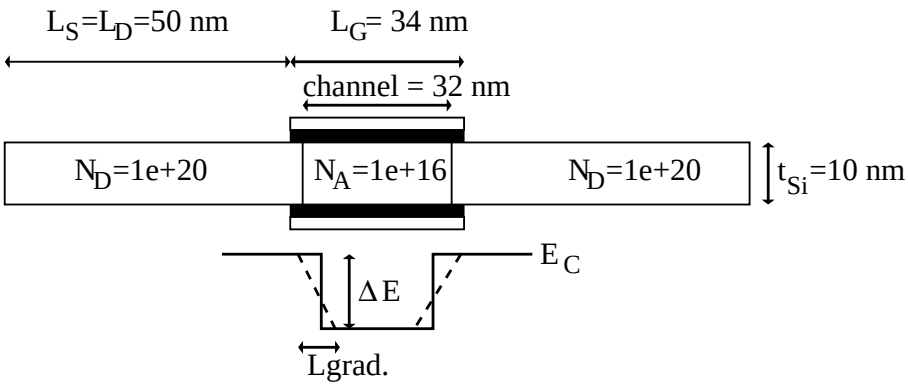

Figura 5: Semplice disegno dei transistori simulati. Nella parte inferiore della figura, è rappresentata la banda di conduzione E_C. La linea continua rappresenta una HJ brusca, la linea tratteggiata una HJ graduale. La banda di valenza é continua lungo l'intera struttura.

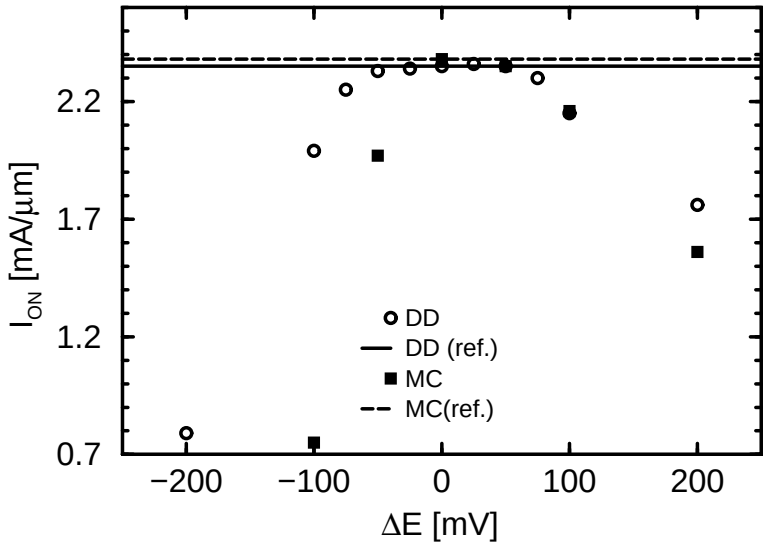

Figura 6: Correnti di drain per V_{GS}=V_{DS}=V_{DD} nei transistori con HJ brusche, per diversi ΔE, ottenute con simulazioni DD et MC. Le discontinuità sono posizionate a x=-15.2 nm et x=15.2 nm. Le righe orizzontali rappresentano le correnti ottenute per il transistore di riferimento (senza HJ).

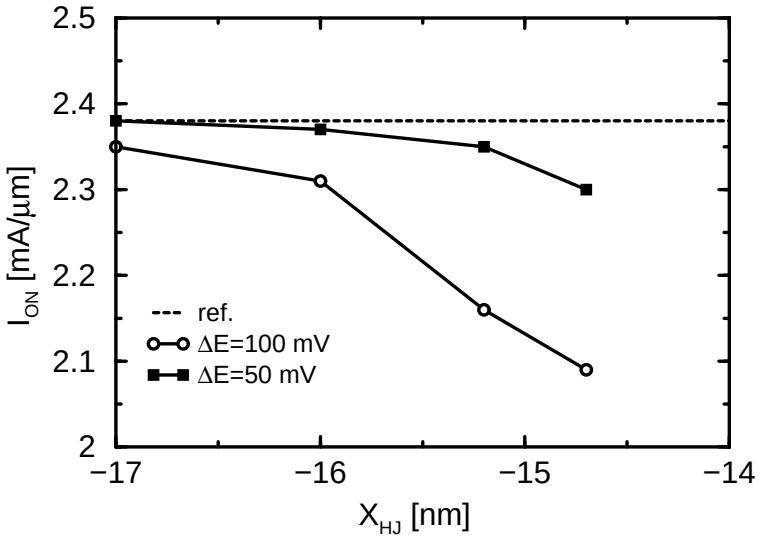

Figura 7: Correnti di drain per V_{GS}=V_{DS}=V_{DD} nei dispositivi con HJ brusche, per diverse posizioni della discontinuitá. ΔE=50 meV, 100 meV. Lungo l'asse x, x_{HJ} indica la posizione su x dove si trova la discontinuità.

Sono stati considerati i casi sia di HJ brusca che HJ graduale: le figure 6 e 7 presentano i risultati per discontinuità brusche, la figura 8 per discontinuità graduali. Le prime serie di simulazioni di transistori MOS hanno mostrato un compromesso tra il guadagno in termini di corrente ottenuto grazie ad una più elevata velocità dei portatori, e la perdita causata da un cattivo controllo elettrostatico legato all'eterogiunzione tra source e canale. Questi problemi possono essere superati utilizzando HJ graduali, anche se il compromesso tra velocità e controllo elettrostatico limita comunque il guadagno di corrente.

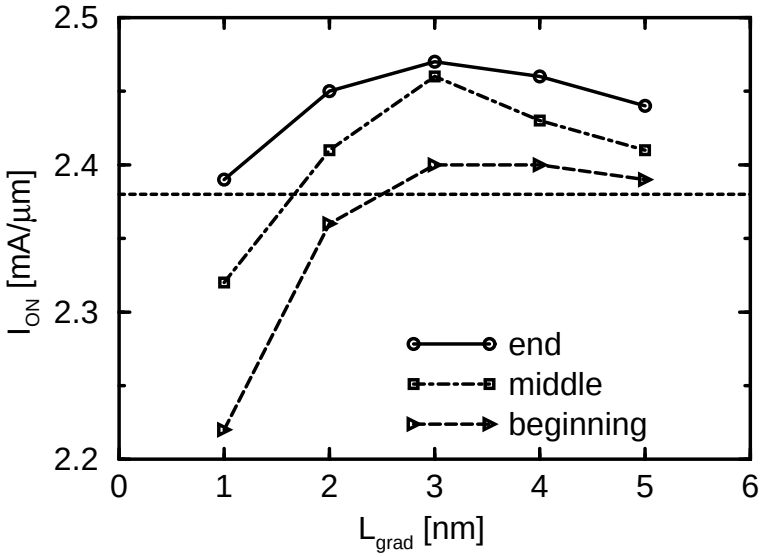

Figura 8: Correnti di drain per V_{GS}=V_{DS}=V_{DD} in funzione dell'estensione della regione graduale. ΔE=100 meV. La linea continua rappresenta la corrente del transistore di riferimento (senza HJ).

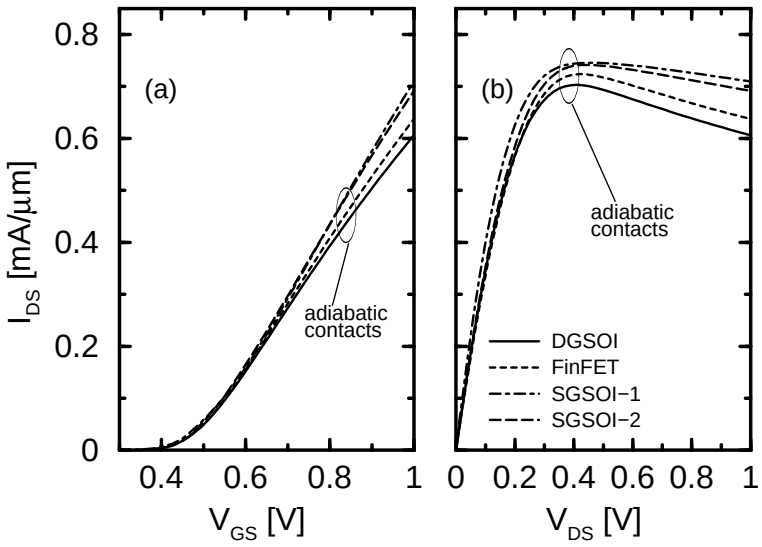

Figura 9: (a) Caratteristica di trasferimento (V_{DS}=1.0 V) e (b) caratteristica di uscita (V_{GS}=1.0 V), calcolate tramite simulazioni 3D elettrotermiche. Il gate ed i contatti di source e drain sono trattati come adiabatici, e la dissipazione del calore può avvenire unicamente tramite lo strato di ossido interrato.

Effetti di autoriscaldamento nei dispositivi SOI

La tecnologia SOI aumenta il problema della dissipazione del calore, sia perchè lo strato sepolto di ossido è caratterizzato da una conducibilità termica che è due ordini di grandezza inferiore rispetto a quella del silicio (e di consequenza esso impedisce la dissipazione del calore generato nella regione attiva), sia perchè la conducibilità termica di strati sottili di materiale semiconduttore è più bassa rispetto a quella del silicio di tipo "bulk", a causa del confinamento dei fononi e dei fenomeni di scattering che avvengono al contorno.

Nella seconda parte di questo lavoro di tesi, si è studiato l'impatto dell'autoriscaldamento in diverse strutture di tipo SOI: transistori a canale n a singolo e doppio gate, e FinFET. Sono state effettuate simulazioni in 3 dimensioni per verificare come l'autoriscaldamento possa ridurre la corrente di drain per i diversi dispositivi. I transistori coinvolti in questo confronto sono stati progettati in modo tale da avere le medesime caratteristiche elettriche in condizioni isoterme a 300 K. Le simulazioni sono state effettuate con un simulatore commerciale [27]; attenzione particolare è stata rivolta alla scelta delle condizioni al contorno per i contatti. I risultati hanno mostrato che l'autoriscaldamento degrada le performance ed è dipendente dal tipo di struttura. In particolare, le resistenze termiche associate alle regioni di accesso di source e drain, così come il rapporto tra la superficie disponibile per la dissipazione di calore attraverso l'ossido interrato ed il volume della regione attiva, cambiano da una struttura all'altra. La

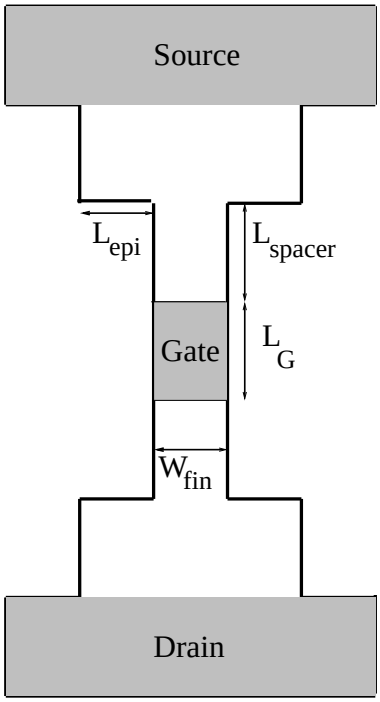

Figura 10: Rappresentazione schematica del FinFET con regioni di source e drain accresciute per epitassia. La figura non é disegnata in scala. Da notare i parametri coinvolti nelle simulazioni: L_{spacer} (distanza tra il bordo del gate e la regione accresciuta) ed L_{epi} (spessore dell'epitassia).

figura 9 presenta le caratteristiche di trasferimento e di uscita ottenute con simulazioni elettrotermiche: essa conferma come l'autoriscaldamento influisca in maniera diversa sui diversi dispositivi considerati.

Successivamente l'attenzione è stata concentrata su dispositivi di tipo FinFET, con lunghezza di gate uguale a 30 nm, al fine di studiare la dipendenza degli effetti di autoriscaldamento dai fondamentali paramentri geometrici quali lunghezza delle regioni di accesso di source e drain, spessore dello strato di ossido interrato e distanza tra "fins" di silicio adiacenti. L'analisi è stata effettuata tramite simulazioni elettrotermiche in tre dimensioni, ottenute con un simulatore commerciale.
I risultati hanno evidenziato come l'autoriscaldamento degradi pesantemente le prestazioni del dispositivo in termini di corrente erogata, anche se la sua dipendenza dai parametri considerati è debole.
Infine, sono state studiate soluzioni tecnologiche alternative: la prima riguarda la realizzazione di regioni di source e drain accresciute epitassialmente (*raised* source e drain), al fine di ridurre le resistenze serie associate alle regioni di accesso (la figura 10 offre una rappresentazione schematica di tale scelta); la seconda riguarda la riduzione del-

l’altezza del fin di silicio, adottata per semplificare il processo di realizzazione del fin stesso ed il suo drogaggio. Le simulazioni evidenziano che la distanza tra fins adiacenti non è un parametro critico dal punto di vista termico, quidi FinFETs ottenuti con altezza limitata e ridotto fin–pitch possono rappresentare una buona soluzione per aumentare la densità di integrazione mantenendo gli effetti di canale corto sotto controllo.

Résumé et présentation génerale des mémoires

Introduction

La vitesse d'intégration des circuits CMOS a été maintenue jusqu'à maintenant, en suivant la loi célèbre de Moore qui stipule que la densité des transistors integrés dans une puce double chaque 24 mois [1]. Même si beaucoup de chercheurs avaient prédié, dans le passé, que cette progression pourrait s'arrêter bientôt, la tendance n'a pas encore rencontré des problèmes majeurs. Les dispositifs MOSFET modernes ont des longueurs de grille plus petites que 100 nanométres, donc on peut dire qu'il s'agit d'un régime deca–nanometrique.

La miniaturisation des MOSFETs ne peut pas être réalisée par une simple réduction des dimensions géométriques du dispositif, car celá causerait des problémes significatifs en termes de fiabilité due à l'augmentation des champs électriques internes aux dispositifs. En fait, une mise à l'échelle efficace des dimensions des MOSFETs peut être obtenue grâce à une approche plus élaborée. Cette approche est appelée, en anglais, *scaling*. Le procédé de scaling concerne tous les paramétres physiques des dispositifs, pas seulement géométriques tels les longueurs, les largeurs et les épaisseurs, mais aussi les concentrations des profils de dopage et les tensions d'alimentations. Le scaling a pour bût d'augmenter le degré d'intégration des transistors et d'obtenir des meilleures performances que les générations précédentes tout en maintenant la même fiabilité des dispositifs.

Différentes techniques de scaling ont été proposées :

- scaling à champ électrique constant (*constant field scaling*), proposé par Dennard en 1975 [4] ;
- scaling à tension constante (*constant voltage scaling*) ;
- scaling généralisé (*generalized scaling*), dans lequel la réduction des dimensions du dispositif et de la tension sont traitées à part [5] ;
- scaling sélectif (*selective scaling*), où la réduction de la longueur de grille est traitée différemment de la réduction des connexions électriques.

Cependant nombreuses nouvelles questions surviennent avec la miniaturisation des dimensions des transistors :

- effets de canaux courts (*Short–Channel Effects*, SCE) ;
- réduction de la barrière d'énergie potentielle entre source et canal à cause de la tension de drain (*Drain Induced Barrier Lowering*, DIBL) ;
- courants de fuite sous–seuil elevés (*high off–state leakage current*) ;
- courants de grille elevés à cause de l'effet tunnel des porteurs à travers l'oxide (*high gate leakage current*) ;
- courants de jonction elevés ;
- resistances parasites elevées ;
- réduction de la resistance de sortie pour les opérations analogiques ;
- réduction de la transconductance g_m ;
- augmentation des capacités liées aux interconnections ;
- augmentation de la production de la chaleur dans la région active du dispositif ;
- fluctuations des procedés technologiques.

Pour surmonter les limitations liées à la réduction des dimensions des transistors MOS, solutions différentes sont à l'étude, entre lesquelles on rappelle :

- dispositifs qui incluent matériaux avec propriétés différentes de celles du silicium pour réaliser les régions de source, de drain et de canal ;
- amélioration de la mobilité des porteurs, obtenue par contrainte ou par exploitation de la dependance de la mobilité par l'orientation de la surface du cristal ;
- ingénierie du dielectric de grille, obtenue grâce aux oxides avec permittivité plus élevée de celle du SiO_2 (*high–k oxides*), pour limiter les courants de fuite à travers la grille ;
- nouvelles technologies de contact pour réduire les resistances parasites.

De plus, comme le transistor de type "bulk" va devenir obsolète pour les prochains noeuds technologiques, nouvelles architectures sont actuellement à l'étude, comme les dispositifs silicium–sur–isolant (*silicon–on–insulator*, SOI), qui permettent de contrôler les effets de canaux courts en maintenant un faible dopage de canale. Ça c'est possible en utilisant l'épaisseur de la couche mince de silicium (t_{Si}), qui est un nouveau paramètre pour la conception du transistor. Entre les différents dispositifs SOI, on peut reconnaître plusieurs familles comme les transistors planars single– ou double–grille, les FinFETs ou les nanofils.

	Affinité électronique χ	[A]	[B]	[C]
Source	χ_1	Relaxed–SiGe	Relaxed–Si	Relaxed–$Si_{1-x}C_x$
Canal	χ_2	Strained–Si	Strained–Si	Si

TAB. 2: Trois possibles candidats pour les matériaux de source et canal pour réaliser transistors MOS avec hétérojonctions entre source/drain et canal. Ces solutions satisfont la condition que $\chi_2 > \chi_1$ et donc le canal a une bande interdite inférieure que la source.

Simulation Monte Carlo des hétérojonctions

Hétérojonctions et modifications du logiciel de simulation

Pour augmenter la vitesse d'injection des porteurs qui entrent dans le canal, et donc obtenir un majeur courant de drain, récemment on a proposé l'utilisation de materiaux différents pour réaliser les régions de source/drain et le canal. Dans ce cas là, si on utilise materiaux avec différente affinité electronique, on réalise une hétérojonction [33] entre les régions d'accés et le canal. La discontinuité de la bande de conduction peut permettre d'augmenter l'énergie cinétique des porteurs qui entrent dans le canal, et donc augmenter leur vitesse. Le tableau 2 présente trois solutions possibles pour obtenir cette discontinuité.

Une approche analogue avait déjà été utilisée dans le passé pour la technologie bipolaire : les transistors bipolaires à hétérojonction (*heterojunction bipolar transistor*, HBT) sont caracterisés par un émetteur avec une bande interdite majeure que la base. Cette solution permit d'avoir un courant de dérive des porteurs entre l'émetteur et la base, et aussi de droguer beaucoup la base elle même, et ces solutions permettent d'augmenter les performances par rapport au transistor bipolaire conventional [36].

Le but de la première partie du manuscript est étudier avec la méthode Monte Carlo les effets que les hétérojonctions ont sur le mouvement de porteurs dans transistors double–grille à canal de type n. Tout d'abord, le logiciel de simulation a été modifié pour pouvoir simuler ces structures. Nouvelles interfaces ont été introduites, pour avoir la possibilité de définir la position et la valeur ΔE de la discontinuité de la bande de conduction.

Dans le cas de discontinuités graduelles, elles sont traitées comme un champ électrique ajouté au champ determiné par la loi de Poisson, en suivant un approche déjà utilisé pour les HBT [53].

Le cas des hétérojonctions abruptes est plus compliqué ; pour le traiter, différents modèles ont été utilisés, en fonction de l'énergie initiale et finale du porteur qui passe la discontinuité :

– énergies inférieures à 75 meV : le porteur n'a pas une énergie cinétique trés élevée, donc le modèle parabolique fournit une approximation suffisante pour la structure

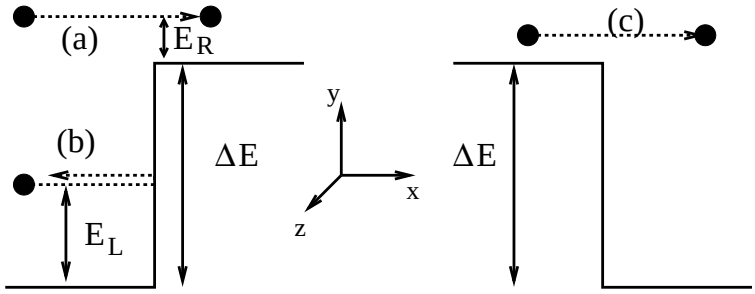

FIG. 11: Cas possibles d'un électron qui marche par la direction x et il rencontre une discontinuité abrupte de la bande de conduction.

des bandes du silicium ;
- énergies entre 75 et 500 meV : dans ce cas-là, le modèle parabolique ne peut pas être utilisé et donc la recherche de l'état final est effectuée sur toute la structure des bandes (*full–band approach*) ;
- énergies majeures que 500 meV : l'état final est choisi par vie aléatoire, avec la seule contrainte de mantenir la même vitesse de group de l'électron qui traverse l'hétérojonction.

Si l'état final existe, l'énergie du porteur est augmentée ou decalée par ΔE et la particule traverse la discontinuité, autrement elle est reflétée. La figure 11 presente les trois cas qu'on peut avoir quand un electron encontre une discontinuité de la bande de conduction. Il faut souligner que le logiciel de simulation ne considére pas l'effet de tunnel des porteurs à travers la barrière d'énergie potentielle qui ressortit quand on crée une hétérojonction.

Vérification du modèle

Pour vérifier les modifications qu'on a introduit dans le logiciel de simulation, on a simulé une structure unidimensionale simplifiée, dopée uniformément, avec deux discontinuités d'amplitude ΔE symmetriques de la bande de conduction. La figure 12 présente la band de conduction et celle de valence pour la structure utilisée. Pour valider les modifications introduites, on a suivi deux prcours :

- simulations non autoconsistantes, sans aucun champ électrique applié. n_L et n_R verifient l'equation n_R/n_L=$\exp(\Delta E/k_B T)$, comme prévu par la théorie. La figure 13 présente n_R/n_L pour ΔE entre 50 et 200 meV ;
- simulations autoconsistantes, sans aucun champ électrique applié. Une région de désertion se forme près de l'hétérojonction et elle origine une chute de potentiel (potentiel de *built–in*) égal à ΔV qui compense ΔE. La figure 14 présente ΔV pour ΔE entre 50 et 200 meV.

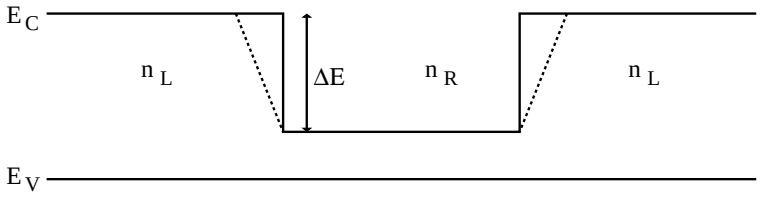

FIG. 12: Bande de conduction (E_C) et de valence (E_V) pour la structure simplifiée qu'on a utilisé pour tester le logiciel. n_L représente la concentration des porteurs dans la région avec grand gap, tandis que n_R représente la concentration des porteurs dans la région avec petit gap. La ligne pointillée représente la discontinuité graduelle.

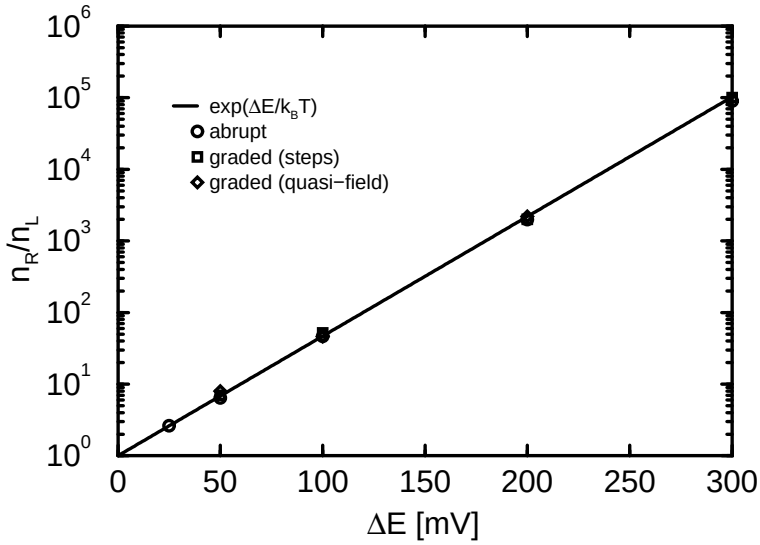

FIG. 13: Rapport entre la concentration des electrons à droite et à gauche de la discontinuité, pour différent ΔE. La théorie est respectée.

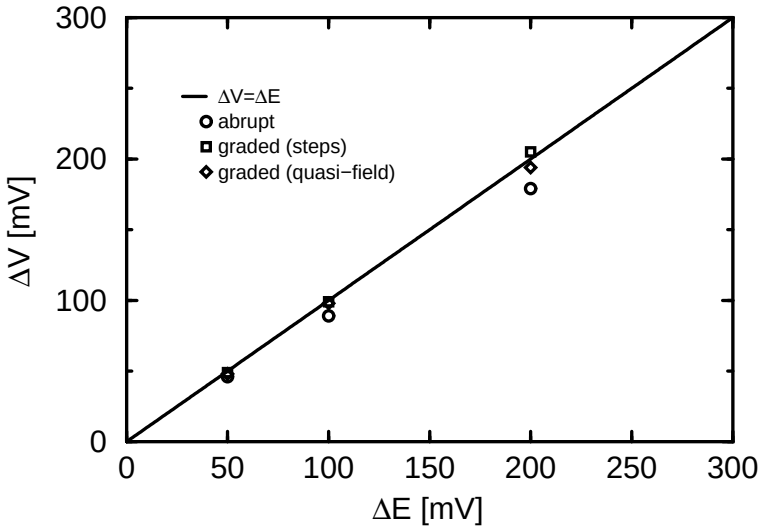

FIG. 14: Comparaison entre le potentiel de built–in ΔV et ΔE. La théorie est respectée pour différentes valeurs de ΔE.

Simulations Monte Carlo des transistors double–grille avec hétérojonctions entre source/drain et canal

On a appliqué les simulations Monte Carlo, effectuées avec le logiciel qu'on a modifié, à l'étude du transport des porteurs dans transistors SOI double–grille, avec une longueur de grille de 34 nm, un épaisseur de l'oxide de grille égal à 1 nm, un épaisseur de la couche de silicium de 10 nm et un dopage de canal très bas. La fonction de travail de la grille a été modifiée pour avoir un courant sous–seuil de 100 nA/μm, et cette opération a été répétée pour tous les dispositifs considerés.
La figure 15 présente un dessin des dispositifs considerés. Pour réduire le poids du calcul, les corrections quantistiques n'ont pas été introduites dans les simulations. On a consideré les cas de HJ soit abruptes soit graduelles : les figures 16 et 17 présentent les résultats pour discontinuités abruptes, la figure 18 pour discontinuités graduelles.
Les premières séries de simulations de transistors MOS avec HJ abruptes ont montré un compromis entre le gain en courant obtenu grâce à une plus grande vitesse des porteurs, et la perte causée par un mauvais contrôle électrostatique lié à l'hétérojonction entre la source et le canal. Ces problèmes peuvent être depassés en utilisant HJ graduelles, même si le compromis entre vitesse et contrôle électrostatique limite en tout cas le gain en courant.

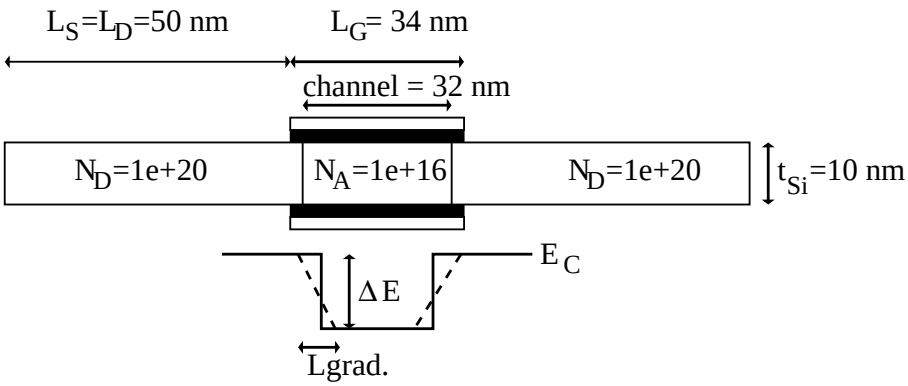

FIG. 15: Simple dessin des transistors simulés. Dans la partie inférieure de la figure, la bande de conduction E_C est presentée. La ligne continue répresente une HJ abrupte, la ligne pointillée une HJ graduelle. La bande de valence E_V est continue à travers la structure entière.

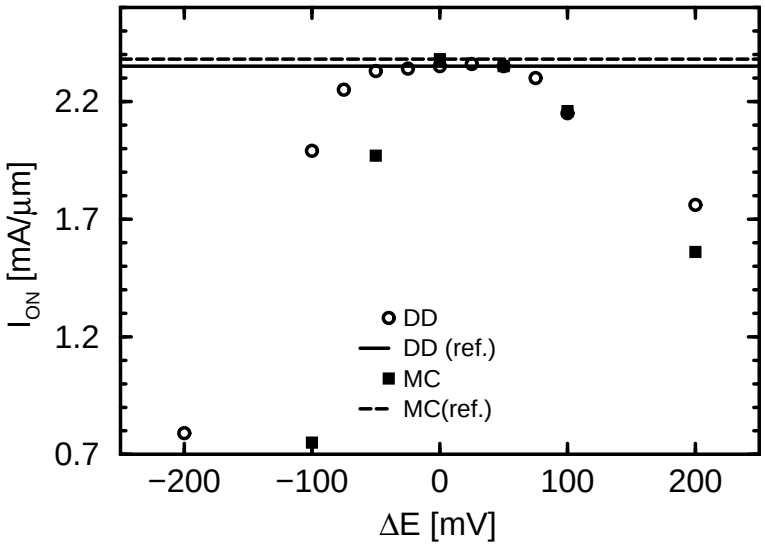

FIG. 16: Courants de drain pour V_{GS}=V_{DS}=V_{DD} dans les transistors avec HJ abruptes, pour différentes ΔE, obtenues avec simulations DD et MC. Les discontinuités sont placées à x=-15.2 nm et x=15.2 nm. Les lignes horizontales répresentent les courants dans le transistor de reference (sans HJ).

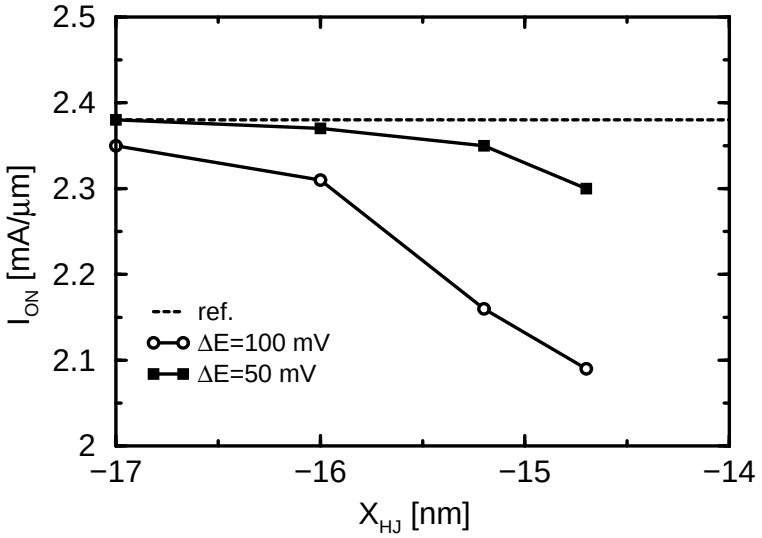

FIG. 17: Courants de drain pour V_{GS}=V_{DS}=V_{DD} dans les dispositifs avec abruptes HJs pour différentes positions de la discontinuité. ΔE=50 meV, 100 meV. Sur l'axis x, x_{HJ} indique la position sur x où la discontinuité se trouve.

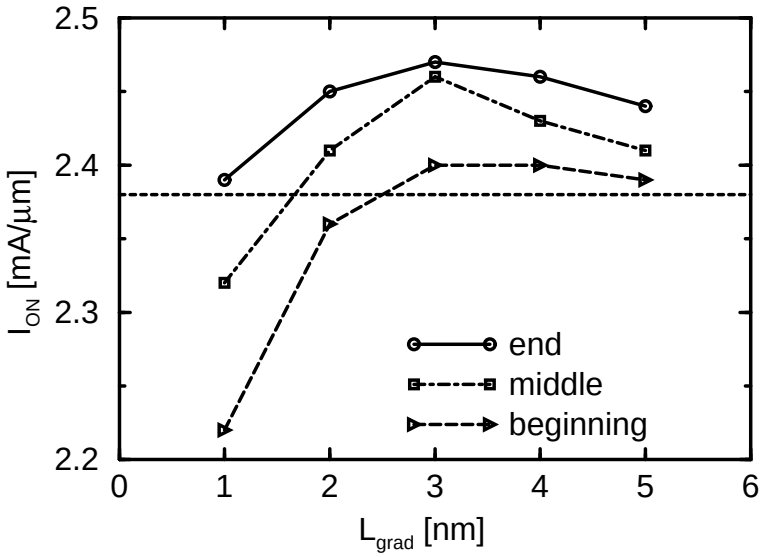

FIG. 18: Courants de drain pour V_{GS}=V_{DS}=V_{DD} en fonction de l'extension de la région graduelle. ΔE=100 meV. La ligne continue répresente le courant du transistor de reference (sans HJ).

