Doctoral Thesis

Development of a Multisubband Monte Carlo Simulator for Nanometric Transistors

Author:	Supervisors:
Cristina Medina Bailón	Dr. Francisco Gámiz Pérez
	Dr. Carlos Sampedro Matarín

A thesis submitted in fulfillment of the requirements
to obtain the International Doctor degree as part of the
Programa de Doctorado en Física y Ciencias del Espacio
in the

Nanoelectronics Research Group
Departamento de Electrónica y Tecnología de los Computadores

Granada, January 9, 2017

Editor: Universidad de Granada. Tesis Doctorales
Autora: Cristina Medina Bailón
ISBN: 978-84-9163-102-6
URI: http://hdl.handle.net/10481/44885

Declaration of authorship

El doctorando / The doctoral candidate, Cristina Medina Bailón, y los directores de la tesis / and the Ph.D Supervisors, Dr. Francisco Gámiz Pérez and Dr. Carlos Sampedro Matarín:

Garantizamos, al firmar esta tesis doctoral, que el trabajo, titulado ***Development of a Multisubband Monte Carlo Simulator for Nanometric Transistors***, ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Guarantee by signing this doctoral thesis: that the research work, entitled ***Development of a Multisubband Monte Carlo Simulator for Nanometric Transistors***, has been done by the doctoral candidate under the direction of the thesis supervisors and, as far as our knowledge reaches, in the performance of the work, the rights of authors to be cited (when their results or publications have been used) have been respected.

Granada, 9 de Enero de 2017 / 9th January, 2017.

Cristina Medina Baión Dotorando / Ph.D Candidate	Dr. Francisco Gámiz Pérez Director de la tesis / Ph.D Supervisor
	Dr. Carlos Sampedro Matarín Director de la tesis / Ph.D Supervisor

To Noa

It's creepy, but here we are, the Pilgrims, the crackpots of our time trying to establish our own alternative reality. To build a world out of rocks and chaos.

Chuck Palahniuk, *Choke*, 2001

Acknowledgements

Me gustaría dedicar unas líneas para agradecer a toda la gente que de algún modo ha contribuido al desarrollo de esta tesis doctoral y que han formado parte de mi iniciación en la carrera investigadora. Por supuesto, en primer y preferente lugar, quiero agradecer a mis directores de tesis D. Francisco J. Gámiz y D. Carlos Sampedro lo mucho que han hecho por mi durante estos 4 años. En primer lugar, por haberme brindado en su momento la oportunidad de formar parte de este proyecto. También he de agradecer su inestimable ayuda y consejo que me han ofrecido siempre que ha hecho falta. Pero sobre todo esto, su incansable apoyo siempre que lo he necesitado. Hemos dedicado muchas horas a una estimulante la discusión científica y espero que podamos continuar esto en el futuro.

Me gustaría reservar un lugar especial en estos agradecimientos a quienes han hecho notables contribuciones porque siempre han estado disponibles para resolver las dudas imposibles. A D. Luca Donetti por su valiosa ayuda tanto por brindarme todo su conocimiento sobre Monte Carlo cómo por enseñarme a usar el cluster para las simulaciones. A mi tutor de doctorado, D. Andrés Godoy, por aconsejarme todas las actividades de doctorado más acordes a mi información y estar siempre dispuesto a ayudarme en cualquier materia. También me gustaría agradecerle las horas que ha pasado revisando mi trabajo de forma tan rápida y eficaz. En este sentido, la contribución de D. José Luis Padilla ha sido sin duda esencial por resolver todas mis dudas física y filosóficas. Siempre me ha ofrecido toda la ayuda posible para allanar mis comienzos como investigadora. Sin su ayuda esta Tesis no habría sido posible.

Me gustaría también agradecer al Departamento de Electrónica y Tecnología de los Computadores y a sus últimos directores, D. Enrique Carceller y Juan Antonio López por poner a mi disposición los medios necesarios para que este trabajo se haya podido realizar. Y también agradecer a todos los integrantes del grupo de investigación *Nanoelectronics Research Group*, con los que he tenido el placer de trabajar y colaborar en estos años.

A D. Karol Kalna, D. Raúl Valín, Dña Natalia Seoane, D. Asen Asenov, D. Vihar Georgiev y D. Toufik Sadi, les tengo que agradecer, no sólo su implicación para hacer posible mi estancia en las Universidades de Swansea y Glasgow en 2014 y 2016, respectivamente, sino también que me hicieran sentir integrada desde el primer momento y el trato personal recibido. I felt like I was at home during the working time of these months.

Agradezco también a todos mis compañeros de despacho durante estos cuatro años lo bien que nos hemos llevado y incluyendo los ratos de descanso tomando café o cervezas. A Pilar y Celso, con quienes he compartido la mayor parte de esta etapa, y a Ernesto (aunque no me deje nodos libres en el cluster), José María y Abdalia, las últimas incorporaciones. A Celso, en particular agradecerle los ánimos diarios y la forma positiva de pensar que contagia. Aunque no esté en el despacho también me gustaría mencional a Carlos ya que comenzamos esta andadura juntos y prácticamnete la vamos a terminar a la vez.

Institucionalmente, también agradezco al Ministerio de Economía, Industria y Competitividad el haber financiado estos años de investigación a partir del proyecto TEC2011-28660. Este esfuerzo público para promover la investigación es esencial, y confiemos en que en el futuro se siga manteniendo esta apuesta por la investigación desde todas las instituciones públicas.

En el ámbito personal, tengo mucho que agradecer a mis padres, Carmina y Antonio, que me han apoyado y animado siempre para llegar a dónde estoy. A Antonio, María José, Carmina, Alberto, Rocio y Fran, por hacerme el trago menos amargo. Y, por supuesto, a mis sobrinos Adriana, Alma, Noa y Leo (la última incorporación) por brindarme esos momentos de juego y alegría que te recuperan de largas semanas de trabajo.

A mis grandes amigos Alba, Pablos, Manolillo, Bruno, Pili, Jesús, Huayqui, Adrián y Fran, por las horas incontables que conversaciones. A Julia, por todas sus consejos de traducción. Y para terminar, a Nacho, por haber estado siempre a mi lado apoyándome (aguantándome) en todo momento, animándome y mimándome. Su apoyo incondicional ha sido mi constante durante estos años y su paciencia mi variable favorita.

Contents

List of Abbreviations and Symbols

Abbreviations

BJT	Bipolar Junction Transistor
BOX	Buried oxide (isolation layer of a SOI structure)
BTBT	Band–to–band tunneling
BTE	Boltzmann Transport Equation
CMOS	Complementary MOS technology
DD	Drift–Diffusion approximation
DIBL	Drain induced barrier lowering
DTMOS	Dynamic Threshold MOS
EMC	Ensemble Monte Carlo
EOT	Equivalent Oxide Thickness
ETSOI	Extremely Thin SOI
FET	Field–Effect Transistor
FDSOI	Fully depleted SOI
GAA	Gate-all-around
GLM	Gate leakage mechanism
HG-EHBTFET	heterogate electron–hole bilayer TFET
ITRS	International Technology Roadmap for Semiconductors
JFET	Junction Field-Effect Transistors
MEMS	Micro–Electro–Mechanical Systems
MIS	Metal–Insulator–Semiconductor
MOSFET	Metal–Oxide–Semiconductor Field–Effect Transistor
MuGFET	Multiple gate field-effect transistor
NEGF	Non-Equilibrium Green Functions

NMOS	n–channel MOSFET
PDSOI	Partially depleted SOI
PMOS	p–channel MOSFET
SCEs	Short Channel Effects
SGSOI	Single Gate SOI
SPMC	Single Particle Monte Carlo
S/D tunneling	Direct Source to Drain tunneling
SoC	System-on-Chip
SOI	Silicon-on-Insulator
SS	Subthreshold swing
TCAD	Technology computer aided design
TFET	Tunneling field–effect transistor
TSV	Through Silicon Via technology
UTB	Ultrathin-Body
VCBM	Voltage-controlled bipolar-MOS
WKB	Wentzel-Kramers-Brillouin approximation

Symbols

α	Non-parabolic parameter
$\vec{B}$	Magnetic field
C_{ox}	Oxide capacitance
$\vec{\varepsilon}$	Electric field
ε_{OX}	Electrical permittivity of the channel
ε_{Si}	Electrical permittivity of the gate dielectric
E_g	Bandgap energy
G	Generation rate function per unit volume
I_{ON}	Higher current of the devices at low drain bias
I_{OFF}	Lower current of the devices at low drain bias
J_n	Electron carrier density
J_p	Holes carrier density
$\vec{k}$	Carrier wave vector
λ_N	Natural channel length
L_G	Gate length
m^*	Carrier effective mass
μ_n	Electron mobility
n	Electron concentration
Ω	Normalization volume
p	Hole concentration
ϕ	Azimuthal angle
ρ	Carrier density
SiO_2	Silicon dioxide
T	Temperature
θ	Polar angle
T_{OX}	Gate dielectric thickness
T_{Si}	Silicon Thickness
u	Sound velocity in the material
V_{DD}	Supply voltage
V_{DS}	Drain–to–source voltage
V_{GS}	Gate–to–source voltage
V_{th}	Threshold voltage

Physical constants

ϵ_0	vacuum permittivity	$8.85418782 \cdot 10^{-12}$ F/m
m_0	electron rest mass	$9.10938291 \cdot 10^{-31}$ Kg
h	Planck's constant	$6.62606957 \cdot 10^{-34}$ J·s
$\hbar$	reduced Planck's constant	$1.05457172 \cdot 10^{-34}$ J·s
k_{B}	Boltzmann's constant	$1.38064881 \cdot 10^{-23}$ J/K
q	elementary charge	$1.60217656 \cdot 10^{-19}$ C

Abstract

Nanoelectronics Research Group
Departamento de Electrónica y Tecnología de los Computadores

Development of a Multisubband Monte Carlo Simulator for Nanometric Transistors

by Cristina Medina Bailón

The ultimate objective of this PhD Thesis is the study of the performance of nanometric transistors, and the importance that quantum effects have on the determination of their behavior. To do so, this work presents a description of the new architectures which are postulated as an alternative for future technological nodes, and the simulation tools employed to achieve an accurate determination of the electrostatic and transport properties of such devices, accounting for the dominant quantum effects which they undergo.

We start with a summary of several technological architectures which are proposed to overcome the downscaling limitations of conventional planar devices. They are required to keep under control the short-channel effects (SCEs), that is, the loss of the control of the channel charge by the gate terminal.

The starting point of the simulation frame which is a Multisubband Ensemble Monte Carlo (MS-EMC) scheme is analyzed. This tool is based on the mode-space approach of quantum transport where the system is decoupled in the confinement direction and the transport plane, where the 1D Schrödinger equation and the 2D Boltzmann Transport Equation (BTE) are solved, respectively. Both equations are coupled to the 2D Poisson Equation to keep the self-consistency of the solution. It has already demonstrated its capabilities in different scenarios keeping a reasonable computational effort with respect to the full-quantum approach.

However, this code has been parallelized in order to allow for the study of more complex devices in a reasonable simulation time. Other techniques for statistical enhancement are included in order to reduce the stochastic noise.

Furthermore, the appearance of the leakage currents modifies the stable performance of the conventional MOSFETs. Accordingly, a deep study of each physical mechanisms responsible for these leakage currents was carried out. One of the main advantages of considering this MS-EMC simulator is that quantum effects can be included in a separate way because of the decoupled approximation allowing for an independent inquiry.

In this regard, we have focused on: *i*) Source-to-Drain tunneling (S/D tunneling), which allows electrons go from the source to the drain through the channel potential barrier; *ii*) the band-to-band tunneling (BTBT), which occurs when the conduction band aligns with the valence band and so electrons/holes from the valence/conduction band tunnel into the conduction/valence band; and finally *iii*) the gate leakage mechanism (GLM), which results in the tunneling of carriers from the substrate to the gate and also from gate to the substrate through the gate oxide. In general, we demonstrate that these quantum effects get even more relevance as the device dimensions are reduced. Using the numerical and analytical approaches, several relevant electrostatic and transport studies of different architectures are accomplished when these phenomena are included in order to determine their influence on the device characteristics.

Resumen

Nanoelectronics Research Group
Departamento de Electrónica y Tecnología de los Computadores

Development of a Multisubband Monte Carlo Simulator for Nanometric Transistors

by Cristina Medina Bailón

Sinopsis

El objetivo principal de esta tesis doctoral es el estudio de las prestaciones de los transistores nanométricos cuando se incluyen mecanismos cuánticos. Para ello se presenta una descripción de las nuevas arquitecturas que se postulan como alternativa para futuros nodos tecnológicos, y las herramientas de simulación empleadas para lograr una determición precisa de las propiedades electrostáticas y de transporte de dichos dispositivos, incluyendo los efectos cuánticos dominantes que sufren.

Comenzamos esta tesis con un resumen de varias tecnologías que se proponen para superar las limitaciones del escalado de los dispositivos planares convencionales. Se les exige mantener bajo control los efectos de canal corto (short-channel effects, SCEs), que implican la pérdida de control de la carga del canal a causa del terminal de puerta.

A continuación, se analiza el simulador Multisubband Ensemble Monte Carlo (MS-EMC) empleado en este trabajo. Se ha demostrado sus capacidades en diferentes escenarios manteniendo un esfuerzo computacional razonable con respecto a la aproximación puramente cuántica. Sin embargo, ha sido paralelizado para permitir el estudio de dispositivos más complejos con un tiempo de simulación

razonable. Asímismo, otras técnicas de mejora estadística se han incluido con el fin de reducir el ruido estocástico.

Además, la aparición de las corrientes de pérdidas modifica el rendimiento estable de los MOSFET convencionales. En consecuencia, debe llevarse a cabo un estudio profundo de cada uno de los mecanismos que inducen estas corrientes de pérdidas. En este sentido, nos hemos centrado en los siguientes fenómenos: *i*) túnel de fuente a drenador (source-to-drain tunneling, S/D tunneling), que permite que los electrones vayan de la fuente al drenador a través de la barrera de potencial; *ii*) túnel banda-a-banda (band-to-band tunneling, BTBT), que se produce cuando la banda de conducción se alinea con la banda de valencia y, por lo tanto, los electrones/huecos de la banda de valencia/conducción pueden realizar túnel hacia la banda de conducción/valencia; y, por último, *iii*) el mecanismo de pérdidas por la puerta (gate leakage mechanism, GLM), que da como resultado el túnel de portadores de sustrato a puerta y también de puerta a sustrato a través del óxido de puerta. En general, demostramos que estos efectos cuánticos cobran aún más relevancia a medida que se reducen las dimensiones del dispositivo. Utilizando aproximaciones numéricas y analíticas, se realizan varios estudios relevantes de la electrostática y del transporte de diferentes arquitecturas cuando se incluyen estos fenómenos para determinar su influencia e impacto en las características del dispositivo.

Objetivos y Metodología

Como se ha comentado anteriormente, la reducción agresiva de las dimensiones del dispositivo ha hecho aumentar los efectos de canal corto y los mecanismos de pérdidas, comprometiendo seriamente el rendimiento de los dispositivos electrónicos. Las soluciones a esos problemas deben ser tajantes y, por tanto, nuevos paradigmas en la estructura del MOSFET y en la composición del material deben ser abordandos. Igualmente, nuevos fenómenos físicos aparecen en el funcionamiento habitual del dispositivo a medida que estos se acercan a las dimensiones nanométricas.

Esta tesis doctoral se dedica al estudio de algunos fenómenos que comprometen el rendimiento de los prometedores dispositivos alternativos para continuar el proceso de escalado mediante un simulador de Monte Carlo. Además, estos efectos aumentan la complejidad computacional provocando la necesidad de desarrollar nuevos modelos técnicos para acelerar las simulaciones.

Nuestro estudio tiene como objetivo principal investigar qué nueva arquitectura es la mejor candidata para implementar futuros nodos tratando de buscar la más robusta contra los mecanismos de pérdidas. Para lograr este objetivo, una vez que se han descrito los efectos cuánticos, se realiza una profunda comparación entre su impacto en diferentes arquitecturas.

Los principales objetivos de este trabajo son:

1. La mejora de nuestro simulador MS-EMC con el fin de reducir el coste computacional y el ruido estocástico, y proporcionar resultados más precisos y realistas.

2. El estudio de varios efectos cuánticos: el túnel de fuente a drenador (S/D tunneling), túnel banda-a-banda (BTBT), y el mecanismo de pérdidas por la puerta (GLM).

3. Análisis del impacto de estos efectos cuánticos en diferentes arquitecturas.

Una vez establecidos los objetivos, se presenta una breve descripción de la metodología seguida en este trabajo. El estudio de los efectos de mejora para el simulator y de los mecanismos cuánticos que se han desarrollado para la herramienta MS-EMC se pueden dividir en tres partes. La primera está centrada en la descripción de las arquitecturas de dispositivos que se están considerando actualmente. A continuación, se proporciona una caracterización del simulador de Monte Carlo. Finalmente, la tercera comprende los mecanismos cuánticos estudiados. Las tres partes se han desarrollado según el esquema siguiente:

- La descripción de las actuales y futuras arquitecturas que se proponen para reemplazar la tecnología estándar y extender el final de la hoja de ruta se presentan en el Capítulo 2. Una vez que el MOSFET convencional se describe teniendo en cuenta su estructura, ventajas y los problemas de escalado, se introducen las alternativas más prometedoras. Especialmente, se han considerado nuevos materiales, óxidos de alta permitividad κ, dispositivos de múltiples puertas y las tecnologías de silicio tenso y silicio sobre aislante (silicon-on-insulator, SOI).

- A continuación, los fundamentos teóricos necesarios para la herramienta Monte Carlo se presentan en el Capítulo 3. El primer paso es comparar los diferentes modelos de simulación para dispositivos semiconductores con especial énfasis en su utilidad y complejidad computacional. A partir de ahí, se comentan las características de los cálculos de transporte utilizadas en el método Monte Carlo, donde las técnicas de vuelo libre y los mecanismos de dispersión se han tenido en cuenta.

- En el Capítulo 4, se ha ilustrado las hipótesis de mejora incluidas en el simulador MS-EMC. El método “ensemble”, la suposición de subbanda múltiple para cada valle y las condiciones de contorno se caracterizan para obtener resultados precisos de acuerdo con las nuevas arquitecturas, mientras que la implementación en paralelo y el cálculo del peso de forma dependiente de energía para superpartículas se incorporan en este código para reducir el tiempo de simulación y el ruido estocástico, respectivamente.

- Haciendo uso de la herramienta avanzada MS-EMC, en el Capítulo 5, se da una descripción detallada de los fenómenos cuánticos incluidos. Después de una visión global de las corrientes de túnel y de pérdidas, se detallan el túnel de fuente a drenador (S/D tunneling), túnel banda-a-banda (BTBT), y el mecanismo de pérdidas por la puerta (GLM), así como su impacto en diferentes arquitecturas de dispositivos.

- Finalmente, las principales conclusiones de esta tesis, junto con algunos hilos de trabajo futuro que naturalmente se derivan del trabajo aquí presentado, se resumen en el Capítulo 6. Una lista de las publicaciones producidas por esta tesis se puede encontrar también al final.

Conceptos Generales

En los últimos años se ha diversificado la arquitectura básica del transistor convencional metal-oxide-semiconductor field-effect transistor (MOSFET) con el fin de mejorar el rendimiento del dispositivo cuando los transistores se aproximan a los últimos límites de escalado. Durante este tiempo se ha pasado de $10\mu m$ al régimen sub - $10nm$. Este dramático escalado ha favorecido tanto la integración de MOSFETs así como enormes avances en nanoelectrónica. La principal motivación del escalado es lograr una mayor densidad de integración y un menor consumo de energía y tiempo de retardo [1]. Las longitudes del canal y las tensiones de alimentación se han reducido para cumplir con dichos requisitos [2–4]. Sin embargo, los dispositivos convencionales se comportan de forma no deseada a una escala tan pequeña a causa de los efectos cuánticos y, por tanto, se puede declarar que la tecnología estándar no puede proporcionar un comportamiento exitoso [5–11].

Prediciendo esta carrera de escalado, la ley de Moore postuló que el número de transistores en un chip de circuito integrado se duplicaría cada 18 meses [12,13]. Un aumento de la velocidad de escalado de los dispositivos permitió que ese régimen fuera alcanzado mucho antes de lo esperado. Pero, precisamente por esa razón, se debe hacer frente a los nuevos problemas tecnológicos ahora, tales como por ejemplo la creciente resistencia parasita entre fuente y drenador, los efectos de canal corto (SCEs) o la corriente de pérdidas a través el óxido de la puerta [14–16].

A medida que nos acercamos al nodo de $7nm$ [1], han aparecido diferentes tendencias para enfrentarse a los problemas específicos que introducen un límite crítico en el escaldado de los dispositivos electrónicos. La Figura 1 muestra las principales arquitecturas que se han tenido en cuenta tanto en la actualidad y como en el pasado. Por esta razón, vale la pena incluir tanto efectos tradicionales, como la movilidad del canal o el control de los efectos de canal corto, y los nuevos problemas asociados con las escalas de dimensiones atómicas, como el confinamiento cuántico y los efectos de dispersión.

Actualmente, hay dos áreas de trabajo principales y paralelas desde el punto de vista de la simulación de dispositivos electrónicos. La primera tendencia se centra principalmente en la búsqueda de nuevas soluciones de ingeniería para crear arquitecturas de dispositivos mejoradas. El segundo es el estudio de los efectos cuánticos que introducen un comportamiento y una variabilidad no deseadas en los dispositivos convencionales con dimensiones nanométricas.

Por un lado, se proponen diferentes arquitecturas tecnológicas para superar las limitaciones de los dispositivos planares convencionales. La aparición de los efectos de canal corto (SCE) y las corrientes de pérdidas modifican el rendimiento estable de los MOSFET convencionales. El control de la electrostática del transistor es la garantía para considerar como aceptables los efectos de canal corto.

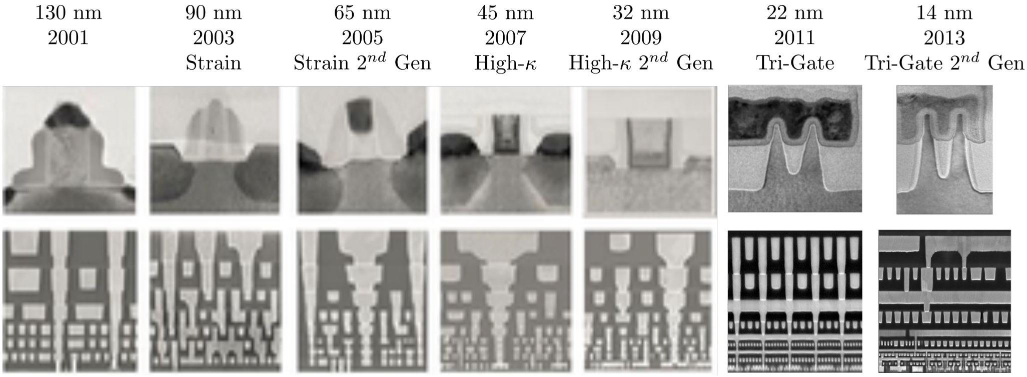


Figura 1: Principales arquitecturas que se han tenido en cuenta tanto en la actualidad y como en el pasado.

Por esta razón, se están estudiando nuevas arquitecturas de transistores basadas en múltiples puertas, diferentes materiales o nuevos modelos de inyección para reemplazar la tecnología estándar y extender el final de la hoja de ruta.

Esta primera tendencia de simulación puede dividirse a su vez en dos áreas principales de trabajo. En primer lugar, las nuevas soluciones de ingeniería se centran en la creación de arquitecturas de dispositivos mejoradas [17, 18], en las que se pretende mejorar las propiedades de transporte de los portadores y mantener bajo control los problemas de canal corto. Por ejemplo, en esta tesis nos hemos centrado en la comparación entre tres distintos dispositivos (Figura 2) y vamos a comentar sus principales diferencias. Cuando se añaden múltiples puertas alrededor del canal, se reducen los efectos de canal corto. Esa es la gran ventaja del Double-Gate Silicon-On-Insulator (DGSOI) y del FinFET frente al Fully-Depleted Silicon-On-Insulator (FDSOI). Además, durante el proceso de fabricación del dispositivo, la oblea estándar para los dispositivos planares es (100), tales como FDSOI o DGSOI, pero puede ser reemplazada por verticales (110), tales como FinFETs.

En segundo lugar, se proponen nuevos mecanismos de inyección y fenómenos físicos aprovechando los nuevos materiales y las dimensiones nanométricas. El transistor túnel de efecto campo (tunnel field-effect transistor, TFET) se ha convertido en una alternativa a los MOSFET convencionales en los últimos años debido a la posibilidad de lograr una elevada pendiente sub-umbral (subthreshold swing, SS) que permite reducir la tensión de polarización, y por tanto el

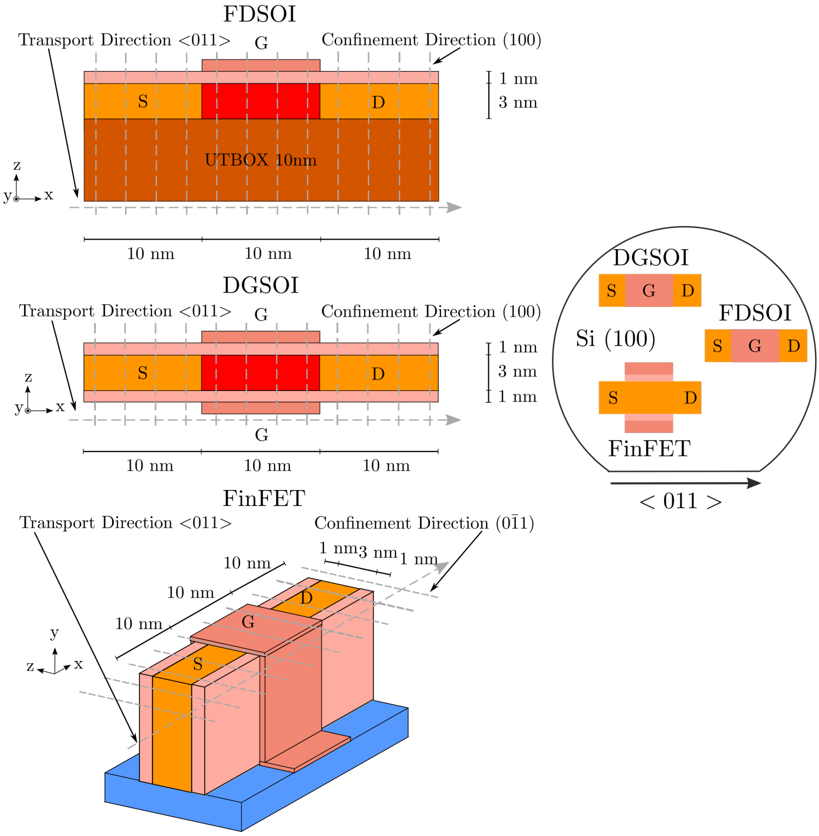

FDSOI
Transport Direction <011>
G
Confinement Direction (100)
S
D
UTBOX 10nm
1 nm
3 nm
z
y
x
10 nm
10 nm
10 nm
DGSOI
Transport Direction <011>
G
Confinement Direction (100)
S
D
G
1 nm
3 nm
1 nm
10 nm
10 nm
10 nm
FinFET
Transport Direction <011>
Confinement Direction (0$\bar{1}$1)
1 nm
3 nm
1 nm
10 nm
10 nm
10 nm
D
G
S
DGSOI
S G D
FDSOI
S G D
Si (100)
S D
FinFET
< 011 >

Figura 2: Estructuras de los dispositivos analizados en este trabajo FDSOI, DGSOI y FinFET con longitud de puerta de $L_G = 10nm$. La ecuación de Schrödinger 1D es resuelta para cada punto del grid en la dirección de transporte mientras que la ecuación de transporte de 2D Boltzmann (BTE) es resuelta mediante em método Monte Carlo en el plano de transporte.

consumo de potencia [19]. Su estructura se puede observar en la Figura 3 dónde se ha representado el dispositivo tipo n basado en silicio TFET analizado en este trabajo. La diferencia más importante entre ambas arquitecturas es el mecanismo de inyección que en los MOSFET convencionales se rige por la emisión termiónica por encima de la barrera de potencial formada entre la fuente y el canal, mientras que en TFETs el mecanismo de inyección de portadora se sustituye por el túnel cuántico a través de la barrera (BTBT).

Por otro lado, el proceso de escalado también conduce a la aparición de efectos

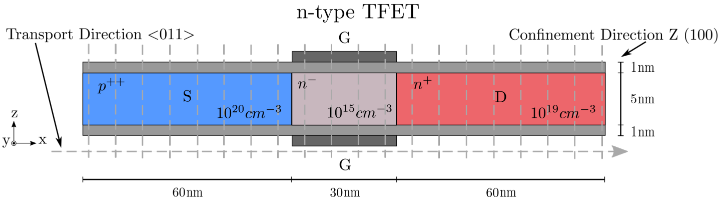


Figura 3: Dispositivo tipo n basado en silicio TFET analizado en este trabajo. La ecuación de Schrödinger 1D es resuelta para cada punto del grid en la dirección de transporte mientras que la ecuación de transporte de 2D Boltzmann (BTE) es resuelta mediante em método Monte Carlo en el plano de transporte.

cuánticos que deben ser cuidadosamente abordados y que están siendo objeto de un profundo estudio [15,20,21]. En consecuencia, ha sido obligatoria su inclusión para explicar los comportamientos del dispositivo a medida que sus dimensiones se reducen. Esta tesis se centra en el estudio de tres efectos importantes que se presentan como un límite del escalado desde el punto de vista de la simulación a partir de la herramienta MS-EMC: el túnel de fuente a drenador (S/D tunneling), túnel banda-a-banda (BTBT), y el mecanismo de pérdidas por la puerta (GLM).

Antes de pasar a comentar la importancia de cada uno de los tres mecanismos, es necesario comentar los principios básicos de la herramienta MS-EMC. Está basada en la aproximación *mode-space* en la que el sistema está desacoplado en la dirección de confinamiento y el plano de transporte, donde se resuelven la ecuación de 1D Schrödinger y la ecuación de transporte de 2D Boltzmann (BTE), respectivamente, tal y cómo se observa en las Figuras 2 y 3 donde cabe resaltar que su comportamiento no se ve afectado por el dispositivo en estudio. Las dos ecuaciones se unen a la ecuación de Poisson 2D para mantener la auto-consistencia de la solución. Asimismo, se ha considerado la aproximación de varias subbandas por cada valle (mínimos de energía en la estructura de bandas) a causa del alto nivel de confinamiento en las nuevas arquitecturas.

Una de las desventajas que tiene este método es el alto tiempo de simulación a medida que se incluyen más efectos cuánticos. Por esta razón, este simulador ha sido paralelizado utilizando el estándar OpenMP con el fin de reducir el coste computacional. Esta mejora tiene dos ventajas principales: el estudio de dispositivos más complejos en un tiempo de simulación razonable y la posibilidad de acelerar la elaboración de nuevos modelos gracias al tiempo reducido en las versiones de prueba. Además, se ha desarrollado un nuevo método para las condiciones de contorno en los contactos ohmicos. En este nuevo procedimiento, se reutilizan las superpartículas que salen por la fuente o el drenador para crear las nuevas a inyectar de forma que se permite el paralelismo y se optimiza la carga computacional. Los parámetros que describen el movimiento de las superpartículas se calculan utilizando las estadísticas de Maxwell-Boltzmann y Fermi-Dirac, pero es esta segunda la que da una imagen más precisa de la función de distribución. Otro efecto de mejora introducido es el modelo de peso dependiente de la energía para reducir el ruido estocástico que las superpartículas de muy alta energía incluyen en el comportamiento del dispositivo. Estas superpartículas de alta energía se pueden definir como eventos raros y por lo tanto su peso debe ser reducido.

Las capacidades del simulador MS-EMC desarrollado en la presente memoria se han demostrado estudiando varios dispositivos nanométricos manteniendo un esfuerzo computacional razonable con respecto a la aproximación puramente cuántica [22–25]. Una de las principales ventajas de considerar este simulador MS-EMC es que los efectos cuánticos pueden ser incluidos de manera separada debido a la aproximación desacoplada que permite un estudio independiente [26–29].

El primer efecto estudiado en esta tesis es el S/D tunneling ya que se ha presentado como un límite para el escalado en estudios basados en simulaciones balísticas usando el método de las funciones de Green de no-equilibrio (non-equilibrium Green's Function, NEGF) [30,31]. Inicialmente, los efectos cuánticos se tuvieron en cuenta para las simulaciones en la dirección de confinamiento para mejorar la escalabilidad. Actualmente, cuando la longitud del canal está cerca de $10nm$, es necesario incluir algunos de ellos en la dirección de transporte. En particular, el *S/D tunneling* introduce un efecto no deseado en dispositivos electrónicos porque los electrones pueden atravesar desde la fuente hasta el drenador a través de la barrera de potencial sin ningún control por parte de la tensión de puerta. Además, el número de electrones afectados por este es aleatorio. Por lo tanto, se produce una desviación en el funcionamiento del dispositivo electrónico introduciendo ruido. Cuando se considera este efecto cuántico, una superpartícula con menor energía que la barrera de potencial cercana a ésta, puede atravesarla o seguír volando en sentido opuesto a la barrera, también conocido cómo backscattering (Figura 4). La probabilidad de realizar este túnel puede calcularse mediante la aproximación Wentzel-Kramers-Brillouin (WKB) [32,33]:

$$T_{dt}(E) = \exp\left\{ -\frac{2}{\hbar} \int_a^b \sqrt{2m^*_{tr}(E_i(x) - E)}\, \mathrm{d}x \right\}. \tag{1}$$

dónde a y b son los puntos iniciales y finales del trayecto, E and m^*_{tr} son la energía y la masa efectiva de transporte, respectivamente, y $E_i(x)$ es la energía de la subbanda *i-th*. A partir de esta aproximación, la probabilidad de túnel para una superpartícula depende principalmente de: la trayectoria del túnel, la barrera potencial y la masa efectiva del transporte.

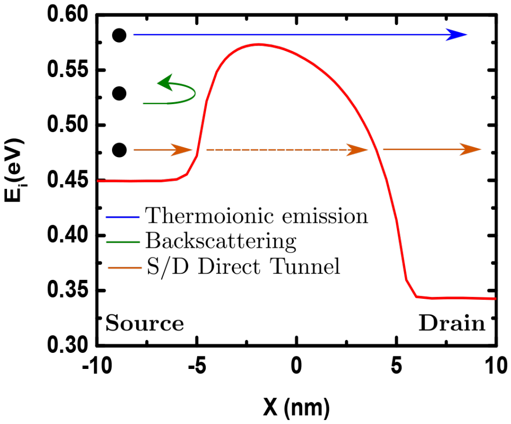


Figura 4: Representación esquemática de las opciones que tiene una superpartícula cuando llega a la barrera de potencial: *i*) si su energía es mayor, realiza una emisión termoiónica volando de fuente a drenador; si su energía es menor que la barrera puede *ii*) sufrir backscattering o *iii*) atravesar la barrera de potencial sufriendo S/D tunneling.

A continuación se ha implementado con éxito en este trabajo un modelo para el mecanismo BTBT. A pesar de que este fenómeno es una fuente de corriente de pérdida en las arquitecturas convencionales [21], se aplica a estudiar TFETs porque es precisamente su principio de funcionamiento, y por lo tanto ya no es considerado como un efecto parásito no deseado. Se produce cuando la banda de conducción se alinea con la banda de valencia y, por lo tanto, los electrones/huecos de la banda de valencia/conducción pueden realizar túnel hacia la banda de conducción/valencia, respectivamente (Figura 5). En particular, se ha desarrollado un modelo BTBT no local que combina la ecuación local para las tasas de generación, la elección de un trayecto de túnel no local y la modificación de la banda de conducción y de valencia según su primer estado energético disponible. Por otra parte, se han comparado diferentes enfoques para la elección de las trayectorias túneles y se observan diferencias relevantes tanto en los niveles de corriente como en la distribución espacial de los portadores generados [34].

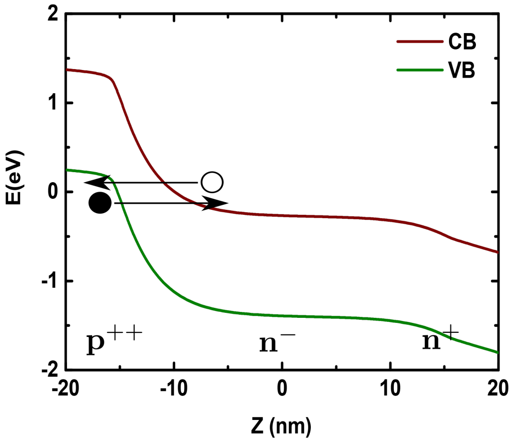


Figura 5: BTBT en una unión pn: se produce túnel de los electrones (huecos) desde la banda de valencia (conducción) de la región p (n) a estados vacíos de la banda de conducción (valencia), respectivamente.

Finalmente, los altos camops eléctricos que aparecen en stas estructuras con dieléctricos muy delgados permiten que se pueda producir túnel desde el sustrato a la puerta a través del óxido [21]. Este fenómeno de túnel se conoce como mecanismo de pérdidas de puerta (GLM) e incluye fenómenos intrínsecos, tales como el túnel directo, así como los extrínsecos, como túnel asistido por trampas (Figura 6). Se han considerado tres suposiciones diferentes para el túnel asistido por trampas: si una superpártícula está situada en el sustrato puede realizar túnel a la trampa, y si está situada en la trampa puede volver al sustrato otra vez o abandonar el dispositivo por el dieléctrico de puerta. La probabilidad de túnel se ha calculado mediante la aproximación WKB en el túnel directo (como para el S/D tunneling) pero, para el túnel asistido por trampa, este coeficiente de transmisión no sólo depende de la estructura de barrera dieléctrica sino también de algunos factores específicos de cada mecanismo. En primer lugar, para la ocupación de la trampa se considera el principio de exclusión de Pauli. En segundo lugar, si el túnel es inelástico, se debe emitir o absorver un fonon. Si la energía de la superpartícula es meyor/menor que la energía de la trampa, un fonón es emitido/absorvido y la probabilidad de túnel es multiplicada por $(1+n(\omega))/n(\omega)$ respectivamente, siendo ω la diferencia de energía entre la superpartícula y la trampa, y $n(\omega)$ el factor Bose–Einstein.

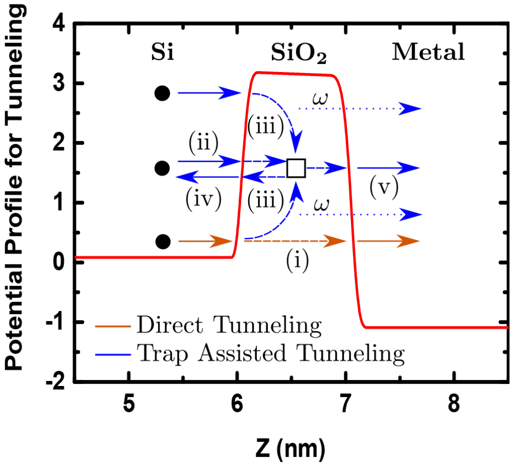


Figura 6: Diagrama esquemático de una estructura MOS con puerta de metal y sustrato de silicio, dónde los mecanismos implementados en el simulador MS-EMC para el mecanismo GLM son: (i) túnel directo, (ii) túnel elástico e (iii) inelástico hacia una trampa emitiendo o capturando un fonón con energía ω,(iv) túnel de la trampa al sustrato, y (v) de la trampa al dieléctrico de la puerta.

Conclusiones y Trabajo Futuro

El objetivo principal de esta tesis doctoral ha sido el estudio de los principales efectos cuánticos en diferentes arquitecturas de dispositivos mediante una herramienta MS-EMC.

En este contexto, las principales contribuciones de este trabajo se enumeran a continuación:

- Se ha proporcionado una descripción amplia de la herramienta MS-EMC que es capaz de tratar fenómenos transitorios y no homogéneos, orientaciones arbitrarias, valles y subbandas diferentes, arquitecturas multipuerta y tecnología SOI. Además, se han tenido en cuenta las dispersiones de fonones y de rugosidad superficial, así como las impurezas ionizadas.

- Como se requieren nuevos modelos debido a la escala de los dispositivos, los recursos computacionales se han mejorado. El código MS-EMC ha sido paralelizado utilizando el estándar OpenMP para reducir el tiempo de simulación. Se ha demostrado que, a pesar de que la simulación del transporte muestra una dependencia entre las distintas partes de la simulación, si la mayoría de los bloques más costosos computacionalmente se dividen en diferentes hilos para un procesador de múltiples núcleos, el tiempo de simulación total se puede reducir sustancialmente. Además, la reducción del coste computacional puede ser beneficiosa para los desarrolladores de código y, por lo tanto, para la elaboración de nuevos modelos.

- Se ha optimizado la condición de contorno para los contactos óhmicos considerando que cuando una superpartícula alcanza el terminal de fuente o de drenador, sale del dominio de simulación pudiendo ser reutilizada sin necesidad de procesos de reorganización de matrices. Haciendo esto, se cumple la condición de neutralidad de los portadores porque se vuelven a usar las superparticulas que salen por la fuente o drenador para crear las nuevas a inyectar. Además, usando este prodecimiento se permite el paralelismo y se optimiza la carga computacional. En cuanto a la distribución de la carga inyectada, se han asumido las estadísticas de equilibrio de Maxwell-Boltzmann y Fermi-Dirac. Hemos demostrado que la segunda da una imagen más precisa de la función de distribución debido a que la inyección de superpartículas depende de la relación entre el ángulo y la energía de inyección. Como resultado, la corriente de drenador es mayor usando las estadísticas de Fermi-Dirac que en la de Maxwell-Boltzmann debido a los parámetros de movimiento iniciales.

- Se ha desarrollado un modelo dependiente de la energía para calcular el peso de las superpartículas (número de electrones por superpartícula) para reducir el ruido estocástico que las superpartículas de alta energía incluyen en el comportamiento del dispositivo. Este modelo ha sido motivado porque este tipo de superpartículas son muy improbables y un peso alto puede modificar la estadística dominante. Se ha demostrado que una elección de peso apropiada de acuerdo con la energía disminuye el ruido en la región subumbral y permite el uso de códigos MS-EMC en estas condiciones de polarización.

- El S/D tunneling se ha presentado como un límite de escalado debido a la modificación de la tensión umbral. Se ha proporcionado una comparación de su impacto en FDSOI, DGSOI y FinFET, siendo su perfil de energía y la orientación de confinamiento las que determinan la degradación física. Se ha concluido que la mayor masa efectiva de transporte en el FinFET hace esta arquitectura sea más fuerte frente a este fenómeno, especialmente en los regímenes sub-umbral y de saturación. Además, nuestras simulaciones han demostrado que el tiempo de túnel no puede ser despreciado porque un túnel instantáneo sobrestima el número de partículas que sufre este efecto aumentando su probabilidad de túnel.

- La implementación del efecto BTBT se ha considerado para el estudio de un dispositivo que usa este mecanismo como su corriente de inyección, en lugar de estudiarlo como un mecanismo de pérdidas en arquitecturas convencionales. Este dispositivo ha sido un TFET cuya ventaja principal en comparación con los MOSFET convencionales es el bajo valor del parámetro SS (subthreshold swing) y una pequeña I_{OFF}. Finalmente, una comparación entre estos resultados y otros obtenidos a partir de simulaciones TCAD ha demostrado que el modelo aquí desarrollado proporciona unos resultados muy similares y por lo tanto valida nuestras hipótesis.

- Un modelo de evaluación del mecanismo de pérdidas por la puerta (GLM), incluyendo túnel directo y asistido por trampas, se ha desarrollado para la herramienta MSB-EMC. Nuestras simulaciones han demostrado que el túnel directo es el fenómeno dominante debido al pequeño espesor de óxido. Por último, se ha estudiado una comparación de su eficacia en dispositivos FDSOI, DGSOI y FinFET. Se ha demostrado que la diferencia principal en nuestras simulaciones entre los tres dispositivos cuando este efecto cuántico se tiene en cuenta es el confinamiento de electrones cerca de la interfaz. La estructura DGSOI es la que muestra más tolerancia en términos de variaciones de corriente de drenaje para cualquier régimen de polarización.

Aunque, algunos resultados preliminares se han presentado en este documento, el potencial completo del simulador desarrollado tiene que seguir siendo explotado en futuros trabajos de investigación. En consecuencia, proponemos las siguientes líneas de trabajo como pasos futuros para continuar el estudio de los efectos de mejora y los mecanismos cuánticos en la herramienta MS-EMC:

1. El estudio de la movilidad de electrones puede ser completado por la inclusión de otros mecanismos de dispersión, como por ejemplo el scattering coulombiano remoto, la ionización por impacto o el transporte de portadoras calientes.

2. En cuanto al simulador de electrones 2D MS-EMC basado en silicio, se pueden considerar tres mejoras: en primer lugar, la introducción del transporte de Monte Carlo para los huecos; en segundo lugar, la extrapolación del código 2D a 3D para simular arquitecturas de dispositivos 3D tales como nanohilos; y en tercer lugar, la inclusión de descripciones numéricas electrostáticas y de transporte con el fin de incluir óxidos de alta permitividad κ y nuevos materiales.

3. Tanto el S/D tunneling como la tecnología strain modifican las características del dispositivo, como la barrera de potencial, y por lo tanto el rendimiento del dispositivo. Por tanto, se abordará un estudio exhaustivo de este efecto en los dispositivos sometidos a tensión.

4. Hemos aproximado el fenómeno no local BTBT mediante el uso de ecuaciones locales y la elección de la trayectoria de túnel no local para el cálculo de la tasa de generación. La aproximación WKB sería útil para ser incluida en la tasa de generación con el fin de describir el movimiento de la partícula en la banda prohibida de acuerdo con el vector de onda imaginaria. Esta tarea debe estar vinculada con un estimador robusto para la corriente de drenador calculada como el promedio espacial de las superpartículas generadas por BTBT a lo largo del canal. De esta manera, la corriente considerará las condiciones iniciales de vuelo libre de las superpartículas BTBT.

5. Una vez que el simulador MS-EMC ha demostrado ser fiable para el manejo de los fenómenos BTBT en TFETs, los próximos pasos estarían orientados a la consideración de geometrías alternativas como el heterogate electron–hole bilayer TFET (HG-EHBTFET). La principal diferencia entre ambos dispo-

sitivos es que los TFETs se basan en túneles puntuales (campos eléctricos inducidos por compuertas y direcciones de túneles principalmente perpendiculares), mientras que HG-EHBTFETs se basan en túneles lineales (campos eléctricos y direcciones de túneles en su mayoría alineados).

6. Un supuesto de movimiento donde la superpartícula realiza túnel a través del óxido de puerta en el GLM debe ser desarrollado con el fin de no ignorar el tiempo de túnel dentro del óxido de puerta.

7. El estudio de los efectos cuánticos aquí desarrollados se ha llevado a cabo independientemente. Sin embargo, los dispositivos reales se ven afectados por un gran número de fenómenos a medida que disminuyen las dimensiones del dispositivo. Por esta razón, se puede considerar un análisis del impacto del mayor número de efectos cuánticos para obtener una solución más precisa.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scaling of MOSFETs and the transition to new device architectures

In recent years, the basic architecture of the conventional metal-oxide-semiconductor field-effect transistor (MOSFET) has been diversified in order to improve the device performance when transistors approach to the ultimate scaling limits. Within this time it has been reduced from about $10\mu m$ to the sub–$10nm$ regime. This dramatic downscaling has favored the integration of MOSFETs as well as tremendous advances in Nanoelectronics. The main motivation of the downscaling is to achieve higher integration density and lower power consumption and delay time [1]. Channel lengths and supply voltages have been decreased to meet these requirements [2–4]. However, unexpected performance appears in conventional devices when quantum effects are taken into account [5–11]. Subsequently, standard technology can not provide successful performance at such a small scale.

Predicting this scaling race, the so-called Moore's law stated that the transistor number on an integrated circuit chip would double every 18 months [12,13]. An accelerated device scaling allowed that regime to be reached much earlier than expected. But precisely for that reason, new technological issues must be now faced, such as the increasing source/drain parasitic resistance, short-channel effects (SCEs) or the gate oxide leakage [14–16].

As we are approaching to the $7nm$ node and beyond [1], different trends have appeared to face up specific issues of the scaling limiting critical dimensions. For this reason, it is worth including both the traditional effects, such as channel mobility or short-channel control, and new issues associated with length scales on the atomic dimensions, such as quantum confinement and scattering effects.

Currently, there are two main and parallel work areas in the simulation frame. The first trend is mainly focused on novel engineering solutions to create improved device architectures. The second one is the study of quantum effects, which introduce undesirable performance and variability, in the nanometric dimensions of the conventional devices produced nowadays.

On the one hand, different technological architectures are proposed to overcome the limitations of conventional planar devices. The appearance of the short-channel effects (SCEs) and the leakage currents modify the stable performance of the conventional MOSFETs. The plain control of the transistor electrostatics is the guarantee for acceptable short-channel effects. For this reason, new transistor architectures based on multiple gates, different materials or new injection models are under consideration to replace standard technology and to extend the end of the roadmap.

This first trend can be in turn divided into two main work areas. Firstly, novel engineering solutions are focused on creating improved device architectures [17, 18]. This development route has been mainly proposed to improve the carrier transport properties, and to keep under control the short channel issues. For example, adding multiple gates around the channel reduces the short-channel effects in Double-Gate Silicon-On-Insulator (DGSOI) or FinFET in contrast with Fully-Depleted Silicon-On-Insulator (FDSOI). Furthermore, during the device manufacturing process, the standard wafer made of planar devices (100), such as FDSOIs or DGSOIs, can be replaced with vertical ones (110), such as FinFETs. Secondly, new injection mechanisms and physical phenomena taking advantage from new materials and nanometric dimensions are also proposed. The tunnel field-effect transistor (TFET) has become an alternative to conventional MOSFETs in the last years due to the possibility of achieving low subthreshold swing (SS) that allows low off current and operation at low V_{DD} [19]. The most important difference between both architectures is the injection mechanism which in conventional MOSFETs is governed by the thermionic emission above the source barrier, whereas in TFETs the carrier injection mechanism is replaced by quantum mechanical tunneling through the barrier (band-to-band tunneling

(BTBT)).

On the other hand, the scaling process also leads to the appearance of quantum mechanical effects that should be carefully addressed and which are currently under deep study [15, 20, 21]. In consequence, it has been mandatory their inclusion in order to explain the device behaviors as their dimensions are scaled down. This thesis is focused on the study of three important effects that are presented as a scaling limit from the point of view of simulating possibilities given by the MS-EMC tool: the Source-to-Drain tunneling (S/D tunneling), the band-to-band tunneling (BTBT), and the gate leakage mechanism (GLM).

First of all, quantum effects were taken into account for simulations in the confinement direction to improve scalability. Currently, when the channel length is near $10nm$, it is necessary to include some of them in the transport direction. In particular, Source-to-Drain tunneling (S/D tunneling) introduces a non-desire effect in electronic devices because electrons fly from source to drain through the potential barrier without any gate voltage control. In addition, the number of electrons affected by this is random. It therefore produces a deviation in the functioning of electronic device, which introduces noise. It has been presented as a scaling limit effect in ballistic non-equilibrium Green's Function (NEGF) approach [30, 31].

Secondly, a BTBT model has been successfully implemented in this work. Despite this phenomenon is considered as a leakage current in conventional architectures [21], it is applied to study ultra-scaled silicon-based n-type TFETs because it is precisely its working principle, and so it is no longer an unwanted parasitic effect. Different approaches for the choice of the tunneling path have been compared and relevant differences are observed in both the current levels and the spatial distribution of the generated carriers [34].

Finally, the high electric field across the ultra-thin insulator in such structures oxide coupled with low oxide thickness leads to the possibility that charge carriers can overcome the barrier for transport set up by the dielectric layer, resulting in the tunneling of carriers from substrate to gate and also from gate to substrate through the gate oxide [21]. This tunnel mechanism is known as the gate leakage mechanism (GLM) and it includes intrinsic phenomena, such as the direct tunneling, as well as extrinsic ones, as trap-assisted tunneling.

1.2 Objectives

As previously introduced, the aggressive reduction of the device dimensions has increased the short-channel effects and the leakage mechanisms, seriously compromising its performance. The solutions to those problems need to be assertive and new paradigms in the MOSFET structure and material composition are being addressed. In addition, new physical phenomena appear in the usual device performance as they approach to nanodimensions.

This PhD Thesis is devoted to the study of some phenomena that compromise the performance of the promising alternative devices to continue the MOSFET downscaling process by means of a Monte Carlo simulator. Also, these effects increase the computational complexity causing the necessity of developing new technical models in order to speed up the simulations.

Our study is particularly aimed to investigate which new architecture is better candidate to implement future nodes trying to look for the more robust one against leakage mechanisms. To achieve this goal, onces the quantum effects have been described, a deep comparison among their impact on different architectures is analyzed.

The main goals of this work are:

1. The enhancement of our MS-EMC simulator in order to reduce the computational cost and the stochastic noise, and to provide a more accurate and realistic results.

2. Study of several quantum phenomena: the Source-to-Drain tunneling (S/D tunneling), the band-to-band tunneling (BTBT), and the gate leakage mechanism (GLM).

3. Analysis of the impact of these quantum effects on different architectures.

1.3 Thesis outline

Once the objectives have been established, a brief description of the methodology followed in this work is presented. The study of the enhancement effects and the quantum mechanisms which have been developed for the MS-EMC tool can be divided in three parts. The first one is consigned to the description on the device architectures which are under consideration nowadays. Then, a characterization of the Monte Carlo simulator is provided in the second part. Finally, the third one comprises the quantum mechanisms. The three parts have been worked out according to the following scheme:

- The description of the present and future booster which are proposed to replace standard technology and to extend the end of the Roadmap are presented in Chapter 2. Ones the conventional MOSFET is described bearing in mind its structure, advantages, and the problems of downscaling, the most promising alternatives are introduced. Specially, it has been considered the introduction of new materials, the high–κ oxides, the addition of multiple gates, and the stain and the silicon-on-insulator (SOI) technologies.

- Then, the necessary theoretical background for the Monte Carlo tool is presented in Chapter 3. The first step is to compare the different simulation models for semiconductor devices with special emphasis on its utility and computational complexity. Accordingly, the Monte Carlo method can be deeply summed up and applied to transport calculations in semiconductors, where the free-flight techniques and the scattering mechanisms are taking into account.

- In Chapter 4, the enhancement assumptions included in the advanced MS-EMC simulator have been illustrated. The synchronous-ensemble method, the multisubband supposition and the boundary conditions are characterized in order to get accurate results according to the novel architectures, whereas the parallel implementation and the energy-dependent weight for superparticles are incorporated in this code in order to reduce the computational time reduction and the stochastic noise, respectively.

- Making use of the advanced MS-EMC tool, in Chapter 5, a detail methodology of quantum phenomena are given. After a global overview of the tunneling and leakage currents, the Source-to-Drain tunneling (S/D tunneling), the band-to-band tunneling (BTBT), and the gate leakage mechanism (GLM) are detailed, as well as their impact on different devices architectures.

- Finally, the main conclusions of this Thesis, along with some future work threads that naturally ensue from the work herein presented, are summarized in Chapter 6. A list of the publications yielded by this thesis can be found at the end.

Chapter 2

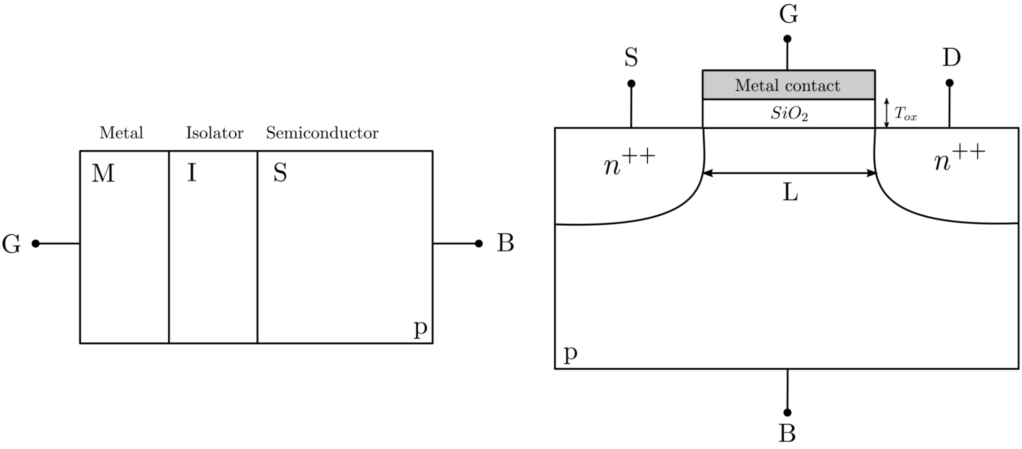
MOSFET devices

The Metal-Oxide-Semiconductor Field-Effect-Transistor (MOSFET) has been the principal component of integrated circuits for more than forty years. Gordon E. Moore realized in the sixties that the number of transistors in a chip followed an exponential growth. And according to this observation he postulated his famous Moore's Law [12, 13]: "the number of transistors in a dense integrated circuit quadruples and its performance doubles approximately every 18 months". This exponential size reduction continues to be nowadays the current trend in the number of transistors per chip. The scaling of the electronic devices improves the performance and the integration density. Moreover, it has an impact on the trends of many other circuit performance metrics and economic indicators. However, it is not clear that the semiconductor industry can keep on making use of the scaling process in the future. The nanometric dimensions of the devices produce undesirable effects and an increase in the variability of the characteristics of the devices, introducing a limit in the scaling of the devices. For this reason, it is necessary the study of alternative technical approaches for electronic devices which fulfill the power consumption and delay time requirements demanded by the International Technology Roadmap for Semiconductors (ITRS) [1].

In this chapter, the conventional MOSFET is described bearing in mind its structure and advantages, as well as the problems and limits of downscaling. Then, some alternatives are proposed to replace standard technology and to extend the end of the Roadmap. Finally, the advantages of the silicon-on-insulator (SOI) technology and a classification of the different structures based on the number of the gates appears at the end of the chapter.

2.1 Conventional MOSFET and downscaling limits

The Field Effect Transistors (FETs) are unipolar devices that mainly transport carriers, electrons or holes, in a parallel layer below a control gate. The most used one is the MOSFET whose principal characteristic is the isolation of the transport layer from a metal gate by including an oxide. These transistors are a four-terminal device based on the Metal-Isolator-Semiconductor (MIS) structure, as shown in Figure 2.1. From a practical point of view, silicon is the preferred semiconductor for conventional technologies and large scale manufacturing because it is very abundant in the nature and it is therefore cheap. It can be easily oxidize to produce in a controllable and reproducible way the silicon dioxide, SiO_2 an insulator with very interesting properties. In addition, the $Si - SiO_2$ interface presents a very high quality due to the low density of interface traps.



MIS structure

n-type MOSFET

Figure 2.1: N-type MIS and MOSFET structures, where L is the channel length, and S, G, D and B are the source, gate, drain and body terminals, respectively.

The operation is based on the control of the transport carriers which flows from the source (S) to the drain (D) terminals by the voltage applied in the gate (G) terminal. The fourth terminal is the body (B) which determines the necessary voltage applied in the gate to start the conduction which is known as the threshold voltage, V_{th}. The body terminal is often connected to the source terminal.

Figure 2.2 shows the operation of an n-type MOSFET. When the voltage applied to the gate increases, the potential barrier to be surmounted by electrons from source to drain decreases, so the conductivity increases. In addition, the inclusion of a drain to source voltage V_{DS} changes the potential between the metal gate and any location in the channel. Thus, even though V_{GS} is defined as the voltage between the gate and the channel in the source terminal, in the drain terminal it corresponds to $V_{GD} = V_{GS} - V_{DS}$. This means that each location in the channel has different voltage between the gate and the channel, and the conduction layer.

When an appropriate positive V_{GS} and a low V_{DS} are applied, the channel can be considered uniform and the device is in the linear regime (Figure 2.2.a). The moment that V_{DS} increases (Figure 2.2.b), a variation of the thickness in the channel is observed until a lack of channel is exhibited near the drain (Figure 2.2.c). This depletion occurs when $V_{th} = V_{GS} - V_{DS}$ and the device reaches the saturation region, hence $V_{DS} = V_{DSsat}$. If V_{DS} continues to increase, the point at which the channel saturates shifts from the drain (Figure 2.2.d).

The theory of scaling was postulated in the seventies and it denotes the possibility of fabricating functional devices with improved performance metrics and reduction in sizes. Due to adverse short channel effects (SCEs), the device dimensions can not be arbitrarily reduced even if allowed by lithography. Consequently, the design of scaled transistors has developed three scaling strategies, defined in Table 2.1, where different reduction factors are introduced for geometrical dimensions (α) and voltages (λ). On the one hand, simple similarity laws maintain unaltered either the maximum internal electric field or the supply voltage [2, 3]. The selection between both scaling rules implies respectively the choice between the device reliability and the system integration capability. On the other hand, these models became inadequate to the design of transistors as the gate length approached to $L_G = 1\mu m$, because the scaling of V_{th} and the supply are very aggressive. For this reason, a more sophisticated criteria was developed, according to which the geometrical dimensions and the voltages are scaling in an independent way [4].

In spite of the large number of innovations introduced to achieve scaled devices

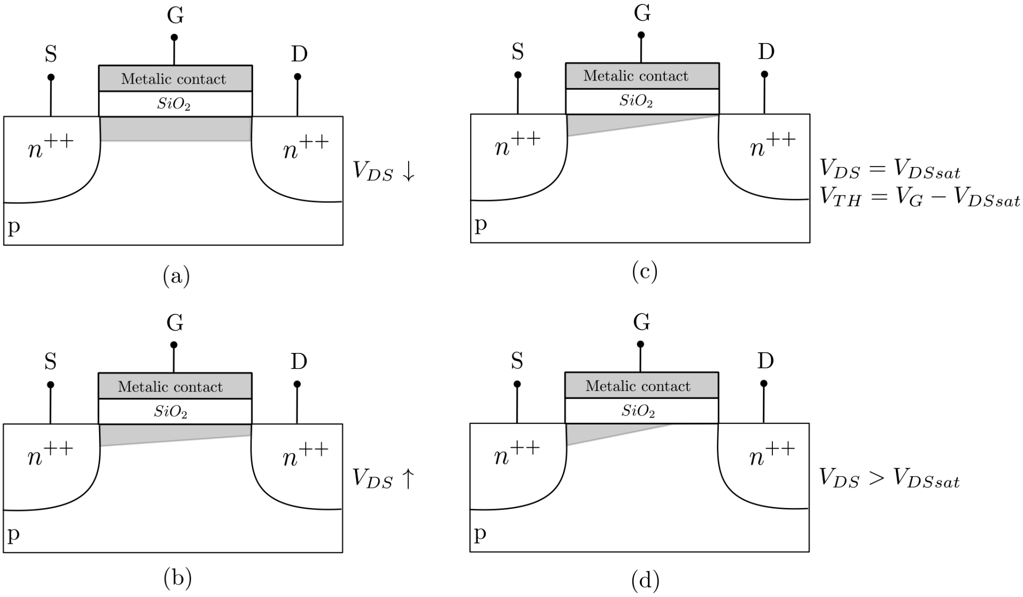


Figure 2.2: Operation of an n-type MOSFET when the gate voltage V_{GS} is positive and constant, and the drain to source voltage V_{DS} is increased.

with improved performance (including new materials and processes), the basic architecture of the MOSFET transistor has not changed dramatically for decades. Nonetheless, the benefit obtained by scaling the devices approaching the end of the Roadmap has some technological difficulties in conventional structures. These issues, such as the variability of the device figures of merit or performance [6,9–11] or the intrinsic unavoidable physical limits when the channel length is under $50nm$ [5, 7, 8], can not be completely overcome. For example, the direct source to drain tunneling (S/D tunneling) will distort the MOSFET operation at transistor channel lengths around $3nm$ [31] becoming this effect a new scaling limit [30].

It is worth understanding why this trend has slowed down in order to be able to continue the MOSFET scaling beyond the $22nm$ technology node. These effects can be exacerbated as the channel length is in the nanometer range. Some of the most remarkable and undesirable effects are:

- A large sensitivity to SCEs. They result in the loss of control of the gate terminal over the channel charge. They arise when the drain-source field controls a large fraction of the charge located below the gate.

Scaling strategies			
Parameter	Constant Field rules	Constant Voltage rules	Mixed rules
Dimensions	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$
Voltages	$\frac{1}{\alpha}$	1	$\frac{1}{\lambda}$
Fields	1	α	$\frac{\alpha}{\lambda}$
Doping	α	α^2	$\frac{\alpha^2}{\lambda}$
Current	$\frac{1}{\alpha}$	α	$\frac{\alpha}{\lambda^2}$
Capacitance	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha}$
Interconnect resistance	α	α	α
Delay	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$	$\frac{\lambda}{\alpha}$
Power delay product	$\frac{1}{\alpha^3}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2 \cdot \lambda}$
Power area-density	1	α^3	$\frac{\alpha^3}{\lambda^3}$

Table 2.1: Scaling strategies for MOSFET technology, where different reduction factors are introduced for the geometrical dimensions (α) and the voltages (λ).

- The increase in the circuit power density due to the impossibility to scale the supply bias and threshold voltage even further, keeping a high I_{ON}/I_{OFF} ratio.

- Unacceptable high device variability, as a consequence of technological fabrication issues. One example is a reduction of the control of the density and location of the dopant atoms in the channel, and the source and drain regions. As a result, this effect implies a reduction of the control of the I_{ON}/I_{OFF} ratio.

These undesirable effects are strongly interrelated because an improvement of one of them can influence to an aggravation of the other. An important parameter that provides information about the performance of a device as mentioned in the list above is the I_{ON}/I_{OFF} ratio, where I_{ON} is the drain current when $V_{GS} = V_{DS} = V_{DD}$ whereas $I_{OFF} = I_D$ when $V_{GS} = V_{th}$ and $V_{DS} = V_{DD}$. Ideally, the best device would be that one with the highest I_{ON}/I_{OFF} ratio. For this reason, it is necessary to achieve a good operation of the device in order to keep a reasonable I_{ON}/I_{OFF}.

On one side, the reduction of I_{ON} is required to downscale the power density. Consequently, it is necessary to decrease V_{DD} in order to reduce the power consumption. When V_{DD} is reduced, V_{th} has to be reduced as well to keep the same gate voltage overdrive, resulting in a exponential increase of I_{OFF}. This influence is caused by the subthreshold current and it is determined by the subthreshold swing, SS. Moreover, the variability, which is the difference in performance for the same electronic device, is another serious limiting factor for the downsizing. Its biggest problem is directly connected to off-current variation, V_{th} and SS. Thus, the most critical limit of downsizing is the off-current increase for the MOSFETs.

On the other side, if the supply voltage reduction trend is halted to maintain higher V_{th}, the electric field increases in every portion of the MOSFET. It therefore leads to a reliability problem and an increase of SCEs, such as the S/D tunneling or the drain induced barrier lowering (DIBL) [16]. At the same time, the impact of DIBL can strongly affect the I_{ON}/I_{OFF} ratio by yielding a punch-through in the channel. DIBL causes the reduction of the source barrier due to the drain voltage, makes the source junction forward-biased, and increases the I_{OFF}. In addition, the presence of relevant traps in the gate dielectric degrades significantly the DIBL. These interface traps can be viewed as dynamic charges which change with applied bias shifting V_{th}.