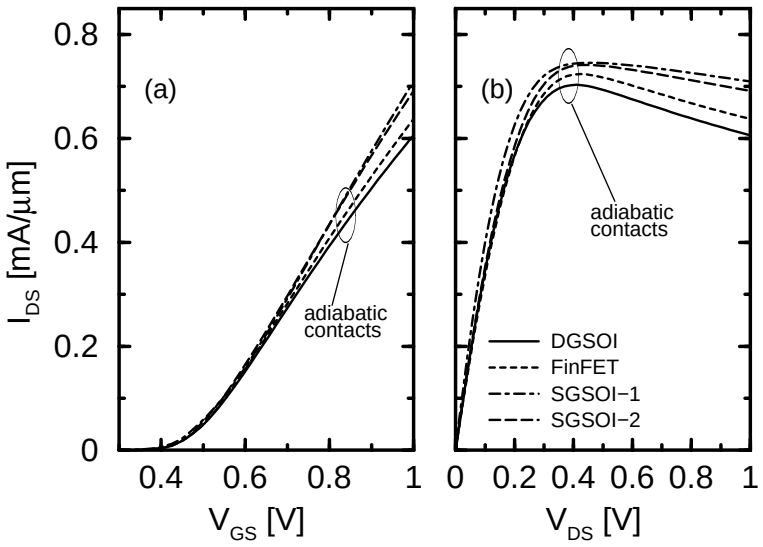

FIG. 19: (a) Caractéristique I_{DS}–V_{GS} (V_{DS}=1.0 V) et (b) caractéristique I_{DS}–V_{DS} (V_{GS}=1.0 V), calculées avec simulations 3D électrothermiques. La grille et les contacts de source et drain sont traités comme adiabatiques, et la dissipation de la chaleur peut se passer seulement à travers la couche d'oxide enterrée.

Effets d'auto–échauffement dans les dispositifs SOI

La technologie SOI augmente le problème de la dissipation de la chaleur, soit parce que la couche enterrée d'oxide est caractérisée par une conductivité thermique qui est deux ordres de grandeur inférieure que celle du silicium (et donc elle bloque la dissipation de la chaleur generée dans la région active), soit parce que la conductivité thermique des couches trés minces de material semiconducteur est beaucoup inférieure que celle du silicium de type "bulk", à cause du confinement des phonons et des interactions aux limites.

Dans la deuxième partie de ces mémoires de thèse, on a étudié l'impact de l'auto–échauffement dans différentes structures de type SOI : transistors à canal n single– et double–grille, et aussi FinFET. On a performé simulations en trois dimensions pour verifier comment l'auto–échauffement peut décaler le courant de drain pour les différents dispositifs. Les transistors impliqués dans cette comparaison ont été dessinés pour avoir les mêmes caractéristiques dans le cas isothermique à 300 K. Les simulations ont été conduites avec un logiciel commercial et elles ont montré que l'auto–échauffement se développe en différent façon entre les différents structures considerées, parce que la chaleur produite dans la région active peut être dissipée soit par les contacts, soit à travers la direction verticale à travers le substrat, soit enfin entre transistors adjacents. Ces parcours pour la chaleur sont différents pour différentes architectures. La figure 19 présente les caractéristiques I–V obtenues avec simulations électrothermiques : elle

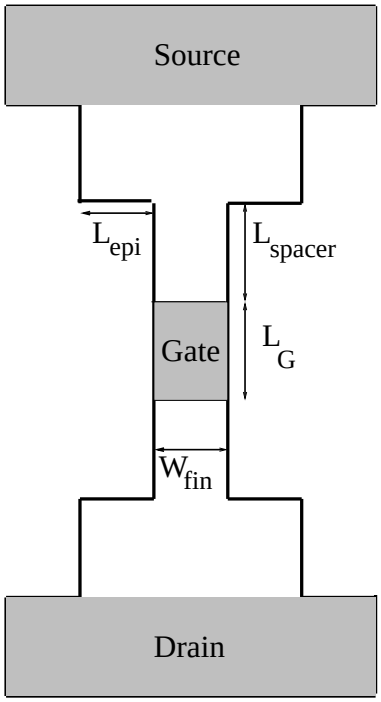

FIG. 20: Vue du FinFET avec régions de source et de drain accrues par épitaxie. La figure n'est pas dessinée à l'échelle. Il faut souligner les paramètres impliqués dans les simulations : L_{spacer} (distance entre le bord de grille et la région accrue) et L_{epi} (épaisseur de l'épitaxie).

confirme que l'auto–échauffement influence différemment les structures considerées.

En suite on a focalisé sur les simulations des dispositifs de type FinFET, avec une longueur de grille de 30 nm ; l'auto–échauffement a été étudié en fonction des différents paramètres technologiques : longueur des accès de source et de drain, épaisseur de l'oxide enterré et distance entre "fins" de silicium adjacents. L'analyse a été conduite avec simulations électrothermiques en trois dimensions, obtenues avec un logiciel commercial.

Les résultats ont montré que l'auto–échauffement dégrade beaucoup les performances du dispositif par rapport au courant de drain, même si sa dépendance aux paramètres considerés est très faible.

Enfin, solutions technologiques alternatives ont été étudiées : la première concerne le réalisation de régions de source et de drain accrues par épitaxie (source et drain surélevés), pour obtenir une réduction des résistances serie (la figure 20 présente un simple dessin pour une telle solution) ; la deuxième concerne la réduction de l'hauteur du fin de silicium, adoptée pour simplifier le procedé de fabrication du fin lui–même et aussi son dopage.

Les simulations mettent en évidence que la distance entre fins adjacents n'est pas un paramètre critique du point de vue thermique, donc FinFETs obtenus avec hauter limitée et fin–pitch réduit peuvent répresenter une bonne solution pour augmenter la densité d'intégration en maintenant les effets de canaux courts bien maîtrisés.

Chapter 1

Introduction

1.1 Trends in microelectronics

The semiconductor industry lived an amazing progress of the performance during the last fifty years. Its impetuous and exponential growth is evident to everyone: electronic products went into our lifes and they have radically changed our habits, how we work and how we communicate. The progresses have been extraordinary in terms of:

- *speed*: modern electronic systems became ever faster. An useful parameter to measure this trend could be the clock frequency of a Central Processing Unit (CPU) of a modern personal computer. This frequency has increased by an order of magnitude in less than ten years (see Fig. 1.1);

- *computational capability*: not only the speed of electronic systems, but also the number of functions that they can offer to the customer has exploded. This is the case of modern cellular phones, that in few years became ”multimedia center” able to perform oral communication, take pictures, play music and videos, exploit features of GPS navigators;

- *dimensions and portability*: electronic circuits are becoming smaller and smaller, and even more portable. The small dimensions reached by cellular phones or notebooks currently on sale, as well as the increasing relevance of the so called *wearable electronics*, could represent good examples.

Moreover, this technological trend soon became a market expectation, so that not only the microelectronics systems are actually smaller and smaller, faster and faster, cheaper and cheaper, but also the customers have been ”well” educated by advertizing such progresses.

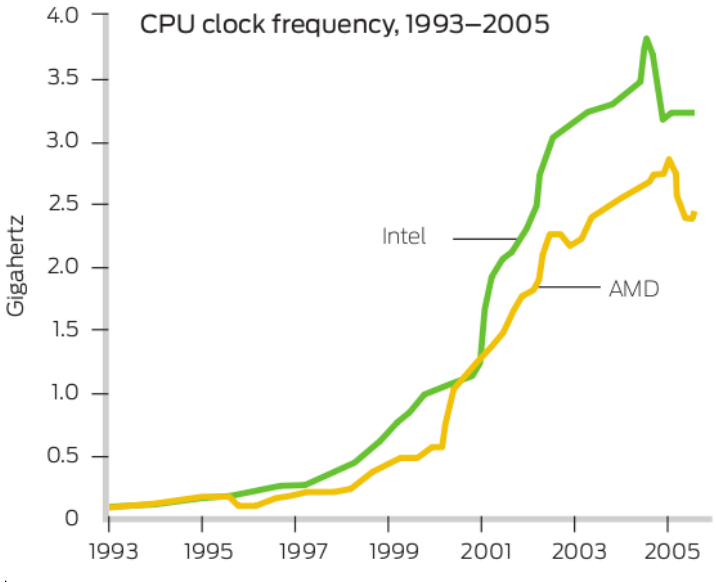

Figure 1.1: Graph of the clock speed development of AMD and Intel processors from 1993 until the end of 2005. Between 1993 and 1999, the average clock speed increased tenfold.

The main actor of this panorama is the Integrated Circuit (IC). Modern Integrated Circuits are present in almost every electronic system, from personal computers to communications systems, from medical equipments to automotive ones. An integrated circuit is a miniaturized electronic circuit that is manufactured on a surface of a thin layer of a semiconductor substrate. ICs consist mostly of *transistors*, even if analog circuits commonly contain resistors and capacitors as well and inductors are used in some high frequency analog circuits.
Semiconductor ICs are fabricated in a layer process which includes these key steps:

- *imaging*: a layer of photoresist is subject to photochemical reactions resulting from irradiation; as the wavelength of radiation is getting shorter, and hence absorption coefficient increases, thickness of imaging resist is gradually reduced;

- *deposition*: deposited species are formed as a results of chemical reaction between gaseous reactants at elevated temperature in the vicinity of the substrate; solid product of the reaction is deposited on the surface of the substrate; used to deposit films of semiconductors (crystalline and non–crystalline), insulators as well as metals. The most common thin film deposition method in advanced semiconductor manufacturing is the Chemical Vapor Deposition (CVD) with its many forms;

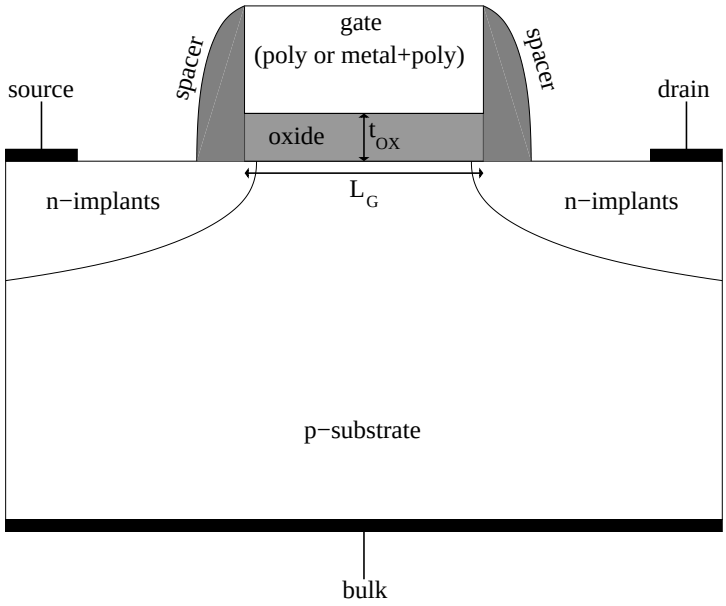

Figure 1.2: 2–D cross–section of a typical Bulk transistor. The case presented is an n–channel MOSFET; the bulk contact is called also substrate contact. The figure is not drawn to scale; in particular, in real devices the gate oxide is much thinner with respect to the others regions.

- *etching*: process of etching through chemical reaction between chemically reactive species and the material to be etched.

The main process steps are supplemented by doping, cleaning and planarization steps. Mono–crystalline silicon wafers are used as the substrate; photo–lithography is used to mark different areas of the substrate to be doped or to have polysilicon, insulators or metal layers deposited on them. Regarding *digital* ICs, the dominant technology is currently the Metal–Oxide–Semiconductor Field–Effect-Transistor (MOSFET, also referred to as MOS transistor) that has completely replaced old technologies like the bipolar one, that is currently adopted only for analog and radio–frequency applications.. Fig. 1.2 presents a typical structure of an n–channel MOSFET. When the device is made directly on a silicon substrate, this is named *Bulk* MOSFET; Fig. 1.2 presents a typical structure of an n–channel Bulk MOSFET. Associated to its p–type counterpart, this is the most common structure that has been adopted in digital circuits.

The progress of electronic circuits has been accomplished mainly by the miniaturization of the MOSFET: the downsizing of the transistor decreases its capacitance without compromising the current. Therefore, it increases the circuit operating speed and decreases its per–transistor power consumption. Moreover, transistor size reduction increases the number of components that can be crammed into a circuit at constant die area (*transistor integration*). Since the fabrication costs are roughly proportional to die

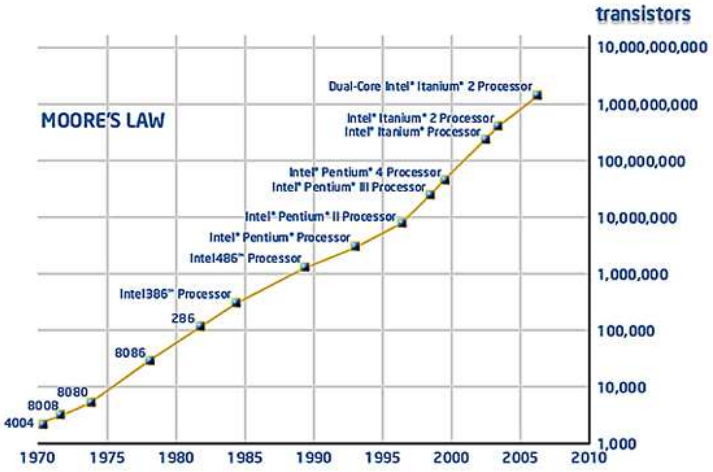

Figure 1.3: Plot of transistor counts against dates of introduction. The curve shows counts doubling every two years (source:*http://web.sfc.keio.ac.jp*)

area, miniaturization also reduces the cost of production per single transistor and even per single function.

1.2 MOSFET scaling

As it has been mentioned in Section 1.1, the shrink of the dimensions of the MOS transistor is the key factor in order to improve performance, increase integration scale and therefore reduce production costs.
The first who noticed and quantified the exponential growth in integration of semiconductor technology was Gordon E. Moore in 1965 [1], at the time director of the R&D of Fairchild Semiconductors, and later co–founder of Intel Corporation. Based on the data collected during his work at Fairchild (unit costs, MOSFET dimensions, device integration, technologies available, performance and power consumption achieved) and extrapolating their future trend in the absence of any evident show–stopper to further downscaling, he predicted the doubling of devices density per year until 1975. About 10 years later, his empirical observations still held true and he updated the forecast [2]. Since then, the approximate yearly doubling of the device density has been known as the Moore's law (see Fig. 1.3) [3].
The reduction of the dimensions of a MOSFET has been dramatic during the last three decades. Starting at a minimum feature length of 10 μm in 1970 the gate length was gradually reduced to 0.15 μm minimum feature size in 2000, resulting in a 13% reduction per year. Hovewer proper downsizing of MOSFET requires not only a size reduction of the gate length and width. Indeed, a simple shrinking of the device geometrical

dimensions would lead to serious reliability issues connected with the increase of the internal electric field. In fact, the steady reduction in MOSFETs dimensions was obtained through a more complex process called *scaling*. The scaling process concerns all physical parameters of the devices, not only the geometrical dimensions length, width and thickness but also doping and supply voltage, and its aim is to increase the degree of integration and achieve higher performance with respect to the previous generation while maintaining device reliability. Scaling requires a reduction of all other dimensions including gate/source and gate/drain overlap, oxide thickness and depletion layer widths, and scaling of depletion layer widths also implies scaling of substrate doping density. The *scaling rules*, paced by a scaling parameter $\alpha > 1$, establish the correct reduction or increase of each device physical parameter in order to achieve the scaled devices with the desired criteria.

Different approaches to scaling have been explored. The *constant field* scaling has been formalized for the first time in early '70 by Dennard [4], with the idea of reducing the device dimensions while still preserving an identical configuration between the scaled device and the unscaled one and maintaining the same current–voltage behavior. The effect of keeping the electric field unchanged in the scaled device is to avoid undesirable high fields effects such as mobility degradation, impact ionization, or hot–carrier effects. The results of this scaling approach are summarized in the third column of Table 1.1.

The constant field scaling results in circuit speed increasing in proportion to the factor α and circuit density increasing as α^2. Two problems quickly appeared: the built–in potentials and the MOSFET subthreshold slope do not scale; consequently the threshold voltage V_t (that is the gate voltage above which the channel is formed, the MOSFET in "ON" and significant current begins to flow from the source to the drain) could not be reduced any more following the scaling without an unacceptable increase in leakage current. In order to overcome these limitations, a second scaling method has been proposed: the *constant voltage* scaling. The constant voltage scaling does not present this problem and has been the preferred scaling method for longtime, since it provides voltage compatibility with older circuit technologies. The disadvantage of constant voltage scaling is that the electric field increases as the minimum feature length is reduced. This leads to velocity saturation, mobility degradation, increased leakage currents and lower breakdown voltages.

In order to accomodate this trend, constant field and constant voltage scaling rules have been updated to *generalized scaling* rules [5], where it became necessary to treat separately the reduction of the supply voltage and the geometrical dimensions: the slower scaling of the supply voltage with respect to the geometry (by a factor $\varepsilon > 1$) leads the electric field inside the device to increase proportionally to ε (see the third column of Tab. 1.2). This factor had to be chosen carefully to balance the trade–off between the increase in leakage current and the increase in electric field, which is a threat for device

Physical Parameter	Symbol	Constant Field Scaling	Constant Voltage Scaling
Gate Length	L_G	$1/\alpha$	$1/\alpha$
Gate Width	W_G	$1/\alpha$	$1/\alpha$
Electric Field	$\mathcal{E}$	1	α
Gate Oxide Thickness	t_{OX}	$1/\alpha$	$1/\alpha$
Wiring width		$1/\alpha$	$1/\alpha$
Doping	N_A,N_D	α	α^2
Gate Capacitance	C_G	$1/\alpha$	$1/\alpha$
Voltage	V_{DD}	$1/\alpha$	1
On–current per device	I_{DS}	$1/\alpha$	1
Power Dissipation	P	$1/\alpha^2$	1

Table 1.1: Comparison of the effect of scaling on MOSFET device parameters. This table compares constant field and constant voltage scaling.

reliability.
Finally, in more recent technology generations, a new scaling approach has been used. This has been called *selective scaling* and did not scale the wiring at the same extent as the gate length in order to improve the wiring yield without degrading the gate delay [6]. The dimensional scaling parameter α has been split into two separate spatial dimension parameters: $\alpha_d > 1$ for scaling the gate length and vertical dimensions; $\alpha_w > 1$ (smaller than α_d) for scaling the device width and wiring (fourth column of Tab. 1.2).
The scaling of MOSFET device parameters is summarized by Table 1.1 where constant field and constant voltage scaling are presented, and Table 1.2 where generalized and selective scaling are compared.

1.3 The crisis of conventional Bulk MOSFET

As detailed in the previous sections, scaling has progressed very successfully in the past: the structure of the MOS transistor has remained relatively unchanged for nearly three decades, and the downsizing of the MOSFET has increased performance and integration level. However, producing MOSFETs with channel lengths much shorter than 100 nm is a challenge, and the difficulties of semiconductor device fabrication are always a limiting factor in advancing integrated circuit technology. In recent years, the small size of the MOSFET, in the order of a few tens of nanometers, has created operational problems, such as:

- *Short Channel Effects*: the Short Channel Effects (SCE) consist in a loss of control

Physical Parameter	Symbol	Generalized	Selective
Gate Length	L_G	$1/\alpha$	$1/\alpha_d$
Gate Width	W_G	$1/\alpha$	$1/\alpha_w$
Electric Field	$\mathcal{E}$	ε	ε
Gate Oxide Thickness	t_{OX}	$1/\alpha$	$1/\alpha_d$
Wiring width		$1/\alpha$	$1/\alpha_w$
Doping	N_A,N_D	$\varepsilon\alpha$	$\varepsilon\alpha_d$
Gate Capacitance	C_G	$1/\alpha$	$1/\alpha_w$
Voltage	V_{DD}	ε/α	ε/α_d
On–current per device	I_{DS}	ε/α	ε/α_w
Power Dissipation	P	ε^2/α^2	$\varepsilon^2/\alpha_w\alpha_d$

Table 1.2: Comparison of the effect of scaling on MOSFET device parameters. This table compares generalized and selective scaling.

of the gate on the potential profile along the channel, due to the reduction of the gate length L_G: the main macroscopic effect is the decrease of the threshold voltage while L_G is scaled down. The expression for the threshold voltage at zero–bias is given by

$$V_{t0} = V_{FB} + 2\Phi_F - \frac{Q_d}{C_{OX}} \tag{1.1}$$

where V_{FB} is the *flat–band voltage* (the voltage that must be applied to the gate to bring the semiconductor energy band), Φ_F is the Fermi potential, Q_d is the depletion charge and C_{OX} is the oxide capacitance.
This equation is accurate in describing large MOS transistor, but it collapses when applied to small–geometry MOSFETs. In fact Eq. 1.1 assumes that the bulk depletion charge is only due to the electric field created by the gate voltage, while the depletion charge near n–type source and drain region is actually induced by pn junction band bending. Therefore the amount of bulk charge supported by the gate voltage is overestimated, leading to a larger V_t than the actual value. A typical parameter adopted by the designers in order to evaluate this effect is the so–called V_t *roll–off*.

- *Drain Induced Barrier Lowering* (DIBL): the source and drain of a MOSFET form pn junctions within the substrate. The width of the depletion regions associated with the junctions increases with applied reverse bias. If the channel is long enough, the application of a drain bias does not modify the potential barrier of the source junction, while in a short channel device the potential barrier at the source can be reduced by a value $\Delta\Phi$ depending on the drain bias itself. This reduction of the potential barrier reduces the threshold voltage, and the magnitude of such

reduction is defined by the DIBL, that is the difference of V_t at low V_{DS} and high V_{DS} [7]. For short channel devices, DIBL values larger than about 150 mV/V are unacceptable. In extreme cases the potential barrier at the source can become so small that the current between source and drain is no longer controlled by the gate: this phenomenon is called *punch–through*.

- *Higher off–state leakage current* (I_{OFF}): as the MOSFET channel length is reduced to 50 nm and below, the suppression of off–state leakage current (i.e. the drain current when the transistor is nominally "OFF") becomes an increasingly difficult technological challenge. It will be extremely difficult for Bulk–Si MOSFET technology to meet industry–specified performance targets for both drive current and leakage current for the forthcoming technological nodes.
  In a conventional Bulk–Si MOSFET, the channel dopant concentration must be increased as the distance between the source and drain junctions decreases, in order to avoid electrostatic coupling between the junctions beneath the channel surface, and therefore to keep under control the I_{OFF}. On the other hand, higher doping results in degraded low–field mobility and hence reduces the transistor drive current.

- *Increased gate leakage current* (I_G): the gate oxide, which serves as insulator between the gate and channel, should be made as thin as possible to increase the channel conductivity and performance when the transistor is "ON" and to reduce subthreshold leakage when the transistor is "OFF". In other words, in order for the gate voltage to effectively modulate the channel electric potential, the gate–to–channel capacitance must be maximized. This has historically been achieved by reducing the thickness of the SiO_2 gate dielectric, but the gate oxide is lower than 2 nm in today state–of–the–art CMOS technology, so that the quantum mechanical phenomenon of electron tunneling occurs through the thin gate oxide. This phenomenon leads to a degradation of key parameters vital for high performance device operation (larger gate leakage current but even oxide breakdown and lower channel mobility as well) and an increased power consumption [8].

- *Increased junction leakage*: in order to minimize short channel effects, the depth of the source and drain junctions must be reduced as the channel length is reduced. An ion implantation technique called "halo", where impurity dopants (p–type for n–channel MOSFETs) are placed next to the junction tip, is often used to control the SCE in planar transistors. The concept of "super–halo", where channel doping is highly localized, was proposed as the ultimate architecture for planar Bulk CMOS transistor scaling.
  It is evident that in order to make devices smaller, junction design has become more complex, leading to higher doping levels, shallower junctions, "halo" doping and so forth. The formation of ultra–shallow junctions is a significant tech-

nological challenge, particularly because low sheet resistances in the MOSFET access region are needed for high transistor drive current. To keep these complex junctions in place, the annealing steps formerly used to remove damage and electrically active defects must be curtailed, increasing junction leakage. Heavier doping is also associated to thinner depletion layers and more recombination centers that result in increased leakage current, even without lattice damage.
Subthreshold leakage (including subthreshold conduction, gate–oxide leakage and reverse–biased junction leakage), which was ignored in the past, now can consume up to half of the total power consumption of modern high–performance VLSI chips [9].

- *Increased parasitic resistances*: one of the key obstacles to effective device scaling is the increasing extrinsic resistance of transistors. Historically, the main components of this parasitic resistance consisted of channel, junction, and silicide–to–junction contact resistance components. However, as device dimensions approach the 45 nm technology node, additional parasitic components related to the contact resistivity start to influence the circuit performance increasingly [10].

- *Lower output resistance*: for analog operation, good gain requires a high MOSFET output impedance, which is to say, the MOSFET current should vary only slightly with the applied drain–to–source voltage. As devices are made smaller, the influence of the drain competes more successfully with that of the gate due to the growing proximity of these two electrodes, increasing the sensitivity of the MOSFET current to the drain voltage. To counteract the resulting decrease in output resistance, circuits are made more complex, either by requiring more devices, for example the cascode and cascade amplifiers, or by feedback circuitry using operational amplifiers.

- *Lower transconductance*: the transconductance g_m of the MOSFET determines its gain and is proportional to hole or electron mobility, μ. As MOSFET size is reduced, the fields in the channel increase and the dopant impurity levels increase. Both changes reduce the carrier mobility, and hence the transconductance. As channel lengths are reduced without a proportional reduction of drain voltage, raising the electric field in the channel, the result is velocity saturation of the carriers, limiting the current and the transconductance (under the assumption, valid for ”long devices”, of diffusive transport characterized by μ).

- *Interconnect capacitance*: traditionally switching time was roughly proportional to the gate capacitance of gates. However, with transistors becoming smaller and more transistors being placed on the chip, interconnect capacitance (that is the capacitance of the wires connecting different components of the chip) is becoming

a large percentage of capacitance. Signals have to travel through the interconnect, which leads to increased delay and lower performance.

- *Heat production*: the ever–increasing density of MOSFETs on an integrated circuit is creating problems of substantial localized heat generation that can impair circuit operation. Circuits operate slower at high temperatures, and have reduced reliability and shorter lifetimes. Heat sinks and other cooling methods are now required for many integrated circuits including microprocessors. Power MOSFETs are at risk of thermal runaway. As their on–state resistance rises with temperature, if the load is approximately a constant–current load then the power loss rises correspondingly, generating further heat. When the heat sink is not able to keep the temperature low enough, the junction temperature may rise quickly and uncontrollably, resulting in destruction of the device.

- *Process variations*: with MOSFETS becoming smaller, the number of atoms in the silicon region that determines many of the transistor's properties is becoming smaller, with the result that control of dopant numbers and placement is more erratic [11]. During chip manufacturing, random process variations affect all transistor dimensions: length, width, junction depths, oxide thickness and so forth, and become a greater percentage of overall transistor size as the transistor shrinks. The transistor characteristics become less certain, more statistical. The random nature of manufacture means we do not know which particular example MOSFETs actually will end up in a particular instance of the circuit. This uncertainty forces a less optimal design because the design must work for a great variety of possible component MOSFETs.

It is evident that the challenges related to scaling electron devices to the nanometric scale are impressive, and many serious problems are expected if scaling is to be pursued until its atomic limits.

With the progressive externalization of production tools to the suppliers of specialized equipment, the need arose for a clear roadmap to anticipate the evolution of the market and to plan and control the technological requests of IC production. The *International Technology Roadmap for Semiconductors* (ITRS) is a set of documents produced by a group of semiconductor industry experts and updated every 2 years. These experts are representative of the sponsoring organizations which include the Semiconductor Industry Associations of the US, Europe, Japan, Korea and Taiwan. This documents represent best opinion on the directions of research into the development of semiconductor industry, including time–lines up to about 15 years into the future; the ITRS involves areas of interest as photolitography, assembly and packaging, modeling and simulation, emerging devices and materials, and more.
The scenario outlined by the ITRS for the forthcoming technological nodes (TN) shows

that it will be extremely difficult for Bulk MOSFET technology to meet industry specified performance targets, because there is a increasing consensus in the scientific community that the conventional Bulk architecture may soon reach its scaling limit.
In order to overcome these limitations, to continue channel length scaling and therefore sustain the performance trend, many solutions have been proposed and are currently under investigations. The most important innovations that have been suggested regard:

- *transistor structure*: Ultra–thin (UT) Silicon–On–Insulator (SOI) MOSFETs are an attractive option for device scaling because they can effectively reduce the SCE and eliminate most of the leakage paths [12]. SOI technology and novel device architectures will be detailed in Section 1.4.

- *mobility enhancement techniques*: mobility enhancement is an attractive option because it can potentially improve device performance beyond any of the benefits resulting from device scaling. The two main approaches being pursued are strain engineering (both process– and substrate–induced) and orientation effects.
  Strain effects induced during the fabrication process can increase the channel mobility; both tensile and compressive stresses can be introduced in any one of three dimensions by process techniques. On one hand, technology scaling will reduce the space available to introduce stress. On the other hand, a larger area will be under higher stress in shorter channel lengths in scaled technology. The scalability of local strain is one of the most important topics for future CMOS performance.
  Inversion layer mobility depends on surface orientations and current flow directions. For p–channel MOSFETs, hole mobility is 2.5 times higher on (110)–oriented surfaces compared with that on standard wafers with (100) surface orientation. However, electron mobility is the highest on (100) substrates. In order to fully exploit the advantage of carrier mobility dependence on surface orientation, new 3–D technologies to fabricate CMOS on hybrid substrates with different crystal orientations have been developed, with the channel of nFETs on Si with (100) surface orientation and pFETs on (110) surface orientation.

- *gate stack engineering*: as mentioned in Section 1.3, starting from the 65 nm technologica node the gate leakage current, due to the tunnel effect of the carriers through the thin gate oxide, has been a major issue. In order to alleviate this undesired phenomenon, a possible solution could be to increase the gate capacitance for a fixed dielectric thickness by using alternative gate–dielectric materials with permittivities higher than that of SiO_2 (referred to as *high–k oxides*). By replacing the conventional SiO_2–based dielectric with a thicker insulator featuring a higher dielectric constant, it is possible to reduce the quantum tunneling current through the dielectric between the gate and the channel [13]. After almost a decade of intense research on different high–k alternatives, the family of hafnium oxide (HfO_2)–based materials has emerged as the leading candidate to replace

SiO_2 gate dielectrics in advanced CMOS applications [14]. In 2007 for the first time transistors for the 45 nm TN featuring high–k oxide and metal gate have been incorporated in an high–volume manufacturing process [15].
On the other hand, CMOS process compatibility and reliability are major issues for almost all known high–permittivity gate dielectrics; furthermore, the barrier height of high–k insulators, that is the difference in conduction band energy between the semiconductor and the dielectric (and the corresponding difference in valence band energy), strongly affects leakage current level. For many alternative dielectrics this value is significantly lower than that one of Silicon dioxide, tending to increase the tunneling current, and therefore somewhat reducing the advantages of higher dielectric constant.
The use of metal gate electrodes, which eliminates poly–Si depletion, and the use of stacks composed by metal gate/high–k dielectric, can result in aggressive scaling. In order to achieve appropriate V_t it is essential to use metal gates with a near–band–edge work function for conventional planar MOSFETs. Research on band–edge dual work function metal gate electrodes has been gaining momentum, as conventional gate stacks run out of steam for sub–65 nm technologies. However, the most critical challenge that remains for metal gate/high–k stacks is the V_t stability of metal gates when in contact with Hafnium–based dielectrics.

- *novel contact technology*: as mentioned in the previous section, one of the key obstacles to device scaling is the increasing extrinsic resistances of transistors. From this point of view, a strong effort is being performed by the scientific community who is exploring solutions such as novel silicides, innovative design for the source and drain extension regions (Raised S/D), or low–resistance filling materials for the contacts.

- *source and drain engineering*: alternative materials are adopted in the source/drain regions, featuring band–gap different from that of silicon. This approach will be largely discussed in Section 3.1 because is the core topic of the first part of this thesis work.

1.4 Silicon–On–Insulator Technology

As discussed in Section 1.3, scaling trends will make in a very near future the conventional Bulk transistor obsolete, and structures alternative to Bulk architecture are thoroughly studied.
A first approach could be represented by the Silicon–On–Insulator Technology, in which the device is not build directly on the silicon substrate. SOI technology refers to the use of a layered silicon–insulator–silicon substrate in place of conventional silicon substrates in semiconductor manufacturing to reduce parasitic device capacitance and there-

fore improve performance. SOI–based devices differ from conventional silicon–built devices in that the silicon junction is above an electrical insulator, typically silicon dioxide. Since the first commercially viable implementation of SOI, announced by IBM in August 1998, many methods to build silicon layers on SiO_2 Buried Oxide (BOX) have been proposed: SIMOX technology (Separation by IMplantation of OXygen), wafer bonding methods (such as *SmartCut*® from SOITEC, *NanoCleave*® from Silicon Genesis corporation or *ELTRAN*® from Canon) or finally *seed methods*. A successful semi–automatic microprocessor design migration, from planar to a double–gate FinFET design translation, has already been reported [16].
A MOS transistor designed on an SOI substrate has a cross–section like the one in Fig. 1.4. The main differences with respect to the conventional Bulk transistor presented in Fig. 1.2 are:

- the vertical isolation protects the active region from many parasitic effects like radiation–induced photo–currents and latch–up;
- the isolation reduces the parasitic capacitances and leakage currents of the pn junctions;
- the lateral inter–device isolation in the SOI case provides very good isolation.

Starting from an SOI wafer, many different types of MOS transistor can be obtained. The first classification is between the planar Single–Gate (SG) SOI MOSFET, and the Multiple–Gate SOI MOSFETs (also referred as MugFETs). While SG devices have already found a commercial application, MugFETs are still confined in the research area.
The planar Single–Gate SOI MOSFETs can be further classified into two categories:

- *Partially–Depleted* (PD): if the active silicon body is thick enough to contain completely the depletion region in strong inversion;
- *Fully–Depleted* (FD): if the thickness of the active area t_{Si} is smaller than the depletion layer width in inversion and therefore the depletion region reaches the top of the buried oxide. The depths of the source and drain diffusions are limited by t_{Si} as well, which is a critical parameter for such devices.

At the moment all the integrated circuits on SOI wafers use the Partially–Depleted structure because the technological process to create a relatively thick body is much easier. The design rules of a PD–SOI MOS transistor are not very different from the Bulk case and the scaling rules are almost the same, for example the doping concentration of the body must be increased while decreasing the gate length L_G in order to control the SCE.

On the contrary, the FD–SOI MOSFET is quite different from the Bulk transistor, and it is attractive for many reasons. The main issues about the use of an ultra thin silicon body are:

1. very large mobility, due to undoped silicon body. Unfortunately this advantage is reduced by the use of a thin silicon layer that enhances the source/drain series resistances and degrades mobility [17]. This is why such architectures normally feature re–grown S/D (silicon is thicker outside the gate areas)[18];

2. the threshold voltage V_t depends on the depletion charge; due to small t_{Si}, this charge is low and V_t is low as well, unless N_A is further increased. A possible solution available for the designers is to increase the gate work–function, by adopting, for example, metal gates [19];

3. the back–oxide thickness t_{BOX} is an important parameter to set the subthreshold behavior: it is necessary to have $t_{BOX} \gg t_{Si}$, t_{OX} or the subthreshold slope S of long channel devices could be too large;

4. if a large N_A is adopted in order to control SCE, the mobility is degraded; moreover the device becomes more sensitive to the dopant fluctuations [20].

Even if SCE are detrimental in short–channel devices with undoped body, this effect can be kept under control by adopting an adequate t_{Si}, designed with appropriate scaling rules. Many L_G–t_{Si} relations have been proposed [21], [22]. Following [23] the body thickness should be chosen

$$6\delta < L_G < 8\delta \qquad \text{where } \delta = \sqrt{\frac{\varepsilon_{Si} t_{Si} t_{OX}}{\varepsilon_{OX}}} \tag{1.2}$$

ε_{Si} and ε_{OX} being the dielectric permittivities of silicon and of the gate oxide, respectively. It can be shown that, for gate length shorter than 50 nanometers, a body thickness smaller than 10 nm is required. The creation of such ultra small t_{Si} is still a major technological problem because of the difficult reproducibility of silicon films with the same small thickness.

Concerning multiple–gate devices, between the MugFETs it is possible to distinguish:

- *Planar Double–Gate (DG)* transistors (see Fig. 1.4): they use conventional planar manufacturing processes to create double–gate devices. In planar DG transistors the channel is sandwiched between two independently fabricated gate/gate oxide stacks. The primary challenge in fabricating such structures is achieving satisfactory alignment between the upper and lower gates [24];

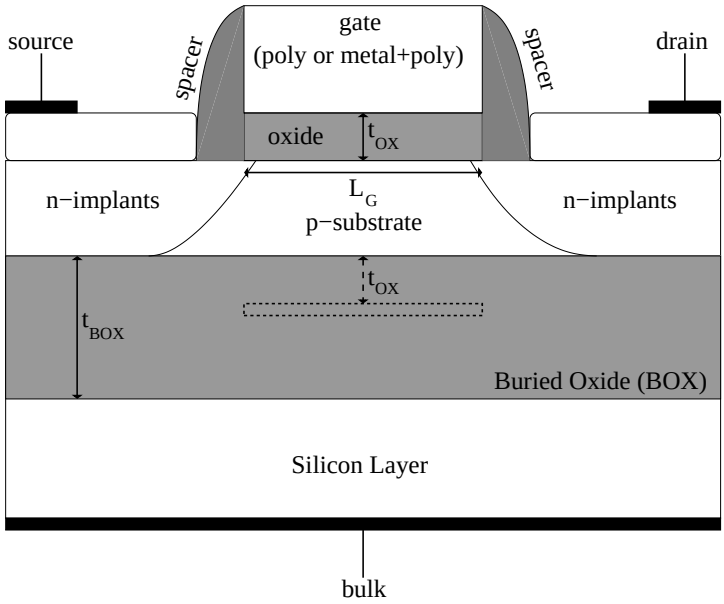

Figure 1.4: 2–D cross–section of a typical planar SOI transistor. The case presented is a n–channel MOSFET, and the figure is not drawn to scale. A thick buried oxide separates the active area from the silicon layer. The dashed–contoured region inside the BOX represents the second gate in the case of a planar Double–Gate transistor. The Bulk contact is called also substrate contact.

- *FinFETs*: non–planar transistor built on an SOI substrate (see Fig. 1.5). The distinguishing characteristic of the FinFET is that the conducting channel is formed in a thin silicon ”fin”, which forms the body of the device and is covered by the gate stack on three sides (see Fig. 1.5). The fin is characterized by two main design parameters: its width W_{fin} and its height H_{fin}.

Fig. 1.4 shows a schematic structure of planar Single– and Double–Gate SOI MOSFET, whereas Fig. 1.5 presents a simple sketch of a trigate FinFET. MugFETs family includes other kinds of multi–gate transistors, as Gate–All–Around FETs or Omega–FETs, that anyway are not the subject of this thesis work. All the MugFETs take advantage of the fully depleted behavior, and the FD concepts that have been presented for the Single–Gate transistor can be extended to the case of multiple–gate architectures; moreover, the FD operating mode has found an interesting application in the case of MugFETs because the presence of more gate contacts relaxes the L_G–t_{Si} relationship dictated by short channel effects and permits the use of thicker bodies. The control of the electrostatic potential within the body by the gate voltage is greatly enhanced and the short–channel effects are reduced in comparison with single–gate devices. The Double–Gate SOI MOSFET is the most common example of multiple–gate transistor and the research about it has increased in the last years. Like the SG MOSFET, the thickness of the active

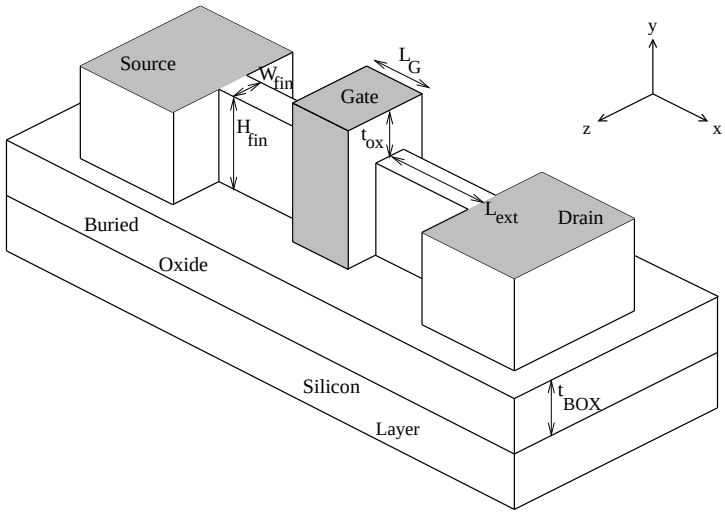

Figure 1.5: 3–D sketch of a typical FinFET, the figure is not drawn to scale. The most important geometrical parameters are reported. Metal gate and contacts are not drawn: only their interfaces to silicon are presented (grey shadowed regions).

area (film thickness t_{Si} in the case of planar DG device or fin width W_{fin} in the case of DG FinFET) is a dominant parameter. Following again [22] the expression for the characteristic length for short channel effects is:

$$\delta = \sqrt{\frac{\varepsilon_{Si} t_{Si} t_{OX}}{2\varepsilon_{OX}}} \tag{1.3}$$

Comparing 1.2 and 1.3 it can be demonstrated that, for the same gate length and the same oxide thickness, the DG transistor can have a larger t_{Si} than the SG case. This is positive from the point of view of the technological process and guarantees a better scalability. Moreover the thicker body reduces the series resistances that are detrimental in the short–channel SG devices.
From the point of view of current drive, the double–gate transistors can exhibit a drain current I_{DS} and transconductance g_m about twice the values of a SG transistor with the same area occupation, because of the two channels in parallel. The DG transistor also has some peculiar mobility enhancements:

1. Increase in the mobility when the channel is in moderate inversion due to *volume inversion* [25]. The name comes from the fact that the inversion charge is not concentrated at the two Si–SiO_2 interfaces but is distributed over the entire silicon layer or fin thickness. This a well–known advantage of the double–gate configuration;

2. higher surface roughness limited mobility. In fact, for fixed inversion charge, the DG case features a smaller effective field than the SG case [26].

The FinFET device has still many technological problems to solve, mainly related to its non–planar process and layout. The vertical dimension is the fin height H_{fin} and it plays the role of the width W for the planar MOSFET. Because it is not possible to build thin fins taller than about 100 nm, many FinFETs must be connected in parallel to obtain a relevant drain current.
While the research is advancing, the DG technology is more and more optimized. Many solutions under study for advanced SG devices, like metal gates, are going to be used in the DG case too. According to the current ITRS predictions, the Double–Gate technology may enter production in 2012.

Chapter 2

Simulation of electron devices

The aim of this chapter is to give some basics about the numerical simulation of electron devices as well as the most important details regarding the simulation tools that were used for this thesis work. First of all, we present the Boltzmann transport equation for transport in semiconductor devices, the drift–diffusion method and the electro–thermal method, and we give some details about the commercial simulation software that we adopted, *Sentaurus* from Synopsys [27]. Then a brief overview on a statistical approach largely adopted to study the carrier transport in electron devices, called *Monte Carlo method*, is given,

2.1 The Boltzmann transport equation

In order to evaluate the behavior of the carriers inside a semiconductor device we need to compute the *distribution function* $\mathcal{F}(\vec{r},\vec{p},t)$. The distribution function depends on the carrier momentum, carrier position and time and assumes values between 0 and 1. $\mathcal{F}d\vec{r}d\vec{p}$ describes the probability to find an electron/hole in a certain position $\vec{r}$ with momentum $\vec{p}$ at the instant t, and it is solution of the *Boltzmann Transport Equation* (BTE):

$$\frac{\partial \mathcal{F}}{\partial t} = -\nabla_{\vec{r}} \cdot \left(\frac{d\vec{r}}{dt} \mathcal{F} \right) - \nabla_{\vec{p}} \cdot \left(\frac{d\vec{p}}{dt} \mathcal{F} \right) + \left(\frac{\partial \mathcal{F}}{\partial t} \right)_C \tag{2.1}$$

The BTE represents a charge balance inside an elementary volume in the space ($\vec{r}$,$\vec{p}$). The first and second term in the right–hand side (RHS) of Eq. 2.1 are the net flux of $\mathcal{F}$ in the $\vec{r}$ and $\vec{p}$ space, respectively. The third term describes the collisions due to perturbations of the carrier motion caused by the interactions with semiconductor lattice. These interactions are called *scattering events* and the collision term can be expressed as

$$\left(\frac{\partial \mathcal{F}}{\partial t} \right)_C = \int_{\vec{p}} \left[\mathcal{S}\left(\vec{r},\vec{p}',\vec{p}\right) \mathcal{F}\left(\vec{r},\vec{p}',t\right) - \mathcal{S}\left(\vec{r},\vec{p},\vec{p}'\right) \mathcal{F}\left(\vec{r},\vec{p},t\right) \right] d\vec{p}' \tag{2.2}$$

where $\mathcal{S}(\vec{r},\vec{p},\vec{p}')$ is the probability of the collision event changing the electron momentum from $\vec{p}'$ to $\vec{p}$.
It is possible to rewrite Eq. 2.1 by considering that

$$\frac{d\vec{r}}{dt} = \vec{v}$$

and

$$\frac{d\vec{p}}{dt} = -q\vec{\mathcal{E}}$$

and by moving every term to the left hand side of the equation, with exception to the collision term:

$$\begin{aligned} &\frac{\partial \mathcal{F}}{\partial t} + \vec{v}\cdot\nabla_{\vec{r}}\mathcal{F} + q\vec{\mathcal{E}}\cdot\nabla_{\vec{p}}\mathcal{F} = \\ &= \int_{p'} \{\mathcal{S}(\vec{p}',\vec{p})\mathcal{F}(\vec{r},\vec{p}',t)[1-\mathcal{F}(\vec{r},\vec{p},t)] - \mathcal{S}(\vec{p},\vec{p}')\mathcal{F}(\vec{r},\vec{p},t)[1-\mathcal{F}(\vec{r},\vec{p}',t)]\} \end{aligned} \tag{2.3}$$

where $\vec{v}$ is the group velocity, $\vec{\mathcal{E}}$ is the electric field and q is the fundamental electronic charge.
The BTE is valid within the semi–classical approach, which assumes a classical description of the particle, while the scattering rates are calculated by quantum mechanics. The closed form solution of the BTE is particularly difficult even for the case of simple device geometries, mainly due to the form of the collision term: in the complete case, the BTE is a function of seven variables (time, three components of $\vec{r}$ and three components of $\vec{p}$).
There are two main methods for the solution of 2.1:

- *approximated methods*, where a set of simpler equations is derived from the BTE and then solved;
- *direct methods*, which need complex numerical calculations

The first approach includes the well–known "moments method" which is described next; the second one includes the statistical Monte Carlo approach.

2.2 The moments method

The moments method is based on:

- Reduction of the number of dimensions of the unknown variables. For this purpose the function $\mathcal{F}$ is replaced by its statistical moments;

- strong approximation of the collision term which is described by a single parameter τ. This parameter represents the characteristic time needed by the system to return to equilibrium.

In this simplified approach, the collision term present in the RHS of Eq. 2.1 can be written as

$$\left(\frac{\partial \mathcal{F}}{\partial t}\right)_C \simeq \frac{\mathcal{F}_{eq} - \mathcal{F}(\vec{r},\vec{p},t)}{\tau} \tag{2.4}$$

where $\mathcal{F}_{eq}$ indicates the distribution function in the equilibrium conditions, and τ is a microscopic relaxation time.
It is useful to express the distribution function in terms of velocity instead of momentum, because this approach facilitates the calculations of electrical currents. In equilibrium conditions, if we neglect the Pauli's exclusion principle, the Maxwell–Boltzmann distribution function should be adopted:

$$\mathcal{F}_{eq}(\vec{r},\vec{v}) = n(\vec{r}) \left(\frac{2\pi k_B T_o}{m^*}\right)^{-3/2} \exp\left(-\frac{m^* \mid \vec{v} \mid^2}{2k_B T_o}\right) \tag{2.5}$$

where $n(\vec{r})$ is the carrier concentration, T_o is the lattice temperature, k_B is the Boltzmann constant and finally m^* is the effective mass. The main feature of Eq. 2.5 is to have *spherical symmetry* in $\vec{v}$ with respect to the origin. The adoption of this Maxwell–Boltzmann distribution function is justified in equilibrium condition and in the absence of degeneracy.
The objective of the moments method is to reduce the dimensionality of the mathematical problem, and to obtain a system of equations where the statistical moments of the BTE appear. A first example of the moments methods is the *drift–diffusion* (DD) model. In order to reduce the number of dimensions, it is necessary to eliminate the dependence of BTE on $\vec{p}$. The dependence of the momentum $\vec{p}$ is eliminated by evaluating the statistical moments of the distribution function up to a given order.

Zero–order moment : It is given by

$$\int_{\vec{p}} \mathcal{F}(\vec{r},\vec{p},t) d\vec{p} = n(\vec{r},t) \tag{2.6}$$

where $n(\vec{r},t)$ is the number of carriers in a volume $d\vec{r}$ at a certain instant t. If we integer the first member of Eq. 2.1 over $\vec{p}$, we can obtain[1]:

$$\int_{\vec{p}} \frac{\partial \mathcal{F}}{\partial t} d\vec{p} = \frac{\partial}{\partial t} \int_{\vec{p}} \mathcal{F} d\vec{p} = \frac{\partial n}{\partial t} \tag{2.7}$$

[1] In the following, two notations will be used indifferently:

$$d\vec{p} = dp_x dp_y dp_z = d^3p$$

First–order moment : $\langle \vec{v} \rangle$ is the average velocity of the carrier population obtained by averaging the group velocity according to

$$\langle \vec{v} \rangle = \frac{\int_{\vec{p}} \vec{v}_G \mathcal{F} d^3p}{\int_{\vec{p}} \mathcal{F} d^3p} \tag{2.8}$$

where $\vec{v}_G$ is the *group velocity*. From 2.6 and 2.8 we obtain

$$\int_{\vec{p}} \vec{v}_G \mathcal{F} d^3p = n(\vec{r}, t) \langle \vec{v} \rangle \tag{2.9}$$

Second–order moment : $\langle v^2 \rangle$ is the *mean squared velocity*, defined as

$$\langle v^2 \rangle = \frac{\int_{\vec{p}} v_G^2 \mathcal{F} d^3p}{\int_{\vec{p}} \mathcal{F} d^3p} \tag{2.10}$$

$\langle v^2 \rangle$ is related to the kinetic energy of the carriers.

Higher–order moments can be obtained but the most common methods for solving the BTE consider only low–order functions. In particular the drift–diffusion model uses only the moments above, as we will describe in the next section.

2.2.1 The drift–diffusion model

The drift–diffusion model (DD) is the most popular approximated method to solve the Boltzmann transport equation. It is widely diffused in commercial device simulation tools.
In modern devices the quantum effects and many non–local effects (like hot–carriers and velocity overshoot) are difficult to track with the approximated methods. However the drift–diffusion model can be calibrated to improve its accuracy in the short–channel regime. This characteristic, together with its robustness and efficiency, explains its success.

This model involves three variables: electron concentration n, hole concentration p and electric potential ϕ. The mathematical system to solve includes five equations which are briefly described in the following paragraphs, without including the mathematical derivation from the BTE.

Poisson equation

It is the simplest equation of the drift–diffusion model and is the master equation of any electrostatic problem. It follows from Maxwell's equations under quasi–stationary

hyphotesis. For a silicon volume with both a donor dopant (concentration N_D) and acceptor dopant (concentration N_A), the equation is expressed as

$$\nabla^2\phi = -\frac{\rho}{\varepsilon} = -\frac{q}{\varepsilon}(N_D - N_A + p - n) \tag{2.11}$$

where ϕ is the electric potential, ε is the dielectric permittivity of the material, p is the holes concentration and n the electrons concentration.

Charge continuity equations

The charge continuity equation for electrons obtained as the 0^{th} order moment of the BTE reads:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \vec{J}_n + \mathcal{G} - \mathcal{R} \tag{2.12}$$

where $\vec{J}_n = -qn\vec{v}_{avg}$ is the electron current density, $\vec{v}_{avg}$ is the average velocity. The last term, $\mathcal{G} - \mathcal{R}$, substitutes the collision term of the BTE, and is the difference between the generation and recombination function. These functions represent the electron–hole pairs that are generated and recombinated in the volume unit and time unit.
In the case of holes, the charge continuity equation is

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \vec{J}_p + \mathcal{G} - \mathcal{R} \tag{2.13}$$

Current density equations

Equations 2.12 and 2.13 require two additional constitutive relations for the density currents $\vec{J}_n$ and $\vec{J}_p$. Here we consider a simple steady–state 1D case, for which the BTE can be written as

$$\frac{q\mathcal{E}_x}{m^*}\frac{d\mathcal{F}}{dx} + v_x\frac{d\mathcal{F}}{dx} = \frac{\mathcal{F}_{eq} - \mathcal{F}(v_x, x)}{\tau} \tag{2.14}$$

where $\mathcal{E}_x$ is the electric field along the x axis and the generation–recombination term is expressed using approximation 2.4. The current density can be expressed from the first–order moment of the BTE (2.14) as

$$J(x) = q\int v_x\mathcal{F}(v_x, x)dv_x = q\frac{q\tau}{m^*}\mathcal{E}_x n(x) - q\tau\frac{dn}{dx}\langle v^2\rangle \tag{2.15}$$

In Eq. 2.15, the second–order moment is evident in the last term. This quadratic dependence on v_x is simplified by approximating the average kinetic energy with the average thermal energy. This approximation assumes that carrier temperature is in equilibrium with the silicon lattice, and no hot–carriers effects are present. Thus, for the 1D carrier gas, we obtain

$$\frac{1}{2}m^*\langle v^2\rangle = \frac{1}{2}k_BT \quad \Rightarrow \quad \langle v^2\rangle = \frac{k_BT}{m^*} \tag{2.16}$$

Then we introduce the concept of *mobility*

$$\mu = \frac{q\tau}{m^*} \tag{2.17}$$

and the *diffusion coefficient* by using the Einstein's relation

$$D = \frac{\mu k_B T_o}{q} \tag{2.18}$$

The final expressions of the electron/hole current densities are

$$J_n = qn(x)\mu_n \mathcal{E}_x(x) + qD_n \frac{dn}{dx} \tag{2.19}$$

$$J_p = qp(x)\mu_p \mathcal{E}_x(x) - qD_p \frac{dn}{dx} \tag{2.20}$$

The drift–diffusion model is composed of equations 2.11, 2.12, 2.13, 2.19 and 2.20. This set of equations must be solved over the entire MOSFET under appropriate boundary conditions to obtain the potential ϕ and the charge densities n and p.

2.2.2 The thermodynamic model

The *thermodynamic* model (also named *non–isothermal* model, or *electro–thermal* model) extends the drift–diffusion approach to account for electro–thermal effects, under the assumptions that the charge carriers are in thermal equilibrium with the lattice. Therefore, the lattice temperature and the carrier temperature are described by a single temperature T, that in this case is not constant throughout the whole simulated structure. The thermodynamic model is defined by the basic set of partial differential equations already presented in Section 2.2.1: the Poisson equation 2.11, the electron continuity equation 2.12 and the hole continuity equation 2.13 with non uniform T. Eq. 2.19 and Eq. 2.20 can be rewritten as a function of the electric potential ϕ:

$$J_n = -qn(x)\mu_n \frac{\partial \phi(x)}{\partial x} + qD_n \frac{dn}{dx} \tag{2.21}$$

$$J_p = -qp(x)\mu_p \frac{\partial \phi(x)}{\partial x} - qD_p \frac{dn}{dx} \tag{2.22}$$

The electron current density 2.21 and the hole current density 2.22 are generalized to include the temperature gradient as a driving term:

$$J_n = -qn(x)\mu_n \left(\frac{\partial \phi(x)}{\partial x} + P_n \frac{\partial T}{\partial x}\right) + qD_n \frac{dn}{dx} \tag{2.23}$$

$$J_p = -qp(x)\mu_p \left(\frac{\partial \phi(x)}{\partial x} + P_p \frac{\partial T}{\partial x}\right) - qD_p \frac{dn}{dx} \tag{2.24}$$

where P_n and P_p are the absolute thermoelectric powers. Moreover, the thermodynamic model includes the lattice heat flow equation:

$$\begin{aligned} c\frac{\partial T}{\partial t} - \nabla \cdot kT = -\nabla \cdot [(P_n T + \phi_n)\vec{J_n} + (P_p T + \phi_p)\vec{J_p}] - \\ - (E_C + \frac{3}{2}k_B T)\nabla \cdot \vec{J_n} - (E_V - \frac{3}{2}k_B T)\nabla \cdot J_p + qR(E_C - E_V + 3k_B T) \end{aligned} \tag{2.25}$$

where k is the *thermal conductivity* and c is the *lattice heat capacity*; E_C and E_V are the conduction and valence band energies, respectively, and R is the recombination rate.

2.3 The Monte Carlo method

As we have already mentioned in Section 2.1, the Boltzmann transport equation is difficult to solve, so that a solution is usually found by using the approximated methods. An alternative way is based on direct methods, like the Monte Carlo approach, based on statistical calculation using random or pseudo–random numbers.
In its present form, the method is attributed to Fermi, Von Neumann and Ulam, who developed it for the solution of problems related to neutron transport [28, 29]. The MC method is currently adopted in a large number of applications, that include physical sciences (as physical chemistry, quantum chromodynamics or molecular dynamics), mathematics (it is useful to obtain numerical solutions to problems which are too complicated to solve analytically), finance and business.
Concerning the application of this method to carrier transport in semiconductor devices, historically the MC simulators were first used to study high–field effects in MOS transistors and transport in III–V devices, which already featured non stationary transport effects for micron and sub–micron gate lengths [30]. Now, this simulation approach is still used in short–channel devices because it can describe the transport within the channel more accurately than the drift–diffusion model. Quasi–ballistic transport, complex scattering models and carrier energy distribution can be properly simulated with the MC method. In this section, we present its basics; in Chapter 3 we will give more details about the MC simulator adopted in this work, as well as a wide description of the modifications that have been brought to the code in order to simulate semiconductor band–gap discontinuities.

2.3.1 Basic concepts of a Monte Carlo Device Simulation

The Monte Carlo method applied to electron transport in semiconductors simulates the motion of all particles inside a lattice, under the effect of a force $\vec{F}$ given by the presence of an electric field $\vec{\mathcal{E}}$ and of perturbations (also called *scattering events* or collisions).

The time between two successive collisions (named carrier *free–flight*) and the scattering events are selected stochastically based on appropriate probabilities. The duration of a scattering event is usually negligible with respect to the flight duration, so it is considered istantaneous. In the most general case, the motion of the carriers during the free–flight is determined by the following equations:

$$\frac{d\vec{r}}{dt} = \frac{1}{\hbar}\nabla_{\vec{k}}\left[E_\nu(\vec{k})\right] \tag{2.26}$$

and

$$\frac{d\vec{k}}{dt} = \frac{q}{\hbar}\nabla_{\vec{r}}\phi(\vec{r}) = -\frac{q\vec{\mathcal{E}}(\vec{r})}{\hbar} = \frac{\vec{F}(\vec{r})}{\hbar} \tag{2.27}$$

where

- q is the magnitude of the electron charge,
- $\vec{r}$ is the electron position vector,
- $\vec{k}$ is the electron wavevector,
- $\phi(\vec{r})$ is the electrostatic potential at the position $\vec{r}$,
- $\vec{\mathcal{E}}(\vec{r})$ is the electric field at the position $\vec{r}$,
- $\vec{F}(\vec{r})$ is the force induced by the electric field at the same position and
- $E_\nu(\vec{k})$ gives the electron kinetic energy of band ν as a function of $\vec{k}$.

The E–$\vec{k}$ relation is called *dispersion relation* and it takes into account the energy band–structure of the silicon lattice. A key advantage of the MC method is the possibility to include realistic models for the dispersion relation, based on look–up tables and obtained, for example, from the local empirical pseudopotential approach of Cohen and Bergstresser [31]. Because this characteristic permits a proper treatment of carriers with large energies and velocities, the analysis of highly out–of–equilibrium transport is one of the best field of application of the MC method.
However, this full–band approach is complex and it involves a large appetite in terms of computational resources, much larger than in deterministic methods as the drift–diffusion one. This aspect is unwelcome in a engineering context, and precludes direct application of the model to engineering device design, therefore simplyfied models are adopted. As an example, near the conduction and valence bands minima the dispersion relation is often approximated by a *parabolic* relationship between the energy E and the wavevector $\vec{k}$:

$$E(\vec{k}) = \pm\frac{\hbar^2 k^2}{2m^*} \qquad \text{being } k = \mid \vec{k} \mid \tag{2.28}$$

where the positive sign is for the conduction band and the negative sign is for the valence band, and m^*, defined as

$$\frac{1}{m^*} \equiv \frac{1}{\hbar^2}\frac{\partial^2 E(\vec{k})}{\partial k^2}$$

is the effective mass of the particle. The energy zero is taken at the band extrema, so $E(\vec{k})$ represents the carrier's kinetic energy. Since $\hbar^2 k^2$ can be shown to be the electron's momentum $\vec{p}$, from Eq. 2.28 the energy and momentum are related as they are for a free electron, so that Eq. 2.26 and Eq. 2.27 can be rewritten as:

$$\frac{d\vec{r}}{dt} = \nabla_{\vec{p}}\left[E(\vec{p})\right] = \vec{v} \tag{2.29}$$

$$\frac{d\vec{p}}{dt} = (-q)\vec{\mathcal{E}} = \vec{F} \tag{2.30}$$

See [32] for a more exhaustive overview on quantum mechanics.
The force $\vec{F}$ accelerates the carrier, thus increasing its energy E and momentum $\vec{p}$. The movement is stopped at a certain instant by a perturbation. The time between two perturbations defines the duration t of the free–flight, during which Eq. 2.29 and 2.30 are integrated for evaluating the variation of $\vec{r}$ and $\vec{p}$:

$$\vec{r}(t) = \vec{r}(0) + \int_0^t \vec{v}(t')\,dt' \tag{2.31}$$

$$\vec{p}(t) = \vec{p}(0) + \int_0^t \vec{F}(t')\,dt' \tag{2.32}$$

At the end of a free–flight, the position and momentum of the particle are updated.

As an example, we can refer to an electron that has moved in a three–dimensional lattice under the influence of a constant electric field directed along the z axis. After moving for a time t under the influence of the field, the electron's momentum and position are obtained from Eq. 2.29 and 2.30:

$$p_x(t) = p_x(0) \tag{2.33}$$
$$p_y(t) = p_y(0) \tag{2.34}$$
$$p_z(t) = p_z(0) + (-q)\mathcal{E}_z t \tag{2.35}$$

$$x(t) = x(0) + \frac{p_x(0)}{m^*}t \tag{2.36}$$
$$y(t) = y(0) + \frac{p_y(0)}{m^*}t \tag{2.37}$$
$$z(t) = z(0) + \left(\frac{E(t) - E(0)}{(-q)\mathcal{E}_z}\right) \tag{2.38}$$

where the second term on the RHS of Eq. 2.38 is due to the fact that $v_z = \partial E/\partial p_z$ (from Eq. 2.29) and $\partial p_z/\partial t = (-q)\mathcal{E}_z$ (from Eq. 2.30), and therefore

$$\int_0^t v_z(t')dt' = \int_0^t \frac{\partial E}{\partial p_z}dt' = \int_0^t \frac{\partial E}{-q\mathcal{E}_z} = \frac{E(t) - E(0)}{(-q)\mathcal{E}_z}$$

In Eq. 2.36, 2.37 and 2.38 we have used a parabolic relation between the energy E and the momentum $\vec{p}$

$$E(t) = \frac{p^2(t)}{2m^*} \tag{2.39}$$

The determination of the free–flight t_{FF} is a very difficult task because it depends on the scattering rate Γ which is the frequency of collisions: the higher the frequency, the shorter the time $\tau = 1/\Gamma$ between two collisions, which is the duration of the free flight. The rate Γ is a function of the scattering probability $\mathcal{S}$, which always depends on the final energy $E(t_{FF})$. Collisions can be due not only to lattice vibrations (phonons), but also to the ionized impurities, to the roughness of the Si–SiO_2 interface, to the other carrier (plasmons). If more sources of perturbation are involved, the total collision rate is

$$\Gamma(E) = \sum_{i=1}^{k} \Gamma_i(E) = \sum_{i=1}^{k} \frac{1}{\tau_i(E)} \tag{2.40}$$

where the sum is done over all sources of scattering.
In order to overcome the difficulties related to the calculation of t_{FF}, a constant Γ scheme is adapted. For a given constant rate $\Gamma = \Gamma_0$, it can be shown that the duration of the free–flight is given by

$$t_{FF} = -\frac{1}{\Gamma_0}\ln(r_c), \tag{2.41}$$

where r_c is a casual number uniformly distributed between 0 and 1.
Once t_{FF} is found from Eq. 2.41), the simulation of the particle free–flight can proceed as follows:

1. the movement of the particle is evaluated from Eq. 2.29 and Eq. 2.30 while the final momentum and position are updated using expressions similar to Eq. 2.33–2.38;

2. we need to find which source of scattering has caused the end of the free flight. From the knowledge of $E(t_{FF})$, we can calculate $\Gamma_i(E(t_{FF})) = 1/\tau_i(E(t_{FF}))$ for each type of scattering. The total scattering rate, given by Eq. 2.40, must be lower than Γ_0;

3. the rates Γ_i are normalized by Γ_0. This will result in a scale of probability from 0 to 1 where each type of collision has a certain probability to happen;

4. because the total rate Γ is less than Γ_0, a certain range of the probability scale does not correspond to any type of collision. This is treated as a new scattering mechanism, denoted as *self–scattering*;

5. a random number is mapped on the probability scale and identifies the collision event that stopped the flight. If a real collision has happened, the state $(E, \vec{p})$ of the particle is updated following the model of the right scattering source. If self–scattering has happened, the state of the particle remains the same as at the end of point 1.