A traditional solution for the SCEs is reducing the vertical dimensions, such as the gate oxide thickness, increasing the electric field between the gate electrode and the channel, so as to enhance the gate bias controllability of the channel potential [15]. However, thinning the equivalent oxide thickness (EOT) also increases the gate leakage.

Finding solutions to these issues has slowed down the shrinking of the device size, leading to the use of new materials and device designs. The semiconductor industry has been able to reinvent itself keeping a continuous improvement of the performance of their product. There is still a long way to go before the MOSFET technology will be exhausted, as it will be detailed in the next section. Meanwhile, the electronic industry has been able to find solutions to these issues, leading to the use of new materials and device designs. In consequence, new transistor architectures are considered to replace standard technology [17, 18] and to fulfill the performance demanded by ITRS for the future technology nodes [1] (Table 2.2). An evidence of the continuous deceleration of the scaling process is also depicted in table 2.2. On average, the shrink rate of the gate length is predicted to vary from 0.7 to 0.85 in 3 years according to Moore's law [12]. But precisely for that reason, new technological issues must be now faced.

Year	2013	2015	2017	2019	2021	2023	2025	2027
Commercial name (nm)	14	10	7	5	3.5	2.5	1.8	1.3
Metal half pitch (nm)	40	32	25.3	20	15.9	12.6	10	8
L_G (nm)	20.2	16.8	14	11.7	9.7	8.1	6.7	5.6
V_{DD} (V)	0.86	0.83	0.80	0.77	0.74	0.71	0.68	0.65

Table 2.2: Parameters for actual and future technology nodes predicted by ITRS 2013 for high-performance devices, where L_G is the physical gate length and V_{DD} is the supply voltage.

2.2 Alternatives to conventional MOSFET for downscaling

In the last decades, the manufacturing industry of semiconductors was mainly based on Silicon. By contrast, the range of application of conventional MOSFETs has been suffering an important boost because a large amount of alternative technologies and materials have been chosen to replace it. As the sub-10nm nodes become closer (Table 2.2), it is worth including traditional effects, such as channel mobility behavior or short-channel control, and new ones associated with dimensions in the range on the inter-atomic distance, such as quantum confinement and scattering effects [13]. Two main work areas may be differentiated in the study of possible alternatives concerning processes and devices:

1. The first one is mainly focused on novel engineering solutions in order to create improved device architectures. For instance, new materials (germanium

and III-V compound semiconductors), high mobility bulk materials (strain Si, SiGe, etc), the introduction of new gate dielectrics (high–κ materials), and novel designs with great electrostatic control, like multigate devices (MuG). Its development has been conceived to improve the carrier transport properties and to keep under control the short channel effects, creating devices that prove to be more "long–channel" like in their behavior [17, 18]. In parallel to the implementation of these new devices, the scaling process also leads to the appearance of quantum mechanical effects that are currently under deep study. As an example, S/D tunneling increases the current because of the number of electrons that flow from source to drain is higher. This effect is of special interest when operating in near-threshold regime, because the leakage current increases and V_{th} decreases [27].

2. The second approach, which is more long-term focused, is that of the "new injection mechanisms". This alternative seeks the exploitation of new device paradigms based on different injection mechanisms and the nanometric regime enforced by scaling. Among some of the most promising solutions of this area, we focus on the Tunneling Field–Effect Transistors (TFETs) [19]. The process of quantum mechanical tunneling through a barrier between energy bands, known as band–to–band tunneling (BTBT), governs the injection of carriers in TFETs, contrary to MOSFETs where the thermionic emission dominates. BTBT is a clear example of novel injection mechanism devices where the tunneling currents are calculated by the Wentzel-Kramers-Brillouin (WKB) approximation to determine the tunneling rate [33]. TFETs exhibit a similar architecture as the MOSFETs except that the source and the drain are oppositely doped, and hence it can be considered as a gated $p-i-n$ diode (Figure 2.3).

   The subthreshold swing SS of the conventional MOSFETs at room temperature under ideal conditions faces a theoretical barrier of $60mV/dec$ at room temperature:

$$SS = \frac{dV_{gs}}{d\left(\log I_{ds}\right)} = \frac{kT}{q}\ln 10\left(1+\frac{C_d}{C_{ox}}\right) > \frac{kT}{q}\ln 10 \approx 60\text{mV/dec.} \tag{2.1}$$

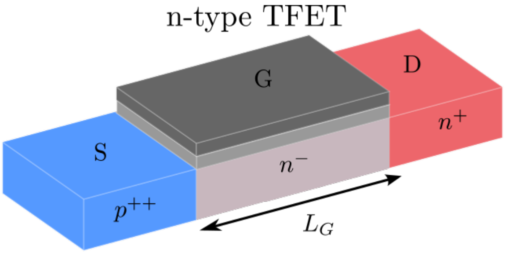


Figure 2.3: Schematic of n-type TFET, where L_G is the channel length, and S,G and D are the source, gate and drain, respectively.

In practice, the best MOSFET implementations cannot bring SS$< 70-80$ mV/dec, as it provokes an even worse situation [36, 37] with the subsequent limitation in the downscaling of V_{DD}. Notwithstanding the above, TFETs should not be constrained by the aforementioned SS limit as depicted in Figure 2.4, where the transfer curves of several MOSFET types (bulk Si, high mobility, and multigate), and TFET are compared. Each of the different alternatives presents a particular enhancement, but despite the lower I_{ON}, TFETs presents the higher difference between the on and off states. They therefore are introduced as a potential substitute for the conventional MOSFET used in low-power applications.

2.3 Mainstream technologies

In recent years, the basic architecture of the conventional MOSFET has been diversified to improve the device performance when transistors approach the ultimate scaling limits. As a consequence, different trends have appeared to face up specific issues of critical dimensions limiting the scaling, such as channel mobility, short-channel control, quantum confinement or scattering effects.

On the one hand, the replacement of the silicon channel with alternative semiconductors has been proposed to improve the carrier transport properties. This is the case of III-V materials, known to have superior electron mobility, whereas germanium is used to enhance the hole mobility. Moreover, the strain technology has been also adopted to improve the carrier mobility and so the I_{ON}.

On the other hand, gate leakage currents must be kept under control to contain power dissipation in the off state and guarantee the device reliability. For

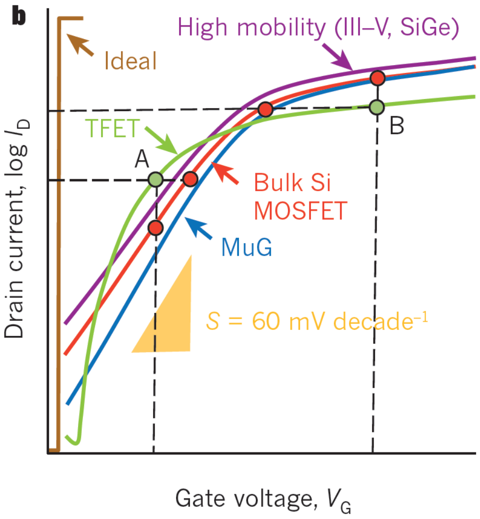


Figure 2.4: Transfer curves for different types of MOSFET (bulk Si, high mobility, and multigate), compared with a TFET transfer curve. Since the slope in a TFET can be smaller than $60mV/dec$, V_{DD} can be scaled down without significally increasing I_{OFF}. Figure extracted from [19].

example, improvements in electrostatics can enable much shorter L_G at constant V_{th} and I_{OFF}, which means density scaling improvements. The inclusion of high-κ oxides enhances the gate control over the channel. An alternative approach to improve the electrostatic confinement is to implement new architectures surrounding the channel with more opposing gates since they reduce S/D interaction: Multiple gate field-effect transistors (MuGFETs) [38]. In this respect, improved control of the threshold voltage roll-off and low values of the SS could be achieved, when the channel length is below $0.1\mu m$, using ground plane architectures such as Silicon-on-Insulator (SOI) technologies [39].

Moreover, these technologies can be mixed to enhance the device characteristics. For instance, the $32nm$ technology uses strained silicon, high–κ and replacement metal gate flow [40].

2.3.1 New materials

One of the main limitations in the conventional technology based on silicon is the degradation of the carrier mobility. Most of the compound semiconductors, such

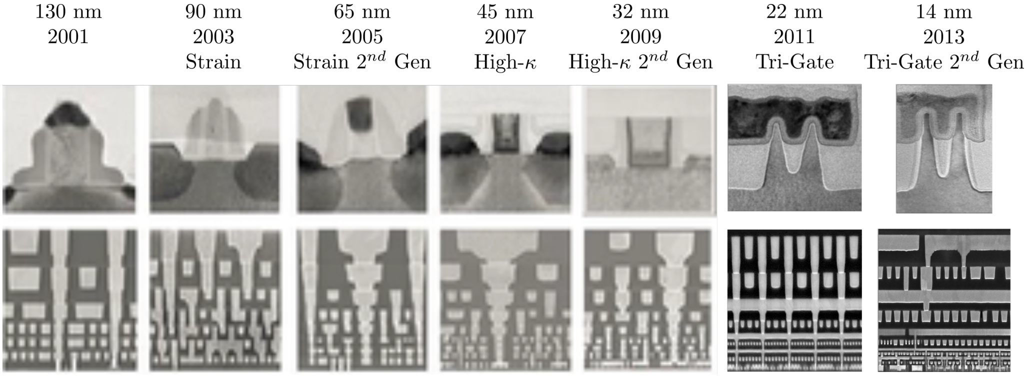


Figure 2.5: Main innovations introduced on the nodes previous to the present and the present one. After [17].

as *GaAs* or *InP*, have higher mobility than silicon. However, it is difficult to manufacture MOSFETs with these materials due to difficult integration with native gate oxides [41]. As a consequence, it is necessary to study the characteristics of these new materials in order to integrate them on silicon substrates.

The introduction of new materials other than silicon in the channel is a promising alternative to improve I_{ON}. The device current can be calculated as $I = Q \cdot v$, where Q is the charge in the channel and v is the velocity of the carriers. If Q is assumed to be constant, the higher the v, the higher the I_{ON}. As the channel length is downscaled, the physical processes which limit v vary. In general, the carrier transport is diffusive due to the carrier scattering. The velocity can therefore be calculated as $v = \mu \cdot E$, where μ is the mobility and E the electric field [42]. However, when the channel lengths approach to nanoscaled dimensions, the carrier transport is expected to be ballistic or quasiballistic [43]. In that case, v corresponds to the injection velocity of the electron at the source. The improvement of the carrier velocity mainly depends on the effective mass (m^*), which becomes a crucial component in both cases [44]. For this reason, low effective mass (m^*) materials are an important factor to focus when choosing new materials.

In addition, its implementation in Complementary MOS (CMOS) technology would imply the use of different materials that improve hole or electron transport properties for p-type and n-type MOSFET, respectively [45]. The germanium has been selected as a highly regarded solution for PMOS [46], while III-V semiconductors are the best materials for NMOS [47,48]. This difference in the optimal material means that novel new integration techniques will also be required.

Other limitation that complicates the introduction of new materials in the CMOS technology is the expensive and weak substrates as well as the difficulty of manufacturing them in massive scale [49]. For this reason, the success of a non-silicon based technology depends on its compatibility to use the same fabrication processes for the overall substrate in mainstream manufacturing.

2.3.2 Strain Technology

Strained channels are used to improve the performance rather than the geometrical scaling because they increase the mobility and therefore, I_{ON} [50–52]. Strain is a relatively old topic in semiconductor physics. The impact of strain on silicon and germanium resistivity has been experimentally observed since the early 1950s [53, 54]. However, strain was not introduced in the semiconductor industry until the $90nm$ technological node in 2004 [17, 55], as depicted in Figure 2.5. Later, the $65nm$ generation introduced in 2005 further improvements in these strain techniques to increase transistor performance.

Strain affects the characteristics of MOS transistors in several aspects. In general, when a strain technique is used, the lattice mismatch compresses or stretches the Si and it induces shifts in the conduction and valence bands of silicon. The carriers are therefore redistributed in the subbands, affecting the threshold voltage of the transistors and changing scattering [56, 57]. The orientation also plays an important role in hole mobility whereas its impact is limited in electron mobility [58].

Global strain techniques produce the same strain condition throughout the entire wafer. The devices are realized in strained silicon layers epitaxially grown on a relaxed $SiGe$ virtual substrate. The lattice mismatch between $SiGe$ and Si yields a biaxial strain in the epitaxial silicon layer. In particular, since the size of the Si lattice is smaller than the one of $SiGe$, it will be stretched in two directions (biaxially). Changing the mole fraction x of the relaxed $Si_{1-x}Ge_x$ virtual substrate, the strain magnitude in the silicon layer can be varied. The main limitation of this technique is that the same strain configuration is provided for all devices. For instance, a compressive biaxial strain improves hole mobility whereas it degrades electron mobility and n-MOS and p-MOS are differently affected. For this reason, local strain techniques are used to introduce strain engineering in mass production, as they have a higher cost-effectiveness and greater design versatility.

Instead of global techniques, local strain techniques are the ones that can induce different strain configuration in selected regions of the wafer. The widely used local strain technique is uniaxial stress induced by source and drain stressors. It is based on the lattice mismatch between the silicon in the channel region of the MOSFET and the locally epitaxial grown $Si_{1-x}Ge_x$ and $Si_{1-x}C_x$ in source and drain regions for p-type and n-type MOSFETs, respectively. Taking into account that the lattice constant of this material is larger than in Si, the S/D regions expand and compress the channel resulting in a uniaxial compressive strain in the channel direction.

Strain is probably the most cost-effective CMOS technology booster developed in recent years. However, not all strain combinations result in higher mobility. For example, an important degradation of the performance can be obtained from the enhancement of the intervalley scattering, as this one causes a loss of the boosting capabilities of strained channels not only because of the scattering itself but also due to the increase in the transport effective mass as a consequence of the repopulation of primed subbands [59]. This effect, which has an intrinsic statistical origin, may be considered as a new source of device variability in future technological nodes, making necessary further optimizations in this respect.

2.3.3 High–κ Materials

As the dimensions of the devices are scaled down, the electronic industry drastically reduced the gate dielectric thickness in each new generation. The problem was introduced after the $65nm$ technological node, since the gate leakage limited the oxide thickness reduction. Additionally, the manufacturing industry started to focus on lower-power products, in spite of improving the intrinsic device performance.

As it is broadly known, the higher the electric field in the insulator, the larger the leakage current. Moreover, higher electric fields may cause an early dielectric breakdown. The equivalent oxide thickness (EOT) provides the thickness that a

SiO_2 film would have to have the same capacitance as the considered dielectric with given thickness (T) and permittivity (κ). Thus, with $\kappa(SiO_2)$=3.9, the EOT of a planar device can be calculated as:

$$EOT = \frac{3.9}{\kappa}T \tag{2.2}$$

The introduction of high–κ oxides enables a return to electrical gate oxide thickness scaling while keeping a reasonable amount of gate leakage currents at low power consumption [60, 61]. The most common high–κ oxide is the HfO_2, with a dielectric constant five times larger than for SiO_2. For this reason, the same dielectric performance can be obtained with a dielectric layer five time thicker. It therefore significantly reduces the leakage current in the gate. However, the height of the potential barrier is smaller for the HfO_2 than for the SiO_2 and consequently, it increases the thermionic current. In addition, a material with larger permittivity will have generally a lower band gap, so it becomes a worse insulator. One of the main drawbacks of high–κ materials is the high and almost unacceptable density of traps when the dielectric is directly deposited on top of the silicon.

The first node using high–κ dielectric was a hafnium-based gate dielectric at the $45nm$ node in 2007. A stack of HfO_2-SiO_2 was used, where the SiO_2 was introduced to guarantee a good interface with the silicon [62] and thus reducing the interface traps. Nonetheless, an extensive research is still being carried out in order to achieve high-quality insulators reducing the EOT below $1nm$ [63]. In spite of benefiting from the reduction of the EOT, the high penetration of the drain field in the channel increases the DIBL and the SS. For this reason, maintaining a good electrostatic performance is not enough. However, this technology dominates the electronic industry at $45nm$ and $32nm$ generations [17] as shown in Figure 2.5.

2.3.4 SOI Technology

In the last decades, Silicon-On-Insulator (SOI) technology, which is based on the introduction of a buried insulator layer, known as the buried oxide (BOX), in the Si substrate, has been recognized as the best alternative to conventional technology (Bulk) [64, 65]. The main objective of SOI technology is the elimination of the drawbacks of conventional Si devices. Its development has been mainly motivated by three different causes [66]:

- A reduced effect of both the ionizing radiation and the parasitic capacitance with respect to bulk technology. Most of the charge generated by the ionizing radiation in the body of the transistor is stopped at the buried oxide (BOX), thus avoiding that the generated charge reaches the channel. The extra generated current is therefore very low.

- The fabrication of this technology is preferable to other alternatives due to the improvement of the performance of devices based on SOI.

- The insulator substantially reduces effects such as DIBL, leakage currents and the punch-through. These effects are the consequence of an unsuitable performance of devices with channel length lower than $32nm$ when they are manufactured using conventional technology due to SCEs.

From the point of view of manufacturing, the compatibility between SOI and conventional technology resulted in a rapid consolidation of both structures, hence it became extensively used in Industry [67] as well as in Academy [68]. In fact, some of the indispensable steps in conventional MOSFETs to guarantee the correct isolation between devices, and to reduce the parasite effects are unnecessary in SOI thanks to the BOX and the lateral dielectric. For this reason, the chips based on SOI are compact and fabricated more easily. Nevertheless, the manufacturing cost of SOI is slightly higher than for conventional devices because wafers have to be pre-processed in order to get the substrate [69]. Another problem of this technology is the influence of Si film thickness on the electron mobility degradation due to phonon confinement, Coulomb scattering, and doping fluctuations [70]. These drawbacks are justified by a large amount of advantages, which are summarized in the following list [66, 71]:

- This technology is fully compatible with existing manufacturing facilities.

- It can significantly reduce the number of steps in the device fabrication process due to the use of an insulator substrate.

- The density of integration is higher due to the easier circuit design.
- SOI technology is tolerant to ionizing radiation.
- A specific operation velocity requires lower bias, hence a higher operation velocity is required for a specific bias.
- Higher control over short channel effects (SCEs).
- Parasitic capacitance is reduced.
- It is more flexible due to the fact that some structures can only be integrated in SOI technology.
- Different type of devices can be integrated in the same chip. As an example, high velocity or power devices, Micro-Electro-Mechanical Systems (MEMS) or optic elements, MOSFETs, Bipolar Junction Transistors (BJT), Junction Field-Effect Transistors (JFET), or diodes can be integrated in the same SOI wafer.
- Both planar and three-dimensional devices can be manufactured. Three-dimensional devices can be integrated thanks to stacking of packages, die stacking, or Through Silicon Via (TSV) technology [72].

The most used device based on SOI technology is still the single gate MOSFET. Nonetheless, a huge number of new devices have appeared due to the buried oxide and the development of novel techniques to manufacture SOI structures. For this reason, it is possible to chose which one is the best candidate to each application. In this section, a classification of SOI devices is detailed regarding the number of gates.

2.3.4.1 Single gate devices

The single gate devices (SGSOI) are the natural continuation of MOSFETs based on conventional CMOS technology. Depending on the existence of a neutral region under the channel, SGSOI can be classified in two different devices:

- Partially depleted SOI (PDSOI). The behavior of this device is similar to conventional MOSFET but it has the advantage of the total isolation thanks to the buried oxide. Nevertheless, the neutral region under the channel can cause different problems for the standard operation such as the floating body, that changes V_{th} according to its history (hysteresis process) [73–75], effects caused by the parasitic bipolar transistor made up for drain, neutral, and source regions, or the kink effect in high bias between source and drain. One way to suppress these effects is connecting the gate with the body. This technique creates the Voltage-controlled bipolar-MOS (VCBM) or the Dynamic Threshold MOS (DTMOS), being their main characteristics both the advantage of using the parasitic bipolar transistor and the reduction of the threshold voltage due to the gate bias. Theses devices have an almost ideal behavior in the subthreshold regime and a drastic reduction of floating body effects.

- Fully depleted SOI (FDSOI). When the thickness of the Si film standing on the insulator is decreased, the neutral region under the gate starts being reduced and it could disappear. When the T_{Si} is thinner than the channel depletion depth, the body is totally depleted which means that the channel takes up all the Si region between both oxides [76, 77]. It can improve the electrostatic confinement even more than PDSOI, in which T_{Si} dimensions are significantly larger that the depletion depth. For this reason, it can be also called extremely thin SOI (ETSOI) or ultrathin-body (UTB) because of its thinner silicon film. The main advantage of FDSOI devices is the insusceptibility to floating body effect and the remarkable improvement of SCEs, DIBL and SS [78]. Moreover, they reduce the random-dopant fluctuations due to the virtually undopped channel. In addition, the body can be biased with a thin BOX, resulting in a benefit to system-on-chip (SOC) designers [79]. The dependence of the inversion charge on the bias and the channel thickness T_{Si} is the main drawback considering that it introduces fluctuation in the V_{th}. Overall, in spite of the steady progression in the minimum achievable T_{Si} [78, 80], quantum confinement becomes critical for $T_{Si} \leq 5nm$ due to the increase of V_{th} and the variation of the scattering effects [81].

2.3.4.2 Multiple gate devices

As channel dimensions are shrunk, the short channel effects (SCEs) appear in the device. These effects mainly cause the loss of control of the gate terminal over the channel charge. Despite the progress in electrostatic confinement and competitive fabrication cost, the addition of multiple gates (MuG) surrounding the channel reduces the short-channel effects [82]. These devices are not required to be used only in SOI technology, although they can be used to simplify manufacturing [83, 84]. For this reason, these devices are going to be generally explained in the next Section 2.3.5.

2.3.5 MuG architectures

Historically, the MOSFET design has been focused on the scaling without losing the control of the gate over the charge. However, the SCEs appear as the channel length is scaled down and consequently they introduce the aforementioned lack of control of the gate over the charge. For this reason, the multiple gate devices (MuG) were proposed by the industry and researchers to continue the downscaling process [38, 85–88].

Fully depleted devices, such as FDSOI and MuG, can control I_{OFF} through architecture by altering the gate work function in order to target V_{th} rather than doping profiles [89]. The channel can therefore remain undoped facing up the random dopant fluctuations and the mobility degradation, which is caused by the Coulomb scattering and doping fluctuations in planar SOIs. For this reason, MuG devices also mitigate the degradation of the device performance.

In addition, MuG can incorporate the boosters proposed for the CMOS technology, including high–κ oxides, strained materials or alternative channel orientations that can affect the transport properties and, hence, the behavior of the device [87].

Each additional gate improves the short channel control and relaxes the manufacturing constraints as opposed to single gate devices. For example, a channel thickness (T_{Si}) is required to be one fourth of the channel length in order to guarantee acceptable short-channel effects in Fully depleted SOI (FDSOI). However, an extremely thinner T_{Si} can represent a critical parameter in the fabrication of electronic devices as they are scaled down. The addition of more gates to the transistors improves the electrostatic control of the channel, and it reduces therefore the SCEs. Thus, this critical T_{Si} of a double gate transistor is approximately twice as wide as the T_{Si} of a single-gate device with the same short-channel behavior. As a result, it alleviates the fabrication problem.

Moreover, other benefits appear resulting from the increase of the gate control such as the reduction of DIBL, SS, puch-through or leakage current [16, 71]. This effect can be easily understandable considering the "natural channel length" parameter (λ_N), which represents the extension of the electric field lines from S/D regions to the channel [90]. It can be calculated as [18]:

$$\lambda_N = \sqrt{\frac{\varepsilon_{Si}}{N\varepsilon_{OX}}T_{OX}T_{Si}} \tag{2.3}$$

where ε_{Si} and ε_{OX} are the electrical permittivity of the channel and the gate dielectric, respectively, N is the number of gates, T_{Si} is the film thickness, and T_{OX} is the gate dielectric thickness. A device will have minimized SCEs if L_G is approximately six times longer than λ_N. For instance, if a single gate (N=1) and a GAA (N=4) devices are compared, the second one can be scaled down without critical SCEs for the same L_G.

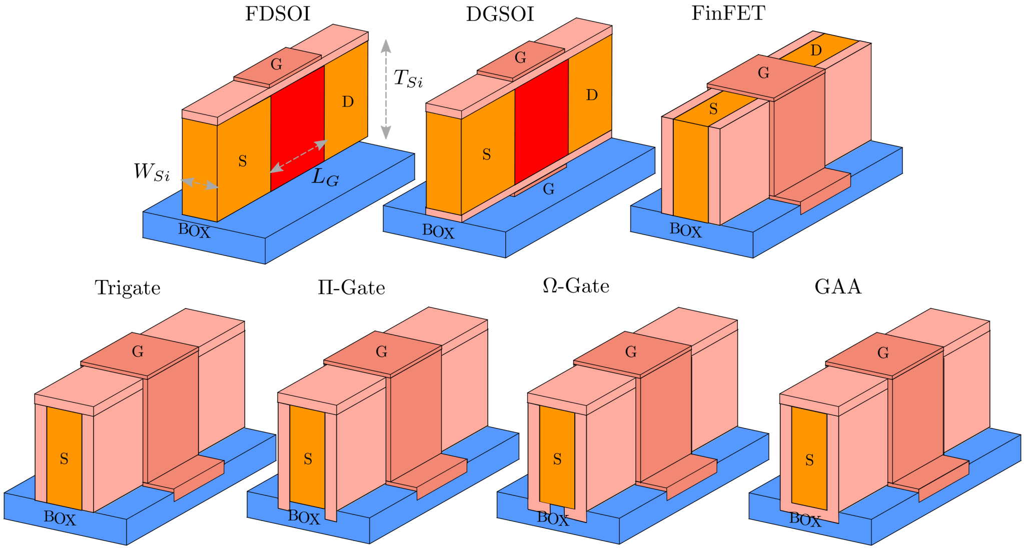


Figure 2.6: Architectures of different Multiple gate devices.

There are two different architectures [82]: vertical or planar MuG transistors. The gates of the first one mentioned are perpendicular to the standard wafer orientation, whereas they are parallels in the planar disposition. One of the main advantages of vertical devices is the excellent control of the channel by the gate due to the diminished leakage current in the off-state. They are also less susceptible to several sources of variability such as V_{th} variability or random dopant fluctuations. FinFET [91–93] or TriGate [94] are two examples of vertical MuG devices; as for planar transistors, they are mainly focused on silicon over insulator (SOI) technology [66].

The first device based on this concept was a double gate SOI (DGSOI) with gates oriented horizontally and a channel forming a 2D layer located below the gate, as depicted in Figure 2.6 [95–98]. In general, PDSOI, FDSOI and DGSOI are considered planar devices due to the reasons described above. Then, the MOSFET design was changed from planar to 3D or vertical devices. The channel and the gates are vertically oriented making its fabrication easier than the horizontal ones. The first fabricated double gate was a vertical device called DELTA [85], which was the predecessor of FinFETs [91–93], in which two opposite faces of the channel are covered by the gate (Figure 2.6). The channel of the FinFET is fabricated using an ultra-thin layer of silicon that sits on top of an insulator. As a result, it reduces the electric field from the gate to the fin on the top. One advantage of the FinFET is the excellent control of the conducting channel by the gate. FinFETs are also less susceptible to several sources of variability such as V_{th} variability or random dopant fluctuations [82].

Once the improvement in SCEs was verified due to the inclusion of the second gate, the next step was to extend the number of gates to more than two as shown in Figure 2.6 including TriGate architecture, where three faces (two sides and the top) of the channel are covered by the gate [94, 99–101]. Then, two more sophisticated devices were developed: Π-Gates featuring the side gates extended below the channel [102], and Ω-FETs, where the gate does not only wrap around the two sides and the top, but also part of the fourth [74, 103].

Devices with two (or more) opposing gates were mainly included in the $22nm$ technology in order to increase the gate control over the channel by enclosing it. In particular, TriGate devices have been implemented successfully into manufacturing on the $22nm$ node (Figure 2.5) [17, 104]. The second generation of trigate transistor ($14nm$) improves the device performance with significant reductions in operational power consumption thanks to the rectangular fin profile and sub-fin isolation optimizations that minimize the channel doping [105].

Finally, the use of a gate wrapped around the semiconductor increases its control over the channel and optimizes the improvement of SCEs when L_G is reduced [106–109]. Gate-all-around (GAA) devices were first reported in the late 1990s [110, 111]. Nanowires (NW) are an extreme case of GAA devices, having height and width dimensions roughly the same (or even cylindrical) and atomically small ($< 10nm$) dimensions [88,112–116]. The main difference between conventional planar or finned devices and nanowires is that the carrier conduction moves from the surface to the center of the device.

Chapter 3

Monte Carlo Simulation

Device simulation is one of the most important tools in the design of novel architectures. The semiconductor industry must face different technological challenges in very wide sectors, such as the computation, the telecommunications, the electronic instrumentation, or the optoelectronics. These challenges can be solved thank to the advantageous results provided by simulation tools, because they can select among a wide range of devices which is the optimal one for a specific set of requirements. In addition, the use of simulation techniques beforehand can save money and time in the fabrication of new electronic devices. Firstly, these tools can asses the performance of new structures and, subsequently, the optimal design can be chosen to deal with a particular technological problem under study. Secondly, the results of these simulations can be compared with experimental data obtained by the characterization of test devices.

The complexity of the model approaches in the simulation of semiconductor devices as well as the computational cost mainly depend on the technological challenge that we want to face up, and on the detail level of the physical mechanisms that are included in the studied device. The first section of this chapter is devoted to different simulation approaches usually employed in the description of semiconductor devices with special emphasis on its utility and computational complexity. The second section will be focused on the principles of the Monte Carlo method applied to the study of electron transport in semiconductors.

3.1 Simulation approaches

The standard flowchart of electronic device design is depicted in Figure 3.1, where the modeling steps and the behavior of the electrical devices based on fundamental physics in Technology Computer Aided Design (TCAD) are described. Theses design tools are more important due to the increased demand of electronic devices from the commercial sector. It therefore involves an improvement of the fabricated products and a cost saving. Broadly speaking, the following steps can be considered: firstly, the market needs demand a device with specific performance and hence the fabrication processes are simulated. Then, its performance is simulated from which the I-V characteristics are obtained as well as the small signal parameters. Thirdly, a process to extract the required parameters is carried out in order to evaluate the device performance at circuit level. Finally, the obtained results are compared to the market specifications. If they do not satisfy the commercial requirements, the process is repeated again changing some fabrication parameters.

Figure 3.2 shows a hierarchy of simulation techniques used nowadays in the design process. Each approach represents a different level of description according to the two main simulation characteristics: the approach complexity and the simulation time [117]. All the simulation levels, with the exception of the compact approach at the bottom of the hierarchy, involve the solution of a set of coupled partial differential equations. Moreover, the carrier transport can actually be described by an integrate-differential equation, such as the Boltzmann Transport Equation (BTE) or through a quantum formulation including different kinds of carrier scattering.

If the model complexity is only keeping in mind, the less strict is the compact approach. It contains a reduced amount of information about the physic of the system and, in general, it mimics the device performance using analytical approximations and empiric parameters. This approach needs very low computational time and provides a simple solution. However, its validity is limited due to the lack of awareness regarding very important details such as the distributed nature of the parameters or the device geometry.

The next level of the hierarchy is the drift–diffusion approach (DD) of BTE [118]. DD represents an approximation obtained from the two first momenta

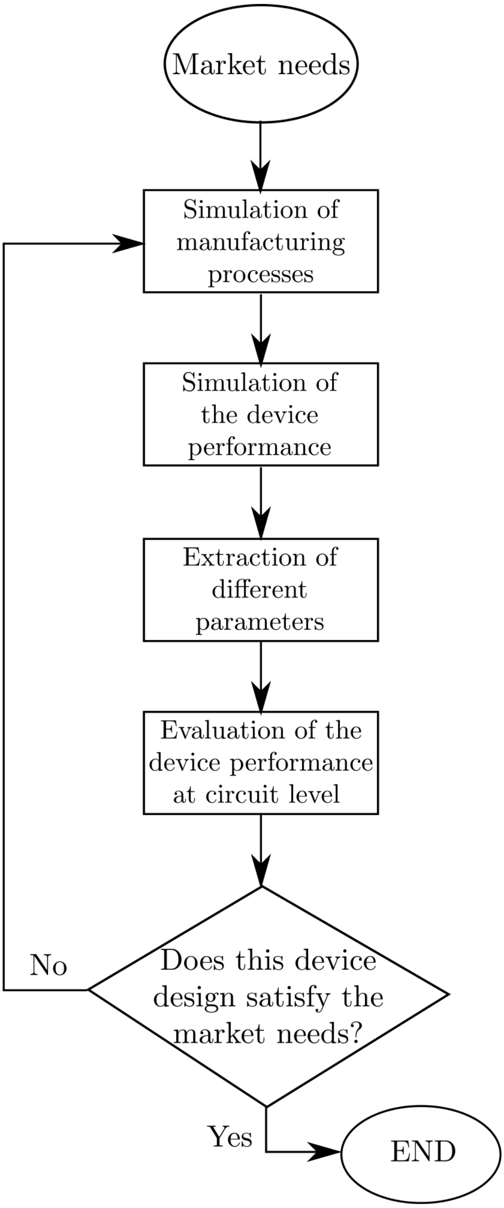

Market needs
Simulation of manufacturing processes
Simulation of the device performance
Extraction of different parameters
Evaluation of the device performance at circuit level
Does this device design satisfy the market needs?
No
Yes
END

Figure 3.1: Standard flowchart of electronic device design, where the modeling of process steps and the behavior of the electrical devices are based on fundamental physics in Technology Computer Aided Design (TCAD).