The cycle is repeated over time and the state of the particle is recorded to create the statistics of energy, velocity, etc. The simulation can be stopped after a reasonable number of steps or after the convergence has been achieved.

2.3.2 Ensemble Monte Carlo

In the previous section we have seen an example of cyclic algorithm to simulate one particle in a semiconductor lattice under the influence of an electric field. In a device many particles are present and must be simulated in the proper way to describe the complete behavior of the device. In general, there are two types of Monte–Carlo simulation for semiconductor devices:

ensemble : it calculates the trajectories of all particles at the same time

incident flux : it simulates one particle for a certain time, builds up the statistic and then considers another particle.

The first approach is the most popular, so we will discuss only the ensemble Monte Carlo.

In the ensemble MC method, the two dimensional MOSFET is divided into cells by a numerical grid. Each cell is populated with carriers with a given charge *weight*, that is the quantity of charge associated to each particle. In fact, because the number of particles in a real device is very large, only a smaller number can be handled by the MC code. This number must be representative of the entire carrier population. Each simulated carrier has not an elementary charge q, like electrons or holes, but has a charge $Q = kq$, where k can be different from one particle to the other. The new *super–particle* represents the charge of many carriers. Together with k, it is necessary to set an initial energy and momentum for each created particle. This assignment can be done by randomly choosing a sample from a Maxwellian or Fermi–Dirac distribution. As a final step the carriers must be distributed over the grid in a “smart“ way. The choice of the

carrier distribution and of the initial electrostatic potentials is usually based on a known solution, which is the initial condition of the MC simulation.
At this point, the simulation can start: the momentum, energy and position of all particles are traced by the techniques discussed in Sec. 2.3.1. Poisson's equation must be solved after each simulation step to update the electric field. At any time during the simulation the average carrier density, velocity, energy versus position can be computed by averaging over the particles within each cell. The whole process is repeated until numerical convergence is achieved.
This type of simulation approach presents many issues that must be handled carefully, like the treatment of the boundary conditions and the rules for creating the grid. The number of particles involved in a single simulation can be very large and limits the computational efficiency. In particular the choice of the scattering events and the calculation of the final state for all particle after each flight can add a lot of simulation time. This last issue is the drawback of the possibility to include very complex models of the scattering events. This is an important characteristic of the MC approach, that does not reduce the entire analysis of the collisions to a single mobility value, like the drift–diffusion model. This observation is a key point to understand why the Monte Carlo is useful to study the transport in MOSFETs with very short channel length. In particular the quasi–ballistic transport regime cannot be accounted properly in a drift–diffusion simulation, while MC can handle particles that experience few collision events within the channel (*ballistic transport*).

Part I

Band–Gap Engineering in DG SOI MOS Transistors

Chapter 3

Monte Carlo simulation of heterojunctions

This chapter outlines the problem concerning the use of different semiconductor materials, the adoption of such a type of structure in microelectronics and how this problem can be handled by simulation.
In Section 3.1 a brief description of heterojunctions is presented, as well as a typical electron device that have been extensively explored in the past, the heterojunction bipolar transistor. Then we give an overview on how this approach has been adopted in the MOSFET case. In Section 3.2 the basics features of our Monte Carlo tool (named *Band.it*) are presented; in Section 3.3 we detail the modifications that have been introduced in the code in order to simulate electron band–gap discontinuities, as well as the simulations of simplified structures that have been performed in order to validate them in Section 3.4.

3.1 Heterojunctions and Band-Gap Engineering

Most of the transistors that are integrated in circuits currently on sale are based on silicon technology. This is true for MOS transistors, where source, drain and channel region are made of silicon, as well as for older technologies as the bipolar one, where the regions that characterize such a type of devices (base, emitter and collector) were made of the same material. As we have detailed in Chapter 1, for the forthcoming technological nodes the Bulk architecture will become inadequate to sustain the scaling trends foreseen by the ITRS. We have already mentioned in Section 1.3 that researchers are making a strong effort in order to detect possible ways to overcome the limitations of the conventional silicon–based Bulk transistor, and one of them is represented by the adoption of new technologies in order to realize the source and drain accesses. According to this approach the transistor could be made not only of silicon: different materials

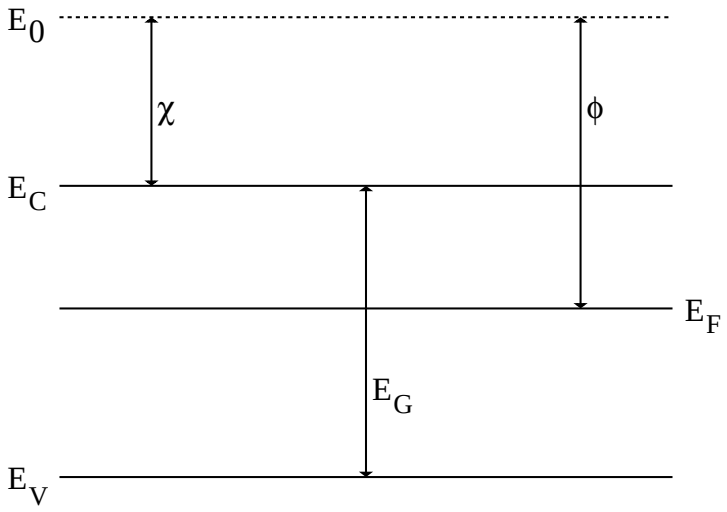

Figure 3.1: Energy bands for an uniform semiconductor. The location of the conduction band is measured with respect to a reference level, E_0, the field–free vacuum level.

may be adopted for the drain, source and channel regions.

When two different semiconductor materials are in close contact, we have a *heterojunction* (HJ). In other words, a HJ is a junction formed between two dissimilar semiconductors. When the two semiconductors have the same type of conductivity, the junction is called an *isotype* heterojunction; otherwise, when the conductivity types differ, the junction is called *anisotype*. In the past fifty years HJ have been extensively studied, and introduced in many important applications; among them the room–temperature injection laser, light–emitter diode, photodetector and solar cell. The main feature that characterizes a HJ with respect to a homojunction is that the materials involved have different energy band–gap; the combination of multiple heterojunctions together in a device is called a *heterostructure*, although the two terms are commonly used interchangeably. Even if an extensive description of HJs and of the models that have been proposed in order to describe them is out of the scope of our work, it could be useful to briefly explain the basic concepts needed to draw energy band diagrams for heterostructure devices.

Let's begin by re–examining the energy band diagram for a uniformly doped, compositionally uniform semiconductor as shown in Figure 3.1. The position of the conduction band E_C and the valence band E_V are determined by the chemical bonding of the atoms and can be measured or calculated by solving the Schrödinger equation. These energy levels must be measured relative to some reference level, that usually is the *vacuum level* E_0, that is the energy of a free electron just outside the neutral semiconductor. The electron affinity χ is the energy needed to remove from the semiconductor an electron located at E_C and make it free. The work–function ϕ is the distance between E_0 and the

Fermi level E_F. For the case presented in Figure 3.1 we have

$$E_C = E_0 - \chi \tag{3.1}$$

$$E_V = E_C - E_G = E_0 - \chi - E_G \tag{3.2}$$

For homostructures, the electron affinity χ and the band–gap E_G are position–independent; in the case of HJ, this statement looses validity. If we consider two different semiconductors 1 and 2 and we keep them in close contact, it is possible to obtain different cases:

1. the conduction and valence bands of the smaller band–gap semiconductor lie completely within the band–gap of the wider band–gap one. This possibility is illustrated in Figure 3.2 and it is known as *type I heterojunctions*. Heterojunction pairs of III–V compounds in which either the group III or the the group V element differ, form this type of HJ. Examples include AlAs/GaAs and GaP/GaAs HJ;

2. the conduction and valence bands of the smaller band–gap semiconductor straddle the valence band of the larger band–gap semiconductor, the interface is known as *type II heterojunctions*. III–V heterojunction pairs in which both the group III and group V elements differ (e.g. GaSb/InAs) form type II heterojunctions; examples of this type include $In_xGa_{1-x}As/Ga_xSb_{1-x}As$ and $Al_xIn_{1-x}As/InP$ systems;

3. E_C of one semiconductor lies below E_V of the other. In this case transport is complicated by the fact that the electron wavefunction changes from electron–like to hole–like as the electrons moves across the heterojunction.

The band diagrams of Fig. 3.1 and Fig. 3.2 ignore electrostatic potentials due to re–arrangement of mobile carriers which occurs near the compositional junction after the semiconductors are placed in contact. When this happens, electrons move from the semiconductor with the higher Fermi level to the other, and an electric field is produced to balance this transfer. The band diagram for the heterojunction is deduced conceptually just as it was for homojunctions; if we consider a type I HJ, with an n–type wider band–gap semiconductor and a p–type smaller band–gap one (as reported in Fig. 3.2), the energy band diagram of the formed junction is reported in Fig. 3.3. Calculating the energy band offsets for an ideal heterojunction is straightforward given the material properties of the two materials using the *Anderson's rule* (also named *electron affinity rule*). The conduction band offset depends only on the electron affinity difference between the two semiconductors:

$$\Delta E_C = \chi_1 - \chi_2 = \Delta\chi \tag{3.3}$$

then using the change in band–gap:

$$\Delta E_G = E_{G2} - E_{G1} \tag{3.4}$$

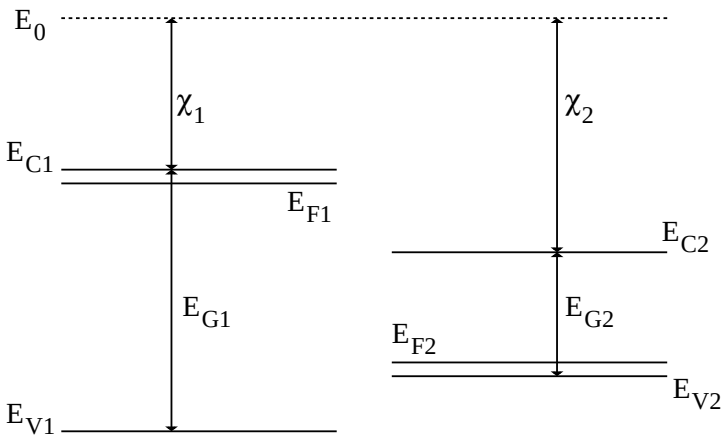

Figure 3.2: Simple plot of a type I heterojunction, before the contact: the smaller band–gap semiconductor is p–type doped and its conduction and valence band lie completely within the band–gap of the wider band–gap one (n–type doped).

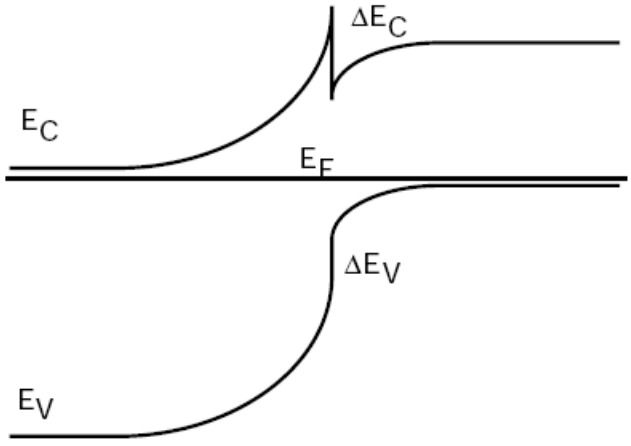

Figure 3.3: Simple plot of a type I heterojunction, after the contact. The involved materials are the same presented in Fig. 3.2.

The valence band offset is simply given by:

$$\Delta E_V = \Delta E_G - \Delta\chi \tag{3.5}$$

Which confirms the trivial relationship between band offsets and band gap difference:

$$\Delta E_G = \Delta E_C + \Delta E_V \tag{3.6}$$

In the frame of the Anderson's idealized model these basic material parameters are assumed unchanged when the materials are brought together to form an interface; therefore, the quantum size effect, defect states and other perturbations which may or may not be the result of imperfect crystal lattice matches are disregarded. When two materials are brought together and allowed to reach chemical/thermal equilibrium, the Fermi level is constant throughout the system. Electrons in the materials leave some regions (depletion) and build up in others (accumulation) in order to find equilibrium. When this occurs a certain amount of band bending occurs near the interface. This total band bending can be quantified with the built–in potential given by:

$$V_{bi} = \phi_1 - \phi_2 = (E_{G1} + \chi_1 - \Delta E_{F1}) - (\chi_2 + \Delta E_{F2}) \tag{3.7}$$

where $\Delta E_{F1} - E_{V1}$ and $\Delta E_{F2} = E_{C2} - E_{F2}$. The built–in potential gives the degree to which band bending occurs, but tells us nothing about the details of spatial–dependence of band–bending. In order to work–out the spatial–dependence of band–bending, we must know the density of states and state occupation given by the Fermi-Dirac distribution.
Finally, Table 3.1 reports E_G and χ for some semiconductor materials. It should be noticed that in real semiconductor heterojunctions, Anderson's model fails to predict actual band offsets, because it ignores the fact that each material is made up of a crystal lattice whose electrical properties depend on a periodic structure of atoms. This periodicity is broken at the heterojunction interface to varying degrees. In cases where both materials have the same lattice, they may still have different lattice constants which give rise to crystal strain which changes the band energies. In other cases the strain is relaxed via dislocations and other interfacial defects which also change the band energies.
A deeper and more extensive description of heterostructures theory could be found in fundamental books on semiconductor physics as [33].

In recent years, HJ have already been exploited in the frame of bipolar technologies. The bipolar–junction transistor (BJT) has been improved to the *heterojunction bipolar transistor* (HBT): HBT technology has become an area of intense resarch in universities and industry worldwide. The main difference between the BJT and HBT is the use of different semiconductor materials, creating a heterojunction, because a wide band–gap emitter is used; moreover, a compositionally graded base or a doping graded base is used to add a drift component to the carrier transport for further improvement.

	E_G [eV]	χ [eV]
Si	1.12	4.01
Ge	0.66	4.13
GaAs	1.43	4.07
AlAs	2.16	2.62
GaP	2.21	4.3
InAs	0.36	4.9
InP	1.35	4.35

Table 3.1: Energy band—gap (E_G) and electron affinity (χ) for some of the most important semiconductors; data from [34].

A wider emitter layer allows the use of a lower doping concentration leading to a low emitter region capacitance. In addition, the emitter–base band discontinuity partially blocks the back–injection of base majority carriers into the emitter: being k_B the Boltzmann constant and T the absolute temperature, the reduction in the reverse injection is, in the first approximation, of the order of the exponential of the ratio of the appropriate band discontinuity (valence band discontinuity ΔE_V for npn transistors, conduction band discontinuity ΔE_C for pnp ones) to $k_B T$. The reduction of emitter capacitance reduces the emitter junction charging time. In HBTs with a graded base, a drift field is present which accelerates injected carriers from emitter to collector, reducing the base transit time. All of the effects mentioned above enhance the current gain and the cut–off frequency beyond those achievable with BJTs. Wide gap emitter also allows very high base doping levels without degrading the emitter injection efficiency, so that a small base resistance can be obtained: in this way the maximum oscillation frequency is increased without degrading the current gain. Figure 3.4 shows a simple sketch of the band diagram for an npn HBT.

Boosting superior performance over silicon bipolar transistors with its combined high speed, high linearity, and high power requirements, the III–V HBT is fast becoming a major player in wireless communication, for application in power amplifiers, mixers, and frequency synthesizers. This solution can handle signals of very high frequencies up to several hundred GHz: few years ago an HBT built from indium phosphide and indium gallium arsenide and designed with compositionally graded collector, base and emitter, was demonstrated to cut off at a speed of 710 GHz [35].

An extensive treatment of HBT is out of the scope of this thesis work; a detailed explanation regarding such a type of devices can be found in the books of Weisbuch [36] or Liu [37].

Recently, attempts have been made in order to replicate the HBT idea in MOSFETs: as the carrier velocity at the source edge is limited by thermal velocity or Fermi velocity, an application of the concept of high–velocity carrier injection in the HBT technology

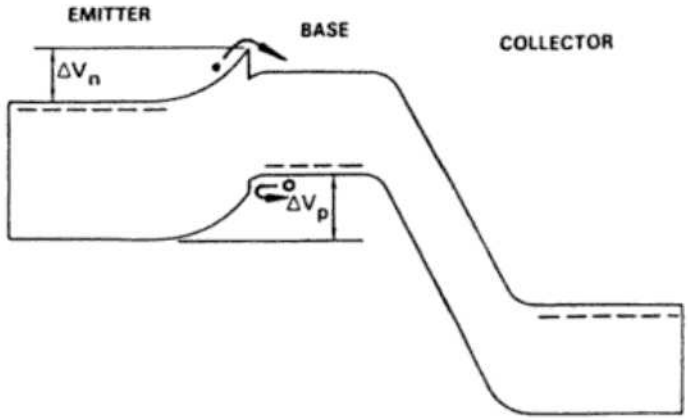

Figure 3.4: Plot of the energy band diagram for an npn HBT. It is evident how the valence band discontinuity plays a key role in impeding the reverse injection.

	Electron affinity χ	[A]	[B]	[C]
Source	χ_1	Relaxed–SiGe	Relaxed–Si	Relaxed–$Si_{1-x}C_x$
Channel	χ_2	Strained–Si	Strained–Si	Si

Table 3.2: Three possible candidates of the source and channel materials for the heterojunction source structures, satisfying the condition that $\chi_2 > \chi_1$ and therefore the channel has a smaller band–gap than that of the source.

to the source/channel edge in MOSFETs can provide the breakthrough of the above physical limitation on the carrier velocity in conventional MOSFET structures, leading to an increase of the injection velocity and therefore of the provided current. Mizuno and his research group proposed a novel SOI MOS structure named SHOT (Source–Heterojunction–MOS–transistor) with an heterojunction source structure for realizing high–velocity electron injection into the channel [38, 39, 40]. In order to realize such a type of devices, the conduction band level of the source region has to be higher than that of the channel, that is the source region must have an electron affinity χ smaller than the channel one. This source band offset structure with the conduction band energy difference between the source and the channel, ΔE_C, allows the injection of higher–velocity electrons into the channel region by using excess kinetic energy due to ΔE_C. Although several HJ structures could be good candidate to satisfy this condition (see Table 3.2), solution [A] proposed in the second column of the table (relaxed–SiGe layers as the source and strained–Si layers as the channel) is the simplest one and the fabrication process to realize it is the easiest one.

3.2 The Monte Carlo tool *Band.it*: basic features

Band.it is a Monte Carlo simulator for the three–dimensional electron gas (3DEG), that treats electrons in the device as a free-carrier gas. It implements the coupling between the Monte Carlo transport and the Poisson equation through a linear coupling scheme [41, 42]. Quantum-mechanical corrections could be introduced by the *effective potential* approach proposed in [43]: the electrostatic potential is corrected in order to include the effect of carrier quantization on the spatial distribution of the inversion charge. Its implementation in the MC code is described in [44], where the authors point out that this technique, applied to a variety of deeply scaled MOSFETs, reproduces with reasonable accuracy integral quantities such as the total inversion charge, even if it fails to reproduce the concentration profile within several angstrom from the Si/SIO_2 interface.
The effective potential is defined as:

$$V_{eff}(x,y) = \int\int V(x',y')G(x'-x,y'-y)dx'dy' \tag{3.8}$$

where the potential energy profile $V = -q\phi + \chi$, including both the electrostatic potential (ϕ) and the electron affinity (χ), is smoothed by a Gaussian function $G(\xi,\zeta)$; the standard deviation of the Gaussian is chosen in order to reproduce the inversion charge density of a coupled Schrödinger-Poisson solver even in devices with very thin gate oxides and over a wide range of voltages. The initial solution of the MC analysis, is calculated from the solution of a drift–diffusion simulations of the same device structure, performed with a commercial simulation tool (Sentaurus, from Synopsys)
Besides the phonon scattering the MC code includes a model for ionized impurities scattering, which follows the usual 3DEG formalism.
Electron-plasmon scattering inside the heavily doped regions could be included as well. This scattering mechanism plays an important role because it thermalizes the particles in the source and drain regions. The carrier-plasmon interaction is a very strong inelastic scattering and has to be included when simulating quasi-ballistic transport, since the amount of back-scattered carriers depends on the balance between elastic and inelastic scattering. Finally a model for the surface roughness (SR) scattering is necessary, since the mobility of MOSFETs in the ”ON” state is limited by this scattering mechanism. In most of the MC simulators, surface roughness scattering is usually modeled with a specular-diffusive reflection of the particles hitting the $Si–SiO_2$ interface, and the percentage of diffused particles is adjusted to fit experimental data [45]. However, the effective potential repels the carriers from the surface and almost none of them can reach the interface; therefore the specular-diffusive approach cannot be used. As a consequence, surface roughness is included as an additional scattering mechanism, whose scattering rate is an increasing function of the effective field component normal to the silicon–dielectric interface [46]. Bandit features an original approach to adapt the SR scattering model for a 2D electron gas to the full–band 3D electron gas MC corrected

by the effective potential. The scattering rate is calculated from the effective field $\mathcal{E}_{EFF}$ since, as documented in [47], the experimental mobility is an unique function of $\mathcal{E}_{EFF}$. The details of this model can be found in [46].
The parameters of the scattering mechanism have been adjusted in order to reproduce the universal mobility curves in unstrained silicon inversion layers [48]. This tool suffers limitations as it does not account for the 2D–gas sub–band structure and for its effects on the phonons and surface–roughness scattering rate. Nonetheless, since the deformation potentials for acoustic phonons and the parameters for SR–scattering have been adjusted in order to fit the mobility curves of both Bulk and SOI devices [17, 46, 48], reasonable accuracy in terms of terminal-currents can be expected. Furthermore, scattering mechanisms that assume an increasingly important role as the silicon film thickness t_{Si} is scaled below 10 nm such as surface optical phonons and the effects of body-thickness fluctuations [49] are not included. For this reason, the simulated current may be overestimated for ultra-thin body SOI MOSFETs. In [50] the results obtained by the MC simulator adopted in this work have been compared with those of a MC simulator for a 2D confined electron gas that explicitly accounts for the effects of quantization on the dispersion relation and on the scattering rates [51]. The results of this comparison confirm that the simulation approach adopted in this work provides terminal currents in good agreement with the more accurate simulator for the 2D electron gas, at least for the devices of interest in this work.

3.3 Implementation of Monte Carlo transport across heterojunctions

In this work the effect of band–gap discontinuity is analyzed with emphasis on the effects on microscopic quantities such as injection velocity, inversion charge in the channel and electrostatic potential profile, without considering the fact that the discontinuity is obtained by using materials with transport properties different than those of pure Silicon. Although a discontinuity of semiconductor composition affect transport properties, such effects are disregarded and we assume the transport properties of pure unstrained silicon within the whole device.
In order to handle heterojunctions, appropriate models are needed to describe a carrier crossing them. In particular, as we worked on n–channel devices, the modifications that have been apported to our MC code regard only electron transport only.
We have considered two cases:

1. *graded* heterojunctions: the conduction band profile varies gradually from a region to another; the presence of the HJ is treated by using a quasi–field, added during the electron free–flight to the electric field determined by the Poisson

equation. In other words, the carriers are moved based on the conduction band edge, instead of using only the electrostatic potential (the driving force is thus $F{=}-qdE_C/dx$). This approach has already been adopted in the past in the case of HBTs [52, 53].

2. *abrupt* heterojunctions: the conduction band profile has a stepwise shape.

The case of abrupt HJ is more complicated than the graded one, and it requires different procedures depending on the energy of the carriers involved. First of all, new features have been added to our mesh generator, named *Inband.it*. It is now possible to define a new interface (called *hetero*) which describes:

- *orientation* of the discontinuity: it is possible to define if the band–gap offset occurs along the x–direction (direction from source to drain) or along the y–direction (direction from the Si–SiO_2 interface to the buried oxide). It should be noticed that in the remainder of this work we will consider only the first case, because the source/drain region are made of materials different than silicon and therefore the discontinuity occurs only along the source–drain direction;

- *position* of the discontinuity: it is possible to define the exact coordinate along x– or y–direction where the offset occurs;

- *value* ΔE of the offset, expressed in meV: when a carrier crosses an HJ its kinetic energy is increased by ΔE if entering a region with larger electron affinity or is decreased by ΔE while entering a region with lower affinity.

The treatment of abrupt HJ requires different procedures depending on the energy of the carrier involved. Let's start by assuming a carrier travelling in the positive direction of the x–axis, and a conduction band offset occurring along the same direction. We name E_L the initial carrier energy, that is the carrier kinetic energy *before* it crosses the abrupt HJ, and E_R the final carrier energy, that is the carrier kinetic energy after the carrier crosses it. We are using the subscripts L and R because in this example the carrier goes from the left to the right.
If both E_L and E_R are lower than 75 meV, a parabolic model provides a reasonable approximation for the band structure, so that it is possible to separate the total carrier energy into components associated with the different transport directions. Three possible cases exist, that are described in Figure 3.5:

1. the electron enters a region with lower electron affinity and its kinetic energy $E_L{=}\hbar^2 k_{xL}^2/(2m_x)$ in the x–direction is larger than ΔE: in this case the particle can enter the lower affinity region and its kinetic energy in the x–direction decreases by ΔE, that is $E_R = E_L - \Delta E$.
   In this case we set $k_{xR}{=}\sqrt{k_{xL}^2 - 2m_x\Delta E/\hbar^2}$; k_y and k_z, the components in the

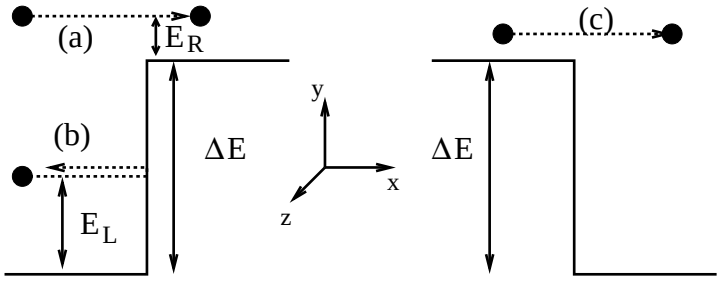

Figure 3.5: Possible cases of electrons traveling in the positive x–direction and crossing an abrupt conduction band discontinuity.

plane normal to the direction of the motion, are not modified. This is the case (a) in Fig. 3.5;

2. the electron enters a region with lower electron affinity, but $E_L{<}\Delta E$; in this case the carrier can not overcome the barrier, therefore it is reflected (k_x is inverted), and the reflection is treated as perfectly elastic; k_y and k_z do not change. This is the case (b) shown in Fig. 3.5;

3. the electron enters a region with higher electron affinity; in this case it can cross the barrier regardless of its energy, and the energy component in the x–direction increases by ΔE, that is $E_R = E_L + \Delta E$.
In this case we set $k_{xR}{=}\sqrt{k_{xL}^2 + 2m_x\Delta E/\hbar^2}$; k_y and k_z are unchanged. This is the case (c) presented in Fig. 3.5.

If one among E_L or E_R, or both E_L and E_R are higher than 75 meV, detailed full–band effects are considered as follows:

- the total kinetic energy is increased or decreased by ΔE depending on the different electron affinities and on the direction of the motion, as already described above: it the carrier enters a region with higher electron affinity, we set $E_R = E_L{+}$ ΔE, otherwise if it enters a region with lower electron affinity, $E_R = E_L - \Delta E$.

- The wavevector on the plane normal to the motion is conserved: if the carrier is travelling in the x–direction, k_y and k_z do not change. k_{xR} is found by searching states in the full–band structure having $k_{yR}{=}k_{yL}$, $k_{zR}{=}k_{zL}$, total energy $E_R{=}E_L{\pm}\Delta E$ and conserving the direction of the group velocity of the incoming electron (v_G). All possible states in the first Brillouin zone (FBZ) are considered. Due to the symmetry of the FBZ, sets consisting of either two or four values k_{xR} are found. Among all possible states, only those conserving the direction of v_G are considered, and one of them is selected randomly. Different approaches for this latter selection have been tried, but no significant dependence on the specific selection methodology has been found.

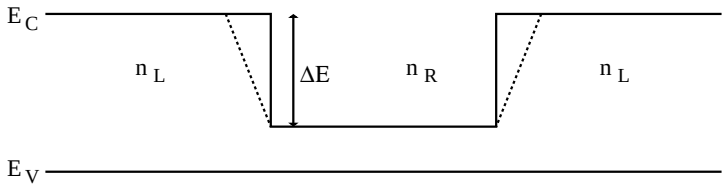

Figure 3.6: Conduction (E_C) and valence (E_V) band profiles for the simplified unidimensinal structure that we have used to validate the code. n_L is the carrier concentration in the wide band–gap region, while n_R is the carrier concentration in the small band–gap one. The dashed line represents the graded HJ, that has been implemented either as a quasi–electric field and as a series of 4 small abrupt HJs.

- If states satisfying the above criteria do not exist (e.g. the case (b) presented in Fig. 3.5) the electron is reflected by an elastic process and k_x is inverted.

At very large energies the conduction band features many branches making the selection of the final state extremely complicated. For this reason, we relax the conservation of k_y and k_z at energies above 500 meV (notice that only a negligible number of electrons is concerned): the total kinetic energy is increased/decreased by ΔE, and the $\vec{k}$–state is randomly selected, with the only constraint of conserving the direction of the group velocity of the incoming electron.
As we have seen at the beginning of this chapter, in presence of heterostructures an energy spike arises, and electrons can normally travel through this energy barrier due to the tunnel effect. In our work this effect is neglected, that is only the current due to thermionic emission is taken into account, while tunnel effect is not considered. In the case of a MOS transistor, this implies that we are considering the worst case in terms of the provided current, because the contribution by the electrons that can enter the channel by tunnel effect are neglected.

3.4 Model verification

In order to validate our model for HJs, we have carried out simulations of a simplified, unidimensional template structure, featuring uniform n–type doping concentration N_D=10^{19} cm^{-3} and two symmetric conduction band offsets with the same ΔE (the electron affinity in the center is ΔE higher than at the two ends). The valence band has been kept constant througout the whole device. Fig 3.6 reports the conduction and valence band profiles of the simulated structure.
First, we have run non–self–consistent simulations with a null electric field and neglecting the Pauli exclusion principle. In the absence of any electric field, at the two sides of the conduction band offsets we should have carriers concentrations n_L (low affinity) and n_R (high affinity) verifying the equation n_R/n_L=$\exp(\Delta E/k_BT)$, where k_B is the

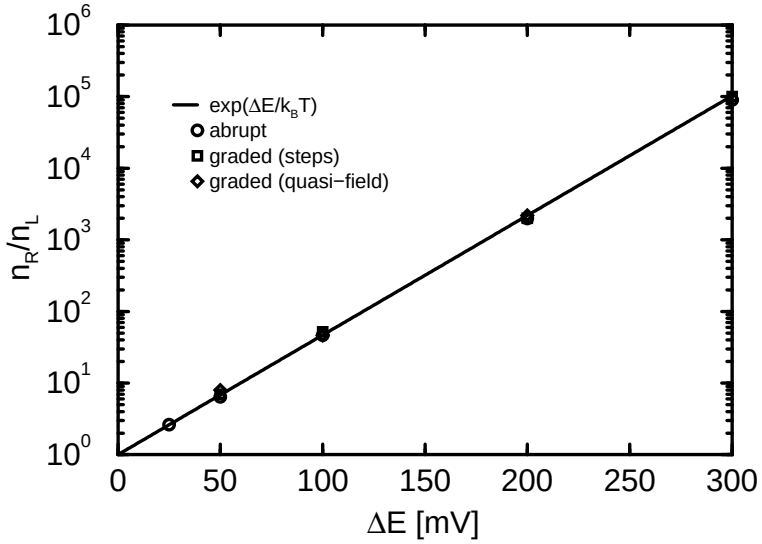

Figure 3.7: Ratio between the electron concentration at the right and left sides of a conduction band offset with height ΔE (affinity in the right side of the structure is higher than in the left side).