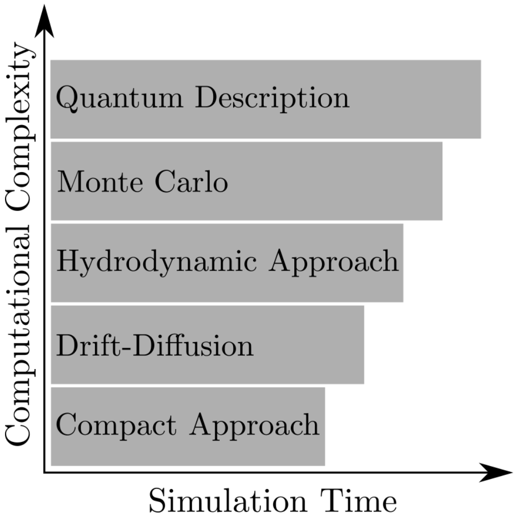


Figure 3.2: Hierarchy of computational approaches used nowadays in the design process according to the two main simulation parameters: the computational complexity and the simulation time.

of BTE, the current continuity equation, and momentum conservation equation. Both equations are coupled to the Poisson one by the electrostatic potential. The velocity and the electric field are related locally in DD causing the impossibility of representing the transport effects outside the equilibrium region.

The hydrodynamic approximation of the BTE is the immediately higher complexity model. The third momentum in BTE and the energy conservation equation are included in this approach complicating the momentum conservation equation. Some transport effects outside the equilibrium region can now be bearing in mind because of the non-local relationship between the velocity and the electric field.

Subsequently, the Monte Carlo technique is in the above complexity level. In this approach, a group of particles is evaluated in a region in which the scattering events are chosen randomly. It is in principle more general, and can go beyond the limits of the BTE. It is mainly used in small devices thanks to its accurate description of the physical processes involved. Nonetheless, due to its statistical nature, the Monte Carlo approach naturally includes statistical fluctuations and noise.

Finally, the highest level in the hierarchy is the quantum description that includes the time-independent Schrödinger equation coupled to Poisson equation, the density matrix, and the Wigner distribution function. Other method that is very popular nowadays is the Non-Equilibrium Green Functions (NEGF). However, theses approaches are computationally too expensive.

In this section, the three most important approaches are described in detail by increasing the computational complexity: Drift-Diffusion, Monte Carlo, and Quantum Description.

3.1.1 Drift-Diffusion

Drift-Diffusion is the simplest approach used nowadays in multidimensional numeric simulations [118–121]. It represents the approximation of the lowest-order in the transport system obtained from the momenta of BTE. In order to attain a self-consistent solution, the equations of the transport approaches are nearly always coupled to the Poisson equation. The general form of the equation is:

$$\nabla \times \epsilon \nabla V = -\rho = q\left(n - p + N_A^- - N_D^+\right), \tag{3.1}$$

where V is the electrostatic potential, q is the magnitude of the electron charge, ϵ is the dielectric permittivity, ρ is the space-dependent charge density obtained from the knowledge of the density of ionized donor (N_D^+) and acceptor (N_A^-) dopants, and the density n and p of mobile electron and hole carriers. Along with the Poisson equation 3.1, only the continuity equations for electron and hole carrier densities enter the model

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot J_n + U_n \tag{3.2}$$

$$\frac{\partial p}{\partial t} = \frac{1}{q}\nabla \cdot J_p + U_p \tag{3.3}$$

where U_n and U_p indicate the net generation–recombination terms. Assuming the 1D case, the current $J(x)$ in general can be redefined as

$$J(x) = q \int v \cdot f(v, x) \tag{3.4}$$

in terms of distribution function with an explicit dependence on velocity. This definition of current can be related to the Boltzmann equation by taking the moment over velocity v. Solving the equations considered above, the current density for electrons (J_n) and holes J_p can be calculated from the sum of two components.

$$J_n = qn(x)\mu_n\varepsilon(x) + qD_n\frac{dn}{dt} \tag{3.5}$$

$$J_p = -qp(x)\mu_p\varepsilon(x) - qD_p\frac{dp}{dt} \tag{3.6}$$

The first one is a drift current governed by the electric fields (ε) with an explicit dependence on velocity and low field mobility (μ). The second one is the diffusion current following the gradient of the carrier density being D_n and D_p the diffusion coefficients. The resulting Drift-Diffusion current is therefore only valid in the region of low fields where the velocity is linear, the low-field mobility being the slope of the curve.

This approximation removes all the thermal effects in the model and it is only valid in the region of low fields where the velocity is linearly-dependent with the electric field. However, the validity of the DD equations is empirically extended by introducing field-dependent mobility models enabling the use of higher electric fields in the simulation process. Despite this extension, this model can only be valid in quasi-equilibrium regions, where the electric field varies softly and the velocity has a local dependence with the field. Nevertheless, the principal advantage of this model is the lower computational time as opposed to others simulation techniques.

3.1.2 Monte Carlo

An alternative method to the simulation of carrier transport, without the discretization of the BTE or its moments, is the Monte Carlo method [122–128].

This technique makes use of a sequence of free-flight interrupted by scattering events that change the momentum and the energy of the particles. The Monte Carlo approximation is widely used in the simulation of semiconductor devices [22–25, 129–134]. Nonetheless, the large amount of particles required to be simulated, the random number generation, and the coupling to the Poisson equation increase substantially the computational time. The Monte Carlo method is the simulation approach used in this work. For this reason, it will be deeply described in Section 3.2.

3.1.3 Quantum Description

The quantum approach describes the transport phenomena in solids with the most complete formulation. However, a full quantum model is highly demanding from a computational point of view. One technique that is gaining popularity in the quantum transport field is the Non-Equilibrium Green Functions (NEGF) formalism [135–137]. The mathematical method known as Green Functions is used in order to solve non-homogeneous boundary value problems. The Green function for a given energy has two input parameters that can be associated with two positions in the real space simulating different transistor areas. This function considers the influence of a perturbation of one input over the other one and, in theory, has the ability of modulate the properties of the system such as electronic density, current density, or density of states.

Full quantum models are very limited due to the higher computational cost as above-mentioned but quantum effects can be included in other approaches such as Drift-Diffusion or Monte Carlo thanks to quantum corrections reducing the simulation time [138]. These corrections allows the inclusion quantum confinement and tunnel effects, which are indispensable as the dimensions approach to nanometric scales. Density Gradient [139] and Effective Potential [140] methods are the more common models in order to include quantum corrections in classic simulations.

The Density Gradient approximation introduces a quantum potential as a additional parameter in the drift expression of the current density. This potential is proportional to the second derivative of the carrier density. It reduces both the electron variation as opposed to the classic potential, and its concentration near the interface.

The Effective Potential technique represents carriers such as Gaussian wave packets with minimum dispersion. This potential and the classic potential are proportional to a convolution integral which represents the quantum-mechanic effect. As for the Density Gradient model, this approach moderates the electron variation and moves the carriers away from the interface.

3.2 Principles of Monte Carlo Method

The Monte Carlo method is a numeric technique that solves the Boltzmann Transport equation (BTE) using a semiclassical transport theory. This technique makes use of a variety of stochastic techniques which use uniform random number sequences in order to solve statistically the Boltzmann equation, without making assumptions about the distribution function. It calculates classical trajectories, known as free-flights, which simulate the motion of particles inside a semiconductor. Then, these flights are interrupted stochastically by scattering events. The scattering rates are calculated according to quantum mechanical rules including electron-phonon, electron-impurity and electron-electron interactions. The motion of carriers is coupled to the Poisson equation updating the force that leads them.

The main magnitude of the classic transport theory is the distribution function $f(\vec{r}, \vec{k}, t)$ which is defined as the probability of finding an electron located in $\vec{r}$ with momentum $\vec{k}$ at an specific time t. Thus, the total number of electrons in the system can be calculated as:

$$\frac{2}{(2\pi)^3} \int d\vec{k} \int d\vec{r} f(\vec{r}, \vec{k}, t) = N, \tag{3.7}$$

where N is the total number of electrons.

The Boltzmann Transport equation characterizes the semiclassical transport and can be obtained from the Liouville's theorem. According to this theorem, the laws of physics using a semiclassical approximation do not provide any tendency for systems initially in different states to accumulate in certain final states in preference to others [141, 142]. A fundamental way to get BTE is based on counting the number of moving particles that go through a small volume in the phase space per unit time. It is given by:

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla_{\vec{r}} f + \vec{k} \cdot \nabla_{\vec{k}} f = \frac{\partial f}{\partial t}|_{coll}, \tag{3.8}$$

where the right hand side of the equation describes the variation of the distribution function because of the collisions, and $\vec{v}$ is the group velocity (below described).

Once the distribution function is known, other magnitudes of interest might be calculated such as the electron drift velocity, the average energy, the diffusion coefficient, the total fields, the grid temperature, or the electron concentration gradient.

In general, if an external electromagnetic field is applied to a semiconductor, the electrons are going to change their momentum $\vec{k}$ according to:

$$\hbar \frac{d\vec{k}}{dt} = -q \left(\vec{\varepsilon} + \frac{1}{c} \vec{v} \times \vec{B} \right), \tag{3.9}$$

where $-q$ is the electric charge, $\vec{\varepsilon}$ and $\vec{B}$ are the electric and magnetic fields, respectively, and

$$\vec{v} = \frac{1}{\hbar} \frac{\partial E(\vec{k})}{\partial \vec{k}} \tag{3.10}$$

is the electron group velocity.

Analyzing equations 3.8, 3.9, and 3.10, it is clear that the band structure of a specific material $E(\vec{k})$ is the required parameter in order to describe any transport phenomenon.

As it is known from Statistical Mechanics, the distribution function in equilibrium state f_0 is the Fermi-Dirac distribution function for a fermion gas. Accordingly, f_0 can be replaced with Maxwell-Boltzmann distribution if the condition $E - E_F \gg k_B T$ is fulfilled, being E_F the Fermi level, and C the normalization constant:

$$f_{MB} = C \exp\left(\frac{-(E - E_F)}{k_B T}\right). \tag{3.11}$$

The collision term in the right part of the equation 3.8 is generated by the imperfections in the ideal crystal lattice, such as phonons, surface roughness scattering, Coulomb scatterings, or, in general, the scattering events included. They can therefore induce transitions to different Bloch states, which are different final states for a particle in a periodically-repeating environment (the crystal). If we define $P(\vec{k}, \vec{k'})$ as the transition probability per time unit from a state $\vec{k}$ to state $\vec{k'}$, taking into account a constant spin without any change caused by the transitions, the collision term can be calculated as the difference of the electrons that undergo scattering between the $\vec{k}$ and the $\vec{k'}$ states.

$$\frac{\partial f}{\partial t}|_{coll} = \frac{V}{(2\pi)^3} \int \left[f(\vec{k'})P(\vec{k'}, \vec{k}) \left(1 - f(\vec{k})\right) - f(\vec{k})P(\vec{k}, \vec{k'}) \left(1 - f(\vec{k'})\right)\right] d\vec{k'}, \tag{3.12}$$

where V is the volume, and $f(\vec{k})$ (or $\left(1 - f(\vec{k})\right)$) the probability of the initial (or final) state to be occupied (or empty), respectively. However, due to the approximation considered in equation 3.11, these coefficients do not contribute and hence we are going to assume $f(\vec{k}) \longrightarrow 1$ or $\left(1 - f(\vec{k})\right) \longrightarrow 0$ from that point on. If equation 3.12 is included in the BTE (equation 3.8), an integro-differential equation is obtained. The complexity of this equation depends on the involved scattering mechanisms and the band structure.

Finding a solution of the BTE is not a simple task. Despite bearing in mind the lineal response and simple scattering mechanisms, it is necessary to include some approximations in order to solve it. From an analytical point of view, transport phenomena in a non-linear regime should be totally described by the BTE, but the resolution of this equation is a very hard-solved mathematic problem. Nonetheless, the analytical techniques applied in simple models for semiconductors can give us a physical interpretation of the non-linear transport problem, introducing simple concepts such as the energy and momentum relaxation times, or the electron temperature.

Undoubtedly, the introduction of numeric approaches has improved the solution of the BTE such as the iterative technique [143] and the more popular Monte Carlo method. The iterative technique solves the BTE using an iterative procedure calculating the distribution function in every step of the process. For this reason, it is a very useful technique when the physical phenomena strongly depends on the distribution function. Nevertheless, the Monte Carlo method is more directly interpretable from the physical point of view. In the case of charge transport in semiconductors, the statistical numerical approach to achieve the solution of the BTE proves to be a direct description of the dynamics of charge carriers inside the crystal. In addition, any required physical information can be easily extracted while the solution of the equations is being built up.

Monte Carlo technique consists of the simulation of one or more carriers inside the crystal subject to the action of electric and magnetic fields and scattering mechanisms. This method considers a semiclassical transport as depicted in Figure 3.3 because the motion of the particles during the free-flight (the time between two successive collisions) is totally classical, whereas the quantum theory describes the scattering rates. The duration of the free-flight and the scattering events are selected stochastically in accordance with some probabilities that describe the microscopic process. As a result, this method depends on the generation of random number sequences.

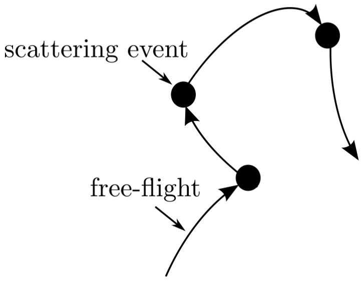


Figure 3.3: Motion of a sample particle as a sequence of free-flights and scattering events.

Figure 3.4 summarizes the structure of a typical Monte Carlo algorithm suited to the simulation of a stationary and homogeneous transport process. This kind of transport process is known as Single Particle Monte Carlo (SPMC) method because the motion of a single electron is considered. If a sufficiently long path of this sample is considered, we can assume thanks to the ergodicity principle that it provides a comprehensive description of the behavior of the entire electron gas.

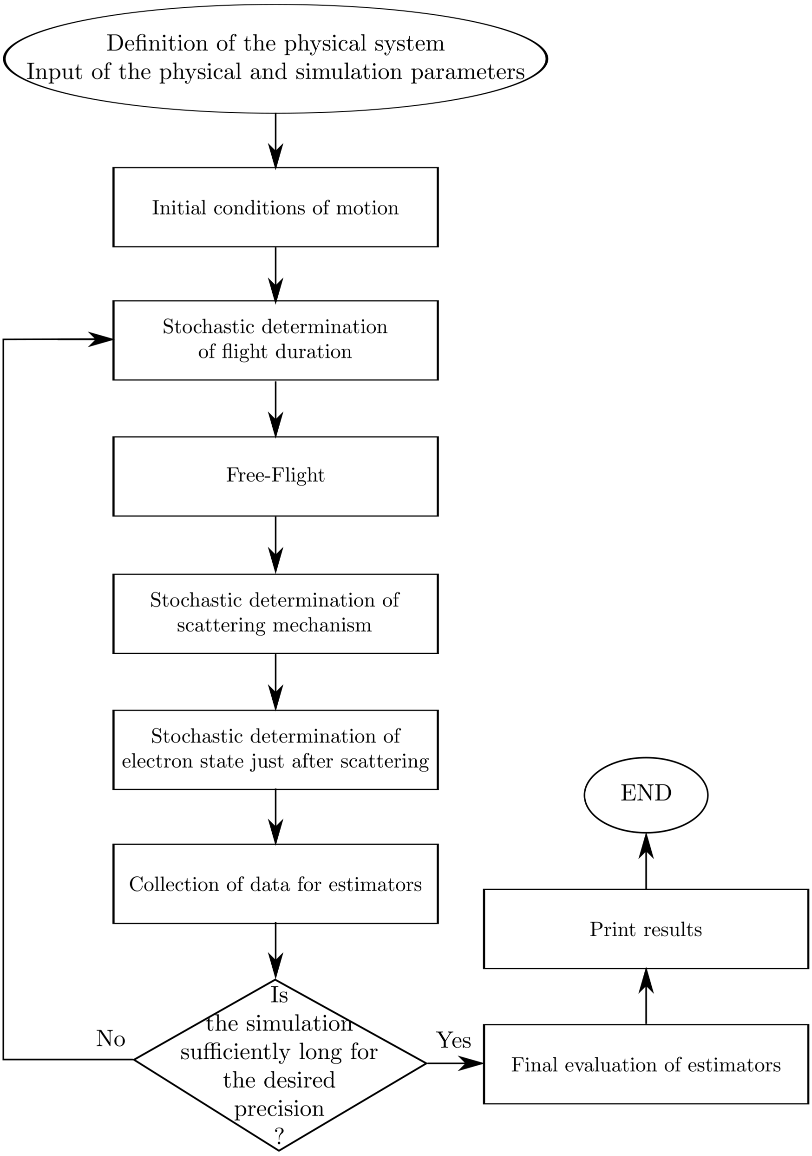

Definition of the physical system
Input of the physical and simulation parameters
Initial conditions of motion
Stochastic determination of flight duration
Free-Flight
Stochastic determination of scattering mechanism
Stochastic determination of electron state just after scattering
Collection of data for estimators
Is the simulation sufficiently long for the desired precision ?
No
Yes
Final evaluation of estimators
Print results
END

Figure 3.4: Flowchart of a typical Single Particle Monte Carlo algorithm.

However, the simulator used in this thesis is the so-called Ensemble Monte

Carlo (EMC) that bears in mind the situation when the transport process under investigation is not homogeneous or not stationary. For the sake of simplicity, the details of the main steps of the SPMC procedure are firstly given in the following subsections and then the EMC is deeply explained in Section 4.1.

Broadly, the starting point of this method is the definition of the physical system where all the inputs and simulation parameters are included. Among all of them we can mention those describing the material, the defects inside the crystal and its lattice, the temperature of the system, the electric field, the total time of simulation (T), or any other simulation details. Thereafter, all the quantities must be initialized and the scattering probabilities must be calculated as a function of the electron energy.

The simulation then starts by analyzing one electron with a wave vector $\vec{k}_0$ that determines the initial conditions of motion. Furthermore, the duration of the first free-flight is stochastically chosen according to the distribution probability, which is calculated by the scattering probabilities. In the interest of simplicity, external electric field ($\vec{\varepsilon}$) is only keeping in mind and, during the free-flight, $\vec{\varepsilon}$ is made to act according to the relation:

$$\hbar \frac{d\vec{k}}{dt} = q\vec{\varepsilon}, \tag{3.13}$$

where $\vec{k}$ is the carrier wave vector, q is the charge ($-q$ for electrons and q for holes), and $\hbar$ is the reduced Planck constant. After that, all quantities of interest, such as velocity, and energy are collected. Subsequently, the end of the free-flight is established by the choice of a scattering mechanism consistent with the distribution probability of all possible scattering mechanisms. Since the selection of this mechanism, a new $\vec{k}$ is randomly chosen as the initial state of the next free-flight. The entire process is iteratively repeated until the magnitudes of interest achieve the targeted error.

Finally, it is worth estimating the statistical uncertainty in order to get the desired precision of the whole simulation (T). Firstly, all the quantities of interest are determined in (N) sub-histories of equal time duration (N/T). These sub-histories must be long enough to be independent one of each other, but short enough to have an appropriate number N of them. Secondly, the average value of each parameter and its standard deviation are calculated to verify the uncertainly. If the simulation is long enough, the magnitudes of the last sub-histories tend to be similar to the average due to the reduction of the statistical nature of their fluctuations [144]. The duration of each sub-history as well as the desired precision are some of the inputs of the system because they control both the total simulation time and the influence of the initial conditions as mentioned in the next subsection.

3.2.1 Initial Conditions

Initially, if we consider the simulation of a stationary state, the total simulation time must be sufficiently long that the initial conditions of the electron have not influence in the final result. The choice of the total simulation time must be a balance between the ergodicity principle ($t \to \infty$) and the requirement of a reduction in the computational time. For instance, when a very low probable initial value for the electron wave vector $\vec{k}$ is chosen, the first iterations will be under the influence of the unappropriated initial decision. Other example is when a very high electric field is applied and the initial energy of the electron is around $k_{\mathrm{B}}T$ (being k_{B} the Boltzmann constant). This energy will be lower than the average energy in stationary state and hence the initial energy will start increasing in every time step until reaching the appropriate value. As a result, the mobility, which shows the electron response to the field, will be much higher than for the steady-state conditions in the first part of the simulation.

If the total simulation time is long enough, the influence of the initial conditions will be negligible in the averaged results. Nonetheless, the elimination of the first iterations of the simulation from the average results may be advantageous minimizing the impact of an unappropriated initial decision. For this reason, if the total simulation is divided into sub-histories and the final of each one is considered as the initial state of the next one, the results can have better convergence. Accordingly, the initial conditions of the first sub-histories only influence in the results.

3.2.2 Free Flight. Self-scattering

The wave vector $\vec{k}$ constantly changes during the free-flight because of the applied electric field according to Equation 3.13. If $P\left[\vec{k}(t)\right] dt$ is the probability that an electron in the state $\vec{k}$ collides during the time interval $(t, t+dt)$, the probability that an electron suffers a collision in $t=0$ and does not suffer another one a time t is [124, 128]:

$$\exp\left[-\int_0^t P\left[\vec{k}(t')\right] dt'\right] \tag{3.14}$$

Consequently, the probability $\mathcal{P}(t)$ that the electron undergoes a new scattering event during a time interval dt around the time t is calculated as:

$$\mathcal{P}(t)dt = P\left[\vec{k}(t)\right] \exp\left[-\int_0^t P\left[\vec{k}(t')\right] dt'\right] dt \tag{3.15}$$

Due to the exponential in the Equation 3.15, it is not practical to generate stochastic free-flights using this expression because it would be necessary to evaluate an integral equation for each scattering event. A simple solution to face this problem is considering that $\Gamma \equiv \tau_0^{-1}$ is the maximum value of $P\left[\vec{k}(t)\right]$ in the region of interest. A new fictitious concept, called self-scattering, is introduced, and the total scattering probability is then constant and equal to Γ. If an electron undergoes a self-scattering, its state $\vec{k'}$ after the collision is the same state $\vec{k}$ that before the event. The electron state is therefore the same, as if it has not undergone any collision.

In this method, $P(\vec{k}) = \tau_0^{-1}$ is constant and the equation 3.15 is simplified:

$$\mathcal{P}(t) = \frac{1}{\tau_0} \exp\left[-\frac{t}{\tau_0}\right] \tag{3.16}$$

The mathematical approximation called direct technique [144, 145] is employed in order to get the free-flight duration according to the Equation 3.16. This technique considers a continuous distribution function $f(x)$ normalized to 1 in the interval (a, b), being $F(x)$ function the integral of $f(x)$. Then, given a random number r uniformly distributed in the interval $[0, 1]$, we correspondingly choose x_r such that

$$r = F(x_r) = \frac{\int_a^{x_r} f(x)dx}{\int_a^b f(x)dx} \tag{3.17}$$

The probability $P(x)dx$ that the obtained value x_r is located within an interval dx around x is equal to dF, since r is a flat distribution.

$$P(x)dx = dF = f(x)dx \tag{3.18}$$

If we apply the aforementioned direct technique to the probability distribution in Equation 3.16, r can be calculated as:

$$r = \int_0^{t_r} \frac{1}{\tau_0} \exp\left[-\frac{t}{\tau_0}\right] dt \tag{3.19}$$

From Equation 3.19, the free flight length can be calculated as:

$$t_r = -\tau_0 ln(1 - r) \tag{3.20}$$

However, since r is a uniform distributed random number between $[0, 1]$, so also is $(1 - r)$ modifying the equation 3.19 to:

$$t_r = -\tau_0 ln(r) \tag{3.21}$$

The computational time consumed by the self-scattering events is, in general, more than compensated by the simplification of the free-flight duration calculation.

Regarding the choice of the constant Γ, it is worth bearing in mind that $P(\vec{k})$ depends on the electron energy E. An accurate choice of Γ can be to consider the maximum value of $P(E)$ in the range of energy that electrons can achieve during the simulation. Nevertheless, this rage of energies is unknown at the beginning of the simulation, when Γ must be chosen. In consequence, it is necessary to estimate beforehand E_{max} trying to avoid a very high value because it can result in a very large Γ and hence a long computation time due to a high number of self-scattering events.

Conversely, it is essential to decide what to do during the simulation if the electron reaches an energy higher than E_{max}. In addition, if the semiconductor model contains several valleys, different values of E_{max} must be provided for each one according to its characteristics.

The method of the Γ choice herein described is one of the possible methods. Broadly speaking, the choice of a specific method depends on the type of simulation and the computational resources.

3.2.3 Scattering Process

During the free-flight, the dynamic of the electron is governed by Equation 3.13. At the end of the electron motion, it is possible to evaluate its wave vector and energy as well as the scattering probabilities $P_i(E)$, being i the ith scattering mechanism. The self-scattering probability can be calculated as the complement to Γ of the sum of $P_i(E)$.

$$\Gamma = \sum_i P_i \tag{3.22}$$

In that moment, a scattering event must be chosen among all the possible mechanisms, as depicted in Figure 3.5. First of all, the successive sums of the P_i are considered

$$P_1, P_1 + P_2, \ldots, P_1 + P_2 + \ldots + P_j, \ldots, \Gamma \tag{3.23}$$

and a uniform distributed random number r is generated. Then, the product $r\Gamma$ is compared with the successive sums making a choice of the jth event when the first of the partial sums $P_1 + P_2 + \ldots + P_j$ is higher than $r\Gamma$.

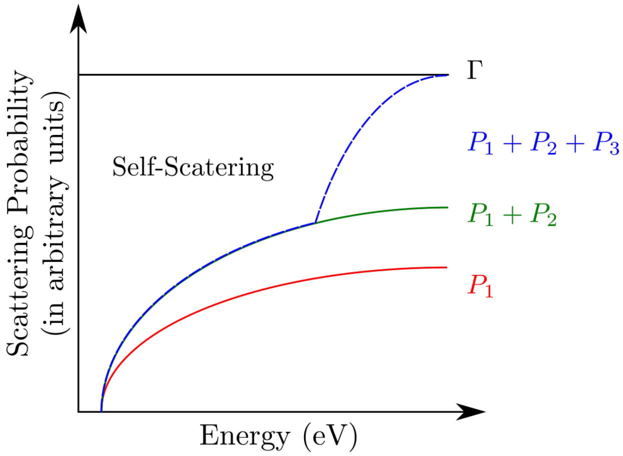


Figure 3.5: Representation of the scattering mechanism choice taking into account three different events, self-scattering and the energy range under consideration.

Finally, if none of all scattering events has been selected, it means that $r\Gamma > P(E)$, and a self-scattering occurs. This event is therefore the most time consuming because all the P_i must be calculated even although we do not use any of them at the end. However, we can avoid this time wasting or time consumption using the fast self-scattering [125]. It consists in setting up a matrix of the scattering probabilities at the beginning of the simulation for each energy range under consideration. In this sense, if the electron energy at the end of the free-flight is inside the nth interval (P_n), the first partial sum to be compared to $r\Gamma$ is P_n. Consequently, if $r\Gamma > P_n$, a self-scattering is chosen, otherwise all P_i must be evaluated. Thanks to this, both the choice of self-scattering event and the calculation of all the P_i only occur when $P(E) < r\Gamma < P_n$.

Moreover, it is necessary to keep in mind that some scattering mechanisms can depend on the distribution function, which is calculated at the end of the simulation and hence it is unknown when the program is configured (at the beginning). For these events, a self-consistent simulation must be performed such as for EMC (Section 4.1). For example, the scattering probability associated with the electron-electron interaction is dependent on the distribution function of electrons. However, the carrier distribution function is one of the results of a SPMC simulation. Accordingly, we have to evaluate this output in order to calculate the scattering probability that is a required parameter to continue the simulation.

Once the scattering mechanism which stops the free-flight of an electron is determined, the new state $\vec{k}_b$ after the scattering must be elected as the final state of the scattering event. If the end of the free-flight is caused by self-scattering, $\vec{k}_b$ must be the same state than before scattering $\vec{k}_a$. The other way around, $\vec{k}_b$ must be calculated randomly, according to the differential cross section of the selected event.

3.2.4 Scattering mechanisms

The electronic transitions of interest for the carrier transport in semiconductors can be classified as intra-valley or inter-valley when the initial and final states belong to the same or different valley of the conduction band, respectively. The most important scattering sources that determine these transitions are: phonons, impurities, and scattering with other carriers.

The interaction between phonons and carriers is a consequence of the crystal deformations obtained from the deformation mechanisms of the potential or the electrostatic forces introduced by the polarization waves of the phonons. The first interaction is very typical in covalent semiconductors, whereas the second one is characteristic of polar materials and compound semiconductor devices. Furthermore, when phonons enable inter-valley transitions, they can be defined as inter-valley phonons. If the initial and final states remain in the same valley, phonons are named as intra-valley phonons.

If the impurities are taken into account, they can be classified as neutral or ionized depending on their interaction strength. If it is weak and the impurity range is short, they are neutral impurities, whereas the ionized ones are Columbian. In general, the Coulomb scattering produces only intra-valley transitions because the effective cross section of this mechanism decreases when the momentum difference between the initial and final state $\Delta\vec{k}$ increases. In intra-valley transitions, $\Delta\vec{k}$ is very large, and so the probability to undergo a inter-valley scattering.

In addition to the scattering mechanisms above-mentioned, other events can be considered as an important phenomenon for specific situations. Few examples are: scattering caused by the roughness in the interface between two surfaces (surface roughness scattering), scattering events produced by composition fluctuations of the crystal potential (alloy scattering), carrier-carrier scattering, or processes in which the carriers are generated or destroyed such as impact ionization and generation-recombination events.

The scattering theory is generally based on Fermi's golden rule, which is derived from the time-dependent perturbation theory of the first order [128,146, 147]. It gives the transition probability between two eigenstates, which are the solution of the Schrödinger equation for the unperturbed Hamiltonian H_0. The scattering rate ($W\left(\vec{k}\right)$) of an electron with wave vector $\vec{k}$ is calculated by:

$$W\left(\vec{k}\right) = \frac{\Omega}{(2\pi)^3} \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} P\left(\vec{k}, \vec{k'}\right) dk' d\theta d\phi \tag{3.24}$$

where $P\left(\vec{k}, \vec{k'}\right)$ is the probability of the electron with initial state $\vec{k}$ undergoes a scattering event and reaches the final state $\vec{k'}$. ϕ is the azimuthal angle, θ is the polar angle, and Ω is the volume of the crystal.

Since the semiconductor crystal is treated as a continuum in the effective mass approximation, $W\left(\vec{k}\right)$ is assumed to be independent of ϕ and it can be calculated randomly between $[0-2\pi]$. In addition, because of the energy and momentum conservation, θ and $\vec{k'}$ are not independent of each other. In accordance, the triple integral in the above equation can be reduced to a single one.

For simplicity, phonon, Coulomb, and surface roughness scattering are only described in this subsection because they are the scattering mechanisms included in the Monte Carlo simulator used this thesis.

3.2.4.1 Phonon scattering

Due to the thermal energy in the crystal lattice at any temperature, electrons are scattered by ion vibration propagating in the crystal lattice with respect to their nominal position. Moreover, such vibrations of the lattice produce a perturbation of the potential. The quantum process, which is called the electron-phonon interaction, therefore determines the influence of these lattice vibrations on electron dynamics. The term phonons quantizes the wave nature of the lattice vibrations [148,149]. Each possible state is characterized by a couple $(\vec{q}, j)$ which is occupied by $n_j\left(\vec{q}\right)$ number of phonons with $\hbar\omega_j\left(\vec{q}\right)$ energy, being $\vec{q}$ the wave-vector variation produced by the scattering $\vec{q} = \vec{k'} - \vec{k}$. The vibration energy can be calculated as:

$$E = \sum_{j,\vec{q}} \hbar\omega_j(\vec{q}) \left[n_j(\vec{q}) + \frac{1}{2} \right], \tag{3.25}$$

fulfilling $n_j(\vec{q})$ the Bose-Einstein statistics

$$n_j(\vec{q}) \frac{1}{\exp\left(\frac{\hbar\omega_j(\vec{q})}{k_\mathrm{B}T}\right) - 1}. \tag{3.26}$$

Diatomic crystal, such as silicon, has six different vibration branches. Three of them are acoustic mode phonons, in which the neighboring atoms displace in the same direction, and hence the changes in lattice spacing are produced by the strain or differential displacement. The other three are optical phonons, in which neighboring atoms displace in opposite directions, and so the displacement produces the changes in lattice spacing directly.

The interaction between the electrons and the lattice ions can be discomposed in two contributions: the interaction between electron and the ions in their equilibrium state, which describes the periodic potential interaction and determines the electron band structure in the crystal; and the lattice vibrations, which correspond to the electron-phonon interactions.

The interaction between an electron (e) and an ion can be calculated using the first order approximation assuming short displacement (α) of the lattice ions (n) about their equilibrium position $(\vec{s}_{n,\alpha})$:

$$V_\alpha\left(\vec{r}_e - \vec{R}_{n,\alpha} - \vec{s}_{n,\alpha}\right) = V_\alpha\left(\vec{r}_e - \vec{R}_{n,\alpha}\right) - \vec{s}_{n,\alpha} \cdot \nabla V_\alpha\left(\vec{r}_e - \vec{R}_{n,\alpha}\right) \tag{3.27}$$

The first term of the right part in the equation 3.27 summed over (n, α, e) is the interaction to the ions in equilibrium and therefore the second term corresponds to the electron-phonon interaction. The corresponding Hamiltonian can then be expressed as:

$$H_{ep} = -\sum_{n,\alpha,e} \vec{s}_{n,\alpha} \cdot \nabla V_\alpha \left(\vec{r}_e - \vec{R}_{n,\alpha}\right) \tag{3.28}$$

Two different electron-phonon interaction can be distinguished: emission, when a phonon with wave vector $-q$ is created; and absorption, when a phonon with wave vector q disappears. Both interactions translate the electron from a state $\vec{k}$ to another $\vec{k} \pm q$. For this reason, the problem is limited to the interaction potential calculation $\nabla V_\alpha \left(\vec{r}_e - \vec{R}_{n,\alpha}\right)$.

- Intra-valley acoustic transitions:
  These transitions are of special interest for the interaction between electrons and acoustic phonons. They are characterized because the atoms in the unit cell move in the same direction and hence an effective displacement of the total cell is produced. When the wavelength is big enough, the vibration amplitude changes not much from one unit cell to another. As a result, the atomic structure is of little importance and the deformation potential can be used in order to consider the continuum [149]. Subsequently, the electron-phonon interaction potential can be written as:

$$H_{ep} = \vec{\epsilon} \nabla \cdot \vec{s} \tag{3.29}$$

where $\vec{\epsilon}$ is a tensor which defines the band shift per deformation unit. In this continuous approximation, the displacement can be characterized as a function of the parameters that describe the creation and the destruction of phonons, $a_q^\dagger$ and a_q, respectively. The scattering probability per unit time for emission and absorption process can be calculated putting into practice the development used in [125] and bearing in mind the elastic case and the approximation of the energy equipartition with non-parabolic bands:

$$\Gamma_{ac}(E) = \frac{\sqrt{2} m_d^{3/2} k_{\mathrm{B}} T_0 \epsilon_l^2}{\pi \hbar^4 u^2 \rho} E^{1/2} \left(1 + 2\alpha E\right) \left(1 + \alpha E\right)^2 \tag{3.30}$$

where u is the sound velocity in the material, α the non-parabolic parameter, ϵ_l^2 is the average value of the deformation tensor along the $\vec{q}$ direction, and ρ is the density.

- Inter-valley transitions:

In these transitions, the initial and final states are located in different valleys and the involved phonons have high momentum. The wavelength of the involved phonon in this type of transitions is similar to the distance between the minimum of the transition valleys, thus $\nabla\vec{k}$ is practically constant and, for a particular branch, so also the energy $\hbar\omega_i$. Accordingly, these transitions can be handled as the intra-valley acoustic ones keeping in mind that the deformation potential and the involved phonon energy must be independent of $\vec{q}$. In consequence, following the same development used in [125], the scattering probability for zero order and the non-parabolic approximation can be determined for the expression:

$$\Gamma_{e,i}(E) = A \left(\begin{array}{c} N_i \\ N_{i+1} \end{array} \right) (E + \Xi_{\mp})^{1/2} \left[1 + 2\alpha \left(E \pm \Xi_{\mp}\right)\right] \left[1 + \alpha \left(E \pm \Xi_{\mp}\right)\right]^{1/2} \tag{3.31}$$

where:

$$A = \frac{m_d^{3/2} \left(D_t K\right)_i^2 \gamma}{\sqrt{2}\pi\hbar^3\omega_i\rho}, \tag{3.32}$$

$$\Xi_{\mp} = \hbar\omega_i \mp \nabla E, \tag{3.33}$$

γ is the number of possibles final valleys that are equivalent in the transition, and the upper and lower parameters are for absorption and emission process, respectively.

As a practical example, if the position of the conduction band valleys in silicon are considered, two different inter-valley transitions should be happen among the six equivalent ∇ minimum of silicon (Figure 3.6):

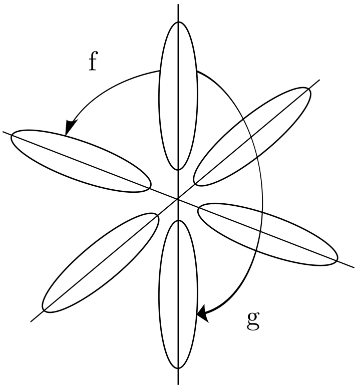


Figure 3.6: G-type and F-type phonon assisted transitions involving the ∇ valleys of silicon.

- g-type processes: the transitions are made along the same direction but opposite directions, such as from $< 100 >$ to $< \bar{1}00 >$.
- f-type processes: the transitions happen among ∇ valleys along different directions, such as from $< 100 >$ to $< 010 >$.

Both the g-type and the f-type are umklapp processes because they transform, by a reflection or a translation, the wave vector to another Brillouin zone as a result of a scattering process. They also involve a reciprocal lattice vector.

The geometrical considerations about the Brillouin zone prove that the phonon mode involved in a f-type process is equal to the distance between Γ and X point in the Brillouin zone of silicon, forming an angle of 11° with the ΓX direction. For the phonon involved in a g-type process, its module is $0.3\Gamma X$ in the ΓX direction. For these processes, the multiplicity of the final valley is 4 for f-type and 1 for g-type.

It can be demonstrated thanks to an analysis of the Group Theory with zero order of interaction that g-type processes can be only assisted by optic longitudinal phonons (LO), whereas the phonon implicated in f-type events can be acoustic longitudinal (LA) and optic transversal (TO). Following this reasoning, the rest of phonons are forbidden. However, some other references [125] suggest that the selection principles must not been fulfilled because the initial and final states do not coincide with the high symmetry points. It is therefore very reasonable consider that these states, which are forbidden when the initial and final states are in high symmetry points, are very limited in comparison to the afforded ones. The effect of the forbidden phonons has been observed experimentally and hence the absorption and emission probabilities can be calculated by means of the first order interaction:

$$\Gamma_{e,i}(E) = A' \begin{pmatrix} N_i \\ N_{i+1} \end{pmatrix} \left[E\left(1+E\alpha\right) + \left(E \pm \Xi_{\mp}\right)\left(1+\alpha\left(E \pm \Xi_{\mp}\right)\right)\right] \\ \left(1+2\alpha\left(E \pm \Xi_{\mp}\right)\right)\left[\left(E \pm \Xi_{\mp}\right)\left(1+\alpha\left(E \pm \Xi_{\mp}\right)\right)\right]^{1/2} \quad (3.34)$$

where:

$$A' = \frac{\sqrt{2} m_d^{5/2} D_1^2 \gamma}{\pi \hbar 4 \omega_i \rho}, \quad (3.35)$$

3.2.4.2 Coulomb scattering

Coulomb centers produce a perturbation potential which can effectively scatter the carriers in a semiconductor device. In addition, the various type of charge in an electronic device can modify its carrier motion diversely. The different origin of these charges is briefly discussed below where the external charges are summarized up:

- Trapped charges at the silicon-oxide interface (Q_{it}): they are due to defects at the $Si - SiO_2$ interface and they generate accessible states in the band gap. These defects can interact with the conduction and valence band emitting or capturing electrons and holes, respectively. The trap density depends on the silicon layer orientation, the thermionic and electric stress which it could be subject to, and the radiation which could receive the interface. All these effects can produce not only the modification of the trap occupation, but also the appearance of new defects and traps degrading the structure electric properties.

- Fixed charges at the silicon-oxide interface (Q_f): These charges, which have positive sign, are located in a narrow layer ($10-20\mathring{A}$) of non-stoichiometric SiO_2 keeping its location fixed regardless of the gate bias. The quantity is dependent on of the oxidation and annealing conditions. Their concentration has to be below $10^{10}cm^{-2}$ in <100> surfaces in order to consider a good quality interface.

- Trapped charges in the gate oxide (Q_{ox}): These charges are associated to present defects in the oxide, such as impurities or broken bonds, and they are distributed all over the insulator volume. As in the previous ones, they are independent of the gate bias and they can be removed by low temperature annealing. The traps generated by any defect can be charged due to the charges injected in the oxide by means of both the hot carrier generation during the device operation or the electron-hole pair generation in the SiO_2 in x-ray processes. Furthermore, these trapped charges are involved in the gate leakage described in depth in section 5.4, in which the direct tunnel through the gate oxide and the trapping and detrapping tunneling are bearing in mind. A trap density for a good quality gate oxide is around $10^{12}cm^{-2}$ in <100> wafer orientation.

- Charged centers by mobile ions (Q_m): these centers are mainly created because of the presence of the alkaline metal ions, such as the sodium or the potassium. Due to its mobile nature, the electric field applied in the gate attracts them to the silicon-oxide interface.

- Ionized impurities (Q_B): the gate voltage induces a band bending and the channel impurities therefore are almost entirely polarized even for low temperatures. This charge significantly affects the carrier mobility.

The ionized impurities have been only taken into account in the parallelized simulator used in the development of this thesis, because they are the principal responsible of the Coulomb scattering due to its influence in the carrier transport study. If the trapped charges in the gate oxide are included, they must be contained in the resolution of the Poisson equation.

The description of the interaction between the electrons and the ionized impurities is based on the Brooks and Herring formalism [150], where the scattering probability per unit time using the non-parabolic approximation is:

$$\Gamma_C = \frac{\sqrt{2} N_I Z^2 q^4 m^{*3/2}}{\pi \varepsilon_{Si} \hbar^4} \frac{(1+\alpha E)^2 (1+2\alpha E)}{L_D^2 \left(4k^4 + L_D^2\right)} \tag{3.36}$$

where N_I is the ionized impurity density, Z is its charge, and L_D is the Debye length (being n_0 the concentration of free carrier)

$$L_D = \sqrt{\frac{q^2 n_0}{\varepsilon_{Si} k_\mathrm{B} T}}. \tag{3.37}$$

This scattering mechanism is anisotropic and hence the azimuthal angle ϕ is randomly chosen between $[0, 2\pi]$ whereas the polar angle θ, which is defined between $\vec{k}$ and $\vec{k'}$ directions, is given by:

$$\cos\theta_r = 1 - \frac{2(1-r)}{1 + 4r\frac{k^2}{L_D^2}} \tag{3.38}$$

where r is a random number between $[0, 1]$.

3.2.4.3 Surface roughness scattering

Interaction of carriers with the microscopic asperities of the semiconductor-oxide interface is one of the scattering in MOSFETs, specially when the device is biased at large gate voltages and densities. If the simulation does not consider quantum effects, the surface roughness scattering can be included as boundary conditions for the transport in the oxide interfaces [133, 151]. In this assumption, the interfaces behave as reflective surfaces where the electrons undergo a specular reflections when they reach the interface of the high mobility channel, whereas the electrons undergo a non-specular reflection if the surface roughness is very large. In the latter case, the non-specular reflection can be modeled on different approximations, such as diffusive reflection or backscattering, but keeping the wave vector module.

However, if quantum effects are considered, the diffusive reflection models become invalids because the maximum of the carrier distribution is located far away from the interfaces [152]. It is therefore necessary to include quantum corrections in order to simulate the surface roughness scattering.

There are two main aims to deal with in the study of this mechanism. Firstly, the interface between the semiconductor and the oxide must be properly modelled. Different studies show its experimental characterization and its consequently modeling thanks to statistic parameters, and finally prove that the interfaces can be generated based on pseudo-random exponential signals [153]. It is also interesting from a variability point of view, the effect of two different interfaces but with the same statistic parameter in the device performance [136]. As a result of that studio, there are similar variations in some important parameters, such as the threshold voltage or the drain current, when the performance of both devices is compared due to the fluctuation in the shape of the interface. Secondly, once the interface has been characterized, it is indispensable to develop a model that replicates the experimental measures and introduces the effects of the imperfect interfaces in the electrostatics and in the transport process.

The particular surface roughness model included in the simulator herein described is based on the method described in [154]. This quantum approximation takes into account a pseudo-2D electron gas where several subbands in each valley can be occupied and it allows a quantitative study of this mechanism.

Chapter 4

Multisubband Ensemble Monte Carlo and Techniques for statistical enhancement

As the channel thickness is reduced in Silicon-On-Insulator (SOI) devices, the carriers undergo a strong confinement in the channel due to the gate insulator and the buried oxide (BOX). This fact gets even more relevance for double gate devices owing to the second gate. When the device thickness is smaller than $50nm$, quantum effects start having a crucial role since there are some behaviors that can only be explained by Quantum Mechanics. For this reason, the study of these devices must be carefully carried out. When the quantum effects are considered, the electron distribution in the channel is seriously affected changing not only its maximum value, but also its distribution shape with respect to the classical simulations. The main effects are a shift in the point of the maximum concentration, which is located around $1nm$ from the interface inside the channel, and a decrease in the channel inversion charge. In order to include these effects, it is needed to self-consistently solve the equations that describe the electrostatic and transport problem.

Nonetheless, new techniques are needed to be developed to appropriately study the advanced architectures as the physical phenomena have more higher complexity level. Accordingly, some high-level models are described in this chapter in order to face this problem and deal with it successfully. In the first place, the synchronous-ensemble method is described in order to obtain an average quantity from the simulation of a steady-state phenomenon using a N number of electrons per particle in the Monte Carlo approach. Subsequently, the Multisubband Ensemble Monte Carlo (MS-EMC) simulator is described in detail where a multiple energy minima in the band structures (valleys) with different energy subbands are considered. In addition, this method analyzes arbitrary orientations in both the transport and confinement directions. Then, the characteristics of the parallel implementation included in this code are discussed in order to understand the necessity of a computational time reduction. The tool herein described has already demonstrated its capabilities studying different advanced nanodevices [22–25].

Once our tool is deeply illustrated, some techniques for statistical enhancement are described in order to develop a more realistic simulator. Firstly, the boundary conditions for superparticles that reach the source and drain terminals are characterized. The new superparticle must be injected in the device obeying the Fermi-Dirac distribution function. Secondly, the energy-dependent weight for superparticles are considered because the electrons tend to be at the less energy levels and so the weight per particle for higher energy range should be lower.

4.1 Ensemble Monte Carlo

As mentioned in Section 3.2, for steady-state and homogeneous transport conditions, it is sufficient to simulate the motion of one single electron (SPMC). In general, by means of the ergodicity principle, a sufficiently long trajectory of this sample electron will give us information on the behavior of the entire electron gas. Nonetheless, when a simulation is carried out to study a transient and/or a non-homogeneous phenomenon, it is necessary to simulate a large number of electrons separately and then follow the successive and simultaneous calculation of their dynamic motions during a small time increment Δt. This procedure is known as Ensemble Monte Carlo (EMC). One example of the necessity of including EMC is electron transport in small devices. Its solution depends on the spatial distribution of the magnitudes of interest, i.e. potential and carrier density [155, 156].

Each Monte Carlo superparticle is assigned a statistical weight which denotes the number of electrons per superparticle. Accordingly, it is considered as a group composed of identical particles because all of them possess identical physical quantity such as energy, velocity, wave vector, or position. Finally, the weight must be adequately adjusted to a suitable number of superparticles since the higher the number of superparticles, the longer the required computation time.

In order to analyze a transient simulation, we are going to consider the time-dependent homogeneous electron gas. One case of especial interest is the study of the dynamic response of the system when the applied field on the electron gas changes. If the number of simulated electrons is large enough, it is possible to calculate the distribution functions $f(\vec{k}, t)$ or $f(E, t)$ based on the histograms which can be obtained from the electron wave vectors and energies in periodic time steps. Broadly speaking, if we want to obtain the average value $< A >$ of the quantity A of interest, it can be estimated according to this sample ensemble as a function of time and it will be representative of the average from the entire gas:

$$< A >= \frac{1}{N} \sum_{i} A(t) \tag{4.1}$$

where N is the total number of particles in the physical system. Figure 4.1 shows the schematic representation of a EMC simulation of a transient system. Each horizontal line represents the trajectories of the N particles and each circle on the horizontal lines shows the specific time at which scattering occurs to this particle. The vertical lines represent the time intervals when the system is observed.

The duration of the transient response (Δt) is not known at the beginning of the simulation. Moreover, it is projected in the order of the largest characteristic time of the electron system because the scattering rates must be updated with the charge of the electron energy with time. This time generally depends on the applied field and the temperature. A possible initial guess of this time in a high-field simulations of the semiconductor transport can correspond to the energy relaxation time or the repopulation time of each subband.

Other simulation of interest in the study of electronic devices is when the system is dependent on the space (non-homogeneous phenomenon). The average value of the quantity of interest in this case can be calculated as the previous one

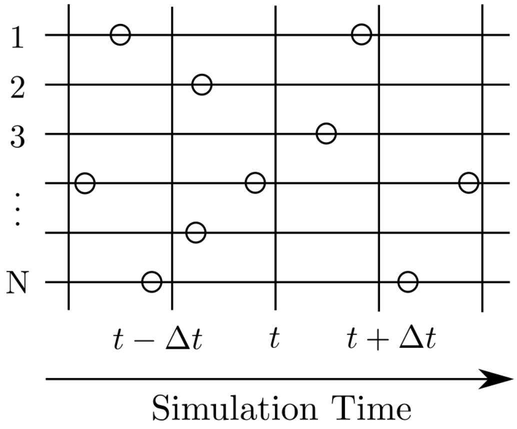


Figure 4.1: Schematic representation of a Ensemble Monte Carlo (EMC) simulation of a transient system where each horizontal line represents the trajectories of the N particles, each circle on the horizontal lines shows the time at which scattering occurs to this particle, and the vertical lines represent the time when the system is observed.

but the ensemble sample must be taken over particles at given positions instead of as a function of time.

Finally, the whole simulation is divided in sub-histories or sub-groups for transient or non-homogeneous systems, respectively, in order to estimate the precision as mentioned for the SPMC (section 3.2).

The simulator used in this thesis is based on the Ensemble Monte Carlo method, in particular it is defined as Multisubband Ensemble Monte Carlo. As a consequence, the main characteristics of this multisubband simulator are going to be deeply defined in the next Section as well as the mechanisms included to enhance its statistical behavior.

4.2 Multisubband Ensemble Monte Carlo

The 2D Multisubbdand Ensemble Monte Carlo simulator is based on the mode-space approximation of quantum transport [157], where the confinement and transport problems are decoupled. Accordingly, the Schrödinger equation is solved in the confinement direction, whereas the Boltzmann Transport Equation (BTE) is solved in a semiclassical way in the transport plane. Both equations are coupled to the Poisson equation to keep the self-consistency of the solution. The main drawback is the decoupling approximation itself because the insertion of coherent phenomena in a direct way in order to study some particular architectures is not possible. Nevertheless, the quantum transport effects can be included in a separate way [26–29]. For this reason, another advantage of the MS-EMC simulator is the ability to switch on and off the phenomena as they are included in a separate routine after each iteration.

At the level of simulation domain, the device structure is divided into slices along the confinement direction, which spreads from the source to the drain regions in the transport direction as shown in Figure 4.2. Then, the 1D Schrödinger equation is solved in each slice for all the valleys and subbands (jth) in the conduction band. From now on, despite different subbands are considered for different valleys, we are going to refer them as ”subbands” in general in order to simplify the discussion (no matter in which valleys they are contained). As a result of this solution, it acquires the gradual change in the transport direction of the energy level $E_j(x)$ and the eigenfunctions $\Psi_j(x,z)$ for each subband. The 2D BTE solves the transport problem in the corresponding plane. The self-consistency of the solution for these carriers is assumed by coupling these two equations to the 2D Poisson equation. An analytical and non-parabolic approximation of the conduction band is evaluated for both the transport and the confinement problem [127]. In particular, the non-parabolic coefficient is chosen as $\alpha = 0.5eV^{-1}$ [124].

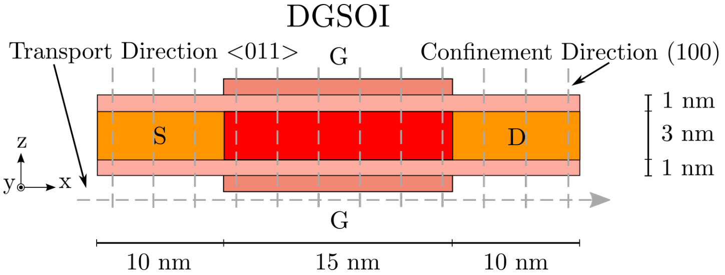


Figure 4.2: DGSOI structure analyzed in this chapter. 1D Schrödinger equation is solved for each grid point in the transport direction and BTE is solved by the MC method in the transport plane.

Figure 4.3 shows the flowchart of the MS-EMC simulator. The initial subband populations and scattering rates are calculated from the eigenfunctions which are the solution of the initial Schrödinger equation. In order to obtain the subband population for each grid point in the transport direction, it is necessary to make a spatial allocation of all the particles in the grid employing the cloud-in-cell method [158,159]. The subband population is compared with the density probability $|\Psi_j(x,z)|^2$ and so the electron concentration $(n(x,z))$ can be calculated:

$$n(x,z) = \sum_{j=1}^{jMax} n_j(x,z) \cdot |\Psi_{vs}(z)|^2, \tag{4.2}$$

where $jMax$ is the total number of subbands considered in all the valleys, and $n_j(x,z)$ is the relative population of each subband.

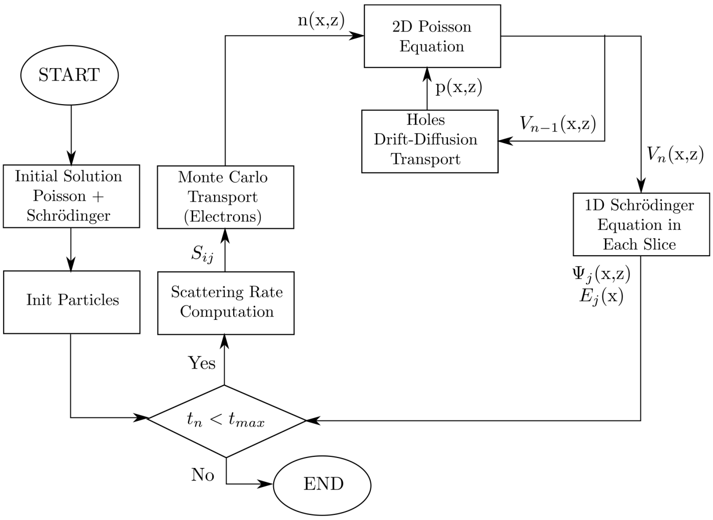

n(x,z)
2D Poisson Equation
START
p(x,z)
Holes Drift-Diffusion Transport
V_{n-1}(x,z)
V_n(x,z)
Initial Solution Poisson + Schrödinger
Monte Carlo Transport (Electrons)
1D Schrödinger Equation in Each Slice
S_{ij}
Ψ_j(x,z)
E_j(x)
Init Particles
Scattering Rate Computation
Yes
$t_n < t_{max}$
No
END

Figure 4.3: Flowchart of the MS-EMC simulator where x is the transport direction, z the confinement direction, $n(x,z)$ and $p(x,z)$ are the electron and hole concentrations, respectively, $V_n(x,z)$ is the potential profile, $E_j(x)$ is the subband energy, $\Psi_j(x,z)$ are the subband eigenfunctions, S_{ij} are the scattering rates, subscript n stands for the iteration number.

In accordance with the mode-space approximation, the drift field that the electrons undergo during the transport motion can be reckoned as the energy variation for each $s-th$ subband and $v-th$ valley in the corresponding direction and it therefore is different for each subband:

$$-q\vec{\varepsilon}_{s,v} = \frac{\partial E_{s,v}(x)}{\partial x}\hat{x}. \tag{4.3}$$

Once the Monte Carlo iterations have been started, it is essential to update the electrostatic potential $V_n(x,z)$ after the transport blocks including the new concentrations $n(x,z)$ and $p(x,z)$ in the Possion equation. Since the tool herein described has been designed focusing mainly on unipolar devices, electrons and holes are treated in a different manner. On the one hand, electrons are calculated using Monte Carlo approach and, hence, the electron concentration $n(x,z)$ is calculated after the electron flight. On the other hand, since holes travel mostly in the low field source region since they are mainly generated in this region, a drift diffusion approach is used to describe them using as an input the potential profile of the last iteration $V_{n-1}(x,z)$. Going into the detail, and following the utilization of the Scharfetter-Gummel discretization scheme [118], the hole concentration $p(x,z)$ comes from the continuity equation:

$$\frac{p_{m+1}-p_m}{d_m} - \operatorname{div}\left(D_p \cdot \nabla p_{m+1} + \mu_p \cdot p_{m+1} \cdot \nabla \Psi_m\right) + R\left(\Psi_m, n_m, p_m\right) = 0, \quad (4.4)$$

where d_m and p_m are the time step and the hole concentration of the m iteration, respectively, and $R\left(\Psi_m, n_m, p_m\right)$ is the net carrier generation/recombination term. The rest of the parameters has their usual meaning. J_p is calculated using the drift diffusion equation 3.6.

The main drawback of MS-EMC simulator is the high computational time because the scattering table must be updated after the solution of the Schrödinger equation in each iteration in order to keep the self-consistency. This updating is required due to changes of the energy levels and the eigenfunctions that modify the scattering probabilities along the whole device. The change in the subband distribution alters the allowed energy levels and the eigenfunctions variations modify the form factors in the scattering processes. Nonetheless, the main advantage of this tool against NEGF approach continues to be the reasonable computational time when scattering mechanisms and quantum effects on the ultrascaled devices are taken into account.

4.3 Parallel implementation of Multisubband Ensemble Monte Carlo

In general, the simulation in the engineering field needs higher computational resources. The average computational time in the Monte Carlo simulator herein developed is around one or two hours per 1ps of simulation time depending on the computer. Parallelization techniques help to face this issue decreasing the computational effort. The reduction of the simulation time has two main advantages in the nanodevice research: first, the possibility of performing parametric studies of the device properties in order to determine its performance because a huge number of simulations are needed and so a reasonable times are wanted; and second, the possibility of speeding up the elaboration of new models thanks to the reduced time in the test versions. Moreover, if we consider the strong advance that multi-core architectures have had in recent years, making available to the user processors of up to 32 cores or even more, it seems logical that parallel simulators are developed to make the most of the available computational resources.

The most common distributed system where the parallel code can be developed is the multicomputers, where each processor has a direct connection to its own local memory. Their main drawback is that the processors have to establish some explicit system of communication with each computer that allows the data exchange. Clusters are an example of this system, where the MS-EMC herein described is simulated. They are basically a collection of desktops or workstations connected through a high-performance network such as Gigabit Ethernet or Myranet. This interconnect network generally does not have special measures that ensure a high bandwidth, or failure protection.

In order to improve the computational time, the multi-core architectures take advantage of the parallelism in desktop computers without the need to use expensive infrastructures. The strong development experienced by this type of architecture in the last 10 years, from the first processors with two cores destined for servers, desktop and laptop computers, up to the present ones that have up to 8 cores per processor [160], has converted the multi-core in the most common processors presented in the high and mid-range equipment. Clusters with multi-core architecture therefore are the best alternative to the expensive MPPs (Massively

Parallel Processors). The typical structure of these processors is based on several cores which process the instructions and several levels of memory that can be dedicated or shared.

Currently, OpenMP is the parallel standard for shared-memory system programming [161–163] and, hence, this language was chosen to parallelize our MS-EMC tool in order to run it on multi-core processors. OpenMP is a collection of directives, libraries and environment variables for programs in Fortran, C and C++.

It is based on the fork-join parallel execution model. The program begins its running with a single thread, called master, which when the first parallel construction is found creates a set of n threads that together with the master distribute the work of execution. Once the parallel region is finished, only the master continues the execution (sequential running) and the threads are deleted.

This directive language is generally used for the parallelization of loops. The procedure is to search the most costly computational loops and then the threads distribute the loop iterations. During the execution of the parallel region, the threads are communicated through shared variables. One has to be very careful when programming because data sharing can lead to poor program behavior due to concurrent access from different threads. The way to avoid this is by using synchronization directives that have the disadvantage of being computationally costly, so we should try to avoid them. For this reason, the OpenMP programming consists basically of three elements: the control of parallelism, the control of the synchronization, and finally, the control of data and communications.

Figure 4.4 shows the flowchart of the MS-EMC simulator including the parallelized code where the most costly computational blocks have been run by different threads. The drawback of this technique is that transport simulation is a sequential process and so the results obtained during the calculation of one module are used in the next one. Additionally, some blocks are not parallelized and so they are executed only by the master node because some of them have dependence on their own iterations, such as disk access subroutines, or other blocks have a very low computational cost and it is not worth paralleling, such as the 2D Poisson equation and the Drift-Diffusion for holes module. Consequently, it is necessary to establish a synchronization element between the threads in order to prevent calls to the following subroutines before the last thread has left the

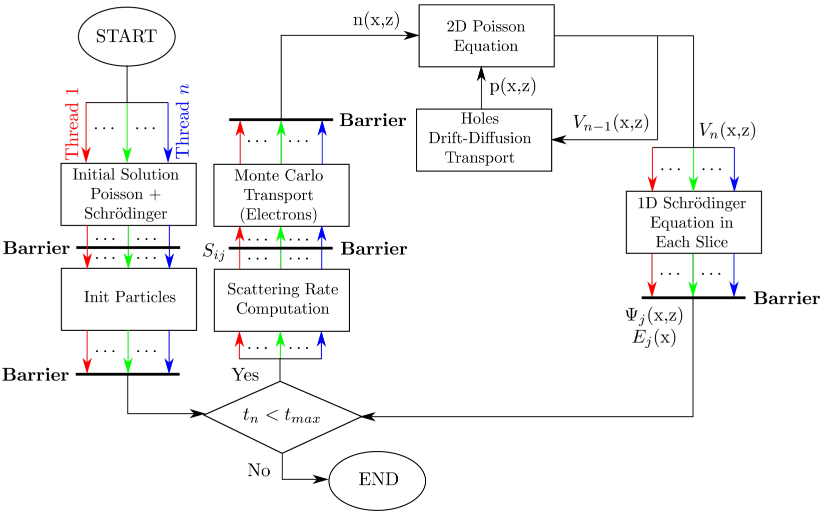


Figure 4.4: Flowchart of the MS-EMC simulator including the parallelized code.

previous one. This element is known as barrier.

In this tool, the solution of the Schrödinger equation and the scattering rate table calculation are the most costly computational block and they are completely parallelized. The solution of the Schrödinger equation in each slice can be calculated in different threads, likewise the scattering probabilities in each device location can be independently obtained after getting the wave function. The parallelization of both blocks is the main advantage of our MS-EMC simulator in a computational efficiency point of view.

Figure 4.5 represents the impact of the parallelization in the simulation time of the simulator 2D MS-EMC as a function of the number of cores for different steady state simulations of $1ps$ duration, with different values of variable *scsc* [35]. There are two main ideas in which the optimization can reduce the computational load of the total simulation time.

On the one hand, the variable *scsc* sets the number of iterations for which the solution of the Schrödinger equation and the update of the scattering table are calculated in a self-consistent way. For low values of *scsc*, such as 1 or 2, the self-consistency is very high and the computational load of the optimized modules is very significant. However, when the value of the *scsc* variable is very high, the computational load of the optimized blocks is greater and the effects of

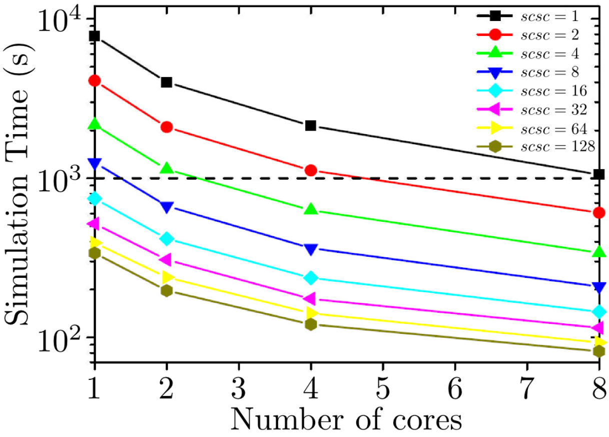


Figure 4.5: Impact of the parallelization in the simulation time of the simulator 2D MS-EMC as a function of the number of cores for different simulations of a steady state of $1ps$ duration, with different values of variable $scsc$. Figure extracted from [35]

the optimization are greater. A very high value of $scsc$ is not desirable because it implies a lower self-consistency which is one of the main characteristics of this simulator.

On the other hand, a sequential simulation for a value of $scsc = 1$ which execution time is approximately 8000 seconds can be reduced to about 1000 seconds when 8 cores are used. On the contrary, if the simulation time is required to be less than 1000 seconds in a sequential machine, the self-consistence would be decreased to values of $scsc > 8$.

These results show the advantages of using parallelism in the simulation design. In addition to the fact that the simulators are faster, they also save a considerable time to the code developers that implement new physical models in the simulator due to the reduction of the execution time in the test simulations.

4.4 Fermi-Dirac injection in source and drain

As already mentioned in Section 3.2, the Boltzmann Transport Equation (BTE) is an integro-differential equation which calculates the carrier distribution $f(\vec{r}, \vec{k}, t)$. The Monte Carlo method solves the BTE by tracing the motion of particles in the simulation space. Imposing the boundary conditions on the BTE in the frame of this tool means that rules must be defined in order to handle the events that occur when the particles reach the boundary of the simulation domain of the interfaces between regions with different physical properties. The choice of incorrect boundary conditions can produce an undesired behavior in the carrier distribution.

The most common boundary conditions are: reflecting, absorbing, injecting, and looping boundaries (Figure 4.6) [147, 164–168]. In the reflecting boundary case, the particles that hit the boundary are reflected. It is often used at the $Si - SiO_2$ interface in classical EMC codes. On the contrary, the particles are removed from the simulation domain when they reach the boundary in the absorbing boundary case. The injecting contacts steadily injects carrier with a given momentum distribution inside the simulation domain at the beginning of each time step. The injection contacts can behave either as reflecting or absorbing for the carriers in the simulation domain. The last case of boundary conditions is the looping contacts, which consist of a pair of a coupled interfaces such that carriers exit the simulation domain from on side and then re-enter the domain at the other side.

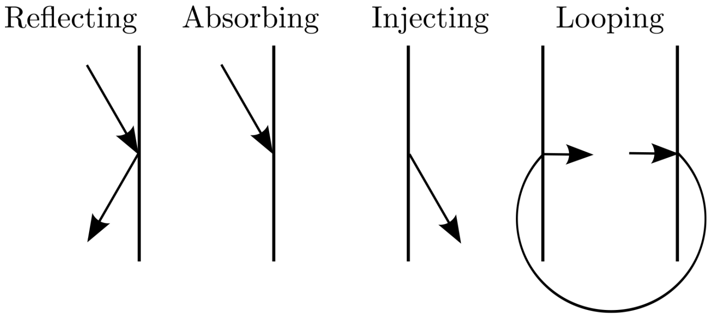


Figure 4.6: Different boundary conditions implemented in Monte Carlo simulator for free-carriers.