Boltzmann constant and T the lattice temperature.
When the simulation starts, electrons are uniformly distributed inside the structure. During the simulation they diffuse and interact with the discontinuities. When the simulation stops we collect n_L and n_R. Results are shown in Fig. 3.7 and Fig. 3.8, considering abrupt and graded discontinuities. The latter have been implemented either as a series of small abrupt HJs (4 steps for each discontinuity) or as a quasi–field. In all cases n_R/n_L follows what is expected from the theory. Fig. 3.7 shows the ratio n_R/n_L as a function of different ΔE, while Fig. 3.8 is a plot of the carrier concentration along the device, for an abrupt HJ and ΔE=100 meV.
On the other hand, when performing a self–consistent simulation, the electron concentration tends to become equal to the doping at all points. A depletion layer forms close to each CBO, producing an electrostatic potential drop that compensates the CBO. Typical E_C and carrier profiles are reported in Fig. 3.9, for ΔE=100 meV.
$\Delta V=\Delta V_1+\Delta V_2$ is the built–in voltage produced by the depletion region; Fig. 3.10 shows that this built–in voltage is exactly equal to ΔE of the CBO for a wide range of ΔE.

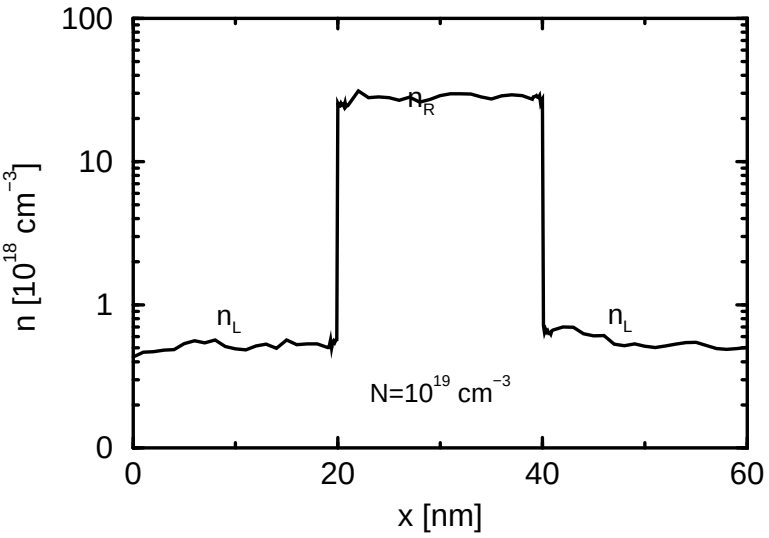

Figure 3.8: Carrier concentration profile for the case ΔE=100 meV, for an abrupt HJ. The simulation is not self–consistent.

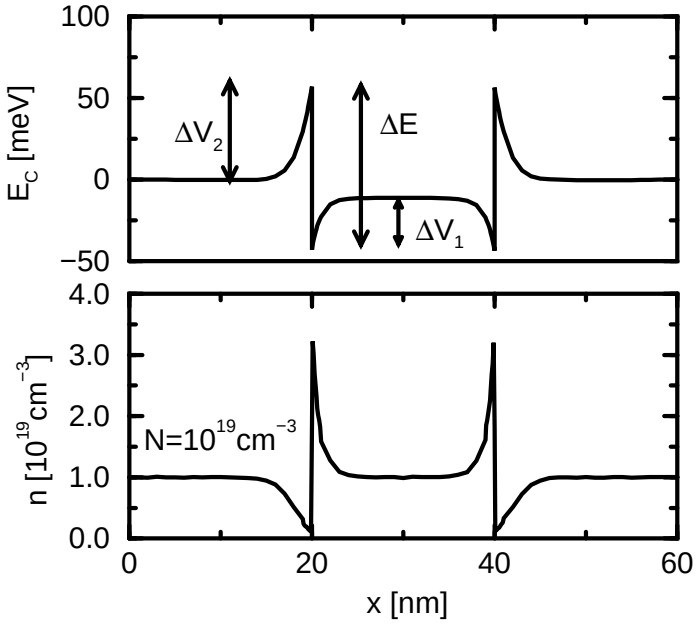

Figure 3.9: Electron concentration and conduction band profile along a structure featuring uniform doping and two abrupt HJs of amplitude ΔE and $-\Delta E$ (i.e., the affinity in the center is larger than at the sides). The simulation is self–consistent.

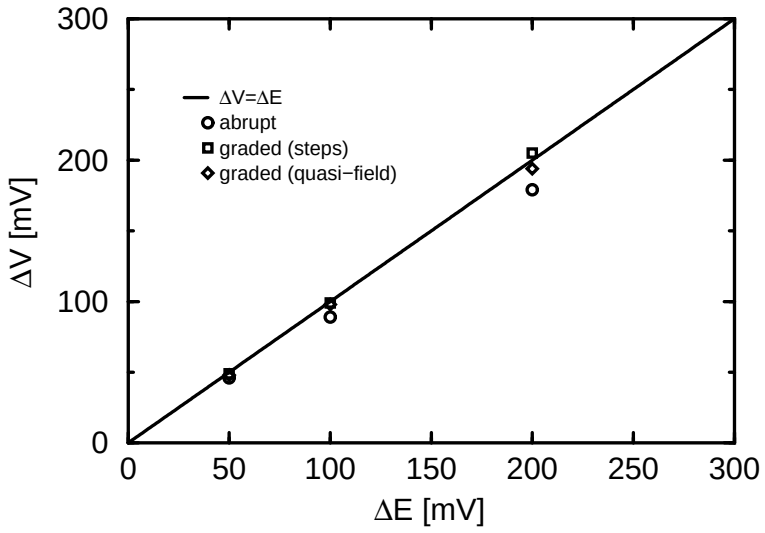

Figure 3.10: Comparison between the built–in voltage $\Delta V = \Delta V_1 + \Delta V_2$ (see top graph in Fig. 3.9) and the CB offset ΔE. The same structure of Fig. 3.9 is considered, but in this case it features abrupt as well as graded HJs.

Chapter 4

Simulation of DGSOI with Heterojunction at Source and Drain

In this chapter we first describe the devices that have been simulated with the Monte Carlo method in order to explore the effect of band–gap discontinuities on electron transport. Then, we present the model that we have adopted in order to interpret the results, and finally we discuss the results themselves.

4.1 Simulated devices

In order to accomplish the work described in this section, we have considered as reference device a double–gate SOI MOSFET. The main geometrical and electric characteristics are listed in Table 4.1. It should be noted that the reported values refer to the reference case, that is the DG MOSFET without any band–gap discontinuity.
The adopted transistor features an abrupt junction between the n–type source/drain regions and the p–type channel; moreover the overlap length between the gate and the S/D region is set to 1 nm for all the considered devices. The gate oxide is a pure SiO_2 dielectric; the gate work–function ϕ_G and the channel doping concentration N_{CH} have been selected in order to have a leakage drain current I_{OFF} equal to 100 nA/μm, in agreement with what is expected by the ITRS. The high ratio between the gate length L_G and the silicon film thickness t_{Si} guarantees a strong immunity to short channel effects, so that the drain–induced barrier lowering is very small (40 mV/V) [54].

Concerning conduction band–gap offsets (CBO), both abrupt and graded heterojunctions have been considered. A description of the different types of discontinuities and of the methods adopted to simulate them can be found in Section 3.3. A simple sketch of the simulated device is presented in Figure 4.1; in the lower part of the figure the conduction band E_C is shown, with both abrupt and graded discontinuities. Moreover, the valence band E_V is kept continous throughout the whole device. In all the cases we

Quantity	Value
Gate Length L_G	34 nm
Gate Oxide thickness t_{OX}	1 nm
Gate Oxide Dielectric Constant ε_{OX}	3.9
Silicon Film Thickness t_{Si}	10 nm
Source/Drain Extension Length $L_{S/D}$	50 nm
Gate Work–function ϕ_G	4.346 eV
Channel Doping Concentration N_{CH}	10^{16} cm^{-3}
Source/Drain Doping Concentration $N_{S/D}$	10^{20} cm^{-3}
Supply Voltage V_{DD}	1 V
Threshold Voltage V_t	0.2 V
Sub–threshold Current I_{OFF}	100 nA/μm
DIBL	40 mV/V

Table 4.1: Main characteristics of the simulated devices.

have assumed that the same offset is present at the source and at the drain: the positions of these discontinuities are symmetric with respect to the center of the channel (x=0), and the source and drain regions are characterized by the same band–gap. For the devices featuring heterojunctions, the gate work–function ϕ_G has been modified in order to have the same leakage current (I_{OFF}=100 nA/μm). This task has been performed by drift–diffusion simulations obtained by the commercial simulation tool Sentaurus from Synopsys [27]. It should be note that in order to evaluate I_{OFF}, a standard DD simulation software is sufficient because I_{OFF} is obtained in a depleted channel regime, and therefore it is defined by electrostatics only (as long as there is no direct tunneling between source and drain).
All the simulated devices feature the same sub–threshold slope, therefore as ϕ_G is set in order to have the same I_{OFF}, the threshold voltage is the same as well (approximately 200 mV).

4.2 The Lundstrom model

Before presenting the results obtained from the simulations of the structures described in the previous section, it is useful to briefly summarize the carrier transport model that we have adopted in order to interpret them. This model has been proposed and largely explored by the group of Purdue University, leaded by Prof. Lundstrom, and it is described in [55, 56, 57].
In this approach, the current provided by the transistor is described by fluxes moving through the channel. Source and drain are treated as reservoirs of carriers at thermal

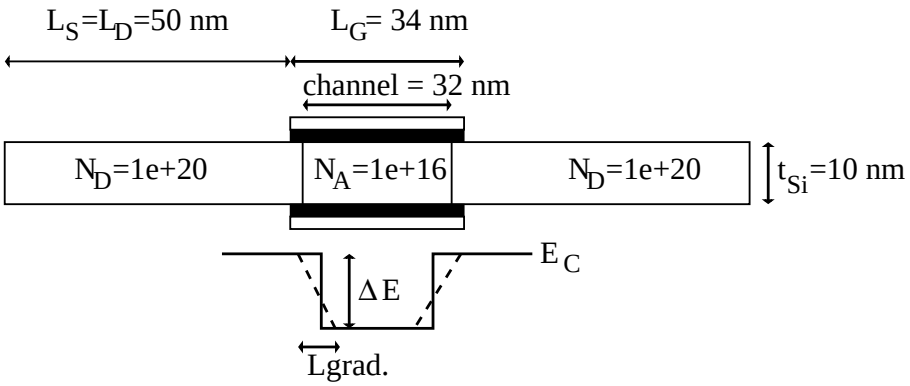

Figure 4.1: Simple sketch of the simulated devices. In the lower part of the figure, the conduction band E_C is presented. Continuous line represents an abrupt HJ, while the dashed line defines a graded one. The valence band E_V is continuous throughout the whole structure.

energy which are injected into the channel, and the total drain current can be modeled in terms of current injected from the source into the channel I^+_{inj}, and current back–scattered from the channel into the source I^-_{inj}, both evaluated at the abscissa x_{inj}, that is the position of the maximum of the potential energy, also referred as Virtual Source (VS). This approach allows to avoid the concept of mobility, that is not easy to define in short–channel devices, while it defines transmission and reflection coefficients to study the transport along the channel. Positive and negative fluxes are related by a *back–scattering coefficient*, $r{=}I^-_{inj}/I^+_{inj}$, that is the ratio between the current injected into the channel and the current back–scattered to the source.
The theory in [55, 56, 57] assumes that the back–scattering coefficient r is determined by the electric field profile near the VS: scattering contributes to I^-_{inj} only by events taking place within the distance L_{kT} from the VS. L_{kT} is the distance that the potential energy takes for a drop equal to k_BT/q with respect to the peak value at the virtual source, where k_B is the Boltzmann constant and T the lattice temperature. Lundstrom and co–workers proposed the following simple expression for r in high–field conditions:

$$r = \frac{L_{kT}}{L_{kT} + \lambda} \tag{4.1}$$

where λ is the mean free path for back–scattering and it is independent of L. L_{kT} (often referred as kT*–layer*) is thus a powerful concept to understand the role of scattering for short–channel devices, where a scattering–based approach is better than a mobility–based one. A deeper understanding of the relevance of the kT–layer in nanoscale transistors may be found in [58].
The drain current can be written as

$$I_{DS} = qN_{inv}(x_{inj})v_x(x_{inj}) \tag{4.2}$$

where $N_{inv}(x_{inj})$ is the inversion charge at the VS, and $v_x(x_{inj})$ is the carrier velocity, averaged over the vertical direction y, at the same point. However, the average velocity v_x^+ of the I_{inj}^+ flux is also a relevant parameter as well, since it is the velocity of the injected carriers. It can be demonstrated [55, 59] that

$$v_x(x_{inj}) \approx v_{inj}^+ \frac{1-r}{1+r} \tag{4.3}$$

where v_{inj}^+ is the velocity of the carriers travelling in the positive direction, that is the carriers that leave the source and enter the channel. It is evident that the quantities r and v_{inj}^+ are important in order to understand the magnitude of the average velocity v_x at the VS, and Eq. 4.2 may be re–written as:

$$I_{DS} = N_{inv} v_{inj}^+ \frac{1-r}{1+r} \tag{4.4}$$

Moreover, the downsizing of the MOSFET dimensions implies that the source–to–drain distance is comparable to the mean free path for phonon scattering, therefore each carrier suffers few scattering events, if no scattering at all, within the intrinsic channel region. This case is referred as *quasi–ballistic transport*, proposed first by Natori [60], and the current that is attainable in the absence of scattering in the channel (*ballistic current*, I_{BL}) may be seen as an upper limit for the drain current provided by the transistor. The ratio of the actual current to the ballistic current

$$BR = \frac{I_{ON}}{I_{BL}}$$

is defined *ballisticity ratio* and it indicates how far we are from this upper limit.
Finally, BR may be defined as a function of the back–scattering coefficient r:

$$BR = \frac{1-r}{1+r} \tag{4.5}$$

In the remainder of this thesis we will focus on the effects of the CBO on N_{inv}, v_x^+ and r at the virtual source, and through equations 4.2, 4.3 on the provided current. Moreover, as we are investigating devices designed on an high–performance concept, we are interested on the so called I_{ON}, that is the current provided by the transistor when V_{GS}=V_{DS}=V_{DD}.

4.3 DG SOI with abrupt heterojunctions

We first explored devices featuring abrupt discontinuities. First of all, we performed MC simulations on the reference device (without any band–gap discontinuity), finding that

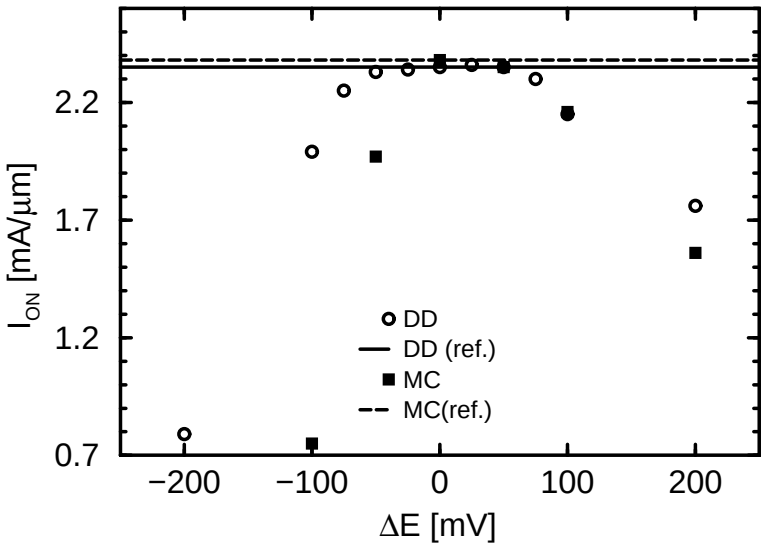

Figure 4.2: Simulated drain current for V_{GS}=V_{DS}=V_{DD} in devices featuring abrupt CBOs with different ΔE. Results obtained with the MC simulator of this work are compared with DD simulations. The discontinuities are placed at x=-15.2 nm (position of the VS in the reference device) and x=15.2 nm. The horizontal lines are the currents in the reference device. The current includes both front and back channel.

the position of the virtual source was at x=-15.2 nm; it may be useful to remember that the simulated devices feature an L_G=34 nm, and that in our reference system the point x=0 nm corresponds to the center of the channel. In the first set of simulations we placed the discontinuities at symmetrical positions with respect to the center of the channel: in other words, we placed an offset ΔE at x=-15.2 nm and an offset $-\Delta E$ at x=15.2 nm (see Fig. 4.1).

Figure 4.2 reports the I_{ON} as a function of ΔE: ΔE values larger than 0 represent MOSFETs featuring a larger affinity in the channel than in the source and drain, while ΔE<0 represent transistors with a smaller affinity in the channel. In all cases the introduction of the conduction band discontinuities reduces the ON–current. The result is not peculiar to the MC simulations and it is predicted also by drift–diffusion one (that have been performed by Sentaurus from Synopsys, as mentioned before). Differences between DD and MC are small for positive ΔE. In the case of negative ΔE the disagreement is large, mostly because in this case the current is mainly controlled by thermionic emission above the CBO of electrons featuring large energies in the source and drain (in order to overcome the barrier, the minimum kinetic energy has to be equal to ΔE). The density–of–states of the full–band MC significantly differs from the DD one at such large energies.

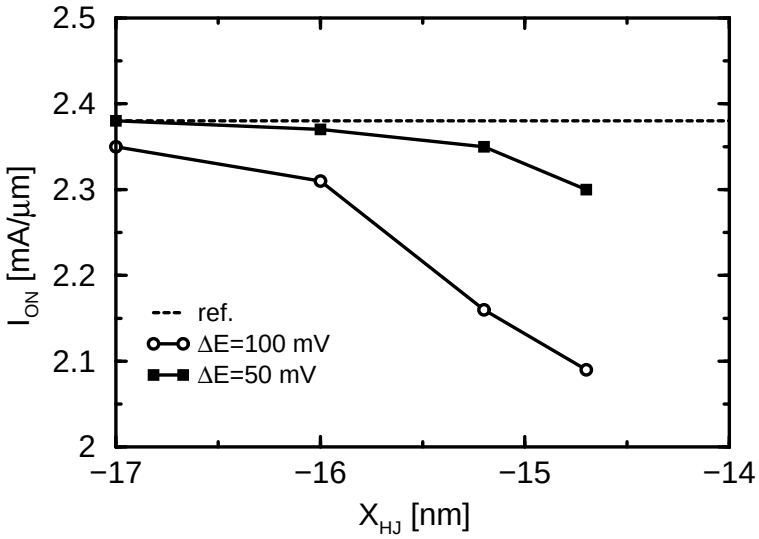

Figure 4.3: Simulated drain current (including both front and back channel) for V_{GS}=V_{DS}=V_{DD} in devices featuring abrupt CBOs for different positions of the CBO at the source (in all the cases the one at the drain is symmetric with respect to the center of the channel, x=0). ΔE=50 meV, 100 meV. On the x–axis, x_{HJ} indicates the position along x where the band–gap discontinuity occurs.

In Fig. 4.3 we have chosen two ΔE values (ΔE=50 meV and 100 meV) and we moved the position of the discontinuity, starting from the gate edge (x=−17 nm) and gradually moving inside the channel. The further current reduction with respect to the reference case, when the conduction band offset is moved, is evident.

In order to understand the origin of the degradation of drain current induced by the conduction band offset, we have analyzed the profile along the channel of some relevant quantities: the conduction band profile E_C averaged over y, the total v_x and positive v_x^+ carrier velocities averaged over the y–direction, and the inversion charge density N_{inv}. We took as example the case where ΔE=100 meV, and the plots are reported in Figure 4.4. Moreover, Table 4.2 reports the values that the quantities plotted in Fig. 4.4 assume at the VS.

The CBO at the source side is placed at the virtual source of the reference case (see top graph, where the HJ is placed exactly at x=−15.2 nm). Even in the structure featuring HJ, the maximum of potential energy is placed at x=−15.2 nm, so that we found that the virtual source is at the same place in the two considered transistors.

The discontinuity acts as a *launcher* for the electrons injected into the channel, as demonstrated by the profile of v_x^+. This latter quantity (see the middle graph) is much larger than in the reference case, in the channel region beyond the VS. However, v_x^+ is essentially the same at the position x=−15.2 nm (the VS in both cases). The device

with the CBO features a lower back–scattering (r in Tab. 4.2), since the electrons moving with negative velocity see an energy barrier and therefore only a fraction of them can go back to the source.
As a result of these two effects (same v_x^+, lower r at the VS) the average velocity v_x just before the offset is larger in the device featuring the CBOs, (see again the middle plot and Eq. 4.3).
However, the back–scattered electrons, that come from the channel and attempt to enter the source, hit the barrier and create an accumulation of charge next to the discontinuity (right side), which tends to prevent further injection from the source. As a result the charge at the VS is lower in the device with $\Delta E \neq 0$ than in the reference (compare the N_{inv} values in Tab. 4.2). This effect overcompensates the enhancement in average velocity, thus reducing I_{ON}. In conclusion we have found that the discontinuity degrades the electrostatics inside the device, overcompensating the advantages provided by an increased injection velocity. This is a very important result, which demonstrates how important it is to use a self–consistent approach.

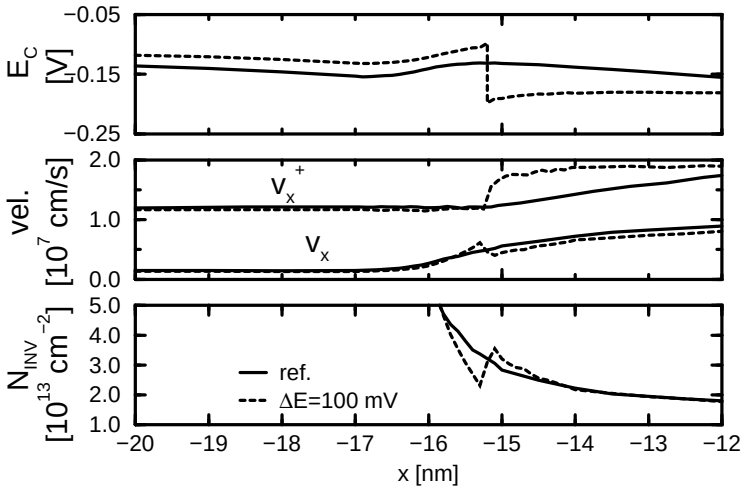

Figure 4.4: Conduction band profile averaged over y (top), velocity profiles averaged over y (middle) and inversion charge density (bottom, including both front and back channel) as a function of the position along the channel. The reference device (solid line) is compared to the case with abrupt HJs at x=$\pm$15.2 nm (dashed line). V_{GS}=V_{DS}=V_{DD}.

@VS	N_{inv} $[10^{13}cm^{-2}]$	v_x $[10^7cm/s]$	v_x^+ $[10^7cm/s]$	r
ref	3.3	0.5	1.20	0.4
ΔE=100 meV	2.3	0.66	1.17	0.24

Table 4.2: Values of the inversion charge N_{inv}, the average velocity v_x, the positive velocity v_x^+ and the backscattering coefficient r at the VS, for the two cases presented in Fig. 4.4.

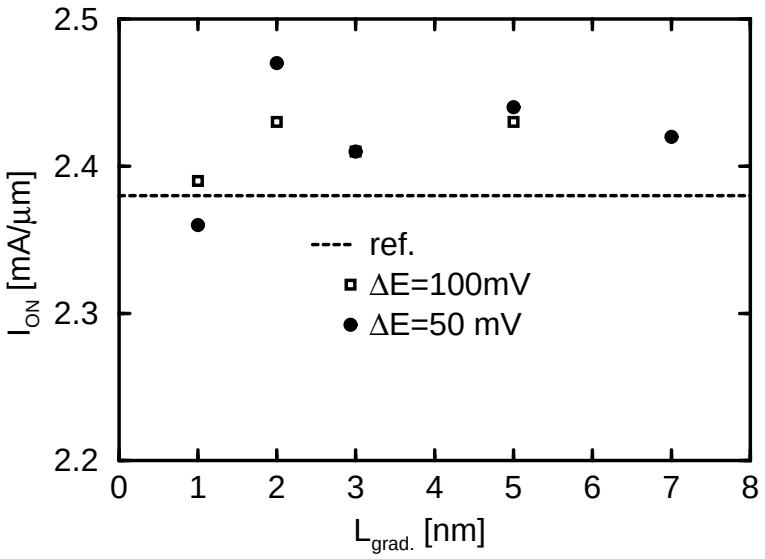

Figure 4.5: ON–current (front and back channel) in devices featuring graded HJs (see Fig. 4.1) as a function of the extension of the graded region. This region begins at the gate edge (x=-17 nm) and extends toward the channel. The dashed line represents the current provided by reference device (without HJ).

4.4 DG SOI with graded heterojunctions

As a second step of our analysis, we considered linearly graded HJs. The idea is to examine whether it is possible to reduce the negative influence of the accumulation of electrons that forms behind the CBO by smoothing the band discontinuity. As detailed in Section 3.3, in this case the heterojunction is treated as an additional electric field, added during the free–flight to the field given by the Poisson equation. Similarly to what has been done in the previous section, in the following we will assume a higher electron affinity in the channel than in the source/drain region: in this case the electric field accelerates the electrons coming from the source and entering the channel. We assume again a device symmetric with respect to the center of the channel (see Figure 4.1 for a simple sketch of the simulated device). In this first set of simulations, the graded region begins at the gate edge (x=-17 nm) and extends for a length L_{grad} towards the channel. It is useful to remember that the gate work–function has been calibrated in order to have the same sub–threshold drain current I_{OFF}=100 nA/μm for all the simulated devices.

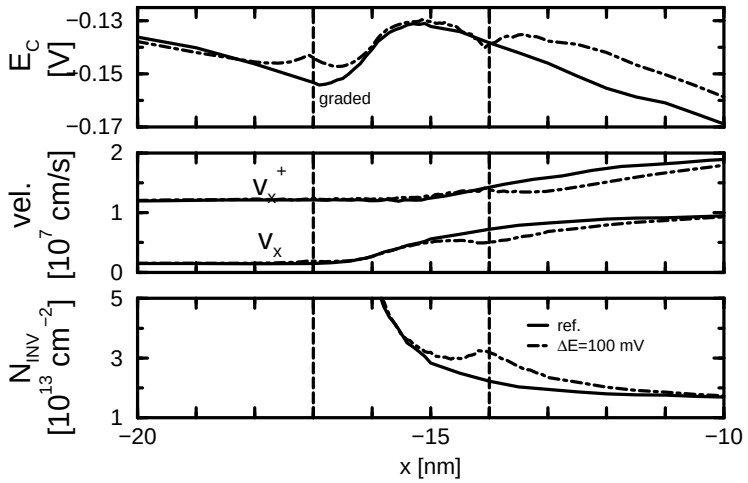

Figure 4.6: Conduction band profile averaged over y (top), velocity profiles averaged over y (middle) and inversion charge density (bottom, including both front and back channel) as a function of the position along the channel. The reference device (solid line) is compared to a case with graded HJs from -17 nm to -14 nm and from 14 nm to 17 nm, corresponding to L_{grad}=3 nm (dashed line). V_{GS}=V_{DS}=V_{DD}.

Fig. 4.5 shows I_{ON} as a function of the extension L_{grad} of the HJ, for ΔE=50 meV and 100 meV. With respect to the abrupt case shown in the previous section, we now see a slight current improvement, in particular when the region with grading includes the position of the VS in the reference case (that is x_{inj}=-15.2 nm). This condition requires that L_{grad} $\geq$2 nm, since the grading starts at the gate edge, -17 nm, and the VS is at -15.2 nm. The improvement is however modest and decreases for large L_{grad} as the quasi field is reduced.

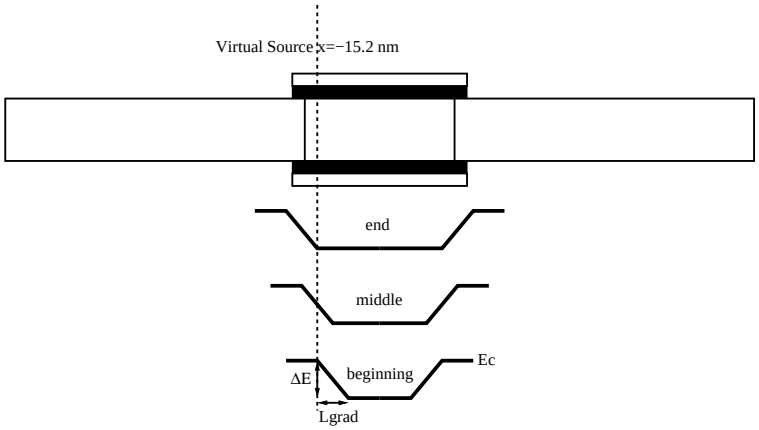

Figure 4.7: Sketch of the different positions of the graded region with respect to the Virtual Source (VS) of the reference transistor, to be considered in Fig. 4.8, 4.9, 4.10, 4.11 and Tab. 4.3.

As reported in the previous section, we can understand this behavior by plotting some internal quantities, shown in Figure 4.6. The virtual source has essentially the same position in the reference and in the graded devices (see the top graph). However, due to an accumulation of charge at the end of the graded region (lower graph), the electric field after the VS is slightly smaller in the graded case. This effect compensates the advantages related to the presence of the quasi–field at the VS, since the average velocities are essentially the same in the two cases. In particular (see middle plot) we have a slightly higher v_x^+ (due to the quasi field at the VS), but the same v_x, since the effect of back–scattering is enhanced due to the lower field next to the VS In other word, the presence of such accumulation of charge at the end of the graded region degrades the kT–layer profile, increasing L_{kT} [55, 59].

In the second set of simulations, we varied the relative position of the graded region with respect to the Virtual Source (see Fig. 4.7), considering three possible cases: the VS can either correspond to the end, the middle or the beginning of the graded region. As previously, higher electron affinity in the channel than in the S/D regions is assumed, and a symmetric structure is considered. Fig. 4.8 shows I_{ON} as a function of L_{grad} for ΔE=100 meV. Even adopting these designs the improvement in terms of provided current is modest, and it decreases rapidly when the graded region is moved towards the channel.

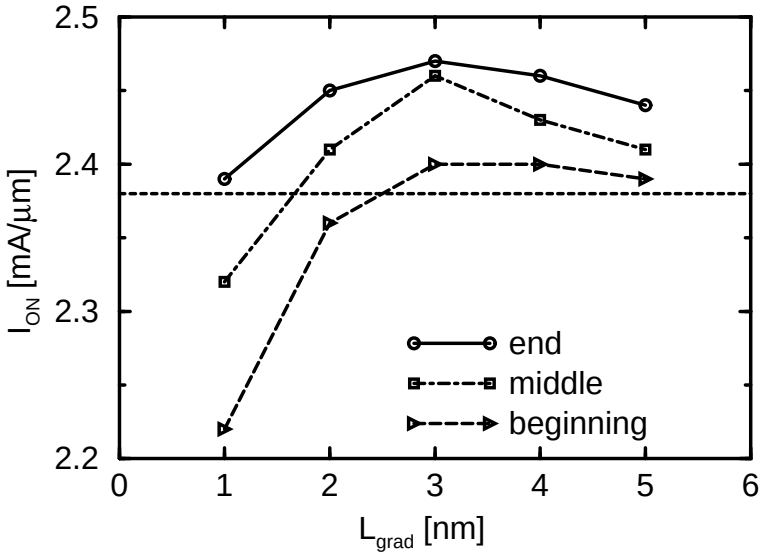

Figure 4.8: Simulated drain current for V_{GS}=V_{DS}=V_{DD} as a function of the extension of the graded region. The three different situations, presented in Fig. 4.7, are considered. ΔE=100 meV. The dashed line represents the current provided by reference device (without HJ). The current includes both front and back channels.

When the Virtual Source is at the end of the graded region (see Fig. 4.9), the position of the VS is slightly moved towards the channel (x_{inj}=-14.5 nm instead of x_{inj}=-15.2 nm). $v_x^+(x_{inj})$, r and thus $v_x(x_{inj})$ are essentially the same (see Tab. 4.3). On the other hand, the inversion charge at the VS is larger than in the reference case, and this effect is responsible for the small current improvement in Fig. 4.8.
When the VS is in the middle of the graded region (see Fig. 4.10), the position of the VS does not change (x_{inj}=-15.2 nm) with respect to the reference, as well as the inversion charge at the injection point (see Tab. 4.3). Even the conduction band profiles are identical, as well as the backscattering coefficient r. On the other hand, the positive velocity v_x^+ is slightly larger than in the reference device, and it is responsible for the small current improvement.
Finally, when the graded region starts at the VS (see Fig. 4.11), it acts as a launcher for electrons. Similarly to the latter case, the position of the VS does not change with respect to the reference device, but now the conduction band profile is steeper. The kT–layer is shorter, and the backscattering coefficient decreases with respect to the previous devices (see Tab. 4.3), i.e. the graded region acts as a barrier for the backscattered carriers.

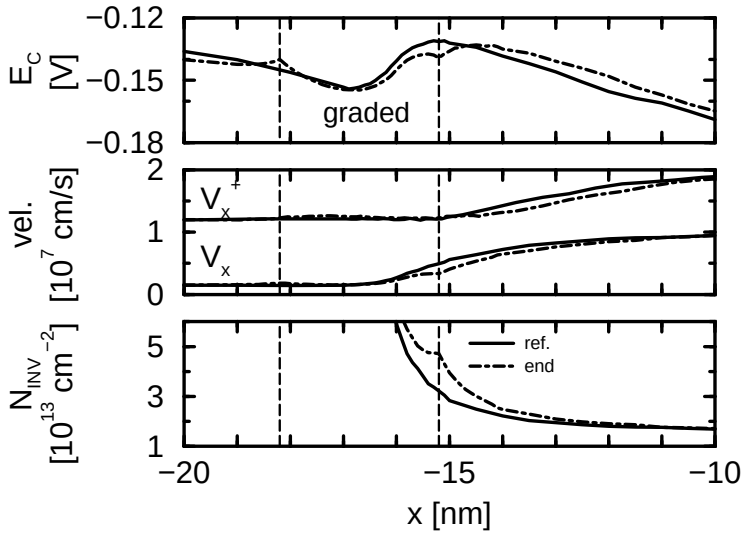

Figure 4.9: Conduction band profile averaged over y (top), velocity profiles averaged over y (middle) and inversion charge density (bottom, including both front and back channel) as a function of the position along the channel. The reference device (solid line) is compared with the device where the VS corresponds to the end of the graded region (see Fig. 4.7). L_{grad}=3 nm, ΔE=100 meV.

However, the inversion charge at the VS is now lower than in the reference case (see Fig. 4.11 and Tab. 4.3), so most of advantages due to the higher velocity are lost. This behavior is similar to the one considered for abrupt heterojunctions (see section 5.1), but now the effects are ”spread” over the graded region.
In summary, there is an evident trade–off between the inversion charge at the VS and the injection velocity, and such trade–off is detrimental and limits the current improvement.
It could be useful to remember that these simulations have been performed in a semi–classical approach, that is the quantum corrections available in our MC tool (see Section 3.2) have been disactivated, in order to speed–up calculations and to allow the analysis of a large number of devices. This approach can lead to an overestimated prediction of the current drive capability of the devices; nonetheless, we checked the impact of quantum corrections on the dependence of the ON–current on the height and position of the heterojunction barrier. In Fig. 4.12 and Fig. 4.13 some results, obtained from quantum corrected simulations have been compared with the fully semi–classical ones. Even if the provided current in the corrected case are lower than in the semi–classical simulations, due to a lower inversion charge inside the channel, the change in current with respect to the reference case is similar, and we have found the same trend between the ON–currents provided by the different structures.
Moreover, as we have already explained in Section 3.3, in presence of heterostructures

@VS	N_{inv} $[10^{13}cm^{-2}]$	v_x $[10^7 cm/s]$	v_x^+ $[10^7 cm/s]$	r
ref	3.3	0.5	1.20	0.4
ΔE=100 meV, L_{grad}=3 nm, end	3.6	0.49	1.26	0.43
ΔE=100 meV, L_{grad}=3 nm, middle	3.3	0.53	1.25	0.4
ΔE=100 meV, L_{grad}=3 nm, beginning	2.4	0.7	1.25	0.28

Table 4.3: Values of the inversion charge N_{inv}, the average velocity v_x, the positive velocity v_x^+ and the backscattering coefficient r at the VS, for the four cases presented in Fig. 4.9, 4.10, 4.11.

an energy barrier arises between the different materials involved, and carriers can travel across this barrier due to tunnel effect. In our simulation tool this phenomenon is not taken into account, so that this contribution to the electrons entering the channel is neglected and therefore we simulate the worst case in terms of provided drain current.

To conclude, we performed Monte Carlo simulations of planar n–channel DGSOI MOSFETs featuring heterojunctions, as those that can be obtained adopting alternative materials for the S/D regions. Although abrupt CBOs between the source and the channel are expected to enhance the injection velocity and thus the current, simulations of nanoscale DGSOI transistors point out that CBOs act as a launcher for the particles but at the same time it creates an accumulation of electrons next to the CBO, at the entrance of the channel. This carrier accumulation influences the device electrostatics, reducing the charge available for transport and overcompensating the velocity improvement, so that the provided current is lower than in the reference case. Due to the same mechanism, only small current improvement are obtained for graded HJs. The performed simulations point out the importance to treat this problem in a self–consistent way.
It should be noticed that these comparisons were made at given I_{OFF}=100 nA/μm, i.e., at given gate overdrive $V_{GS}-V_t$. We have focused on the effect of the band offsets alone. In many practical cases, the source/drain material induces strain in the channel, that is the main responsible of the I_{ON} improvement, whereas, according to our simulations, the effect to the offset alone seems to be modest.

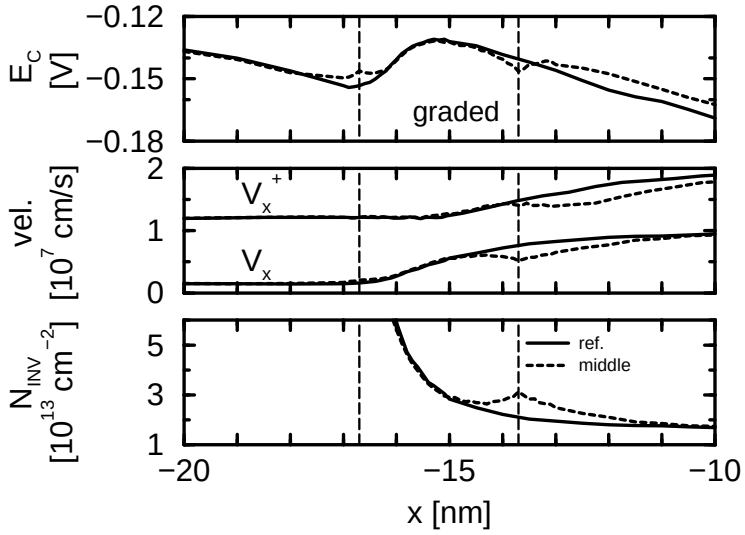

Figure 4.10: Same as Fig. 4.9 but the reference device (solid line) is compared with the device where the VS corresponds to the middle of the graded region (see Fig. 4.7). L_{grad}=3 nm, ΔE=100 meV.

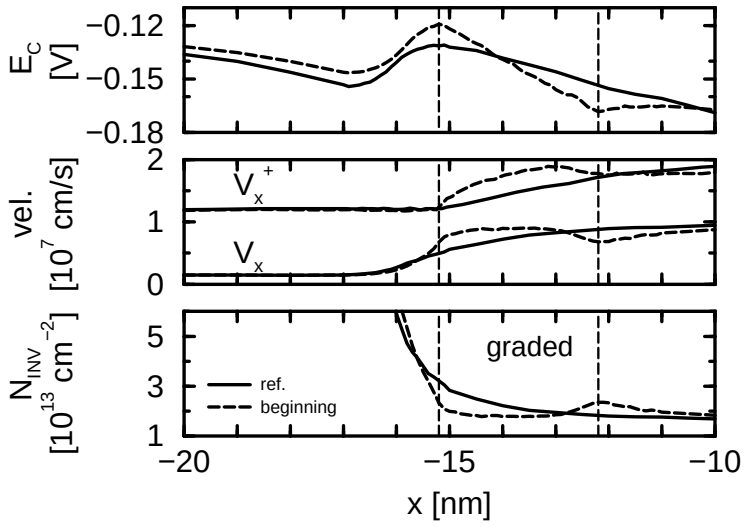

Figure 4.11: Same as Fig. 4.9 and 4.10 but the reference device (solid line) is compared with the device where the VS corresponds to the beginning of the graded region (see Fig. 4.7). L_{grad}=3 nm, ΔE=100 meV.

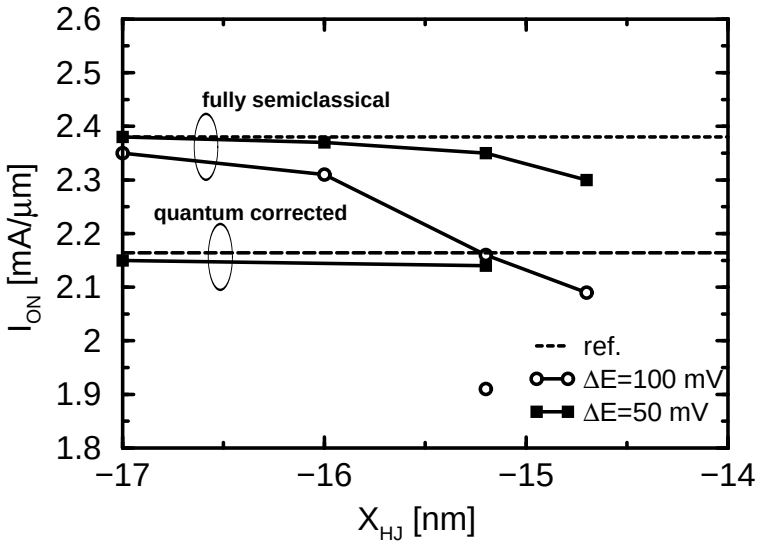

Figure 4.12: Simulated drain current for V_{GS}=V_{DS}=V_{DD} in devices featuring abrupt CBOs. The quantum corrected case is compared to the fully semi–classical one. The figure is the extended version of Fig. 4.3.

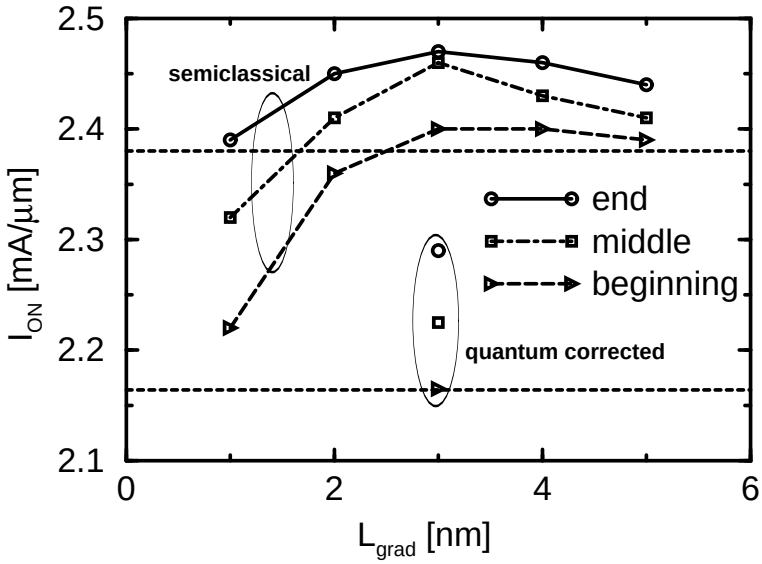

Figure 4.13: Simulated drain current for V_{GS}=V_{DS}=V_{DD} in devices featuring graded CBOs. The quantum corrected case is compared to the fully semi–classical one. The figure is the extended version of Fig. 4.8.

Part II

Self–Heating Effects in SOI structures

Chapter 5

Self–heating in electron devices

In this chapter we will review the problem of Self–Heating Effects (SHE) in Silicon–On–Insulator technology, focusing on the causes that will make it detrimental for the forthcoming technological nodes, the solutions that have been adopted in order to include these effects in models and simulation tools, and finally the impact that Self–Heating has on the performance of different types of SOI transistors.

5.1 Self–Heating Effects in SOI devices

In Chapter 1 we explained that the miniaturization of electron devices in the past decades has allowed the semiconductor industry to perform amazing progresses in terms of speed and integration density of the digital circuits. The scaling of the transistor dimensions is expected to be the main way to continue this trend even in the future years. However, the conventional Bulk structure will become inadequate: this perspective has accelerated the adoption of new materials with transport properties different than those of pure silicon and the introduction of alternative device architectures as the SOI ones. SOI planar single– and double–gate transistors as well as FinFET architectures are good candidates to substitute the Bulk one for the forthcoming technological nodes, despite the added process complexity needed to build such devices [61]. SOI architectures feature an near ideal turn–off slope, low OFF–current and a good control of the short channel effects, as we have seen in Sec. 1.4.
On the other hand, the downsizing of device dimensions and the adoption of innovative structures have consequences on the power dissipation as well: power densities, heat generation inside the device and chip temperatures will reach levels that can prevent the reliable operation of integrated circuits if they are not proprerly handled. Chip–level power densities are currently on the order of 100 W/cm^2. If the rates of integration and miniaturization continue to follow the ITRS guidelines, the chip–level power density is likely to increase even further, as illustrated in Figure 5.1. Increasing power density

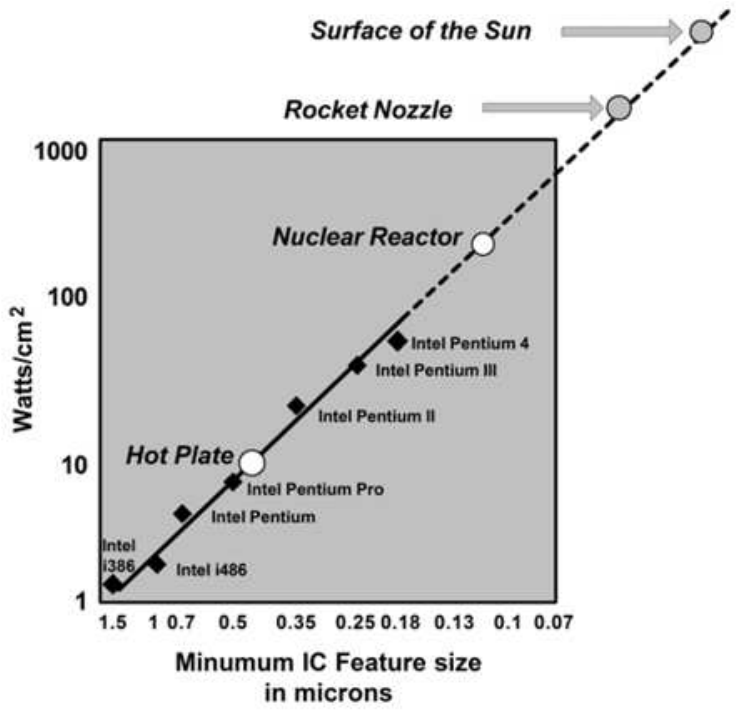

Figure 5.1: Trends of on–chip power density as a function of the minimum IC feature size over the past years. It should be noticed that the y–axis is logarithmic while the horizontal one is linear: the trend is exponential. Data from *http://www.cs.clemson.edu/ mark/330links.html*.

levels will quickly drain batteries in portable devices and may render many electronic systems unusable without significant advances in cooling technology, or without fundamental shifts in design. These trends involve that thermal device design is becoming an important part of microprocessor engineering: while chip–level hot spots are troubling circuit designers, device designers are beginning to encounter thermal management problems at nanometer–length scales within individual transistors [62, 63].

If we look at the self–heating in the context of a field–effect transistor, the applied voltage leads to a lateral electric field which peaks near the device drain and it is maximum at the channel–to–drain junction. This field accelerates the charge carriers (e.g., conduction band electrons in a n–type FET) which gain energy and heat up.
Heat generation in MOSFETs during operation occurs through the interaction of electrons that can scatter with lattice vibrations, with each other, with material interfaces, imperfections or impurity atoms. Some of these interactions only redistribute energy and momentum inside the electron population; however, the electron population also loses net energy by scattering with phonons, consequently heating up the lattice through the mechanism known as *Joule heating*. Other scattering mechanisms chiefly affect the

electron momentum (see [64] for an exhaustive overview on the different scattering mechanisms and their scattering rates).
Regarding the heat generation in MOSFET, the Joule heating rate per unit volume H is typically computed starting from the dot product of the electric field $\vec{\mathcal{E}}$ and current density $\vec{J}$ vectors:

$$H = \vec{J} \cdot \vec{\mathcal{E}} + (R - G)\,(E_G + 3k_B T) \tag{5.1}$$

where the second term on the RHS side represents the heating rate due to non–radiative generation (G) and recombination (R) of electron–hole pairs, E_G is the semiconductor band–gap, k_B is the Boltzmann constant ant T the lattice temperature [65].
It could be useful to mention that this model presents some limitations, because this approach to the analysis of heat generation is strictly local in its nature and fails to take into account the non–local characteristics of carrier heating and phonon emission. In fact, it predicts that the maximum of heat–generation rate takes place at the drain–to–channel junction, where the peak of electric field is located; however, although electrons gain most of their energy at the location of the peak lateral electric field, they must travel several inelastic mean free paths before releasing all of it to the lattice, in decrements of (at most) the optical phonon energy.
In silicon transistors, for example, electrons can gain energies that are a significant fraction of an eV, while the optical phonon energy is only about 50–60 meV. Assuming an electron velocity of 10^7 cm/s (the saturation velocity in silicon) and an electron–phonon scattering time around 0.05–0.10 ps in the high–field region, the inelastic mean free path is then on the order of 5–10 nm. The full electron energy relaxation length is therefore even longer, on the order of several mean free paths. In other words, the highly localized electric field in such devices leads to the formation of a nanometer–sized region (hot spot) displaced inside the the drain diffusion, that is spatially displaced by several mean free paths from continuum theory predictions, and presents a lower peak value. While such a discrepancy may be neglected on length scales of micrometers, or even tenths of a micrometer, it must be taken into account when analyzing heat generation on length scales of the order 10 nm: non–local effects on carrier heating and phonon emission become more relevant as the device channel length approaches the mean free path for phonon emission.
In addition, the $\vec{J} \cdot \vec{\mathcal{E}}$ formulation of the Joule heating also does not differentiate between electron energy exchange with the various phonon modes, and does not give any spectral information regarding the types of phonons emitted. The mechanism through which Joule heating occurs is that of electron scattering with phonons, and consequently only a simulation approach which deliberately incorporates all such scattering events will capture the complete microscopic, detailed picture of lattice heating. In order to satisfy all these requirements, the MC method has been adopted to compute sub–continuum and phonon frequency–specific heat generation rates, with applications at nanometer–length scales [66, 67].

In spite of its inherent local approximation, the conventional model for Joule heating proposed in Eq. 5.1 is still largely adopted because of the ease of implementation in the frame of device simulators, allowing an efficient electro–thermal (ET) simulation by the self–consistent coupling of carrier transport, heat generation, and heat transport, thus providing the possibility to include the self–heating effects in the analysis of the impact of technological options on device performance.
The transport of heat in semiconductors is due to the propagation of phonons, whose net motion is governed by gradients in their density: the contribution of electrons (that is dominant in the case of metals) is lower than 1% and negligible even in the case of high doping concentration.
Different temperatures at two positions in a semiconductor device imply different distributions of phonons. Since the change in phonon distribution may only occur due to scattering, the temperature may vary only over a length larger than the phonon mean free path (approximately 200–300 nm in undoped bulk silicon at room temperature [68]).
The most widely approach for modeling the heat transport is based on the *Fourier's law* of heat diffusion:

$$c\frac{\partial T}{\partial t} = \nabla \cdot (k_S \nabla T) + H(\vec{r}, t) \tag{5.2}$$

where H is the heat–generation rate per unit volume as defined in Eq. 5.1, c is the heat capacity per unit volume, k_S is the thermal conductivity of the semiconductor, that is related to c by the following relationship:

$$k_S = cv\Lambda_S/3 \tag{5.3}$$

where v represents the average phonon velocity and Λ_S is the phonon mean free path.

Thermal transport in bulk transistors has traditionally been modeled in the classical limit, as sub–continuum thermal effects can be neglected for device dimensions larger than the phonon mean free path. Modern device technologies operate at length scales comparable to or lower than the phonon mean free paths, and this leads to sub–continuum transport effects; moreover, future technologies are going to forge deeper into this sub–continuum regime because of scaling. In this case two sub–continuum effects are expected to play a role in bulk transistor thermal transport:

1. the small region of high electric field near the drain gives rise to a strongly localized hot spot, only a few tens of nanometers across, and hence much smaller than the bulk phonon mean free path. This leads to ballistic phonon transport in the vicinity of the heat source, and higher temperatures than those predicted by classical diffusion theory. In this situation, a solution of the phonon Boltzmann Transport Equation is more accurate than the classical heat diffusion equation;

2. the second sub–continuum thermal effect to be expected in ultra–scaled bulk FETs has to do with the non–equilibrium interaction between the generated optical and

acoustic phonons. Since nearly–stationary optical phonons form the majority of the vibrational modes generated via Joule heating, they tend to persist in the hot spot region until decaying into the faster acoustic modes. This non–equilibrium scenario may become particularly relevant when device switching times approach the optical–acoustic decay times, on the order of several picoseconds. A careful transient solution of the phonon populations may be necessary to properly account for the non–equilibrium distribution

These first two issues challenge the continuum diffusion theory of heat transport represented in Eq. 5.1 and 5.2. A higher order treatment of heat transport, which is able to cope with the hot–spot–related issues aforementioned, would require the solution of several phonon Boltzmann transport equations (one for each phonon mode) coupled with each other by the phonon scattering. This approach is difficult due to both the complexity of the solution of the Boltzmann transport equation when applied to realistic structures, and the limited knowledge about the selection rules and transition rates for phonon–phonon interactions. Several simplified approaches for the simulation of heat transport based on the phonon BTE have been proposed, and most of them are presented in [69].
A third sub–continuum effect is caused by the adoption of thin silicon layers in SOI technology, because thermal conductivity in thin films is substantially reduced with respect to bulk crystals due to the enhanced scattering of phonons with the film boundaries, causing a large reduction of the phonon mean free path. For example, the thermal conductivity of a 10 nm thin silicon film is expected to be reduced by an order of magnitude from that of bulk silicon [70].

The enhanced boundary scattering in thin films leads to a reduced phonon mean free path, thus reducing the effects of hot–spot ballistic phonon emission and making the limitations of Eq. 5.2 less critical. The enhanced scattering can be taken into account in the frame of the simple diffusion theory by appropriately modifying the thermal conductivity in order to account for the enhanced boundary scattering.
Moreover, advanced non–traditional device fabrication introduces a number of new materials with thermal conductivities lower than that of bulk silicon. The thermal properties of these materials are therefore expected to play a more significant role in device design and thermal behavior: bulk germanium–based transistors, for example, would suffer from increased operating temperatures due to a substrate thermal conductivity approximately 60% lower than bulk silicon FETs; strained silicon channel devices grown on a graded $Si_{1-x}Ge_x$ buffer layer benefit from an increased mobility, but their thermal behavior is adversely affected by the lower thermal conductivity of the $Si_{1-x}Ge_x$ alloy layer. In SOI technologies the channel is thermally insulated from the underlying substrate by a SiO_2 buried oxide layer; this SiO_2 layer features a thermal conductivity that is two order of magnitude lower than the pure silicon one (as we can see in Table 5.1, where the thermal conductivity for the most frequent semiconductors is reported), and it

Material	Thermal Conductivity [W/mK]
Silicon	148
Germanium	58
GaAs	46
GaP	110
InAs	27
InP	68
SiO_2	14

Table 5.1: Thermal conductivity for some of the most important semiconductors and for silicon dioxide, at 300 K and for bulk materials.

impedes the dissipation of the heat generated in the active region. Alternative dielectrics with higher thermal conductivities to be used as buried oxide for SOI technology are currently under examination [71]
When the heat is generated, the lattice absorbs the extra electron energy and warms to a higher temperature T, and in return affects the electronic transport properties of the material. The electron mobility in undoped bulk silicon decreases approximately as $T^{-2.4}$ around room temperature owing to higher phonon populations and increased scattering rates. When other scattering mechanisms come into play, the electron mobility is more weakly dependent on temperature: it decreases approximately as $T^{-1.7}$ in highly doped silicon and $T^{-1.4}$ in nanometer–thin silicon layers, where boundary scattering becomes of importance [72].
As we have already seen Eq. 4.4 in Section 4.2, the MOSFET drain current could be written as

$$I_{DS} = N_{inv} v_{inj}^{+} \frac{1-r}{1+r} = C_{eff}(V_{GS} - V_t) v_{inj}^{+} \frac{1-r}{1+r}$$

where C_{eff} is the effective oxide capacitance and $BR{=}(1-r)/(1+r)$ represents the ballisticity ratio. The temperature dependence of the drain current can be expressed as

$$\frac{\partial I_{DS}}{\partial T}\frac{1}{I_{DS}} = \frac{\partial v_{inj}^{+}}{\partial T}\frac{1}{v_{inj}^{+}} + \frac{\partial BR}{\partial T}\frac{1}{BR} - \frac{\partial V_t}{\partial T}\frac{1}{V_{GS}-V_t} \tag{5.4}$$

A variation in terms of T impacts the threshold voltage through the intrinsic carrier concentration: V_t is a decreasing function of the temperature. Moreover under non–degenerate quasi–equilibrium conditions v_{inj}^{+} is well approximated by the thermal velocity v_{th} of an equilibrium half–Maxwellian electron distribution, and it is an increasing function of the temperature ($v_{inj}^{+}{\approx}v_{th}{\propto}T^{0.5}$). These two factors promote a larger ON–current at increasing T. On the contrary, the negative temperature coefficient of the ballistic ratio, due to the increasing phonon scattering at increasing T, leads to a degradation of I_{ON}. As the gate overdrive is reduced, the dependence of threshold voltage

on the temperature becomes more and more relevant, and the degradation of the current is therefore expected to be reduced substantially until a critical gate voltage is reached, corresponding to a zero–temperature–coefficient condition. Below such critical value, due to the temperature dependence of V_t and v_{inj}^{+}, the current is expected to rise due to SHE [73]. On the other hand, when the gate overdrive increases, the contribution due to the ballistic ratio turns to be dominant and it degrades the ON–current provided by the MOSFET. As in the following of this work we will be interested in the I_{ON} of high–performance transistors, this will be the most frequent case.

5.2 Simulation approach

3D electro–thermal (ET) simulations have been performed using the Sentaurus device simulator. The Canali model [74] is employed to describe high–field transport; the mobility degradation at the Si–SiO_2 interface, due to surface roughness scattering, and the mobility dependence on the doping concentration are included as well.
As we have seen in Chapter 2 the standard drift–diffusion models underestimate above–threshold drain current in ultrashort devices operating in a quasi ballistic regime, due to its inability to properly account for the off–equilibrium phenomena that occur when the gate length is scaled down. For this reason, the mobility model parameters have been calibrated in order to reproduce Monte Carlo calculated $I_{DS}{-}V_{DS}$ characteristics, according to the approach suggested by Bude [75]. In particular, the conventional model for high–field–dependent carrier mobility proposed in [74] is:

$$v_{sat} = v_{sat0}\left(\frac{T}{T_0}\right)^{\alpha} \tag{5.5}$$

The saturation velocity v_{sat} and the exponent that describes the temperature dependence of the saturation velocity have been modified in order to achieve a good agreement between drift–diffusion and MC results over a wide temperature range. The calibration of the DD transport model has been performed at different temperatures by comparison with the Monte Carlo tool that has been widely described in Section 3.2. We followed the approach already employed in [76] in order to calibrate the high–field–dependent mobility model for fully–depleted SGSOI transistors featuring 25 an 18 nm gate length. All the remaining mobility model parameters related to the dependence on doping concentration and surface roughness are kept at their default values.
3D ET simulations require a large computational burden. In order to perform simulations that satisfy the requirements in terms of CPU time and allocated memory, quantum corrections (even if available through a density gradient approach) are not taken into account. Moreover, in order to minimize the node count of the structures, the simulation domain is only one–half of the complete devices (that is we simulated only $W_{fin}/2$ in the case of FinFET and only $W^{ch}/2$ in the case of planar devices). This is possible by

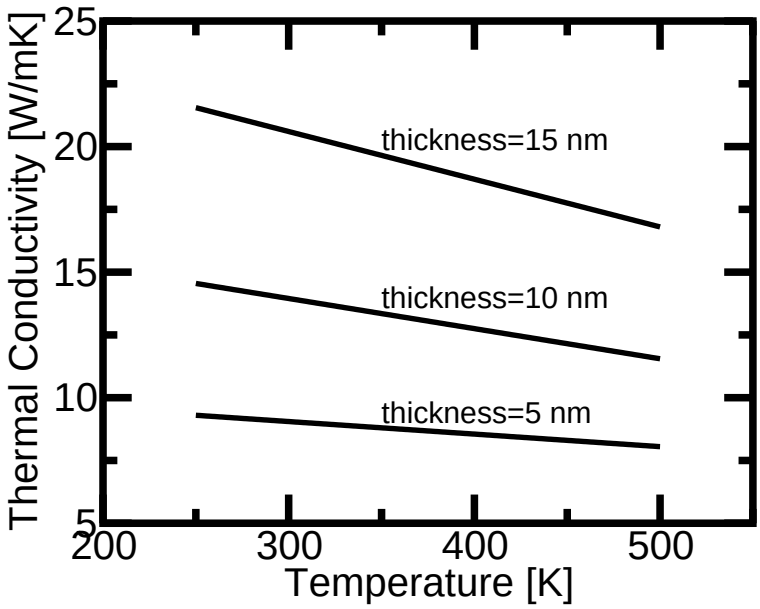

Figure 5.2: Silicon thermal conductivity k_{Si} as a function of temperature in thin Silicon layers calculated according to the model described in [77], that has been extrapolated for thicknesses down to 5 nm.

exploiting the symmetries of the simulated structures.
Concerning ET simulations, the silicon thermal conductivity k_{Si} presents a dependence on film thickness and temperature, as we have seen in Sec. 5.1; in particular, k_{Si} is a decreasing function of these two quantities. Liu *et al.* [77] have presented experimental data of thermal conductivity in ultra–thin silicon layers. In the same article, a model for the effect of thickness–dependent boundary scattering on the thermal conductivity is proposed and validated for a silicon thickness down to 20 nm. Adopting this model, which includes the dependence on temperature, in our paper we extrapolated k_{Si} values for silicon thicknesses down to 5 nm and for temperatures ranging between 250 and 500 K, as shown in Figure 5.2.

Regarding the thermal boundary conditions (BC) that have been adopted in our simulations, an isothermal (IT) 300 K heat sink is placed at the bottom of the 1.8 μm–thick Silicon layer on which the simulated SOI devices have been build. Considering the gate and S/D contacts, they could be treated either as

- *adiabatic*: no heat or work can flow through the contact;
- 300 K IT boundaries.

These two solutions represents the upper lower bound between the easiest condition for heat dissipation (300 K IT) and the most diffult one (no dissipation at all).

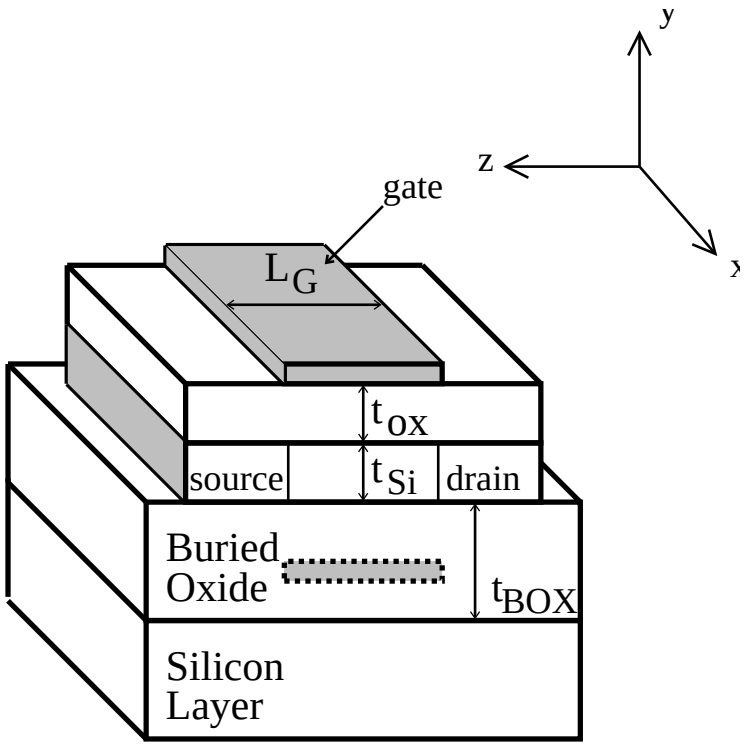

Figure 5.3: 3D sketch of the simulated planar SGSOI and DGSOI. The figures is not drawn to scale. The grey–shadowed dot–contoured region represents the second gate for the DGSOI transistor.