The boundary condition used in the source and drain contacts, also known as ohmic contacts, in the MS-EMC simulator herein described is the mixing of the absorbing/injecting ones. In particular, this condition is used in the free-flight subroutine inside the *Monte Carlo Transport* block in Figure 4.3.

The initial version of this tool is based on Tomizawa's description [128] where the development of a Monte Carlo simulator in semiconductor devices is explained. In the proposed description, the free-flight subroutine simulates the carrier motion and selects the particles absorbed by the source and drain contacts as erasable. Once finalized, other subroutine evaluates the number of particles inside the device erasing all those that have left the contacts. This latter subroutine is also responsible for maintaining carrier neutrality in the surrounding of the source and drain ohmic contacts, absorbing or injecting the required number of particles.

Nonetheless, both subroutines have been joined with the intention of optimizing the code during the parallelization of the 2D simulator. In this way, they can be implemented in a single loop in two steps reducing the computational cost. Firstly, the carrier free-flights are putting into practice in the same way that in the previous version. The difference now is that the superparticles exiting the drain are used again for injection with the corresponding new values into the source and, correspondingly, those leaving the source are injected back into the drain. The index of a exiting superparticle can be reused without any matrix reorganization process by a injected superparticle with new transport properties. This procedure also allows the use of parallel code and the optimization of the computational cost. When a superparticle is injected by a contact, the neutrality condition of the carrier is evaluated. In case the injected superparticle does not fulfill this condition, it is marked to be erased at the end of the loop. Thanks to this improvement, when the free-flight is completed, we will only have to erase the marked superparticles and evaluate again the charge neutrality condition in the contacts. If the number of superparticles injected into the contacts is not enough to fulfill the carrier neutrality condition, the remaining superparticles are injected to fulfill that condition. Finally, each superparticle should be given a proper weight before the injection to satisfy the conservation of total charge as described in the next Section 4.5.

Regarding the carrier distribution, two statistics can be assumed: the equilibrium Maxwell-Boltzmann or the Fermi-Dirac statistics. On the one hand, the equilibrium Maxwell-Boltzmann statistic describes a thermal isotropic injection. The carrier distribution function $f_{MB}(E_k)$ using this statistics is given by:

$$f_{MB}(E) = \frac{1}{\exp\left(\frac{E-E_F}{k_{\mathrm{B}}T}\right) - 1}, \tag{4.5}$$

being the -1 commonly neglected. The superparticle must be absolutely defined before being injected in a particular subband at the source or drain contact in the simulation space by means of three parameters: the angle in which the superparticle is going to be injected φ, the kinetic energy E_k, and the superparticle velocity in the direction normal to the interface v_x. The angle is randomly calculated in the interval $\varphi \in [0, 2\pi]$. In addition, due to the thermal condition, the kinetic energy can be selected using a random number $r \in [0, 1]$ by a uniform distribution $E_k = -k_{\mathrm{B}}Tln(r)$. In general, the velocity of a superparticle in the device is calculated from the kinetic energy and the wave vector. The wave vector in a superparticle for a non-parabolic subband in the crystal coordinates is:

$$\vec{k'} = \frac{\sqrt{2E_k(1+\alpha E_k)q}}{\hbar} \tag{4.6}$$

where α is the non-parabolic parameter. Then, the wave vector $\vec{k'}$ is divided into its transport and confinement coordinates, which effective transport and confinement masses are m'_x and m'_z, respectively, according to the angle φ:

$$k'_x = \vec{k'}cos(\varphi)\sqrt{m'_x};\ k'_z = \vec{k'}sin(\varphi)\sqrt{m'_z} \tag{4.7}$$

Thus, the superparticle velocity in the crystal is given by:

$$v'_x = k'_x\frac{\hbar}{m'_x}\frac{1}{1+2\alpha E_k};\ v'_z = k'_z\frac{\hbar}{m'_z}\frac{1}{1+2\alpha E_k} \tag{4.8}$$

Finally, a change in the coordinates between the crystal and the device orientation are required being θ the angle between them. Assuming that the direction normal to the interface is the x direction in the transport plane, the velocity of a superparticle in the device is calculated as:

$$v_x = v'_x cos(\theta) + v'_z sin(\theta) = \frac{\sqrt{2E_k(1+\alpha E_k)q}}{1+2\alpha E_k}\left[\frac{cos(\varphi)cos(\theta)}{\sqrt{m'_x}} + \frac{sin(\varphi)sin(\theta)}{\sqrt{m'_z}}\right] \tag{4.9}$$

On the other hand, injecting particles according to an equilibrium distribution may be inconsistent with the out-of-equilibrium regime inside the device. Accordingly, the Fermi-Dirac statistic is a more adequate guess but it introduces higher computational cost. This statistic also is an equilibrium distribution but it is a better conjecture in degenerate or quasi-degenerate systems such as the source and drain region due to their high doping. In these regions, the Fermi levels is very close or even inside the conduction band. The probability of injecting superparticles with a given state is required in order to determine the superparticle parameters. It is proportional to the group velocity in the direction normal to the interface and the Fermi-Dirac carrier distribution function $f_{FD}(E)$. The probability is different for each subband and so it can be calculated more generally according to the kinetic energy (E_k) instead of the total energy (E):

$$p_{FD}(E_k) = f_{FD}(E_k)v_x(E_k) \tag{4.10}$$

where $f_{FD}(E_k)$ has been reformulated according the kinetic energy instead of the total energy which include the subband energy E_s:

$$f_{FD}(E_k) = \frac{1}{1+\exp\left(\frac{E_k+E_s-E_F}{k_\mathrm{B}T}\right)} \tag{4.11}$$

In this approximation, the superparticle angle depends on the angle between the crystal and device orientation θ: $\varphi \in [\theta - \frac{\pi}{2}, \theta + \frac{\pi}{2}]$. In consequence, the probability of a superparticle crossing a surface depends on its kinetic energy and its angle of motion and it can be calculated by introducing the equation 4.9 and the φ dependence in $p_{FD}(E_k)$:

$$\begin{aligned} p_{FD}(E_k) &= f_{FD}(E_k)\frac{\sqrt{2E_k(1+\alpha E_k)q}}{1+2\alpha E_k}\int_{\theta-\frac{\pi}{2}}^{\theta+\frac{\pi}{2}}\left[\frac{cos(\varphi)cos(\theta)}{\sqrt{m'_x}}+\frac{sin(\varphi)sin(\theta)}{\sqrt{m'_z}}\right]d\varphi \\ &= f_{FD}(E_k)\frac{\sqrt{2E_k(1+\alpha E_k)q}}{1+2\alpha E_k}\left[\frac{1+cos(2\theta)}{\sqrt{m'_x}}+\frac{1-cos(2\theta)}{\sqrt{m'_z}}\right] \end{aligned} \tag{4.12}$$

Once this probability is determined, the kinetic energy and the angle φ are randomly chosen using the direct technique in the generation of random numbers with given probability distribution function [144, 145] as described for the self-scattering in section 3.2.2.

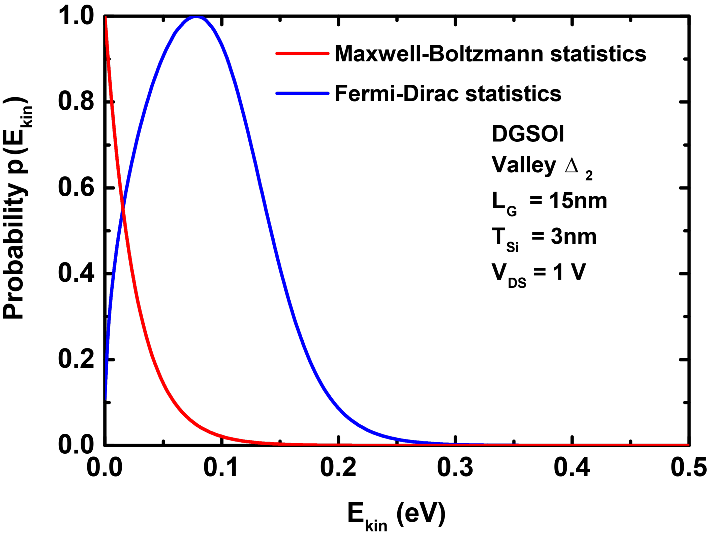


Figure 4.7: Probability p_{FD} of injecting superparticles given by the equation 4.12 and both the Maxwell-Boltzmann and Fermi-Dirac statistics as a function of the kinetic energy. The device analyzed is a DGSOI with $L_G = 15nm$, $T_{Si} = 3nm$, and $V_{DS} = 1V$. The first subband of the Δ_2 valley is considered.

Figure 4.7 shows the probability p_{FD} of injecting superparticles given by the equation 4.12 and both the Maxwell-Boltzmann and Fermi-Dirac statistics as a function of the kinetic energy for the first subband (valley Δ_2) of a DGSOI with $L_G = 15nm$, $T_{Si} = 3nm$, $V_{DS} = 1V$, a gate oxide with Equivalent Oxide Thickness $EOT = 1nm$, and gate work function of 4.385eV, which is depicted in Figure 4.2. The most likely kinetic energy of a superparticle that is injected in the device is around $0.8eV$ using Fermi-Dirac statistics, whereas the maximum value is around $0eV$ using Maxwell-Boltzmann statistics. Subsequently, there is a difference in terms of kinetic energy of a superparticle when choosing between both carrier distribution functions especially highly doped regions as source and drain. As a result, the injection velocity is higher for higher kinetic energy and, hence, the superparticles velocity in the direction normal to the interface (v_x) is also higher for the Fermi-Dirac statistic than for the Maxwell-Boltzmann one in the channel (Figure 4.8).

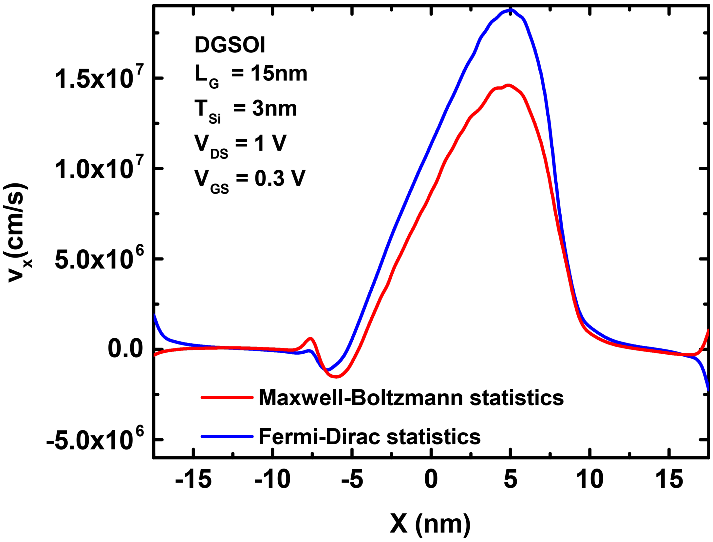


Figure 4.8: Average velocity in the direction normal to the interface (v_x) for both statistics the Maxwell-Boltzmann and the Fermi-Dirac. The device analyzed is a DGSOI with $L_G = 15nm$, $T_{Si} = 3nm$, $V_{DS} = 1V$, and $V_{GS} = 0.3V$.

In general, the drain current in Monte Carlo simulators is calculated by multiplying the velocity by the number of superparticles in the center of the channel. Subsequently, another consequence when choosing between both carrier distribution functions is that the resulting I_D is higher for the Fermi-Dirac statistic than for the Maxwell-Boltzmann one as shown in Figure 4.9. This effect is even more remarkable for lower gate bias.

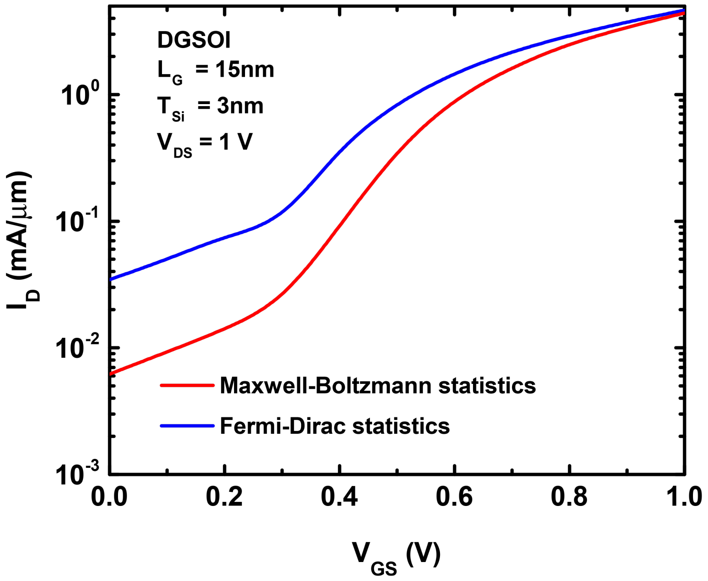


Figure 4.9: $I_D$vs.V_{GS} in the DGSOI device with $L_G = 15nm$ and $T_{Si} = 3nm$ at saturation regime taking into account for both statistics the Maxwell-Boltzmann and the Fermi-Dirac.

4.5 Energy-dependent weight for superparticles

As mentioned in Chapter 2, electronic devices have been scaled down in order to meet Social requirements. Device dimensions and supply voltages have been also decreased according with the scaling strategies. Consequently, small channel devices present generally higher electric fields and so higher electron energy. These carriers, which are also known as tail electrons, have higher energy than a critical energy level corresponding to the voltage bias. They can also create reliability problems and have influence over the carrier distribution. Despite the frequency of these tail electrons is very poor compared to the common ones, they introduce a stochastic noise in the MS-EMC (like any other method using random numbers) in predefined regions by maintaining a given number of particles in these high energy regions.

Subsequently, a weight redistribution technique [169–171] is needed to improve the statistics of the high energy tails for the simulated particles. In the MS-EMC tool, each superparticle is assigned a weight which denotes the number of electrons comprising the superparticle. In order to increase the number of high energy electrons, when a new superparticle in going to be included in the simulation domain with a energy level higher than a critical energy, a lower weight than for the low energy superparticles is assigned. In this case, the simulation has two type of superparticles which are energy-dependent for each subband: heavy and light weighted superparticles. The difference between both superparticles can be quantified by several orders of magnitudes.

In this hypothesis, a higher number of superparticles but with lower electrons per particle is included in the tail area and so the stochastic noise is solved because the statistical average is calculated including a major number of samples. By way of explanation, several superparticles fill the high energy area instead of containing only one superparticle with a particular kinetic energy and velocity, as depicted in Figure 4.10 where an example of an energy tail area is represented. In the left part of this Figure 4.10, a large number of superparticles with light weight fills completely the area and, hence, the statistical average can include different kinetic energy and velocities. In the right part, only one heavy weighted superparticle should bear in mind in the statistical calculation introducing the stochastic noise.

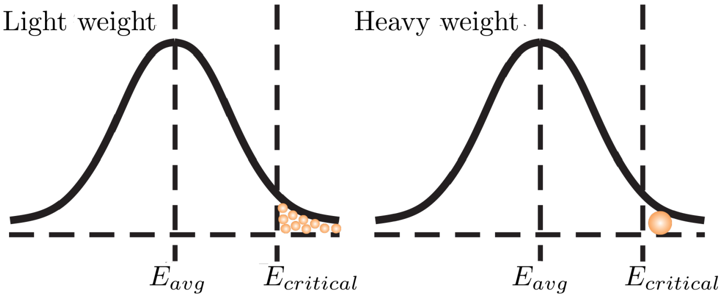


Figure 4.10: Schematic representation of a energy tail area which is filled by light (left) and heavy (right) weight superparticles.

The critical energy is calculated in consonance with the voltage bias for each subband. The energy subbands are modified by the gate and drain voltages and so the superparticles are divided into heavy or light if the barrier is high enough. In the case the potential barrier is tiny, the thermionic current is very hight and all the new superparticles can be considered as heavy weight because they do not have influence over the stochastic noise. Thus, this assumption therefore is only required at low voltage bias.

It is important to highlight in this point that the weight assignment is only considered in two scenarios. Firstly, when the superparticles are initialized in the simulation (*Init Particles* block in Figure 4.3), their weight is assigned according to the comparison between their initial energy and the subband energy. Secondly, when the superparticle is going to be injected in the device because it has reached either the source or the drain, the weight is calculated according to the energy resulting to the Fermi-Dirac statistics for this particular subband (as aforementioned in Section 4.4). Otherwise, the weight change is a forbidden action owing to the conservation of the total weight. In other words, the superparticle weight remains constant during its flight in the device.

Additionally, the conservation of the total charge has been regarded in the Monte Carlo iterations as the conservation of the total weight in the whole device. For this reason, the removed electrons, which are originated from injecting a new light weight superparticle instead of a heavy one, are injected in other superparticles.

Apparently, the main drawback of this method is that the computational time is wasted by simulation particles with low weight. However, as the injection or initialization of superparticles with high energy are rare events, the number of them are much lower than the heavy weight superparticle. It does not therefore introduce a significant modification of the simulation time, whereas dramatically enhance the statistics of the system.

In order to study how this new model affects the device performance, a DGSOI transistor has been simulated (Figure 4.2). The device parameters are the same values as in the previous Section 4.4: $L_G = 15nm$, $T_{Si} = 3nm$, a gate oxide with Equivalent Oxide Thickness $EOT = 1nm$, and gate work function of 4.385eV. The subtle but relevant difference is, how this effect is more remarkable for low bias, the study has been carried out at low drain regime ($V_{DS} = 100mV$) in this Section instead of using saturation regime as in Section 4.4.

The effect of the superparticle weight distribution can be easily understood in Figure 4.11 where the electron distribution from the fundamental subband as a function of total energy is represented for both models. Nevertheless, there is no difference in the total number of carriers in the whole device since the energy-

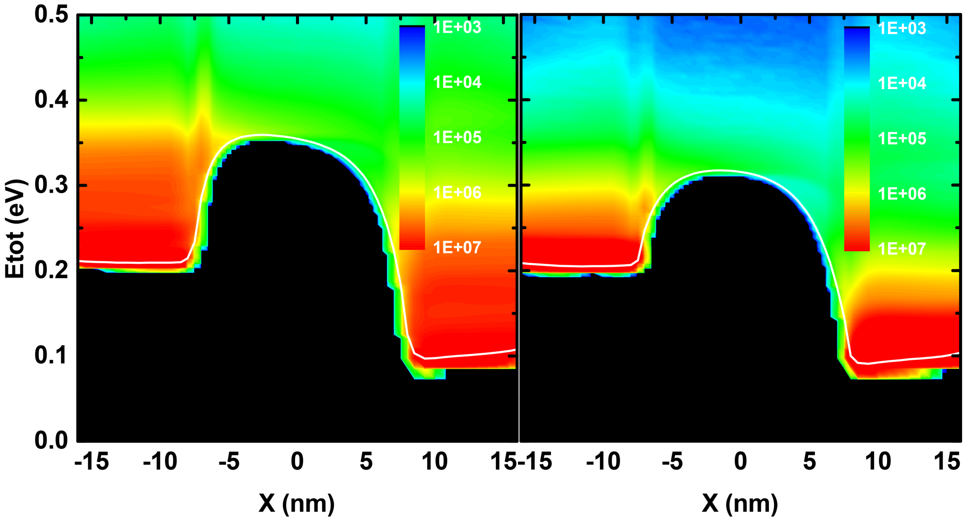


Figure 4.11: Electron distribution in arbitrary units in the lower energy subband as a function of total energy in the DGSOI device with $L_G = 15nm$, $T_{Si} = 3nm$, $V_{GS} = 0.3V$, and $V_{DS} = 100mV$ including both models: when the superparticles have the same weight (left) and when energy-dependent weight for superparticles is considered (right).

dependent weight model satisfies the conservation of total charge as depicted in Figure 4.12. In other words, the electrons are located in the same region but with a more realistic energy.

Finally, the influence of this model on the drain current (I_D) is analyzed in Figure 4.13. This model dramatically reduces the stochastic noise in most of the relevant energy range because the statistic is dominated by the superparticle with the higher weight. For this reason, the drain current is lower when the superparticle weight is calculated according to its energy due to the increase of the rare events. Furthermore, the difference in the I_D between both models is mainly observable for lower gate bias as predicted.

4.6 Conclusions

In this Chapter we have deeply illustrated a description of the MS-EMC tool and some techniques for computational time reduction and statistical enhancement. As the electronic devices have been scaled down, the development of new models

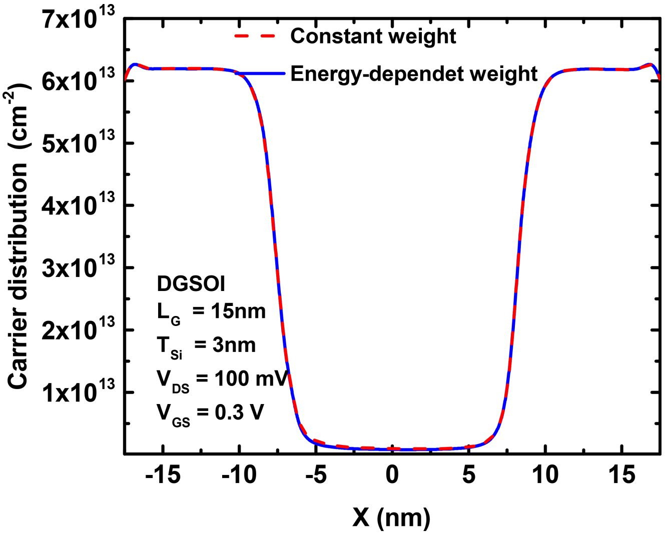


Figure 4.12: Electron distribution in the DGSOI device with $L_G = 15nm$, $T_{Si} = 3nm$, $V_{GS} = 0.3V$, and $V_{DS} = 100mV$ taking into account both models: when the superparticles have the same weight and when energy-dependent weight for superparticles is considered.

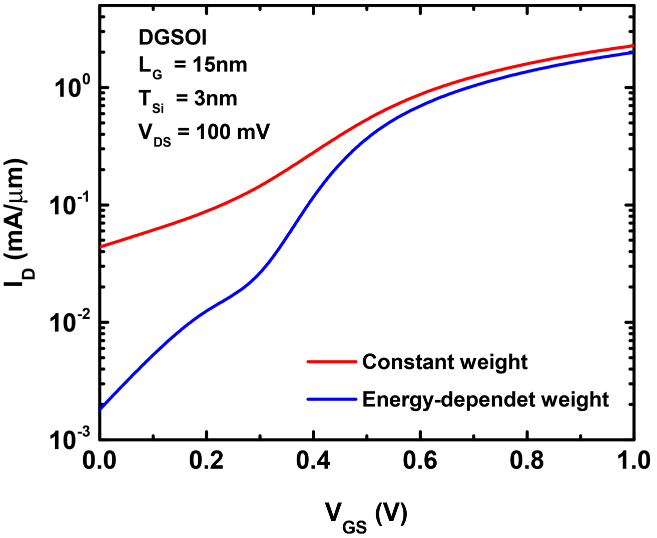


Figure 4.13: $I_D$vs.V_{GS} in the DGSOI device with $L_G = 15nm$ and $T_{Si} = 3nm$ at low drain bias regime taking into account both models: when the superparticles have the same weight and when energy-dependent weight for superparticles is considered.

is necessary in order to study their physical phenomena. In particular, when a simulation is carried out to study a transient and/or a non-homogeneous phenomenon, it is necessary to simulate a large number of electrons separately. This procedure is known as Ensemble Monte Carlo (EMC). Moreover, the inclusion of several subbands in each valley is required due to the high electron confinement in the novel architectures. This assumption has been developed in our MS-EMC tool allowing us to study different advanced nanodevices including new scattering models. In addition, this simulator has been parallelized using the OpenMP standard in order to reduce the computational cost. This enhancement has two main advantages: the study of more complex devices in a reasonable simulation time, and the possibility of speeding up the elaboration of new models thanks to the reduced time in the test versions. Then, a new method for the boundary conditions in the ohmic contacts is introduced. The superparticles are considered to enter in the device space for the source/drain if they reach the other contact. The parameters that describe the superparticle motion are calculated using the Maxwell-Boltzmann and the Fermi-Dirac statistics being the second one which give a more accurate picture of the distribution function. Finally, the energy-dependent weight model is described in order to reduce the stochastic noise that the superparticles with very high energy include in the device performance. These high energy superparticles can be defined as rare events and so their weight should be reduced.

Chapter 5

Tunneling Mechanisms

5.1 Introduction

As mentioned in Chapter 2, conventional CMOS devices have been scaled for more than 40 years in order to achieve higher density and performance, and lower power consumption. In particular, the supply voltage V_{DD} has to be scaled down in order to keep the internal electric fields and power consumption under control. Accordingly, the transistor threshold voltage (V_{th}) has to be adequately scaled to maintain a high drain current and achieve the performance improvement. Nonetheless, a V_{th} reduction can result in a substantial increase of the subthreshold leakage current [172].

Furthermore, as the device dimensions decreases, the MOSFET performance is degraded due to the appearance of the short-channel effects (SCEs). They involve the loss of channel control over the gate terminal and the increase of the drain effect, being the variation of the threshold voltage as a function of the channel length reduction one of the main effects to study.

SCEs not only have influence on the V_{th}, but also on the subthreshold characteristics. The variation of the subthreshold leakage current can be indicated by two parameters: the subthreshold swing (SS) and the transistor off-state current (I_{OFF}). Firstly, SS is the inverse of the slope of V_{GS} versus I_D in the weak inversion state that is, the reciprocal value of the subthreshold slope (S_{th}). In addition, the SS can describe the device efficiency of reaching the off-state when the gate voltage is decreased below V_{th}. As introduced in Section 2.2, the downscaling of V_{DD} requires an increase in the SS which means that the shift between the off and on state in a device is higher and so the efficiency is decreased. It therefore implies a scaling limit for the supply voltage. Another parameter is the transistor off-state current, I_{OFF}, which matches with the drain current when the gate voltage is zero. I_{OFF} is dominated by leakage from the drain-well and well-substrate reverse-bias pn junctions in the long-channel devices. However, the reduction of the power supply and, hence, the V_{th} in short-channel devices forces a large increase of I_{OFF} owing to a weak inversion state leakage. As a conclusion, the main goal is to optimize the device structure in order to minimize the off-current leakage while maximizing the linear and saturation drain current.

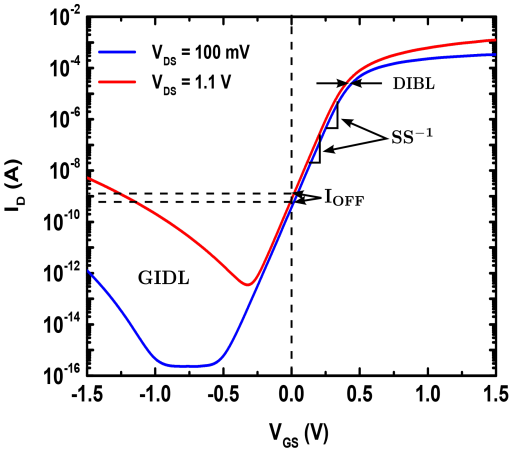


Figure 5.1: $I_D$vs.V_{GS} for a conventional MOSFET with gate length $L_G = 50nm$ at low and high drain bias showing some device parameters: the gate-induced drain leakage (GIDL), the transistor off-state current (I_{OFF}), the subthreshold swing (SS), and the drain-induced barrier lowering (DIBL).

Figure 5.1 shows a typical $I_D$vs.V_{GS} curve for a conventional MOSFET with gate length $L_G = 50nm$ at low and high drain bias where SS and I_{OFF} are represented as well as other two effects which are influenced by the scaled characteristics and they are also particular to the small geometries in the short-channel devices. These effects are more noticeable for higher drain current. On the one hand, the gate-induced drain leakage (GIDL) increases the drain current in the off-state due to the depletion edge at the surface below the gate-drain overlap region. On the other hand, drain-induced barrier lowering (DIBL) results in a notable increase in the drain current due to the extension of the drain to channel depletion region as the drain voltage increases. Moreover, the I_{OFF} increase is typically due to the channel surface current caused by DIBL effect.

As a introduction of this chapter, Figure 5.2 shows different leakage current components and mechanisms in deep-subnanometer transistors, contributing to the off-state current (not just the current from the drain terminal). They are divided into six short-channel leakage mechanisms [15], namely, the subthreshold leakage (I_1), the reverse-bias pn junction leakage (I_2), the gate oxide tunneling current (I_3), the gate current due to the injection of hot carriers from the substrate to gate oxide (I_4), the gate-induced drain leakage (GIDL) (I_5), and finally the channel punchthrough current (I_6). Currents I_2 and I_3 are both on-state and off-state current mechanisms, whereas I_1, I_5 and I_6 occur only in OFF state. I_4 happens typically during the transition between the transistor bias states.

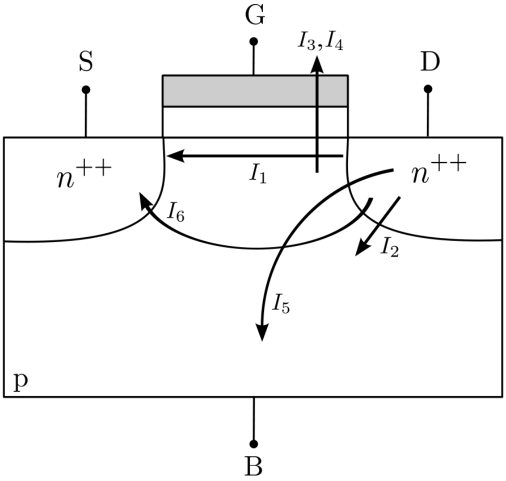


Figure 5.2: Summary of leakage current mechanisms of deep-subnanometer transistors: the subthreshold leakage (I_1), the reverse-bias pn junction leakage (I_2), the oxide tunneling current (I_3), the gate current due to the injection of hot carriers from the substrate to gate oxide (I_4), the gate-induced drain leakage (GIDL) (I_5), and the channel punchthrough current (I_6).

The subthreshold leakage (I_1) occurs between the source and the drain region in weak inversion [173]. When the gate voltage is below V_{th}, the minority carrier concentration is very small and thus some carriers can undergo a thermionic emission owing to the diffusion current.

Some other effects modify this current mainly due to the change in the required attributes in a short-channel device in order to turn it on and, consequently, threshold voltage varies. DIBL provokes the reduction of the barrier height when a high drain bias is applied, resulting in further V_{th} decrease [173]. Then, the body effect increases V_{th} because the reverse biasing well-to-source junction of a conventional MOSFET extends the bulk depletion region [174]. Other example is the V_{th} roll-off which is caused by the reduction of the threshold voltage when the channel length decreases [173]. As the distance between source and drain decreases, their depletion regions can penetrate into the channel length. Subsequently, part of the channel is depleted in the off-state and, hence, the gate voltage has to invert less bulk charge to turn the transistor on.

In general, none of the above change the subthreshold swing, but there are another effects that modulate this parameter. One of them is the narrow-width effect because the manufacturing processes of the small gate width alter both the V_{th} and the SS [14, 175, 176]. In addition, the effect of temperature increases the heat generation and thereby SS linearly increases with temperature and V_{th} decreases [173]. This dependence is of especial interest in digital very large scale integration (VLSI), since circuits operate at high temperature.

Another phenomenon that increases the subthreshold leakage current (I_1) is mainly motivated by the narrow potential barrier between the source and the drain as the dimensions of the electronic devices are reduced. In general, electrons with lower energy than the potential barrier at this position must undergo a backscattering in long-channel devices, whereas the ones in the short channel can also go through the barrier. This quantum effect is known as the Source-to-Drain tunneling (S/D tunneling) and it reduces the number of electrons that surmount the potential barrier because they can go through it, resulting in a modification of the potential barrier and so a V_{th} reduction. The inclusion of this quantum effect in the transport direction plays an important role in the extensive research of scaled electronic devices when transistors approach to nanometer scales since S/D tunneling is presented as a scaling limit [30, 31]. This phenomenon is thoroughly described in section 5.2 as well as a study of how it affects different technological architectures with the goal of determining the degradation included in each one.

Drain and source to well junctions are typically reverse biased, causing the reverse-bias pn junction leakage current (I_2). This current has two main components: one is minority carrier diffusion-drift near the depletion region edge, the other one is due to the electron-hole pair generation in the depletion region of the reverse-biased junction [20]. If both n and p regions are heavily doped and the doping profiles are abrupt, band-to-band tunneling (BTBT) dominates the pn junction leakage [21]. In conventional MOSFET scaling, tunneling phenomena from heavily doped junctions results in parasitic leakage currents and, as the device dimensions are scaled down, the BTBT current is even more significant. However, this process is precisely the working principle of some attractive devices such as the tunnel field-effect transistor (TFET) and it therefore is no longer an unwanted parasitic effect. This leakage mechanism is meticulously detailed in section 5.3 where a model for this phenomenon has been successfully implemented into the MS-EMC simulator. Then, it is applied to ultra-scaled silicon-based n-type TFETs taking into consideration that the BTBT current in silicon involves the emission or absorption of phonons, since it is an indirect band gap semiconductor.