Moreover, a lumped thermal resistance R_{TH}=$2*10^{-4}$ cm^{2}K/W is connected between the gate and a 300 K IT BC. This lumped resistance takes into account the thermal resistance due to the gate dielectric and to the gate–SiO_2 interface and it does not depend on the gate insulator thickness [78].
Finally, the vertical $x-y$ and $z-y$ planes (see Fig. 5.3 and 5.4) are treated as adiabatic boundaries.

5.3 Comparative analysis of SHE in different SOI architectures

In this section we are interested to explore how SHE impact the performance of different SOI architectures: single– and double–gate planar SOI MOSFETs as well as FinFET.

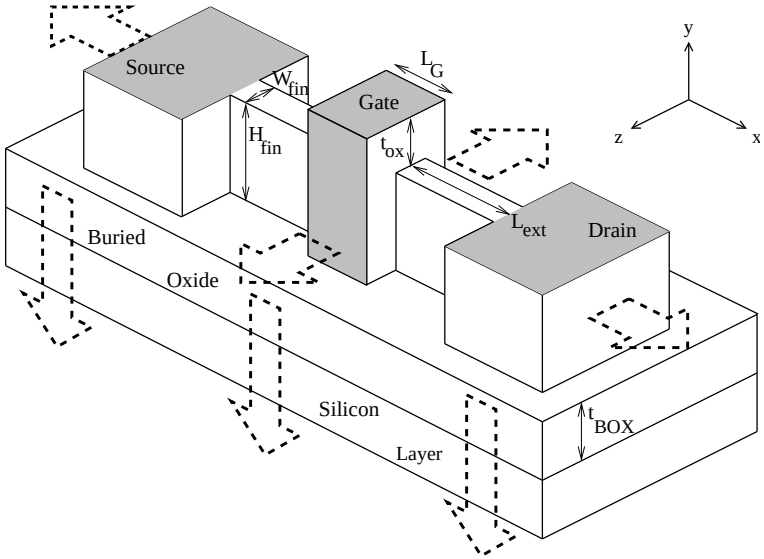

Figure 5.4: Simple sketch of the simulated FinFET. The main cooling paths, through which heat is dissipated, are reported (dashed contoured arrows). The figure is not drawn to scale.

5.3.1 Simulated devices

Fig. 5.3 provides a simple sketch of the simulated planar devices and Fig. 5.4 presents the simulated FinFET as well.

When heat is generated in the active region of the device, there are different paths through which it can be dissipated. In Fig. 5.4 the dashed arrows indicate these paths:

- through the S/D contacts: this possibility is indicated by the arrows along the x–axis;
- through the vertical direction (y–axis): the heat could be dissipated through the buried oxide and the dielectric passivation layer that covers the whole chip;
- between adjacent devices (z–direction).

Tab. 5.2 reports the values assumed for the most important device parameters. The simulation domain is 14 μm wide in the direction along the channel; we have performed simulations featuring a domain 8 μm and 20 μm wide as well, and we have found that the differences in terms of provided current and temperature are less than 1%. This trend is common to all the considered structures, and shows that the simulation domain adopted does not impact the results.
A 1.8 μm thick Silicon substrate layer is placed beneath the BOX. An abrupt junction is assumed between the S/D regions and the substrate.

	DGSOI	SGSOI–1/2	FinFET
Gate Length L_G (nm)	50	64/50	50
Gate Work–func. ϕ_G [eV]	4.6	4.62/4.61	4.6
Gate Ox. Thick. t_{OX} [nm]	1	1/1	1
Gate Oxide Dielectric Constant ε_{OX}	3.9	3.9/3.9	3.9
Silicon Thickness t_{Si} [nm]	10	10/5	n.a.
Channel Width W_{ch} [nm]	250	250/250	n.a.
Fin Width W_{fin} [nm]	n.a.	n.a.	10
Fin Height H_{fin} [nm]	n.a.	n.a.	60
Top–gate Ox. Thick. t_{OX}^{top} [nm]	n.a.	n.a.	5
S/D access length [nm]	35	35	35
S/D Doping Conc. [cm^{-3}]	10^{20}	10^{20}	10^{20}
Ch. Doping Conc. [cm^{-3}]	10^{15}	10^{15}	10^{15}
Buried Ox. Thick. t_{BOX} [nm]	50	50	50
Supply Voltage V_{DD} [V]	1.0	1.0/1.0	1.0

Table 5.2: Key parameters adopted in the comparison presented in this section. Third column: two different SGSOI architectures, defined to have the same IT electrical characteristics as DGSOI and FinFET, are reported.

As we are interested in the impact of self–heating, care has been taken in order to ensure the same IT electrical characteristics for all the devices, that is the same threshold voltage V_{th}, the same transconductance and a good tolerance to the SCE (i.e. we have designed the structures in order to have drain induced barrier lowering values lower than 100 mV/V).
A pure SiO_2 gate dielectric with t_{OX}=1 nm has been chosen. Regarding the FinFET, the top gate features an oxide much thicker than the lateral ones (t_{OX}^{top}=5 nm), so that the contribution to the drain current due to the top gate is negligible.
The DGSOI and FinFET transistors feature the same gate length (L_G=50 nm) as well as the same silicon film thickness (t_{Si}=W_{fin}=10 nm).
The SGSOI architecture is strongly affected by SCE. In order to overcome this problem, two different solutions in designing the SGSOI transistor have been adopted: the first one (named SGSOI–1) features t_{Si} equal to 10 nm but a gate slightly longer than the DGSOI and FinFET (64 nm); the second one (named SGSOI–2) features the same gate length as DGSOI and FinFET but much thinner silicon body (5 nm).
While the FinFET's fin width is set to 10 nm, the planar SOI devices feature a channel width W^{ch}=$5L_G$, a realistic minimum width for planar devices.

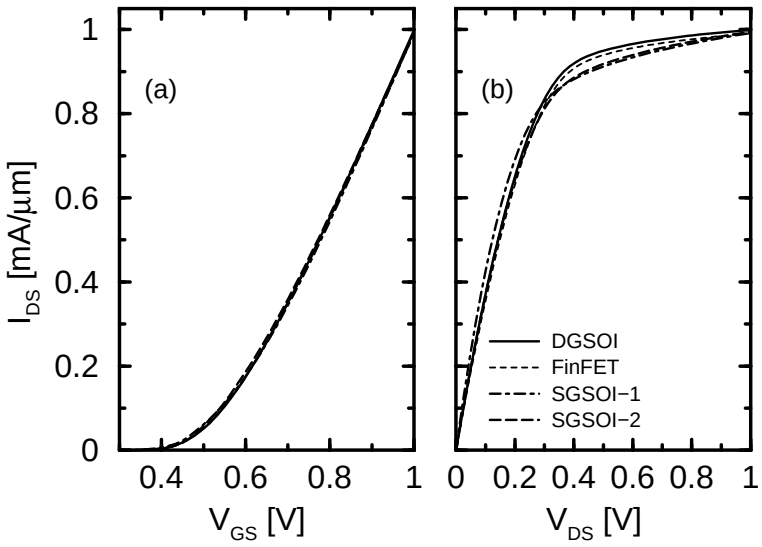

Figure 5.5: (a) Transfer characteristics (V_{DS}=1.0 V) and (b) output characteristics (V_{GS}=1.0 V), calculated by 3D DD IT simulations. The figure confirms that the different devices feature similar simulated I–V curves.

5.3.2 Results

Fig. 5.5 shows the transfer and output characteristics, obtained from IT simulations of the transistors described in the previous subsection. The devices features almost the same IT currents per unit width as is required for a consistent comparison.

Fig. 5.6 shows the transfer and output characteristics, obtained from ET simulations. In these simulations, the gate and S/D contacts are assumed adiabatic, therefore the power generated in the active region can be dissipated only through the buried oxide.

The SHE–related degradation of I_{DS} strongly depends on the device structure: it is maximum for the DGSOI, less critical for the FinFET, and minimum for SGSOI–1. This trend is confirmed by Fig. 5.7, where the maximum temperature rise ΔT_{MAX} (defined as $T_{MAX}-300$ K, T_{MAX} being the maximum temperature reached inside the device) is plotted as a function of the dissipated power per unit width P. The slope of each line can be interpreted as the thermal resistance R_{TH} associated to the corresponding device; the table inset in the same figure reports R_{TH} values for V_{GS}=V_{DS}=V_{DD}. DGSOI presents the highest thermal resistance, SGSOI–1 the lowest one.

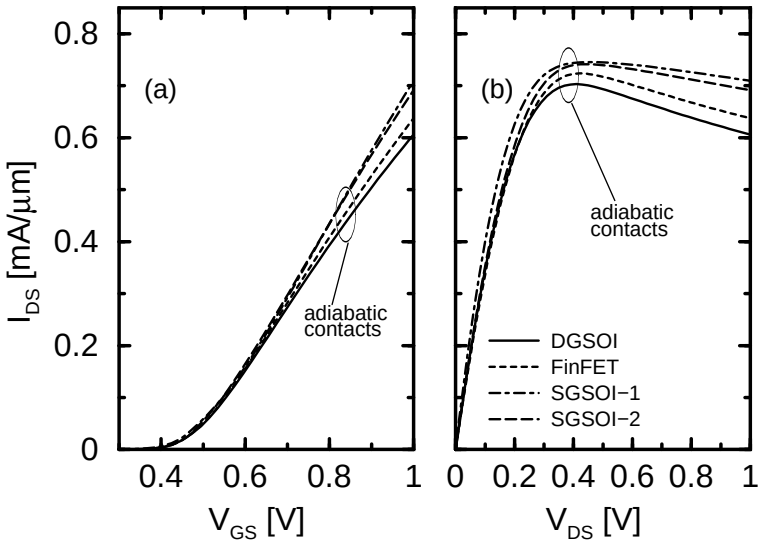

Figure 5.6: (a) Transfer characteristics (V_{DS}=1.0 V) and (b) output characteristics (V_{GS}=1.0 V), calculated by 3D DD ET simulations. The gate and S/D contacts are treated as adiabatic, and the heat flux can occur only through the BOX.

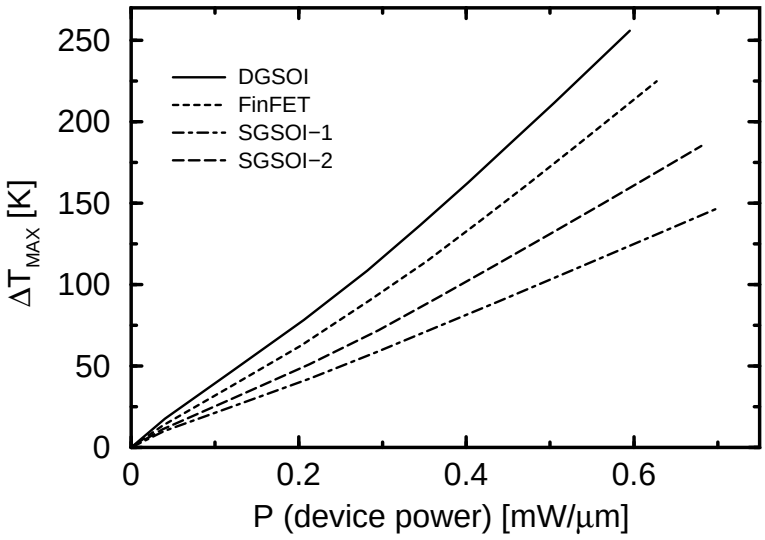

Figure 5.7: ΔT_{MAX} (i.e. $T_{MAX}-300$ K) as a function of the dissipated power per unit width, for the four considered devices; the gate and S/D contacts are treated as adiabatic. Consistently to Fig. 5.6, SGSOI–1 presents the lowest temperature rise and equivalent thermal resistance R_{TH} (i.e. $\Delta T_{MAX}/P$)

	DGSOI	SGSOI–1/2	FinFET
R_{TH} [Kμm/mW]	430	210/272	359

Table 5.3: Thermal resistance R_{TH} values, evealuated for V_{GS}=V_{DS}=V_{DD}, for the simulated devices under the assumption of no heat dissipation through the contacts.

The differences between SGSOI and the other two structures are related to the wider overlap area between the Silicon body and the underlying BOX, through which most of the heat flux occurs. This area is smaller in DGSOI and FinFET (in particular, it is halved in DGSOI, with respect to SGSOI), and this leads to a degraded heat dissipation. Furthermore, the larger SHE of SGSOI–2 compared to SGSOI–1 is due to the degradation of k_{Si} occurring as t_{Si} is scaled down (it should be remembered that t_{Si} is 10 nm in SGSOI–1 and 5 nm in SGSOI–2, and that the thermal conductivity is a decreasing function of the layer thickness, see Figure 5.2). FinFET presents lower SHE than DGSOI because W_{fin} is much lower than W^{ch} (Tab. 5.2). In this case, the impact of the heat dissipation occurring in the direction of device width is larger in the FinFET than in the DGSOI, because its relevance increases as the device width decreases.

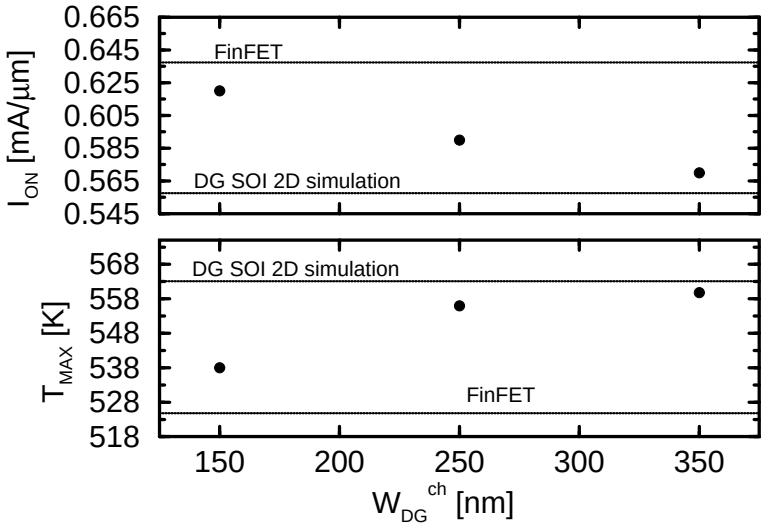

Figure 5.8: I_{ON} (i.e. I_{DS} for V_{GS}=V_{DS}=V_{DD}) and T_{MAX} vs. DGSOI channel widths (W_{DG}^{ch}). The gate and S/D BC are assumed adiabatic, therefore the heat dissipation can occur only through the BOX. As W_{DG}^{ch} decreases, the values depart from the 2D-DGSOI simulation case (infinite W_{DG}^{ch}), approaching the FinFET case.

Fig. 5.8 confirms this behavior: as W^{ch} of the DGSOI is scaled down, the values of I_{ON} and T_{MAX} depart from those obtained from the 2D–DGSOI simulation (corresponding to infinite W^{ch}), and approach the FinFET's ones.

Fig. 5.9 shows the transfer and output characteristics, obtained from ET simulations, when the S/D and gate contacts are treated as 300 K BC. In this case they act as cooling contacts, and they contribute to dissipate the heat generated in the active region. The differences between the devices in terms of drain current degradation become less evident compared to Fig. 5.6: most of the cooling occurs through the S/D contacts and the cross-sectional area of the S/D access regions become the most relevant parameter, leading to larger SHE in the SGSOI-2 and DGSOI cases, compared to the SGSOI-1 and FinFET. For the forthcoming technological nodes the distance between the active region and the S/D contacts is expected to be reduced with the device scaling; this trend could exploit the cooling occurring through the contacts and contribute to keep SHE under control. The dependence of SHE on S/D access length will be investigated in the next chapter for a 30nm gate length FinFET.

To conclude, we performed 3D fully–coupled electro–thermal simulations of different silicon–on–insulator structures. Planar single– and double–gate as well as FinFET transistors have been analyzed, in order to explore how self–heating effect acts on the device performance for these different architectures. It should be noted that the devices

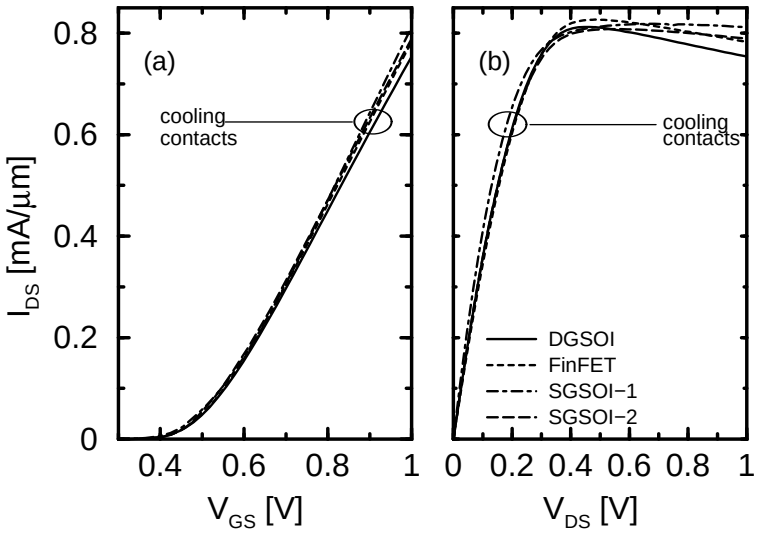

Figure 5.9: (a) Transfer characteristics (V_{DS}=1.0 V) and (b) output characteristics (V_{GS}=1.0 V), calculated by 3D DD ET simulations. 300 K IT BC are assumed at the gate and S/D contacts.

involved in this comparison have been designed in order to have the same isothermal electrical characteristics as threshold voltage, transconductance and DIBL.
Simulations show that SHE detrimentally impacts the device performance in terms of provided current, and its effect is dependent on the device structure. In particular, the thermal resistance associated to the S/D access regions, as well as the ratio between the surface available for the vertical heat dissipation (heat dissipation through the buried oxide) and the volume of the active region, change from a structure to another.

Chapter 6

SHE in 30 nm gate length FinFET

In this chapter, we apply 3D electro–thermal simulations to the analysis of state–of–the–art FinFET transistors, whose a simple sketch is presented in Fig 5.4. We studied the impact of self–heating on the performance of FinFET devices and its dependence on the main geometrical parameters, as source and drain extension length, buried oxide thickness and inter–fins distance. Moreover, technological solutions adopted in FinFET technology, as epitaxially grown source/drain or fin height reduction, have been explored as well.

6.1 Simulated devices

The analysis described in this chapter has been developed by electro–thermal simulations performed with Sentaurus from Synopsys. The adopted simulation approach has been already detailed in Section 5.2 for the analysis of SHE between different SOI architectures.
The main characteristics of the reference FinFET are listed in Tab. 6.1. It should be noted that the FinFET adopted for the comparison reported in Section 5.3.1 features a gate length equal to 50 nm, while in the current case the gate length is equal to 30 nm; moreover, the gate oxide is slightly thicker (t_{OX}=1.2 nm) and it is the same for the lateral and top channels. Finally, the whole device is covered by a SiO_2 passivation layer, whose thickness t_{PAS} is kept constant and equal to 200 nm (this passivation layer is not shown in Fig. 5.4).
Because of the small current provided by a single ”fin”, real FinFET devices consist of multiple fins in parallel (from at least 20 up to 50) with connected source and drain; from a thermal point–of–view, the temperature distribution inside the transistor is a function of the fin position: the inner fins are hotter than the outer ones [79]. In our case, due to the symmetries and the adopted boundary conditions, the simulated structure is representative of an inner fin, therefore we consider the worst case in terms of heat dissipation.

Quantity	Value
Gate Length L_G [nm]	30
Gate Work–function ϕ_G [eV]	4.6
Gate Oxide Thickness t_{OX} [nm]	1.2
Gate Oxide Dielectric Constant ε_{OX}	3.9
Fin Width W_{fin} [nm]	10
Fin Height H_{fin} [nm]	60 , 40
S/D Extension Length L_{ext} [nm]	35
Channel Doping Conc. N_{CH} [cm^{-3}]	10^{15}
S/D Doping Conc. $N_{S/D}$ [cm^{-3}]	10^{20}
Passivation layer thickness t_{PAS} [nm]	200
Supply Voltage V_{DD} [V]	1.1

Table 6.1: Key parameters of the nominal 30 nm gate length FinFET involved in the analysis.

In order to allow a fair comparison between different structures, FinFET currents are normalized by the effective device width (given by H_{fin}+$W_{fin}/2$, because only one–half of the FinFET is simulated, as mentioned in Section 5.2). In the following of this analysis, the S/D and gate contacts are treated as 300 K boundary conditions for all the simulated devices.

6.2 Results

Fig. 6.1 shows the output characteristics I_{DS}–V_{DS} for V_{GS}=V_{DD} obtained from IT and ET simulations of the reference device, with t_{BOX}=50 nm and fin–pitch (Δ_{fin}, that is the distance between adjacent fins) equal to 60 nm. It should be noticed that SHE detrimentally impacts the device performance and causes a negative differential output conductance g_{out} in the saturation region. Data regarding drain current reduction ranging from about 10% to 22% could be found in [80] where the authors compare different experimental I–V curves, obtained from conventional DC measurements and from pulsed measurements, in which the devices remain at room temperature. Moreover, phenomena as drain current degradation and negative output conductance have been reported even in [81].
Fig. 6.2 shows the simulated ON–current and the maximum temperature inside the device T_{MAX} as a function of t_{BOX} varying from 10 nm to 200 nm, for given Δ_{fin}=60 nm and L_{ext}=35 nm. The degradation of I_{ON} is larger than 30% with respect to the standard 300 K DD IT simulation. On the other hand, the simulations reveal a very weak dependence on t_{BOX}, and the ON–current seems to be almost insensitive to this parameter. This is due to the cooling action provided by S/D contacts: since they are treated

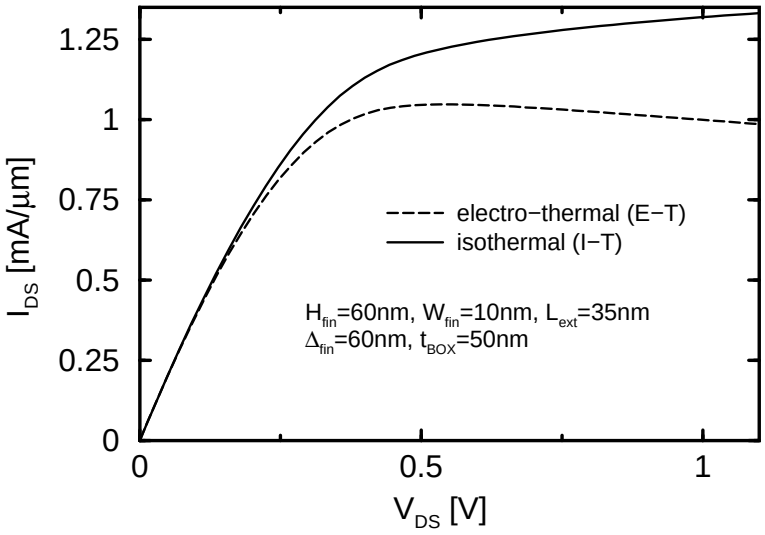

Figure 6.1: Output characteristics (V_{GS}=1.1 V) calculated by 3D IT and ET simulations of the nominal 30 nm gate length FinFET detailed in Tab. 6.1.

as 300 K boundaries and are placed close to the intrinsic device (L_{ext}=35 nm), their impact on the maximum temperature is much larger compared to that of the substrate contact.

Fig. 6.3 presents I_{ON} and T_{MAX} as a function of Δ_{fin} varying from 15 nm to 200 nm, for given t_{BOX}=50 nm and L_{ext}=35 nm. By relaxing the inter–fins spacing, T_{MAX} reduces and I_{ON} increases. This occurs because the thermal interaction between adjacent fins decreases when Δ_{fin} increases: farther is a fin from the adjacent one, easier is the dissipation of the heat generated in the active region. A 10% difference in terms of I_{ON} between the cases corresponding to the lowest and highest Δ_{fin} values should be noted. In Fig. 6.4 the peak temperature rise ΔT_{MAX} is plotted as a function of the dissipated power per unit width, for some cases presented in Fig. 6.3. A more compact layout (smaller Δ_{fin}) leads to larger heating, although the dependence on Δ_{fin} is not dramatic. Fig. 6.5 presents I_{ON} and T_{MAX} as a function of L_{ext}, for a given Δ_{fin} and for t_{BOX}=20 and 50 nm. A maximum in T_{MAX} can be noted for L_{ext} between 50 nm and 60 nm. This is due to the competition of two concurrent effects that occur when S/D contacts are placed farther and farther from the active region:

- source and drain are treated as thermal boundaries at T=300 K, so when L_{ext} increases, the cooling effects due to S/D reduces and T_{MAX} increases;
- the heat flow through the BOX can exploit a larger available area (therefore reducing the temperature of the active region).

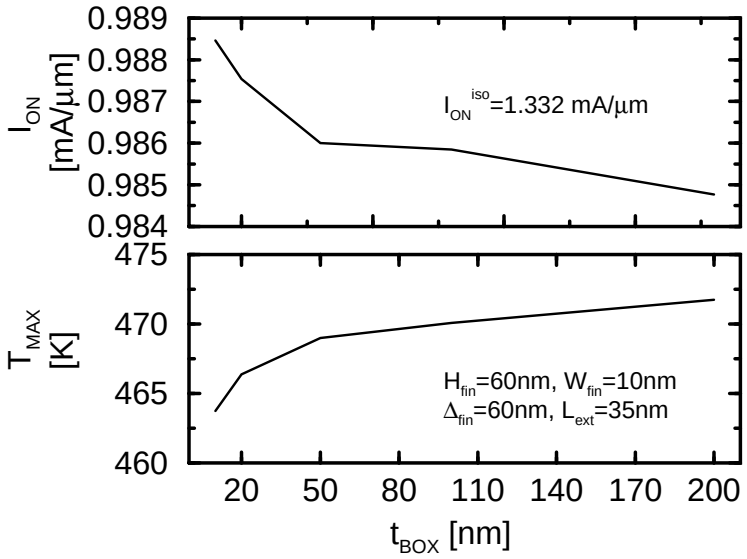

Figure 6.2: I_{ON} (top) and T_{MAX} (bottom) as a function of t_{BOX} for the FinFET under analysis; the results show a weak dependence on the BOX thickness.

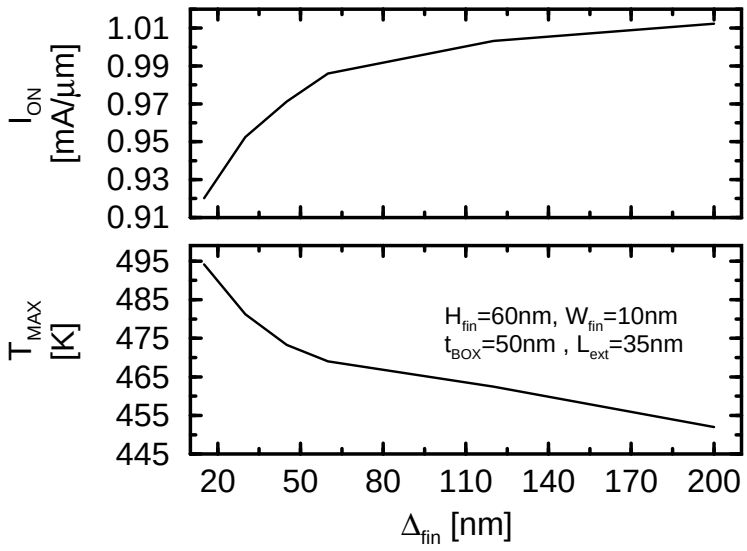

Figure 6.3: I_{ON} (top) and T_{MAX} (bottom) as a function of Δ_{fin} for the simulated structure. The lateral heat dissipation improves increasing the fin–pitch (therefore reducing the thermal interaction between adjacent fins); as a consequence, T_{MAX} reduces.

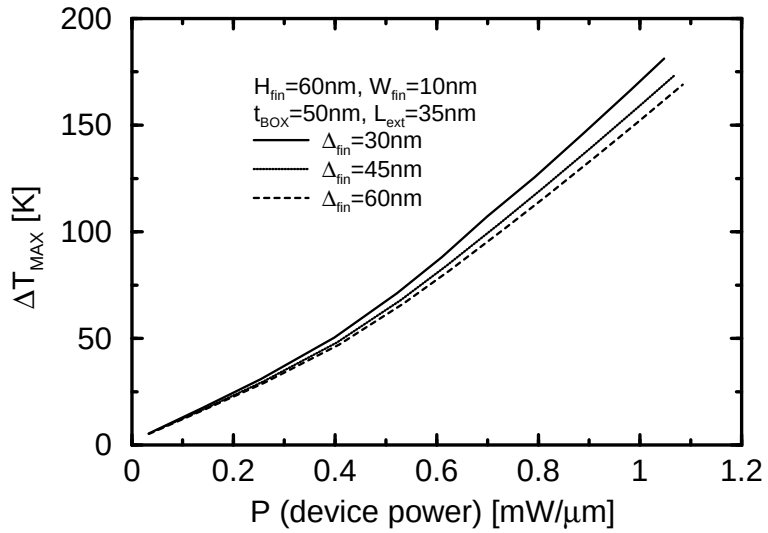

Figure 6.4: ΔT_{MAX} (i.e. T_{MAX}–300 K) as a function of the dissipated power per unit width, for some Δ_{fin} values presented in Fig. 6.3.

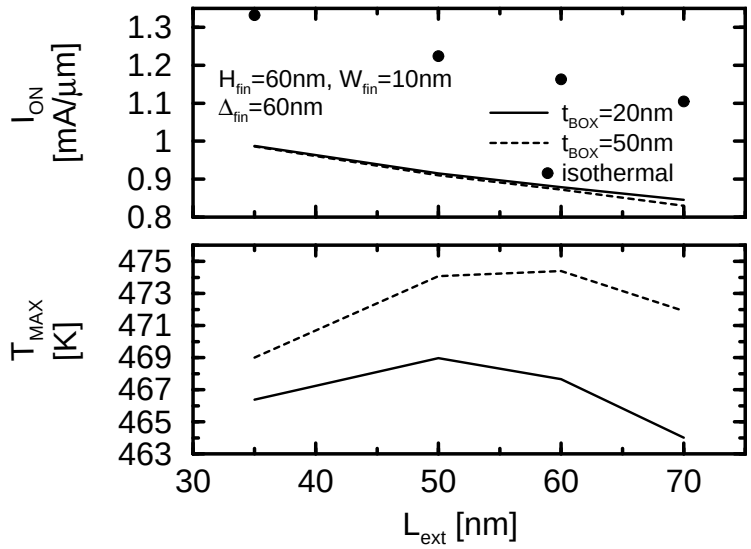

Figure 6.5: I_{ON} (top) and T_{MAX} (bottom) as a function of L_{ext}, for two different t_{BOX} values. I_{ON} shows a monotone reduction, related to the larger series resistance, while T_{MAX} presents a maximum between L_{ext}=50 and 60 nm, due to the competition between longer S/D access regions (larger associated thermal resistances) and wider available area for the heat flux between the silicon fin and the BOX.

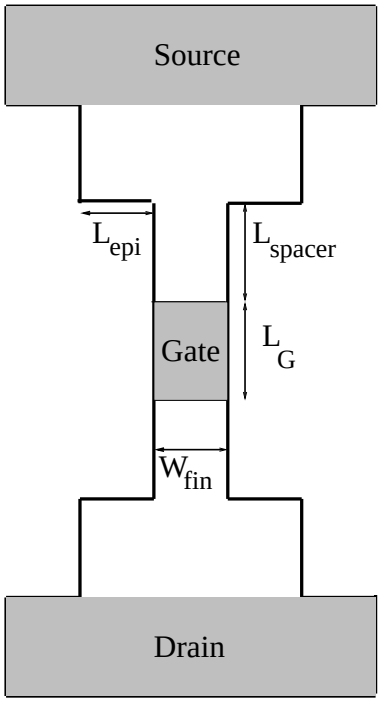

Figure 6.6: Top–view of the RSD FinFET. The figure is not drawn to scale. Note the parameters involved in the simulations of RSD: L_{spacer} (length of the spacer between the gate edge and the epitaxially grown silicon region) and L_{epi} (thickness of the silicon epitaxy).

Therefore, two opposite trends occur: the reduced cooling capability due to the longer L_{ext}, and the increased cooling capability due to the larger overlap between the silicon fin and the buried oxide. Nevertheless, I_{ON} decreases with L_{ext}, due to the impact of larger series resistance. A possible solution to reduce the parasitic series resistances is represented by raised S/D (RSD), obtained by increasing the fin thickness outside of the gate with silicon epitaxy [82]. In order to study the impact of such technological approach on SHE, the dependence on the spacing between the gate edge and the RSD (L_{spacer}), as well as an RSD epi growth (L_{epi}) has been analysed. Fig. 6.6 provides a top view of the modified devices.
In Fig. 6.7 T_{MAX} and I_{ON} are plotted as a function of L_{spacer} for L_{epi} equal to 10 and 20 nm. It should be noticed that the gain in terms of I_{ON}, with respect to the device without RSD, is larger for ET simulations than IT ones: in particular, for a device featuring L_{spacer}=5 nm and L_{epi}=20 nm, ET simulation shows an I_{ON} that is 25.8% larger than the one provided by the FinFET without RSD, while for the same case IT simulation shows an improvement equal to 17.5%.
This is because in the ET case two trends concur: the reduced parasitic series resistances,

and the enhanced dissipation of heat generated inside the device. In fact in the case of RSD, the larger cross sectional area of the access region reduces the thermal resistance towards the S/D 300 K BC; furthermore a larger overlap of the fin over the BOX is exploited by the vertical heat flux toward the bulk. This effect is larger for thicker RSD and reduced L_{spacer}.

In real FinFETs a reduction of the fin height can be useful, because it simplifies the fin definition process and allows to avoid doping shadowing effects. In Fig. 6.8 I_{ON} and T_{MAX} are plotted as a function of t_{BOX}, for H_{fin}=60 nm and H_{fin}=40 nm. The ON–currents per unit width obtained from IT simulations are obviously the same. However, SHE is less detrimental in the device featuring H_{fin}=40 nm, that reaches lower T_{MAX} and therefore provides a larger I_{ON}. This occurs because reducing H_{fin}, the ratio between the surface available for the vertical heat flux (that is the overlap areas with the BOX and the passivation layer) and the volume of the active region increases, allowing better cooling; moreover, this effect becomes more evident reducing t_{BOX}.

To conclude, the analysis has been focused on state–of–the–art 30 nm gate length FinFETs, in order to explore the dependence of SHE on geometrical parameters as buried oxide thickness, S/D extension length, fin pitch and fin height. Even if SHE implies a severe degradation in terms of ON–current, our results show that its dependence on the main layout parameters is weak. Finally, different technological solutions as raised S/D (adopted in order to reduce series resistance) and H_{fin} reduction (adopted in order to reduce doping shadowing) have been explored. The simulations show that from a thermal point–of–view Δ_{fin} is not a critical parameter, therefore multi–finger FinFETs with limited fin height H_{fin} and small fin–pitch Δ_{fin} seem to represent a good option in order to increase the integration density keeping SHE under control, with limited reduction of effective gate area and current, for given device footprint. On the other hand, the adoption of a large number of fins is likely to give rise to large parasitic capacitances, larger area occupation and larger production cost due to more complex layout.

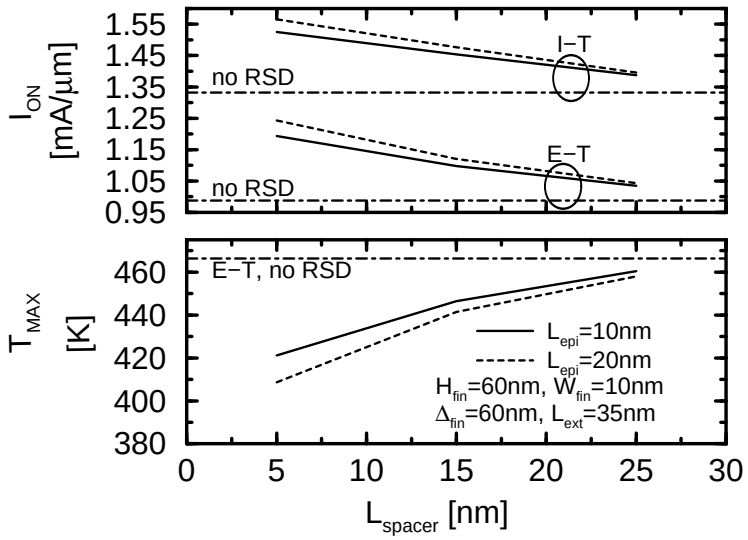

Figure 6.7: I_{ON} (top) and T_{MAX} (bottom) vs. L_{spacer} for different L_{epi}. In the ET case, performance gets the advantage of both lower S/D parasitic series resistance and larger available area for the heat dissipation.

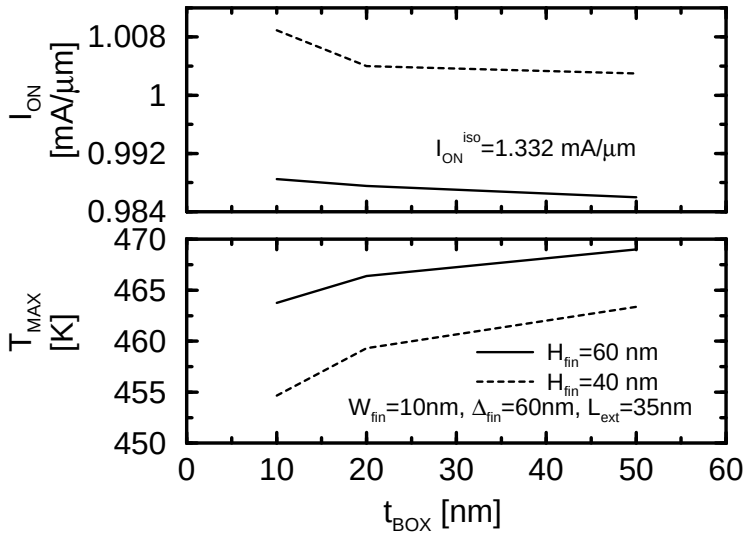

Figure 6.8: I_{ON} (top) and T_{MAX} (bottom) vs. t_{BOX}, for two different H_{fin} values. The ratio between the surface available for the heat flux and the volume of the active region is larger in the device featuring H_{fin}=40 nm, that therefore reaches lower temperatures.

Chapter 7

Conclusions

In this thesis we have discussed some aspects concerning the adoption of novel architectures as Silicon–On–Insulator transistors. Our research has aimed to provide useful results about the transport properties of new devices fabricated with materials alternative to silicon, as well as to offer new insights of the heat generation and dissipation in SOI devices.

In the first part of the manuscript, we have presented a simulation study of nano–scale SOI devices by using the Monte Carlo method. Our 3DEG full–band MC tool has been modified in order to simulate the electron transport across conduction band discontinuities: different approaches have been implemented, depending on the type of discontinuity (abrupt or graded) and the kinetic energy of the particles crossing such discontinuities. Accurate simulations on simplified n–doped unidimensional structures have demonstrated the validity and correctness of the introduced modifications.
Afterward the code has been tested, we have applied it to explore the impact of conduction band offsets on the device performance, in terms of provided current. The study has been performed on n–channel double–gate SOI transistors, featuring conduction–band offsets between the source and drain regions and the channel. Such configuration can be obtained by adopting for the source and drain regions a material different from that adopted for the channel; the purpose of this approach is to increase the injection velocity of electrons entering the channel, and as a consequence to increase the provided drain current. Both abrupt and graded discontinuities have been considered; the gate work–function has been modified in order to have the same I_{OFF} for all the simulated cases. The quantum–mechanical phenomenon of tunnel effect through the energy barrier that arises in presence of heterojunctions, is not taken into account.
Although abrupt conduction band offsets between the source and drain region are expected to enhance the injection velocity of electrons coming from the source and entering the channel, and thus the current provided by the transistor, simulations of double–gate SOI MOSFETs pointed out that the charge accumulation next to the discontinuity

influences the device electrostatics reducing the charge available for transport, overcompensating the velocity improvement. Due to the same mechanism, only small current improvements are obtained in the case of graded discontinuities. Quantum–mechanical corrections to the electrostatic potential, available in order to account for carrier quantization on the spatial distribution of the inversion charge, are not taken into account: this approach can lead to an overestimated prediction of the current drive capability of the devices, but the change in current with respect to the reference case is similar, and we have found the same trend between the ON–currents provided by the different structures.
It should be noted that we have focused on the impact on the performance of the discontinuity alone. In many practical cases, the materials used to fabricate the source and drain regions induce strain in the channel, that is the main responsible of the I_{ON} improvement, whereas according to our simulations, the effect of the band offset alone seems to be modest.

In the second part of this thesis, we performed 3D fully–coupled electro–thermal simulations of different silicon–on–insulator structures. Planar single– and double–gate as well as FinFET transistors have been analyzed, in order to explore how self–heating effect acts on the device performance for these different architectures. It should be noted that the devices involved in this comparison have been designed in order to have the same isothermal electrical characteristics.
Simulations show that self–heating detrimentally impacts the device performance, and it is dependent on the device structure. In particular, the thermal resistance associated to the source/drain access regions, as well as the ratio between the surface available for the vertical heat dissipation (heat dissipation through the buried oxide) and the volume of the active region, changes from a structure to another.
Afterwards the analysis has been focused on state-of-the-art 30 nm gate length FinFETs, in order to explore the dependence of SHE on geometrical parameters as buried oxide thickness, S/D extension length, fin pitch and fin height. Even if SHE implies a severe degradation in terms of ON–current, our results show that its dependence on the main layout parameters is weak. Finally, different technological solutions as raised S/D (adopted in order to reduce series resistance) and fin height reduction (adopted in order to facilitate the fabrication and to reduce doping shadowing) have been explored. The simulations show that from a thermal point–of–view the inter–fin distance is not a critical parameter, therefore multi–finger FinFETs with limited fin height and small fin–pitch seems to represent a good option in order to increase the integration density keeping SHE under control, with limited reduction of effective gate area and current, for given device footprint. On the other hand, the adoption of a large number of fins is likely to give rise to large parasitic capacitances, larger area occupation and larger production cost due to more complex layout.

Bibliography

[1] G. E. Moore, ”Cramming more components onto integrated circuits”, *Electronics*, 1965, vol. 38, no. 8 (republished in *Proceedings of the IEEE*, 1998, vol. 86, no. 1, p. 82).

[2] G. E. Moore, ”Progress in digital integrated electronics”, *IEEE IEDM Technical Digest*, 1975, vol. 21, p. 11.

[3] P. K. Bondyopadhyay, ”Moore’s law governs the silicon revolution”, *Proceedings of the IEEE*, 1998, vol. 86, no. 1, p. 78.

[4] R. H. Dennard *et al.*, ” Design of ion–implanted MOSFET’s with very small physical dimensions”, 1974, vol. 9, no. 5, p. 256.

[5] G. Baccarani, M. R. Wordeman, R. H. Dennard, ”Generalized scaling theory and its application to a $1/4$ micrometer MOSFET design”, *IEEE Transactions on Electron Devices*, 1984, vol. 31, no. 4, p. 452.

[6] D. J. Frank, Y. Taur, H.–S. P. Wong, ”Generalized scale length for two–dimensional effects in MOSFETs”, *IEEE Electron Device Letters*, 1998, vol. 19, no. 10, p. 385.

[7] R. R. Troutman, ”VLSI limitations from drain–induced barrier lowering”, *IEEE Transactions on Electron Devices*, 1979, vol. 26, no. 4, p. 461.

[8] D. A. Buchanan, ”Scaling the gate dielectric: Materials, integration, and reliability”, *IBM Journal of Research and Development*, 1999, n. 49, p. 245.

[9] M. Pedram, S. Nazarian, ”Thermal Modeling, Analysis, and Management in VLSI Circuits: Principles and Methods”, *Proceedings of the IEEE*, 2006, vol. 94, no. 8, p. 1487.

[10] A. Topol *et al.*, ”Lower Resistance Scaled Metal Contacts to Silicide for Advanced CMOS”, *Symposium on VLSI Technology Digest of Technical Papers*, 2006, p. 116.

[11] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer–scale MOSFETs", *IEEE Transactions on Electron Devices*, 2003, vol. 50, no. 9, p. 1837.

[12] B. Doris *et al.*, "Extreme scaling with ultra–thin Si channel MOSFETs", *IEEE IEDM Technical Digest*, 2002, p. 267.

[13] G. D. Wilk, R. M. Wallace, J. M. Anthony, "High–k gate dielectrics: Current status and materials properties considerations", *Journal of Applied Physics*, 2001, n. 89, p. 5243.

[14] E. P. Gusev *et al.*, "Ultrathin high-k gate stacks for advanced CMOS devices", *IEEE IEDM Technical Digest*, 2001, p. 451.

[15] K. Mistry *et al.*, "A 45 nm Logic Technology with High–k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193 nm Dry Patterning, and 100% Pb–free Packaging", *IEEE IEDM Technical Digest*, 2007, p. 247.

[16] T. Ludwig, I. Aller, V. Gernhoefer, J. Keinert, E. Nowak, R. V. Joshi, A. Mueller, S. Tomaschko, "FinFET technology for future micropocessors ", *IEEE International SOI Conference Proceedings*, 2003, p. 33.

[17] D. Esseni, M. Mastrapasqua, G.K. Celler, C. Fiegna, L. Selmi and E.Sangiorgi, "An Experimental study of mobility enhancement in ultrathin SOI transistors operated in double-gate mode", *IEEE Transactions on Electron Devices*, 2003, vol. 50, no. 3, p. 802.

[18] H. van Meer, K. De Meyer, "Ultra–Thin Film Fully–Depleted SO1 CMOS with Raised G/S/D Device Architecture for Sub–100 nm Applications", *IEEE International SOI Conference Proceedings*, 2001, p. 45.

[19] J. Widiez, M. Vinet, B. Guillaumot, T. Poiroux, D. Lafond, P. Holliger, O. Weber, V. Barral,B. Previtali, F. Martin, M. Mouis, S. Deleonibus, "Fully Depleted SOI MOSFETs with WSi_x metal gate on HfO_2 gate dielectric", *IEEE International SOI Conference Proceedings*, 2006, p. 161.

[20] T. Ohtou, N. Sugii, T. Hiramoto, "Impact of Parameter Variations and Random Dopant Fluctuations on Short–Channel Fully Depleted SOI MOSFETs With Extremely Thin BOX", *IEEE Electron Device Letters*, 2007, vol. 28, no. 8, p. 740.

[21] R. H. Yan, A. Ouzmard, K. F. Lee "Scaling the Si MOSFET: from bulk to SOI to bulk", *IEEE Transactions on Electron Devices*, 1992, vol. 39, no. 7, p. 1704.

[22] J. P. Colinge, "Multiple-gate SOI MOSFETs", *Solid State Electronics*, 2004, vol. 48, no. 6, p. 897.

[23] Y. Taur, ”An Analytical solution to a Double-Gate MOSFET with undoped body”, *IEEE Electron Device Letters*, 2000, vol. 21, no. 5, p. 245.

[24] J. Widiez, F. Dauge, M. Vinet, T. Poiroux, B. Previtali, M. Mouis, S. Deleonibus, ”Experimental gate misalignment on double gate SOI MOSFETs”, *IEEE International SOI Conference Proceedings*, 2004, p. 185.

[25] F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, T. Elewa, ”Double–gate silicon on insulator transitor with volume inversion: a new device with greatly enhanced performance”, *IEEE Electron Device Letters*, 1987, vol. 8, no. 9, p. 410.

[26] W. Chaisantikulwat, M. Mouis, G. Ghibaudo, C. Gallon, C. Fenouillet–Beranger, D. K. Maude, T. Skotnicki, S. Cristoloveanu, ”Differential magnetoresistance technique for mobility extraction in ultra–short channel FDSOI transistors”, *Solid State Electronics*, 2006, vol. 50, no. 4, p. 637.

[27] TCAD User Manual, *http://www.synopsys.com*.

[28] N. Metropolis, S. Ulam, ”The Monte Carlo method”, *Journal of the American Statistical Association*, 1949, vol. 44, p. 335.

[29] R. Eckhardt, ”Stan Ulam, John Von Neumann and the Monte Carlo method”, *Los Alamos Science Special Issue*, 1987, p. 131.

[30] M. Mouis, M. Dollfus, B. Mougel, J. F. Pône, R. Castagné, ”Monte Carlo simulation of the effects induced by real–space transfer in a HEMT”, *High speed electronics*, Springer–Verlag, Berlin, 1986, p. 35.

[31] M. L. Cohen, T. K. Bergstresser, ”Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc–blende structures”, *Physical Review*, 1966, vol. 141, p. 789.

[32] D. Bohm, ”Quantum theory”, Prentice–Hall, Englewood Cliffs, NJ, 1951.

[33] R. F. Pierret, ”Semiconductor Device Fundamentals”, Addison–Wesley, 1996, ISBN–10:0201543931, ISBN–13:9780201543933.

[34] P. Bhattacharya, ”Semiconductor Optoelectronic Devices”, 2^{nd} edition, Prentice Hall, 1997, ISBN–10:0134956567, ISBN–13:9780134956565

[35] W. Hafez, W. Snodgrass, M. Feng, ”12.5 nm base pseudomorphic heterojunction bipolar transistor achieving f_T=710 GHz and f_{MAX}=340 GHz”, *Applied Physics Letters*, 2005, vol. 87, p. 252109.

[36] C. Weisbuch, B. Vinter, ”Quantum Semiconductor Structures: Fundamentals and Applications”, Academic Press, ISBN–10:0127426809, ISBN–13:9780127426808.

[37] W. Liu, ”Handbook of III–V Heterojunction Bipolar Transistors”, Wiley–interscience, 1998, ISBN–10:0471249041, ISBN–13:9780471249047.

[38] T. Mizuno, N. Sugiyama, T. Tezuka, Y. Moriyama, S. Nakaharai, T. Maeda, S. Takagi, ”High–Speed Source Heterojunction-MOS-Transistor (SHOT) Utilizing High–Velocity Electron Injection”, *IEEE Transactions on Electron Devices*, 2005, vol. 52, no. 12, p. 2690.

[39] T. Mizuno, T. Irisawa, S. Takagi, ”Device Design of High–Speed Source–Heterojunction–MOS Transistors (SHOTs): Optimization of Source Band Offset and Graded Heterojunction”, *IEEE Transactions on Electron Devices*, 2007, vol. 54, no. 10, p. 2598.

[40] T. Mizuno, Y. Moriyama, T. Tezuka, N. Sugiyama, S. Takagi, ”Experimental Study of Single–Heterojunction MOS Transistors (SHOTs) under Quasi–Ballistic Transport”, *Symposium on VLSI Technology Digest of Technical Papers*, 2008, p. 22.

[41] P. Palestri, N. Barin, D. Esseni and C. Fiegna, ”Stability of Self-Consistent Monte-Carlo Simulations: Effects of the Grid Size and of the Coupling Scheme”, *IEEE Transactions on Electron Devices*, 2006, vol. 53, no. 6, p. 1433.

[42] P. Palestri, N. Barin, D. Esseni and C. Fiegna, ”Revised Stability Analysis of the Non - Linear Poisson Scheme in Self-Consistent Monte-Carlo Device Simulations”, *IEEE Transactions on Electron Devices*, 2006, vol. 53, no. 6, p. 1443.

[43] D. K. Ferry, R. Akis, D. Vasileska, ”Quantum effects in MOSFETs: use of an effective potential in 3D Monte Carlo simulation of ultra-short channel devices”, *IEEE IEDM Technical Digest*, 2000, p. 287.

[44] P. Palestri, D. Esseni, A. Abramo, R. Clerc, L. Selmi ”Carrier quantization in SOI MOSFETs using an effective potential based Monte-Carlo tool”, *ESSDERC 2003 Proceedings*, p. 407.

[45] E. Sangiorgi, M.R. Pinto ”A semi-empirical model of surface roughness scattering for Monte Carlo simulation of silicon N-MOSFETs”, *IEEE Transactions on Electron Devices*, 1992, vol. 39, no. 2, p. 356.

[46] P. Palestri, S. Eminente, D. Esseni, C. Fiegna, E. Sangiorgi, L. Selmi, ”An Improved Semi-Clasical Monte-Carlo Approach for Nano-Scale MOSFET Simulation”, *Solid State Electronics*, 2005, vol. 49, no. 5, p. 727.

[47] S. Takagi, M. Iwase, A. Toriumi, ”On the universality of inversion-layer mobility in n– and p– channel MOSFETs”, *IEEE IEDM Technical Digest*, 1988, p. 398.

[48] S. Takagi, A. Toriumi, M. Iwase, H. Tango, ”On the Universality of Inversion Layer Mobility in Si MOSFET’s: Part I–Effects of Substrate Impurity Concentration”, *IEEE Transactions on Electron Devices*, 1994, vol. 41, no. 12, p.2357.

[49] D. Esseni, A. Abramo, L. Selmi, E. Sangiorgi, ”Physically based modeling of low field electron mobility in ultrathin Single- and Double-Gate SOI n-MOSFETs”, *IEEE Transactions on Electron Devices*, 2003, vol. 50, no. 12, p. 2445.

[50] P. Riolino M. Braccioli, L. Lucci, D. Esseni, C. Fiegna, P. Palestri and L. Selmi, ”Monte-Carlo simulation of decananometric Double-Gate SOI devices: multi-subband vs. semiclassical with quantum corrections”, *ESSDERC 2006 Proceedings*, p. 162.

[51] L. Lucci, P. Palestri, D. Esseni, L. Selmi ”Multi-subband Monte Carlo modeling of nano-MOSFETs with strong vertical quantization and electron gas degeneration”, *IEEE IEDM Technical Digest*, 2005, p. 631.

[52] K. Tomizawa, Y. Awano, N. Hashizume, ”Monte Carlo Simulation of AlGaAs/GaAs Heterojunction Bipolar Transistors”, *IEEE Electron Device Letters*, 1984, vol. 5, no. 9, p. 362.

[53] M. Mouis, J. F. Pône, P. Hesto, R. Castagné, ”Aspect ratio phenomena in the High Electron Mobility Transistor”, *Proceedings IEEE/Cornell Conference on Advanced Concepts in High Speed Semiconductor Devices and Circuits*, July 1985, p. 144.

[54] J.–T. Park, J. P. Colinge, ”Multiple–gate SOI MOSFETs: device design guidelines”, *IEEE Transactions on Electron Devices*, 2002, vol. 49, no. 12, p. 2222.

[55] M. Lundstrom, Z. Ren, ”Essential Physics of Carrier Transport in Nanoscale MOSFETs”, *IEEE Transactions on Electron Devices*, 2002, vol. 49, no. 1, p. 133.

[56] M. Lundstrom, ”Elementary Scattering Theory of the Si MOSFET”, *IEEE Electron Device Letters*, 1997, vol. 18, no. 7, p. 361.

[57] S. Datta, F. Assad, M. Lundstrom, ”The Si MOSFET from a transmission viewpoint”, *Superlattice and microstructures*, 1998, vol. 23, p. 771.

[58] R. Clerc, P. Palestri, L. Selmi ”On the Physical Understanding of the kT–layer Concept in Quasi–Ballistic Regime of Transport in Nanoscale Devices”, *IEEE Transactions on Electron Devices*, 2006, vol. 53, no. 7, p. 1634.

[59] P. Palestri, S. Eminenente, D. Esseni, C. Fiegna, E. Sangiorgi, L. Selmi, "Understanding Quasi–Ballistic Transport in Nano–MOSFETs: Part I–Scattering in the Channel and in the Drain", *IEEE Transactions on Electron Devices,* 2005, vol. 52, no. 12, p. 2727.

[60] K. Natori, "Ballistic metal–oxide-semiconductor field effect transistor", *Journal of Applied Physics*, 1994, vol. 76, no. 8, p. 4879.

[61] C.–T. Chuang,, K. Bernstein, R. V. Joshi, R. Puri, K. Kim, E. J. Novak, "Scaling Planar Silicon Devices", *IEEE Circuits and Devices Magazine,* Jan./Feb. 2004, vol. 20, no. 1, p. 6.

[62] L. T. Su, K. E. Goodson, D. A. Antoniadis, M. I. Flik, J. E. Chung, "Measurement and modeling of self–heating effects in SOI nMOSFETs", *IEEE IEDM Technology Digest*, 1992, p. 357.

[63] L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson, M. I. Flik, "Measurement and Modeling of Self–Heating in SOI NMOSFET's", *IEEE Transactions on Electron Devices*, 1994, vol. 41, no. 1, p. 69.

[64] M. Lundstrom, "Fundamentals of carrier transport", 2^{nd} Edition, Cambridge University Press, 2000.

[65] G. K. Wachutka, "Rigorous thermodynamic treatment of heat generation and conduction in semiconductor device modeling", *IEEE Transactions on Computer–Aided Design of Integrated Circuits and Systems*, 1990, vol. 9, no. 11, p. 1141.

[66] E. Pop, R. W. Dutton, K. E. Goodson, "Monte Carlo simulation of Joule heating in bulk and strained silicon", *Applied Physics Letters*, 2005, vol. 86, p. 082101.

[67] E. Pop, R. W. Dutton, K. E. Goodson, "Analytic band Monte Carlo model for electron transport in Si including acoustic and optical phonon dispersion", *Journal of Applied Physics*, 2004, vol. 96, p. 4998.

[68] Y. S. Ju, E. Goodson, "Phonon scattering in silicon thin films with thickness of order 100 nm", *Applied Physics Letters*, 1999, vol. 74, p. 3005.

[69] S. Sinha, K. E. Goodson, "Review: Multiscale Thermal Modeling in Nanoelectronics", *International Journal for Multiscale Computational Engineering*, 2005, vol. 3, no. 1, p. 107.

[70] E. Pop, R. Dutton, K. E. Goodson, "Thermal Analysis of Ultra–Thin Body Device Scaling", *IEEE IEDM Technology Digest*, 2003, p. 883.

[71] N. Bresson, S. Cristoloveanu, C. Mazuré, F. Letertre, H. Iwai, ”Integration of buried insulators with high thermal conductivity in SOI MOSFETs: Thermal properties and short channel effects”, *Solid–State Electronics*, 2005, vol. 49, no. 9, p. 1522.

[72] D. Esseni, M. Mastrapasqua, G. K. Keller, C: Fiegna, L. Selmi, E. Sangiorgi, ”Low field electron and hole mobility of SOI transistors fabricated on ultrathin silicon films for deep submicrometer technology application”, *IEEE Transactions on Electron Devices*, 2001, vol. 48, no. 12, p. 2842.

[73] B. M. Tembroek, M. S. L. Lee, W. Redman–White, R. J. T. Bunyan, M. J. Uren, ”Self–heating effects in SOI MOSFET’s and their measurement by small signal conductance techniques”, *IEEE Transactions on Electron Devices*, 1996, vol. 43, no. 12, p. 2240.

[74] C. Canali, G. Majni, R. Minder, G. Ottaviani, ”Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation to Electric Field and Temperature”, *IEEE Transactions on Electron Devices*, 1975, vol. 22, no. 11, p. 1045.

[75] J. D. Bude, ”MOSFET Modeling Into the Ballistic Regime”, *Proceedings of the International Conference on SISPAD*, 2000, p. 23.

[76] C. Fiegna, Y. Yang, E. Sangiorgi, A. G. O’Neill, ”Analysis of Self–Heating Effects in Ultrathin–Body SOI MOSFETs by Device Simulation”, *IEEE Transactions on Electron Devices*, 2008, vol. 55, no. 1, p. 233.

[77] W. Liu, K. Etessam–Yazdani, R. Hussin, M. Asheghi, ”Modeling and Data for Thermal Conductivity of Ultrathin Single–Crystal SOI Layers at High Temperature”, *IEEE Transactions on Electron Devices*, 2006, vol. 53, no. 8, p. 1868-

[78] P. He, L. Liu, Z. Li, ”Measurement of thermal conductivity of buried oxides of silicon–on–insulator wafers fabricated by separation by implantation of oxygene technology”, *Applied Physics Letters*, 2002, vol. 81, no. 10, p. 1896.

[79] W. Molzer, Th. Schulz, W. Xiong, R. C. Cleavelin, K. Schrüfer, A. Marshall, K. Matthews, J. Sedlmeir, D. Siprak, G. Knoblinger, L. Bertolissi, P. Patruno, J. P. Colinge, ”Self Heating Simulation of Multi–Gate FETs”, *ESSDERC 2006 Proceedings*, 2006. p. 311.

[80] K. A. Jenkins, R. L. Frank, ”Impact of Self–Heating on Digital SOI and Strained–Silicon CMOS Circuits”, *IEEE International SOI Conference Proceedings*, 2003, p. 161.

[81] J. Jomaah, G. Ghibaudo, F. Balestra, ”Analysis and Modeling of Self–Heating Effects in Thin–Film SOI MOSFETs as a function of Temperature”, *Solid–State Electronics*, 1995, vol. 38, no. 3, p. 615.

[82] J. Kedzierski, M. Ieong, E. Nowak, T. S. Kanarsky, Y. Zhang, R. Roy, D. Boyd, D. Fried, H.–S. P. Wong, ”Extension and Source/Drain Design for High Performance FinFET Devices”, *IEEE Transactions on Electron Devices*, 2003, vol. 50, no. 4, p. 952.

Publications of the author

M. Braccioli, S. Eminente, P. Palestri, D. Esseni, C. Fiegna
"Comparison of Bulk and Ultra–Thin Double Gate SOI MOSFETs for the 65 *nm Technology Node: A Monte Carlo Study"*
MSED 2005 Proceedings

N. Barin, M. Braccioli, C. Fiegna, E. Sangiorgi
"Scaling the High–Performance Double–Gate SOI MOSFET down to the 32 *nm Technology Node with* SiO_2*–based Gate Stacks"*
IEDM Technical Digest, 2005, pp. 609–612

E. Sangiorgi, N. Barin, M. Braccioli, C. Fiegna
"32 *nm technology node Double–Gate SOI MOSFET using* SiO_2 *gate stacks"*
IEEE 2006 International Workshop on Nano CMOS, pp. 38–42

N. Barin, M. Braccioli, C. Fiegna, E. Sangiorgi
"Analysis of two alternative scaling strategies for sub–30 *nm Double Gate MOSFETs"*
IEEE 2006 Silicon Nanoelectronics Workshop Proceedings, pp. 69–70

D. Riolino, M. Braccioli, L. Lucci, D. Esseni, C. Fiegna, P. Palestri, L. Selmi
"Monte–Carlo Simulation of Decananometric Double Gate SOI devices: Multi–Subband vs 3*D–Electron Gas with Quantum Corrections"*
ESSDERC 2006 Proceedings, pp. 162–165

M. Braccioli, P. Palestri, T. Poiroux, M. Vinet, G. Le Carval, M. Mouis, C. Fiegna, E. Sangiorgi, S. Deleonibus
"Monte–Carlo simulation of MOSFETs with Band–Offsets in the Source and Drain"
IEEE ULIS 2007 Proceedings, pp. 39–42

C. Fiegna, M. Braccioli, *et alii*
"Comparison of Monte Carlo Transport Models for Nanometer–Size MOSFETs" Proceedings of the International Conference on SISPAD, 2007, pp. 57–60

M. Braccioli, C. Fiegna, E. Sangiorgi
"Comparative analysis of self–heating in different SOI architectures" EuroSOI Proceedings, pp. 31–32

M. Braccioli, G. Curatola, Y. Yang, E. Sangiorgi, C. Fiegna
"Simulation of Self–Heating effects in 30 *nm gate length FinFET"* IEEE ULIS 2008 Proceedings, pp. 71–74

N. Barin, M. Braccioli, C. Fiegna, E. Sangiorgi
*"Analysis of Scaling Strategies for Sub–*30 *nm Double–Gate SOI N–MOSFETs"* IEEE Transactions on Nanotechnology, vol. 6(4), Jul. 2007, pp. 421-430

I. Riolino, M. Braccioli, L. Lucci, P. Palestri, D. Esseni, C. Fiegna, L. Selmi *"Monte–Carlo Simulation of decananometric nMOSFETs: Multi–subband vs.* 3*D–electron gas with quantum corrections"*
Solid–State Electronics, vol.51(11), Nov. 2007, pp. 1558-1564

M. Braccioli, P. Palestri, M. Mouis, T. Poiroux, M. Vinet, G. Le Carval, C. Fiegna, E. Sangiorgi, S. Deleonibus
"Monte Carlo simulation of MOSFETs with Band–Offsets in the Source and Drain"
Solid–State Electronics, vol.52(4), Apr. 2008, pp. 506–513

M. Braccioli, G. Curatola, Y. Yang, E. Sangiorgi, C. Fiegna
"Simulation of Self-Heating effects in 30nm gate length FinFET"
Solid–State Electronics, article in press