Subsequently, the oxide thickness has to be also reduced proportionally to the channel length in order to maintain a reasonable SCEs immunity in accordance to the scaling strategies. Nonetheless, the oxide thickness reduction involves a higher electric field across the ultra-thin insulator. This fact leads to the possibility that charge carriers can overcome the barrier for transport set up, resulting in the tunneling of carriers through the gate oxide [15, 21]. The increase of the leakage current through the insulator limits the use of SiO_2 as the common insulator material for oxide thicknesses below $2nm$. Consequently, the employment of high–κ materials has been recognized as the best alternative and they are currently being kept in mind in the commercialization of the last generation devices. The gate leakage current can be divided into two major contributions: the gate oxide tunneling current from the substrate to the gate contact (I_3), and the injection of hot carriers from the substrate to gate oxide (I_4). A gate leakage mechanism (GLM) model is deeply described in Section 5.4 including direct and trap assisted tunneling by means of a MS-EMC code. Moreover, it is necessary to bear in mind this phenomenon in conventional devices at such a small scale to understand their performance. For this reason, a comparative of different transistor architectures is provided in Section 5.4.3.

The aforementioned gate-induced drain leakage (GIDL) (I_5) is the immediately leakage current in Figure 5.2. This phenomenon is caused by the high field effect in the drain-substrate junction of a MOS transistor. Figure 5.3 shows a schematic example of the GIDL current in a conventional MOS. Firstly, when the gate is biased to form an accumulation layer at the silicon surface, it behaves like a p-region more heavily doped than the substrate due to the presence of the accumulated holes at the interface. Hence, the depletion layer is much narrower near the surface, increasing the local electric field in that region. Then, if the drain bias is increased, the drain region under the gate can be depleted and even inverted as shown in Figure 5.3 (right). As a consequence of all these effects, minority carriers are emitted in the drain region under the gate. Eventually, they are swept laterally from the depletion layer owing to the lower potential in the substrate for minority carriers, producing the GIDL current [15, 21]. This phenomenon is enhanced for a thinner oxide thickness and a higher drain voltage because a higher potential between the gate and the drain is created and so the electric field increases. The most promising solution to minimize GIDL is the choice of a very high and abrupt drain doping [177].

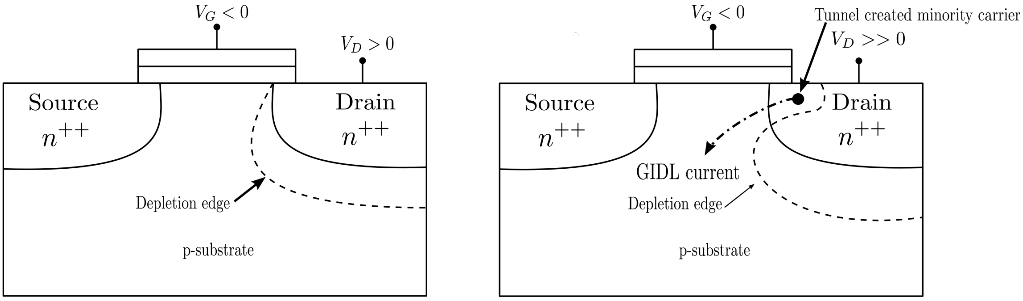


Figure 5.3: Condition of the depletion region near the drain-gate overlap region of MOS transistor when surface is accumulated with negative gate bias (left), and n^+ region is depleted or inverted with high negative bias (right).

Finally, the channel punchthrough current (I_6) completes the leakage current components in Figure 5.2. The depletion regions under the source and the drain extend into the channel due to the proximity between the source and the drain in short-channel devices. In addition, if the drain bias is increased, the distance between these regions decreases and they can be even merged. When both depletion regions are mixed, some majority carriers in the source cross the energy barrier and, hence, the drain can collect them. Subsequently, the punchthrough current results in an increase of the subthreshold leakage current and a degradation of SS. It usually occurs under the surface because of the depletion region shape [14] as depicted in Figure 5.3 (right).

5.2 Direct Source to Drain tunneling

5.2.1 Introduction

The study of quantum effects in conventional and new architectures becomes mandatory to determine which is the best candidate to extend the end of the roadmap. Quantum effects in the confinement direction were taken into account in a first step as the channel thickness was reduced to improve scalability. Nonetheless, due to the channel length dimensions of the devices produced nowadays, quantum effects in the transport direction must be included in order to improve predictability.

In particular, Source-to-Drain tunneling (S/D tunneling) has been presented as a scaling limit effect in ballistic non-equilibrium Green's Function (NEGF) studies [30]. Furthermore, it will distort the MOSFET operation at transistor channel lengths around $3nm$ [31]. When an electron with total energy smaller than the S/D barrier reaches this region, there is a possibility of going through the barrier instead of rebounding as classical dynamics predicts. This phenomenon introduces population under the potential barrier of the subbands, which is a forbidden region for electrons if a classical description is used. It therefore modifies the height of the potential barrier and increases the subthreshold current. In addition, the number of electrons affected by this effect is random. In other words, it produces a deviation in the performance of the electronic device and introduces noise. This effect is of special interest when the operation regime is near-threshold because the leakage current increases and the threshold voltage (V_{th}) decreases [27].

The following section gives a detailed overview of the code developed to carry out our research. Specifically, a deep description of the S/D tunneling algorithm is provided to explain the transmission probability and the motion of an electron which undergoes this effect. Subsequently, a meticulous comparison among FDSOI, DGSOI and FinFET when S/D tunneling is included by means of a MS-EMC simulator is analyzed. This study is very relevant in determining the impact of this quantum effect on these different architectures. It will be shown that the addition of multiple gates in the device architectures and its orientation have different influence on the S/D tunneling and, consequently, on the device characteristics.

5.2.2 Simulation set-up

The fundamentals of the free-flight of an electron in Monte Carlo algorithms establish that the motion of an electron in the transport direction finishes because of the random choice of a scattering event. After each flight of an electron, its new position is calculated. In a semiclassical approximation (Figure 5.4(a)), if the total energy of this electron is higher than the potential barrier at this position, the electron goes from source to drain by thermionic emission. However, if the electron is located at the same position but its energy is lower than the maximum of the potential barrier, it would rebound from it suffering a backscattering or would go from the source to the drain through it due to S/D tunneling.

To establish whether an electron is going to undergo this quantum effect or not, it is necessary to determine the probability of tunneling through the barrier at a specific energy (Figure 5.4(b)). This probability is equivalent to the transmission coefficient because it calculates the number of electrons that is going to undergo S/D tunneling respect to the total number of electrons with lower energy than the top of the potential barrier in the same region. The transmission coefficient $T_{dt}(E)$ is defined as the ratio of the quantum mechanical current density due to a wave impinging on a potential barrier and the transmitted current density. By way of explanation, it is a measure for the probability of a particle, described by a wave package, to penetrate a potential barrier.

In this work, the Wentzel-Kramers-Brillouin (WKB) approximation [32, 33] has been applied in order to calculate the tunneling probability of the electron T_{dt}. The huge advantage of the WKB transmission coefficient is that its numerical evaluation involves only small computational effort because it is based on a transfer-matrix method [178, 179]. Firstly, the calculation of $T_{dt}(E)$ through an arbitrarily shaped barrier is carried out by approximating the barrier as a series of piecewise constant rectangular barriers, for which the transmitted quantum current density can be determined. Then, it is integrated over this series of infinitesimal barriers.

As a result, the tunneling probability of the electron T_{dt} using the WKB approximation [32, 33] is given by:

$$T_{dt}(E) = \exp\left\{-\frac{2}{\hbar}\int_a^b \sqrt{2m_{tr}^*(E_i(x) - E)}\,\mathrm{d}x\right\}. \tag{5.1}$$

Where a and b are the starting and ending points, E and m_{tr}^* are the energy perpendicular to the potential barrier and transport effective mass of the electron, respectively, and $E_i(x)$ the energy of the *i-th* subband. In general, a particle that tunnels through the band gap is described by a wave function with complex wave vector $\vec{k}$, whereas the imaginary part of $\vec{k}$ describes the dampening of the evanescent wave within the tunneling barrier. It enters the above formula for the transmission coefficient, Equation 5.1, and is given by $\vec{k} = 2m_{tr}^*(E_i(x) - E)$. Thus, $T_{dt}(E)$ strongly depends on the relation $\vec{k}(E)$, which can be extracted from the dispersion relation in the band gap. For the S/D tunneling, this relation is the potential barrier structure.

This approximation has already been used to study S/D tunneling in other electronic devices [180]. In this work, the MS-EMC simulator offers a detailed description of subband structure. Consequently, T_{dt} has been calculated for each electron bearing in mind the minimum of the energy of its subband instead of the conduction band [181].

Once the tunnel probability is known, a rejection technique is used to determinate whether the electron will tunnel or not. A uniformly distributed random number r_{dt} between 0 and 1 is generated and compared to T_{dt} (Figure 5.4(b)). On the one hand, if $r_{dt} > T_{dt}$, it will turn back and fly into the device with $v(x) = -v(x)$ suffering a backscattering (Figure 5.4(c)). On the other hand, if $r_{dt} \leq T_{dt}$, the electron will go through the barrier and the particle will be marked to show that it is undergoing S/D tunneling (Figure 5.4(c)).

Several assumptions have been considered in the aforementioned method to enhance the calculation of T_{dt}. The exact starting and ending points when the electron goes through the barrier are calculated in order to reduce the mathematical uncertainty and the rounding errors coming from the discretization. In addition, a maximum tunneling rejection length is also introduced (L_{max}) to consider a realistic tunneling path. In this work, L_{max} has been chosen according to the channel length dimensions, $L_{max} = 10nm$, and it is consider as a constant in all the simulations. Moreover, the comparative between r_{dt} and T_{dt} has been

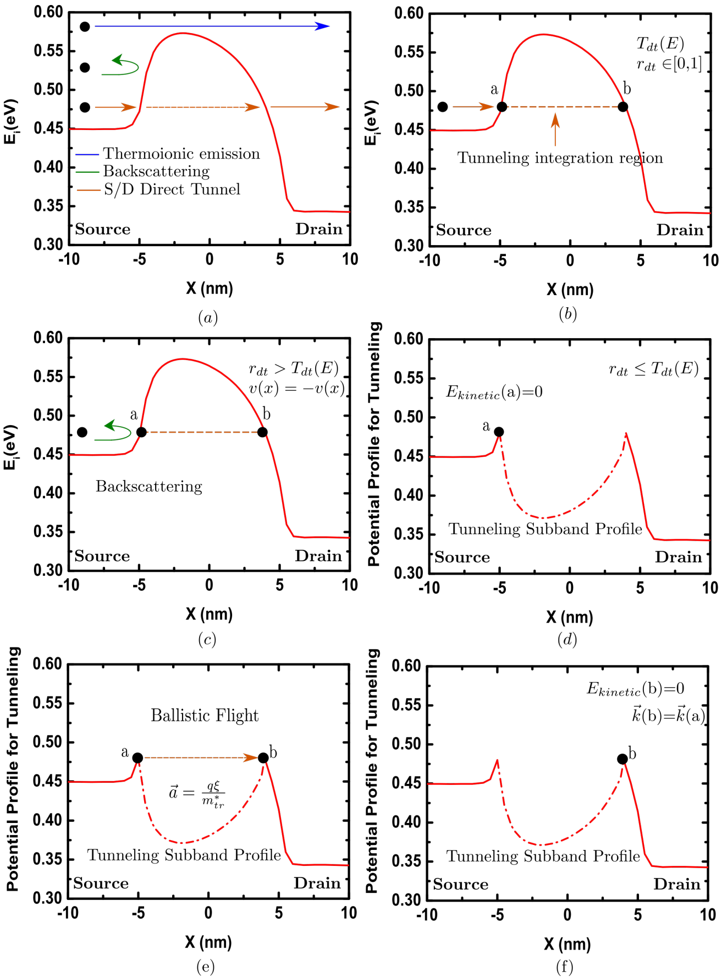


Figure 5.4: Representation of the tunneling model: The potential barrier (a) is inverted and the particle is placed at the starting point a (b), it follows a classical path obeying Newton's second law of motion (c) until it reaches the ending point b (d).

included after each integration step in order to decrease the computational effort. Despite the inclusion of these benefits in the integral calculation, the limitation of this approximation is the integration time of the tunnel probability.

Finally, the tunneling path must be described. The first step is to determine a realistic model for the motion of the particle. Two assumptions have been kept in mind to determine several scenarios. The first one establish that the electron goes directly from the starting point to the ending point in the same time step. In other words, the electron is not going to fly into the potential barrier. This instantaneous tunnel is an unrealistic and a non self-consistent model because it assumes depreciable tunneling time, whereas steady-state full-quantum simulations show charge inside the barrier. Nevertheless, it has been also considered in this work as a limit of this tunneling mechanism. For this reason, a more realistic model is used in this work taking into consideration that electrons are going to fly through the potential barrier during a period of time. The same idea has been employed in several studios which consider the possibility of electrons flying into forbidden regions [182]. Assuming that electrons reach the potential barrier in the starting point in a direction perpendicular to the barrier, they are going to move using Newton mechanics in an inverted potential profile $V(\vec{r}) \rightarrow -V(\vec{r})$.

This foregoing model has been chosen because it represents the motion of an electron in a forgiven region with imaginary $\vec{k}$. In particular, it is an extension for the non-local band-to-band tunneling algorithm (BTBT) [183]. In that work, the same classical path and tunneling probability were considered, whereas the starting and ending point in the tunneling path belong to valence and conduction band, respectively. The main advantage of this choice is that, once it has been implemented in the simulator, it is possible to extend it from the study of BTBT to that of S/D tunneling because the description of both mechanisms is based on the same assumptions.

Before starting its tunneling path, it is necessary to highlight some aspects of the electron motion. Inside the barrier, the carrier is considered as drifted in a conservative field. By mean of this, the angle, which determines k_x-k_y relationship, is maintained before being marked in the starting point a.

This classical trajectory could be found by the following steps [27] as shown Figure 5.4. Firstly, an imaginary particle is placed at the starting point a with zero kinetic energy (Figure 5.4(d)). Then, it accelerates in this system according to Newton's second law of motion, where ξ is the electric field (Figure 5.4(e)). Furthermore, the electron is going to undergo ballistic transport inside the barrier. Lastly, it reaches the ending point b with zero kinetic energy (Figure 5.4(f)).

5.2.3 Description of simulated devices and results

5.2.3.1 Device Structures

Different technological architectures are proposed to overcome the limitations of conventional planar devices. Fully-Depleted Silicon-On-Insulator (FDSOI) devices have been recognized as an alternative to bulk devices. The plain control of the transistor electrostatics is the guarantee for acceptable short-channel effects. In spite of its progress in electrostatic confinement and competitive fabrication cost, the addition of multiple gates surrounding the channel reduces the short-channel effects (SCEs) [82]. Furthermore, the increased electrostatic confinement provided by multiple gates relaxes the manufacturing constraints in comparison to conventional planar devices, as explained in Section 2.3.5. If we consider a double gate device, these gates can be oriented horizontally, Double-Gate Silicon-On-Insulator (DGSOIs), or vertically, FinFETs. Ideally, both channels are activated simultaneously and feature identical characteristics. The gates are parallel to the surface of the wafer for DGSOIs whereas they are perpendicular in FinFETs as depicted in Figure 5.5. One of the main advantages of vertical devices is the excellent control of the channel by the gate due to the diminished leakage current in the off-state. FinFETs are also less susceptible to several sources of variability such as V_{th} variability or random dopant fluctuations [184–186].

It should be highlighted in this point that the FinFET is a 3D structure whereas our MS-EMC simulator makes use of a 2D description. However, it was demonstrated that FinFETs with a big enough aspect ratio show similar behavior in all transport regimes when 2DMS-EMC and other 3D codes are used [187].

The performance of FDSOI, DGSOI and FinFET devices is herein analyzed when S/D tunneling is included in order to determine the impact of such phenomenon. The considered confinement direction of these devices on standard wafers changes from (100) for both planar FDSOI and DGSOI to $(0\bar{1}1)$ for FinFET, whereas the transport direction remains constant <011> as depicted in Figure 5.5.

The difference in the confinement direction modifies the electron distribution in the subbands, and, consequently, the potential profile. Moreover, the carrier transport effective mass is modified [188]. Table 5.1 summarizes the values of the effective masses for each devices and Table 5.2 shows their numerical values.

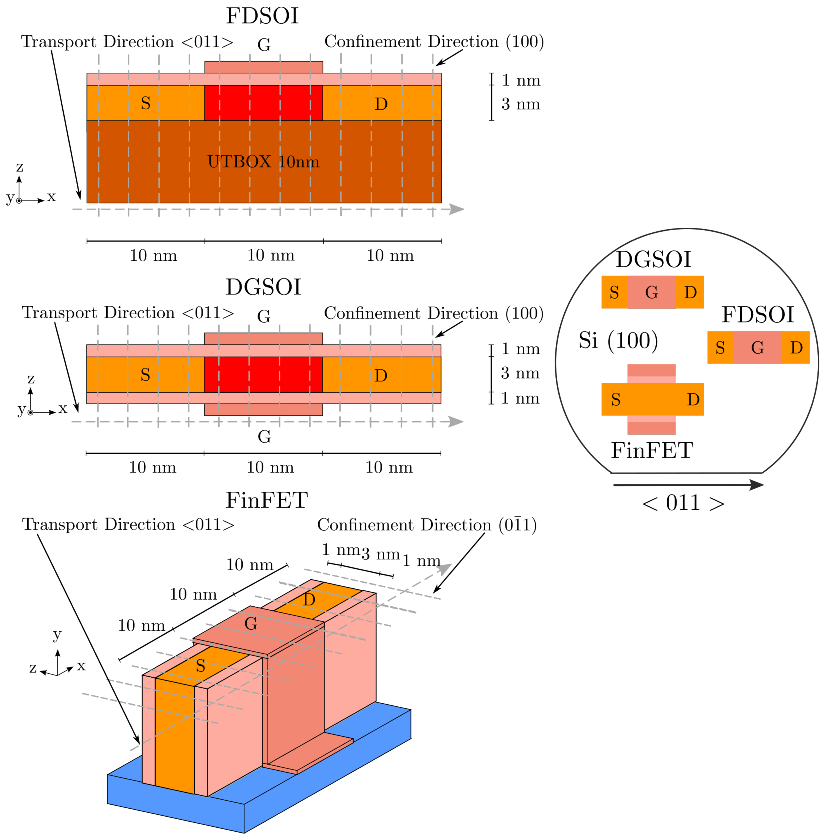

FDSOI
Transport Direction <011>
G
Confinement Direction (100)
1 nm
3 nm
S
D
UTBOX 10nm
z
y
x
10 nm
10 nm
10 nm
DGSOI
Transport Direction <011>
G
Confinement Direction (100)
1 nm
3 nm
1 nm
S
D
G
z
y
x
10 nm
10 nm
10 nm
FinFET
Transport Direction <011>
Confinement Direction (0$\bar{1}$1)
1 nm
3 nm
1 nm
10 nm
10 nm
10 nm
D
G
S
y
z
x
DGSOI
S
G
D
FDSOI
Si (100)
S
G
D
S
D
FinFET
< 011 >

Figure 5.5: FDSOI, DGSOI and FinFET structures analyzed in this work with $L_G = 10nm$. 1D Schrödinger equation is solved for each grid point in the transport direction and BTE is solved by the MC method in the transport plane.

Taking into account that in silicon, $m_l = 0.916m_0$ and $m_t = 0.198m_0$ are the longitudinal and traverse effective masses, respectively, m_0 is the electron free-mass, m_x is the transport mass, m_z is the confinement mass, and Δ_2 and Δ_4 represent the degeneration factors of each valley Δ.

These devices have been parametrized for gate lengths ranging from $5nm$ to $20nm$ and for two values of channel thickness $T_{Si} = 3nm$ and $T_{Si} = 5nm$. The rest of the technological parameters remains constant, gate oxide with Equivalent Oxide Thickness $EOT = 1nm$, and gate work function of 4.385eV. A Back-Plane with a $UTBOX = 10nm$, Back-Bias polarization of $V_{BB} = 0V$, and Back-Plane work function of 5.17eV have been chosen for the FDSOI device since a higher

Device	Valley	m_x	m_y	m_z
FDSOI and DGSOI	Δ_2	m_t	m_t	m_l
(100)<011>	Δ_4	$\frac{2m_l m_t}{m_l+m_t}$	$\frac{m_l+m_t}{2}$	m_t
FinFET	Δ_2	m_t	m_l	m_t
$(0\bar{1}1)$<011>	Δ_4	$\frac{m_l+m_t}{2}$	m_t	$\frac{2m_l m_t}{m_l+m_t}$

Table 5.1: Effective mass in silicon for FDSOI, DGSOI and FinFET devices studied in this work where m_x is the transport mass and m_z is the confinement mass.

Device	Valley	m_x	m_y	m_z
FDSOI and DGSOI	Δ_2	0.198	0.198	0.916
(100)<011>	Δ_4	0.326	0.557	0.198
FinFET	Δ_2	0.198	0.916	0.198
$(0\bar{1}1)$<011>	Δ_4	0.557	0.198	0.326

Table 5.2: Numerical values of effective mass in silicon for FDSOI, DGSOI and FinFET devices studied in this work where m_x is the transport mass and m_z is the confinement mass.

Back-Plane work function improves the electrostatic control [189].

5.2.3.2 Results

A set of simulations including lineal and saturation bias conditions has been performed to determine the importance of S/D tunneling on each device. The energy profiles of the lower energy subband for both tunneling path assumptions of each device when this quantum mechanism is included are shown in Figure 5.6. Both double gate devices (DGSOI and the FinFET) present similar energy profiles whereas the FDSOI transistor presents higher potential barrier (Figure 5.6) along the channel. In addition, the lower energy subband changes from Δ_2 in both FDSOI and DGSOI transistors to Δ_4 in the FinFET.

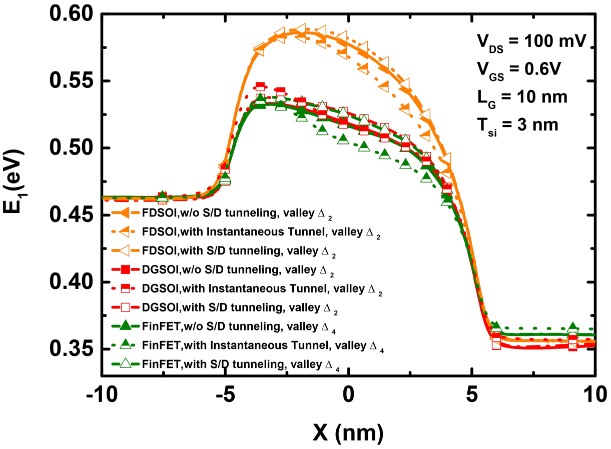


Figure 5.6: Energy profile of the lower energy subband in the 10nm device for DGSOI (valley Δ_2), FDSOI (valley Δ_2), and FinFET (valley Δ_4) with instantaneous tunnel, considering the motion of the electrons inside the potential barrier and w/o S/D tunneling with $V_{GS} = 0.6V$ and $V_{DS} = 100mV$.

For the sake of comparison, Figure 5.6 also shows the difference in the subband profiles instantaneous tunnel is considered in the simulation instead of the model herein developed for the three devices. The potential barrier becomes thinner because of the omission of the electron motion inside the forbidden region, whereas it increases its height when the electrons fly inside it.

The average effective transport mass of the electrons as a function of the total energy and the total population which undergoes from S/D tunneling is depicted in Figure 5.7, corresponding to m_x in Table 5.2. It is higher for the FinFET orientation due to its lower energy subband than for both the FDSOI and the DGSOI. This statement can be extended for the less populated valleys; Δ_2 in FinFET and Δ_4 in FDSOI and DGSOI.

As it can be extracted from Equation 5.1, the higher exponential part, the

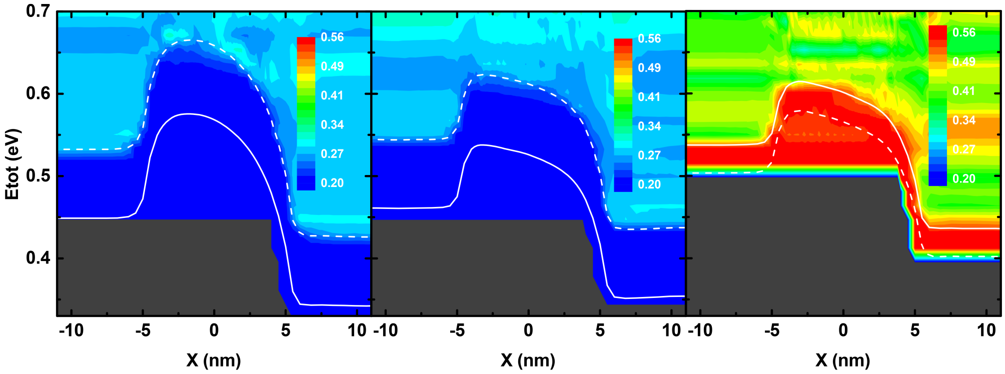


Figure 5.7: Average effective mass of the electron distribution with the lower energy subband of the valley Δ_2 (solid) and of the valley Δ_4 (dashed) as a function of the total energy and the total population in the 10nm device including S/D tunneling for FDSOI (left), DGSOI (middle), and FinFET (right) with $V_{GS} = 0.6V$ and $V_{DS} = 100mV$.

smaller T_{dt}. In summary, it is necessary longer tunneling paths, higher potential barriers or bigger m^*_{tr} in order to obtain smaller tunnel probability. On the other hand, the thinner potential barrier obtained when the instantaneous tunnel model is considered provides higher transmission probability because the particle undergoes a shorter tunneling path.

A comparison between FDSOI and DGSOI devices with the same confinement direction and, consequently, the same m^*_{tr}, shows that the higher energy profile of the FDSOI (Figure 5.6) decreases the transmission probability of an electron. For this reason, a larger number of electrons rebounds at the potential barrier in the FDSOI compared to the DGSOI and the FinFET. However, in spite of the similar energy profile between the DGSOI and the FinFET (Figure 5.6), which means similar tunneling length and height of the potential barrier at a specific starting point a, the higher m^*_{tr} for the FinFET orientation reduce the effectiveness of this phenomenon compared to DGSOI orientation. As a consequence, the number of particles affected by S/D tunneling is higher for the DGSOI than for both the FinFET and the FDSOI. This effect is depicted in Figure 5.8 where the electron distribution from the fundamental subband as a function of total energy is represented.

The reason why the maximum tunneling rejection length (L_{max}) has been chosen to be $10nm$ can be understood thanks to Figure 5.8. Electrons with low energy must go through the potential barrier with longer tunneling paths, which means that the transmission probability of these electrons is smaller. The population inside the barrier which tunneling length is near to $10nm$ in Figure 5.8 decreases substantially. Accordingly, the maximum tunneling rejection length has been chosen in consonance with the steep reduction of the population.

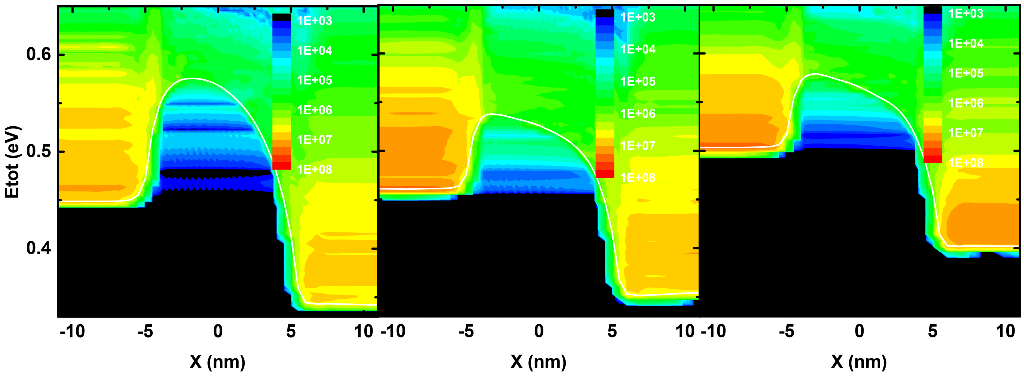


Figure 5.8: Electron distribution in arbitrary units in the lower energy subband as a function of total energy in the 10nm device including S/D tunneling for FDSOI (left), DGSOI (middle), and FinFET (right) with $V_{GS} = 0.6V$ and $V_{DS} = 100mV$.

In Figures 5.9 and 5.10, a meticulous study of how the different assumptions of the electron motion inside the potential barrier can modify substantially the device performance is depicted. When the instantaneous tunnel is considered the tunneling path becomes shorter as mentioned above and so the probability of tunneling increases. In accordance, the percentage of electrons near the potential barrier affected by S/D tunneling with respect to the total number of electron with lower energy than the top of the barrier in the same region as a function of V_{GS} (Figure 5.9) is much higher that when the model described in this work is taking into account. This quantum effect produces a noticeable modification of the $I_D - V_{GS}$ characteristics (Figure 5.10) for both assumptions being the instantaneous tunnel which overestimates the drain current. Despite the increase of the potential barrier when S/D tunneling is included (Figure 5.6), a higher current level is observed. The number of electrons that flows from source to drain is higher due to the possibility of tunneling through the barrier. Because of the overestimation of the number of particles caused by the instantaneous tunnel, we have focused hereafter on the motion model in which electrons are allowed to fly through the potential barrier.

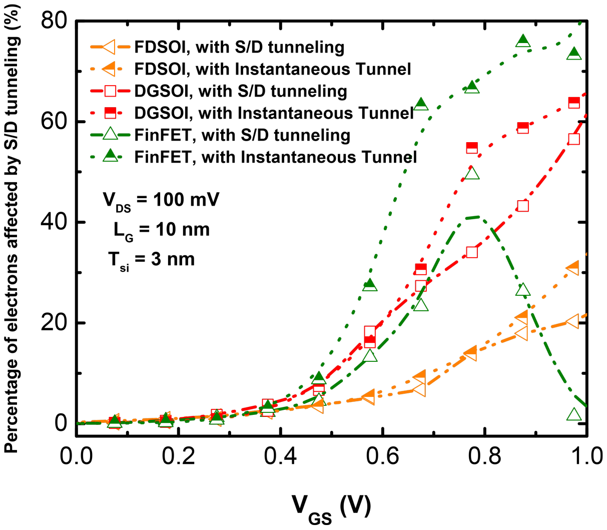


Figure 5.9: Percentage of electrons affected by S/D tunneling near the potential barrier in the 10nm device and $T_{Si} = 3nm$ at low drain bias taking into account the instantaneous tunnel and the model of motion inside the potential barrier developed in this work as a function of V_{GS} for FDSOI, DGSOI, and FinFET.

The percentage of electrons near the potential barrier affected by S/D tunneling with respect to the total number of electrons with lower energy than the top of the barrier in the same region is represented in Figure 5.11. This percentage is higher for DGSOI than for FDSOI and FinFET at low drain bias, whereas it increases for FDSOI at saturation regime. This effect is exacerbated at low drain biases as the dimensions of the device are reduced (Figure 5.11 top). Nevertheless, there is a L_G value which presents a maximum in the percentage of electrons affected by S/D tunneling at saturation regime (Figure 5.11 bottom). The explanation is the following. The Drain Induced Barrier Lowering (DIBL) introduces a reduction of the potential barrier at higher drain bias, which increases with the downscaling. For this reason, the number of electrons affected by S/D tunneling decreases for smaller L_G at saturation regime. In addition, this percentage decreases when T_{Si} increases for both drain bias conditions. This reduction is

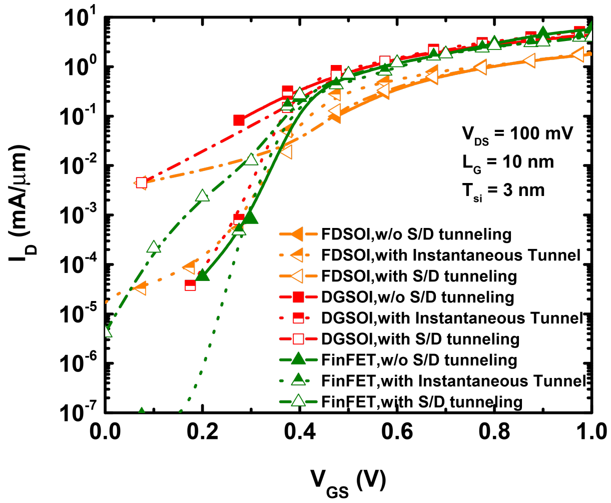


Figure 5.10: $I_D$vs.V_{GS} in the 10nm device and $T_{Si} = 3nm$ at low drain bias taking into account the instantaneous tunnel and the model of motion inside the potential barrier developed in this work as a function of V_{GS} for FDSOI, DGSOI, and FinFET.

motivated by two factors. Firstly, the number of electrons affected by S/D tunneling increases for higher drain bias because of the DIBL. Secondly, these particles decrease for higher devices because the difference between $(E_i(x) - E)$ increases the potential barrier height and, hence, T_{dt} decreases. As opposite to this, there is an increase of particles for FDSOI at lower T_{Si} and low drain bias owing to the longer tunneling path in FDSOI for smaller T_{Si}.

When the aforementioned percentage of electrons affected by S/D tunneling is represented as a function of V_{GS} at saturation regime, a maximum value appears due to the reduction of available carriers for tunneling near the potential barrier as the gate bias increases (Figure 5.12). Three different sections can be easily distinguished. Firstly, the number of particles that undergoes S/D tunneling is insubstantial for low gate bias because of the high and thick potential barrier. Secondly, this percentage increases significantly due to the reduction of the potential barrier until, for high gate bias, the number of available carriers for tunneling near the small potential barrier with lower energy is negligible. This maximum percentage appears for the FinFET but it is shifted to higher gate voltages for the DGSOI. By the way of contrast, the maximum percentage re-

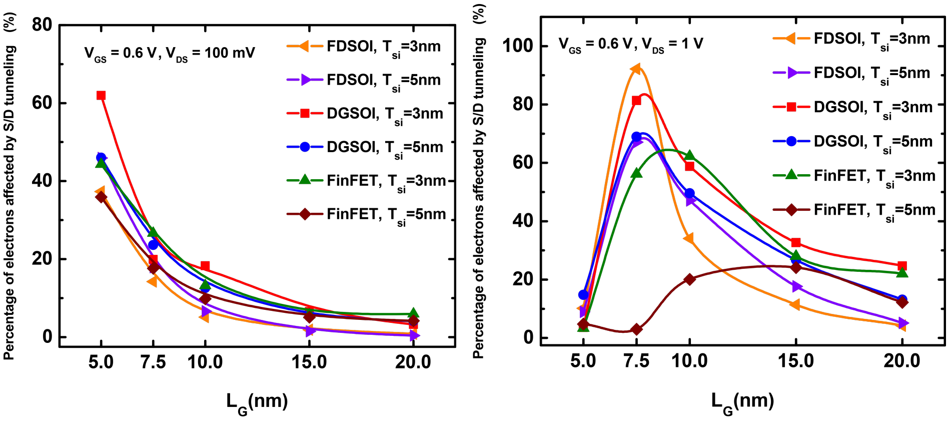


Figure 5.11: Percentage of electrons affected by S/D tunneling near the potential barrier as a function of L_G for FDSOI, DGSOI, and FinFET at low drain bias (left) and saturation (right) conditions with $V_{GS} = 0.6V$.

mains constant in FDSOI as a consequence of the shape of its higher potential barrier. The same behavior in the percentage of electrons as a function of V_{GS} is also shown in Figure 5.9 for low drain bias.

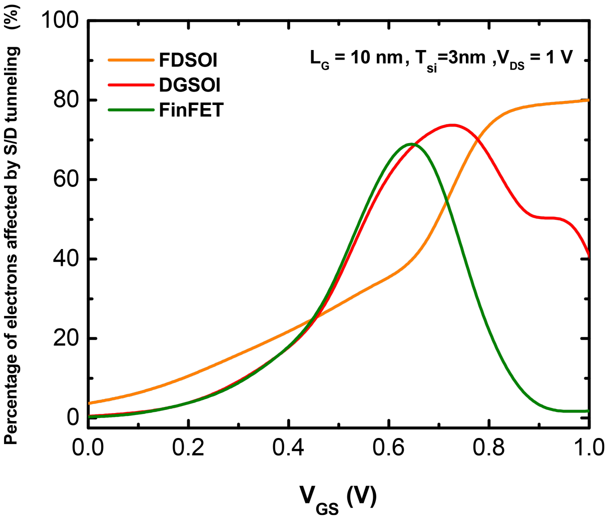


Figure 5.12: Percentage of electrons affected by S/D tunneling near the potential barrier in the 10nm device as a function of V_{GS} for FDSOI, DGSOI, and FinFET at saturation conditions.

It is necessary to highlight that the self-consistent calculation considered in this model leads to variations in the potential profile in comparison with the case without tunneling. The number of electrons with enough energy to surmount the barrier is lower because they can go through it in a lower energy state changing the potential barrier when S/D tunneling is considered as shown Figure 5.6. An increase in current is also observed when this effect is included because of the number of electrons that flows from source to drain is larger (Figure 5.10). It therefore introduces an important reduction in the threshold voltage.

The impact of the S/D tunneling on the threshold voltage variation (ΔV_{th}) in a simulation considering S/D tunneling and another without taking it into account can be observed in Figure 5.13. The percentage of electrons affected by S/D tunneling near the threshold voltage is higher for saturation conditions than for low drain bias owing to electrostatic changes as the drain bias is increased. Accordingly, the reduction of the V_{th} when S/D tunnel is taken into account is intensified for higher V_{DD}. This effect gets even more relevance when the device size is reduced and a higher T_{Si} is considered but the influence of this quantum effect is lower in the FinFET.

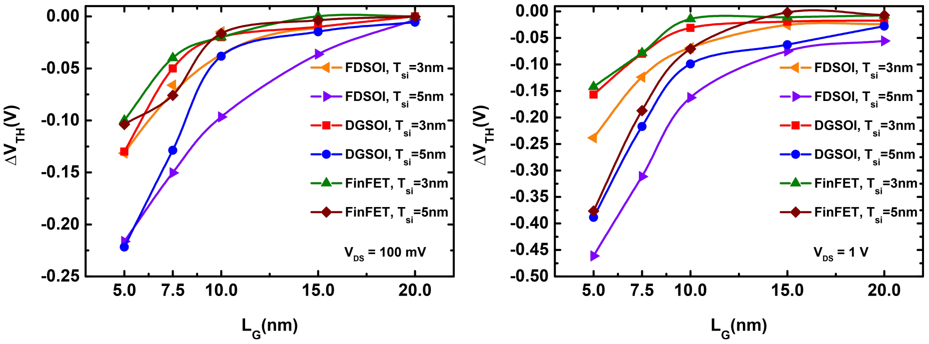


Figure 5.13: Difference between the threshold voltage (ΔV_{th}) of a simulation considering S/D tunneling and a simulation w/o taking it into account as a function of L_G for FDSOI, DGSOI, and FinFET at low drain bias (left) and saturation (right) conditions.

The DIBL is one of the main parameters used to determine the impact of short channel effects (SCEs) when the devices are scaling down. This parameter can be calculated as:

$$DIBL = -\frac{V_{th}^{DD} - V_{th}^{low}}{V_{DD} - V_{DS}^{low}}, \tag{5.2}$$

where V_{th}^{DD} and V_{th}^{low} are the threshold voltage at high drain voltage (V_{DD}) and at low drain voltage (V_{DS}^{low}), respectively. Figure 5.14 shows the DIBL dependency with the channel length when S/D tunneling is included. The DIBL is higher when this quantum effect is kept in mind in the three devices being the DGSOI which presents the lower degradation. The difference between V_{th} with and without S/D tunneling is more pronounced for higher drain biases (Figure 5.13). For this reason, the reduction in V_{th}^{DD} is more noticeable than in V_{th}^{low}. It therefore increases DIBL when S/D tunneling is taking into account. The difference between both simulations becomes more remarkable as the channel length of the devices decreases and T_{Si} increases. Lower DIBL means that the degradation produced in the device because of the drain voltage is lower. In consequence, this effect can be damaging in the device performance as the dimensions of the conventional devices are scaled down, especially for applications where an increase of I_{OFF} can be very harmful.

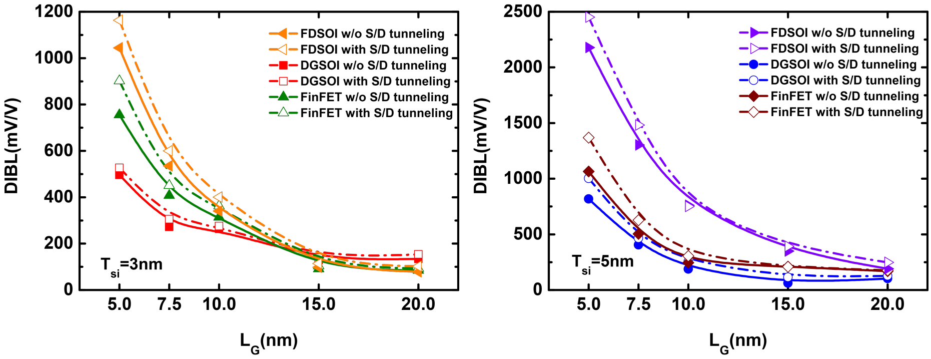


Figure 5.14: DIBL as a function of L_G considering S/D tunneling for FDSOI, DGSOI, and FinFET with $T_{Si} = 3nm$ (left) and $T_{Si} = 5nm$ (right).

Another important parameter that provides information about the performance of a device is the I_{ON}/I_{OFF} ratio, where I_{ON} and I_{OFF} are the higher and lower current of the devices, respectively, when $V_{DS} = V_{DD}$. Ideally, the best device would be that one with the highest I_{ON}/I_{OFF} ratio. This parameter as a function of the channel length is depicted in Figure 5.15 for both simulations being the FinFET which presents higher ratio. Because of the increase of the particles that flow from source to drain at low gate bias when S/D tunneling is considered, I_{OFF} increases. However, the number of electrons near the potential barrier with lower energy is reduced at higher gate bias. Subsequently, there is almost no difference in the I_{ON} current between the simulations where S/D tunneling is included and the ones where it is not bearing in mind. As a conclusion, the I_{ON}/I_{OFF} ratio decreases when this quantum phenomenon is incorporated being the DGSOI which presents the higher difference between both simulations.

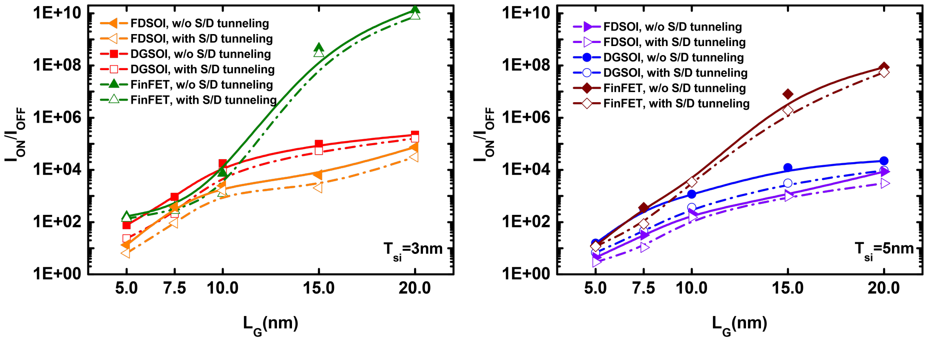


Figure 5.15: I_{ON}/I_{OFF} as a function of L_G considering S/D tunneling for FDSOI, DGSOI, and FinFET with $T_{Si} = 3nm$ (left) and $T_{Si} = 5nm$ (right).

5.2.4 Conclusions

This section presents the implementation of S/D tunneling in a MS-EMC tool for the comparison of its influence in ultrascaled FDSOI, DGSOI and FinFET devices. Two different assumptions about the particle motion when it undergoes this quantum effect are also described. The first hypothesis disregards the tunneling time called as instantaneous tunnel, whereas the second one bears in mind that the particle flies inside the potential barrier. Our calculations show that the instantaneous tunnel decreases the length of the potential barrier in comparison with a model that considers the flight of an electron inside the forbidden region.

It therefore increases the transmission probability and the current in an unrealistic way. The major results of our simulations among the three devices show important differences when S/D tunneling is taken into account fully caused by the energy profiles and the confinement directions. On the one hand, the difference in the energy profile among FDSOI and the both double gate devices FinFET and DGSOI shows that the higher potential barrier in FDSOI reduces the tunneling probability. On the other hand, the change in the confinement directions among FinFET and the two planar devices, FDSOI and DGSOI, modifies the subband distributions and the variation of transport effective mass. In conclusion, the number of particles that undergoes from this quantum effect is lower in FDSOI and FinFET devices, but the FinFET shows less degradation in their subthreshold characteristics and at saturation regime. For this reason, FinFETs are better candidates to implement future nodes, especially for ultra low power applications.

5.3 Band to Band tunneling

5.3.1 Introduction

In recent years, the basic architecture of the conventional MOSFET has been diversified to improve the device performance when transistors approach to the ultimate scaling limits. As the sub-10nm nodes become closer, different trends have appeared to face up specific issues related to the scaling limits concerning critical dimensions [13]. For this reason, it is worth including new effects associated with dimensions in the range on the inter-atomic distance apart from the traditional ones.

At the level of simulation research, two main working areas may be differentiated in the study of possible alternatives concerning processes and devices as it has been mentioned in Section 2.2. The first one is mainly focused on novel engineering solutions in order to create enhanced device architectures. Its development has been conceived to improve the carrier transport properties and to keep under control the short channel effects [17, 18]. The second approach explores new device paradigms based on different injection mechanisms and the nanometric regime enforced by scaling.

This section deals with one of the physical phenomena that improves the performance of electronic devices in the latter area: band–to–band tunneling (BTBT) [21]. One of the most popular devices among some of the most promising solutions based in the BTBT is the tunnel field-effect transistor (TFET) [19]. The process of quantum mechanical tunneling through a barrier between energy bands (BTBT) governs the injection of carriers in TFETs, contrary to MOSFETs where thermionic emission dominates. In this phenomenon, high electric fields ($> 10^6 V/cm$) across a reverse-biased pn junction causes significant currents to flow through the energy barrier due to tunneling of electrons (holes) from the valence (conduction) band of the p (n) region to empty states in the conduction (valence) band of the n (p) region, respectively, as shown in Figure 5.16 [21].

One of the main problems arising in the scaling of conventional MOSFETs is the scaling of their power supply voltage in order to reduce power density. The subthreshold swing (SS) limit of 60mV/dec present in this devices imposes a severe roadblock to reduce the supply voltage and maintain high ON-state currents along with low OFF-state leakages. In practice, the best MOSFET

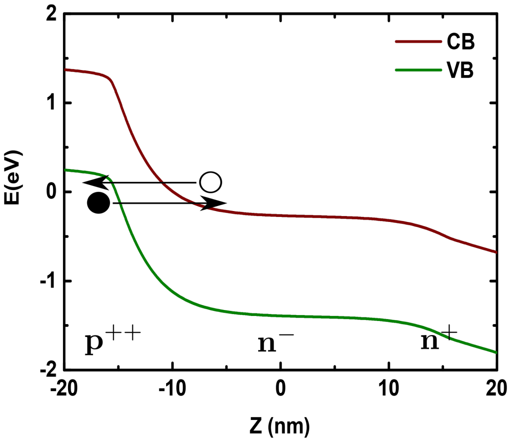


Figure 5.16: BTBT in a pn junction: tunneling of electrons (holes) from the valence (conduction) band of the p (n) region to empty states in the conduction (valence) band of the n (p) region, respectively.

implementations cannot bring SS below 70-80mV/dec, which leads to even worse situations [36,37] with the subsequent limitation in the reduction of V_{DD}. TFET should not be constrained by the aforementioned 60 mV/dec limit and, when the transistor is off, the tunneling barrier keeps the leakage current extremely low. It therefore arises as potential substitute for conventional MOSFETs especially in low-power applications.

In this section, we present the study of the model proposed to introduce BTBT in the MS-EMC simulator bearing in mind the aforementioned TFET. We assess the impact on the BTBT generation rate distribution of two different tunneling path choices. The first one estimates dynamically the BTBT path according to the criterion of the valence band maximum gradient trajectory; and the second one searches for the path featuring the minimum length trajectory. Furthermore, the results obtained with this model are compared with a customized TCAD-based approach including quantum effects that has extensively been used in the last years [190–192].

The section is organized as follows. The first Subsection 5.3.2 outlines the methodology to account for field-induced subband discretization, provides a detailed overview of the BTBT model used in the MS-EMC code as well as describes the TCAD-based simulation approach. We have considered two different criteria for the choice of the tunneling path followed by the carriers when crossing the potential barrier, which leads to different distributions of the generated electron-hole pairs. Subband discretization due to field–induced quantum confinement has been taken into account. Subsequently, the simulation results and discussion for the TFET is presented in Subsection 5.3.3.1. Finally, TCAD simulations accounting for quantization effects are considered for comparison purposes providing very accurate agreement with MS-EMC results.

5.3.2 Simulation set-up

5.3.2.1 Methodology in the MS-EMC simulator

The algorithm developed in this study implements the non-local direct and phonon assisted BTBT considering quantum confinement effects through discretization of conduction and valence bands into energy subbands [193]. Figure 5.17 depicts the flowchart of the modified MS-EMC simulator where the additional blocks are represented. The spatial variation of the energy bands has been taken into account for implementing BTBT thanks to the non-local model, so that the tunneling process can be more accurately described. However, since holes travel mostly in the low field source region, a drift-diffusion approach is used to describe them, whereas the electrons deal with the Monte Carlo method. The semiclassical model herein developed translates the tunneling current into suitable generation rates functions, G_{BTBT}, for electrons and holes in the conduction and valence bands, respectively [181]. This simplifies the treatment of the BTBT and allows low computational cost. Moreover, these blocks can be calculated at a different time step than the Monte Carlo loop allowing a better optimization of the computational load.

The first block in this model corresponds to the quantum corrections for the conduction and valence bands. In MS-EMC simulators the carriers are placed into subbands and, at given x, their position distribution in z is set by the solution of the Schrödinger equation. Thus, a mapping procedure between 2D conduction and valence bands and their respective subbands is required. If this correction is not considered, the generated particles by BTBT can reach a location whose energy is lower (resp. higher) than that corresponding to the first subband in the conduction band (resp. valence band). This procedure would imply a violation of the energy conservation principle and would therefore result into significant

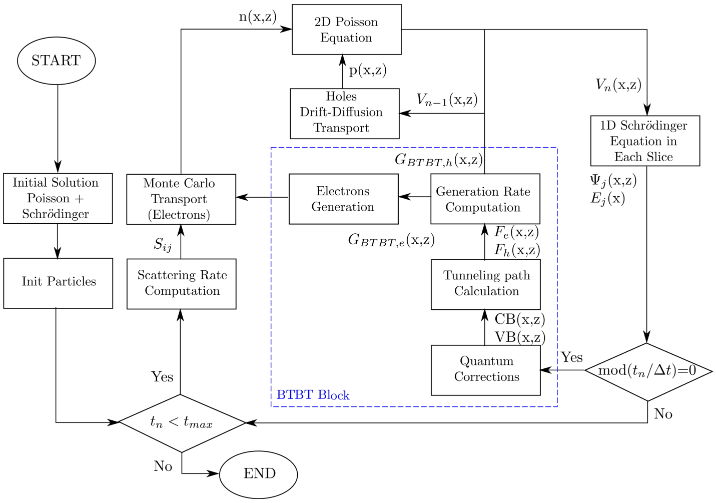


Figure 5.17: Flowchart of the MS-EMC simulator with the additional blocks of the BTBT code where x is the transport direction, z the confinement direction, $n(x, z)$ and $p(x, z)$ are the electron and hole concentrations, respectively, $V(x, z)$ is the potential profile, $E_j(x)$ is the subband energy, $\Psi_j(x, z)$ are the subband eigenfunctions, S_{ij} are the scattering rates, subscript n stands for the iteration number, Δt is the time step where BTBT is calculated, $F_e(x, z)$ and $F_h(x, z)$ are the electric fields of electrons and holes associated to the selected tunneling path, $G_{\mathrm{BTBT},e}(x, z)$ and $G_{\mathrm{BTBT},h}(x, z)$ are the electron and hole generation rates, respectively.

errors. So as to guarantee that BTBT takes place between first bound states, both simulators need to include a band profile modification algorithm allowing the reshaped conduction and valence bands to match their first subbands, E_{e1} and E_{h1}, respectively. Figure 5.18 illustrates a particular example of band modification obtained from the MS-EMC for a vertical cut taken at $X = -10$nm.

Let us now describe the procedure to do so. Since the 1D Schrödinger equation is solved for electrons, their first subband, E_{e1}, is known. Then, at every vertical slice (fixed x), for those points in the conduction band verifying $E_{\mathrm{C}}(z) \leq E_{e1}$, we set $E_{\mathrm{C}}(z) = E_{e1}$. As for the hole treatment, the valence band profile can be accurately approximated along z by a parabolic well as inferred from Figure 5.18. A rectangular well approximation can also be chosen but the parabolic one is known to better adjust the valence band profile [194]. Therefore, for fixed x, the modified valence band would read as

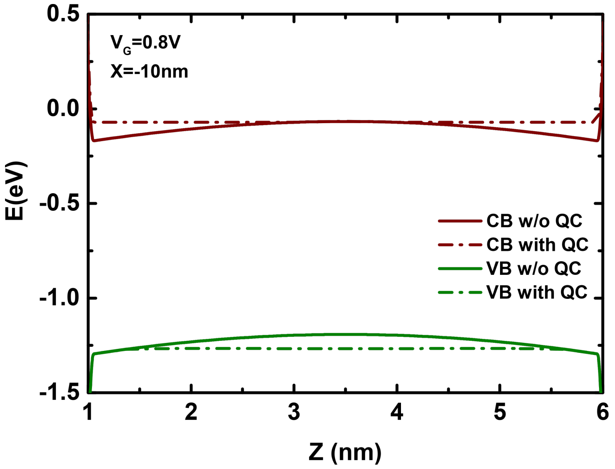


Figure 5.18: Quantum corrections included in the MS-EMC for the conduction and valence bands at $V_{\mathrm{GS}} = 0.8$V, $V_{\mathrm{DS}} = 1$V and $X = -10$nm. Recall that $X = 0$nm corresponds to the center of the device.

$$E_{\mathrm{V,modif}}(z) = E_{\mathrm{V}}(z) - \Delta E_{h1}(z), \tag{5.3}$$

where $\Delta E_{h1}(z)$ is a position dependent quantity that makes use of the analytical resolution of the parabolic well profile as

$$\Delta E_{h1}(z) = \begin{cases} \hbar\sqrt{\frac{|k(z)|q}{m_h^*}} - \frac{1}{2}|k(z)|d^2 & \text{if } \Delta E_{h1}(z) > 0 \\ 0 & \text{if } \Delta E_{h1}(z) \leq 0 \end{cases}, \tag{5.4}$$

with q the electron charge, $k(z) = V''(z)$ and m_h^* the hole effective mass. For each vertical slice, d is the distance from the center of the parabola. Notice how this treatment for holes is more realistic than simply considering for them a rectangular well approximation [181].

Once the bands have been appropriately corrected to match their first subbands, the next step is the determination of the starting and ending points for the tunneling processes. These points are estimated by calculating the path followed by the carriers when they tunnel across the forbidden energy barrier. One on the main advantages of our MS-EMC code is that it allows dynamical determination of the 2D tunneling path according to the up-to-date electrostatic configuration at each simulation step.

Essentially, the idea of determining a certain path associated with the process of inter-band tunneling is somehow an artifact that tries to mimic in a "semiclassical way" the quantum mechanical phenomenon of traversing a potential barrier through the band gap. The estimation of these tunneling trajectories for the carriers is nothing more than an educated guess for being able to include in the MS-EMC simulator a feasible BTBT model.

Two different and acceptable assumptions (from a physical point of view) have been considered here [34] for calculating the tunneling path in the corresponding box of Figure 5.17. The first one estimates the path following the criterion of the valence band "maximum gradient trajectory" (i.e. maximum electric field, F_{max}). In this case, the tunneling path is dynamically computed based on the self-consistent potential obtained during the simulation. It is then easy to understand that for the F_{max} method, as the calculated paths vary with the applied bias, so does the G_{BTBT} distribution. The second way of determining the tunneling path follows the "minimum length" (L_{min}) criterion. If we had to privilege one of the two assumptions, we would be more tempted to tend toward the F_{max} method given that, as the tunneling region is a zone of high electric field, the valence band electrons allowed to tunnel would be more likely expected to do it in a direction equal to the force that is pushing them towards the potential barrier. For that reason, this tunneling path would be in principle more probable from a purely physical point of view. Nonetheless, as the tunneling process is not an intuitive classical event, it should not be regarded as such. Therefore, if one considers that carriers could potentially tunnel in many directions, it could be also reasonable to assume that, for example, electrons choose a tunneling path in which they jump from the valence band to the nearest equi-energetic point in the conduction band (i.e. L_{min} method).

The choice of the non-local tunneling path is very relevant in the calculation of the BTBT rates because we need it to identify the starting and ending points of the tunneling process to estimate an "a posteriori" local effective electric field (whose determination only requires those two points and the distance between them), which, in turn, is used to mimic the actual non-local scenario through the utilization of local equations (later shown). The non-locality is therefore greatly captured by this method since, by determining our tunneling path, we are dynamically accounting for the self-consistent modification of the conduction and valence bands throughout the successive Monte Carlo iterations. As an example, Figure 5.19 shows some examples of different tunneling paths estimated using the two criteria considered in our work.

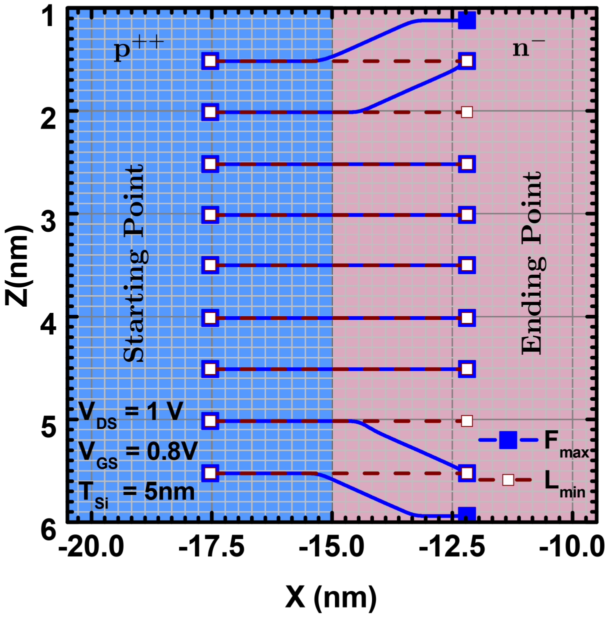


Figure 5.19: Some examples of different tunneling paths estimated using the two criteria considered in our work (F_{max} and L_{min}). Recall that $X = -15$nm corresponds to the limit between the source and the channel.

Still in the box corresponding to the tunneling path calculation, notice that two types of electric fields will need to be considered. One of them for electrons, $F_e(x, z)$, and the other one for holes, $F_h(x, z)$. It is important to highlight the difference between them, as their corresponding carriers effectively follow independent paths. In other words, two generated electrons can reach the same ending point in the conduction band, whereas their counterpart holes could have been generated at different starting points in the valence band. As a result of this, the tunneling generation rates are entirely dependent on: *i*) the local electric field at each point, *ii*) on the full tunneling barrier profile, and *iii*) on the selected tunneling path.

The next step is to obtain G_{BTBT} for electrons and holes. Our calculation is based on the Kane's model applied to the BTBT generation rate per unit of volume [195–197]:

$$G_{\mathrm{BTBT},e}(x,z) = A\left(\frac{F_e(x,z)}{F_0}\right)^P \exp\left(-\frac{B}{F_e(x,z)},\right) \tag{5.5}$$

$$G_{\mathrm{BTBT},h}(x,z) = A\left(\frac{F_h(x,z)}{F_0}\right)^P \exp\left(-\frac{B}{F_h(x,z)},\right) \tag{5.6}$$

where $F_0 = 1\mathrm{V/m}$, P=2.5 for the phonon assisted tunneling process. Equations 5.5 and 5.6 show the quasi-exponential dependence of the generation rates on the barrier width as the electric fields introduced in this equations depend mainly on the selected tunneling path. The prefactor A and the exponential factor B for indirect transitions read as [198]

$$A = \frac{g\,(m_v\,m_c)^{3/2}\,(1+2N_{TA})\,D_{TA}^2\,(qF_0)^{5/2}}{2^{21/4}h^{5/2}m_r^{5/4}\rho\,\epsilon_{TA}\,[E_g(300K)+\Delta_c]^{7/4}}, \tag{5.7}$$

$$B = \frac{2^{7/2}\pi m_r^{1/2}\,[E_g(300K)+\Delta_c]^{3/2}}{3qh}, \tag{5.8}$$

where g is a degeneracy factor, m_c and m_v are the conduction and valence band density of states effective masses, respectively, m_r is the reduced tunneling mass, N_{TA} is the occupation number of the transverse acoustic phonons at temperature T, D_{TA} is the deformation potential of transverse acoustic phonons, and ϵ_{TA} is the transverse acoustic phonon energy. The rest of the parameters takes their usual meaning. In this approach, only the transverse acoustic phonons are taken into account because they have the highest phonon occupation number and the smallest phonon energy [199].

Once the tunneling paths and G_{BTBT} are calculated, the last box inside the BTBT block of Figure 5.17 translates them into generated charges in the MS-EMC in a self-consistent way [181]. As mentioned above, a drift diffusion approach is considered to describe the hole transport as they are mainly generated in the source. They are not treated as individual particles and thus a correction in the hole concentration is required. According to Equation 4.4 in

Section 4.2 and Equation 3.6 in Section 3.1.1, the holes generated by BTBT are simply added to the corresponding drift diffusion equation by multiplying the generation rate and the time step in which BTBT module is calculated ($\Delta p_{\mathrm{BTBT}}(x, z) = G_{\mathrm{BTBT},h}(x, z) \cdot \Delta t$).

Due to the fundamentals of the free-flight of an electron in Monte Carlo algorithms, the position and energy of the superparticles are known. In consequence, a different methodology applies for electrons since a number of superparticles, N_e, is generated in the BTBT process. This N_e generation depends on the time step, Δt, the statistical weight, w, and the electron generation rate $G_{\mathrm{BTBT},e}(x, z)$ at the position of the superparticle. The procedure is as follows: first, these superparticles are generated in the fundamental subband with randomly chosen x inside the considered grid cell with a probability distribution given by $P(x)$ due to the 2D MS-EMC approximation:

$$P(x) = \frac{\displaystyle\int_0^{T_{\mathrm{Si}}} G_{\mathrm{BTBT},e}(x, z)\, dz}{\displaystyle\int_{L_x}\int_0^{T_{\mathrm{Si}}} G_{\mathrm{BTBT},e}(x, z)\, dx\, dz}. \tag{5.9}$$

Second, the number of particles is calculated taking into account the corresponding generation rate for the considered slice x_i:

$$N_e = \frac{\Delta t}{w} \int_0^{T_{\mathrm{Si}}} G_{\mathrm{BTBT},e}(x_i, z)\, dz. \tag{5.10}$$

The weight denotes the number of electrons per superparticle and, hence, it must be adequately adjusted to a suitable number of superparticles (since the higher the number of superparticles, the longer the required computation time). Accordingly, the number of these superparticles is herein calculated seeking to maximize the weight (thus minimizing the number of superparticles), instead of considering a fixed number of superparticles with very low weight (as done in [34]). As a result, the obtained computational time is reduced due to the lower number of superparticles with higher statistical weight. This approximation turns out to be admissible because the weight of a superparticle generated by BTBT is much lower than the one in the source and drain regions.

It is necessary to highlight that we only include in this integral the generation rate contribution of the selected grid cell. This is because the generated superparticle should only include the charge calculated through the integration of the generation rate of the slice in which it is going to emerge in the conduction band, otherwise it could certainly lead to an overestimation of the involved charge.

Note that the fact of determining in a certain way (either via $F_{\max}$ or $L_{\min}$) the BTBT path inside the forbidden energy barrier implies the consideration of the spatial variation of the bands across that path. This is why the methodology described in this section accounts for non-locality phenomena.

In order to optimize the computational load, the time step Δt where BTBT is calculated can be chosen independently of the Monte Carlo time step (t_n). The inclusion of new superparticles in the MS-EMC increases the simulation time because we need to solve the free-flight block of each one regardless of its weight (w). For this reason, it is worth choosing t_n so as to assign an adequate number of electrons per superparticle. Finally, a maximum tunneling rejection length, $L_{\max}$, is also introduced so that, if up to an given integration step the tunneling length of a particle turns out to be higher than $L_{\max}$, the calculation of its tunneling path stops. $L_{\max}$ has been chosen to match the channel length, $L_{\max} = L_G = 30$nm.

Apart from the blocks listed above, a robust estimator for the drain current is needed due to the low current values associated to the generated charge in comparison with the one in the source and drain regions. In MS-EMC, the drain current is generally calculated by the spatial average of the electron current along the channel. This method is known as the terminal current because it bears in mind the electron motion between the source and drain terminals. It can be extended to BTBT but only calculating the average in the generated superparticles from the point of the maximum BTBT generation to the drain. Nonetheless, it proves to be very sensitive to the choice of the averaging domain due to the noisy electron motion near the drain for the simulated device. For this reason, a new technique has been developed in this BTBT block. The drain current is calculated by computing the integral of the BTBT generation rate along the whole device. Moreover, only generation rate of the most likely slice is included in each time step in order to get an accurate result.

5.3.2.2 Methodology in the TCAD simulation scheme

Finally, the simulation results of this BTBT model have been compared with a TCAD simulation scheme in order to validate its implementation. The TCAD approach that we follow using Synopsys [198] is similar to that previously discussed and tested in recent works [190, 200]. However, the nature of the device herein analyzed allowed us to modify it in order to speed up the simulations execution. The particularity of this device is that it features band profiles easily approximated by triangular wells (for electrons in the conduction band) and parabolic wells (for holes in the valence band) as observed in Figure 5.18. Therefore, we can replace in this case the Schrödinger-Poisson solution by an analytical estimation of the subbands with great accuracy and significant saving in simulation time.

The analytical handling for holes in this TCAD approach is exactly the same as that described by Equations 5.3 and 5.4. It is included in Synopsys through the so-called physical model interface (PMI) that allows to access and modify certain models of the simulator by adequate C++ subroutines. As for electrons, an analogue technique with its corresponding subroutine can be implemented [201] for each x as

$$E_{\mathrm{C,modif}}(z) = E_{\mathrm{C}}(z) + \Delta E_{e1}(z), \tag{5.11}$$

where $\Delta E_{e1}(z)$ is a position dependent quantity that makes use of the analytical solution of the triangular well profile

$$\Delta E_{e1}(z) = \begin{cases} |a_1| \left[\frac{q^2\hbar^2 F_\perp^2(z)}{2m_e^*}\right]^{\frac{1}{3}} - d_{\mathrm{ox}}\, F_\perp(z) & \\ & \text{if } \Delta E_{e1}(z) > 0 \\ 0 & \\ & \text{if } \Delta E_{e1}(z) \le 0 \end{cases}, \tag{5.12}$$

where a_1 is the first zero of the Airy function, m_e^* is the electron effective mass, $F_\perp(z)$ is the local electric field normal to the nearest interface, and d_{ox} is the distance from the oxide interface.

Regarding BTBT, we account for it by means of the dynamic nonlocal BTBT model of Sentaurus [198] which dynamically calculates the tunneling paths based on the energy band profiles. The idea is that prior to injecting the carriers, we modify the profiles of the conduction and valence bands (making use of the aforementioned subroutines) which makes them coincident with their first bound states. By doing so, we manage BTBT to occur between first subbands, and not between band edges as it happened semiclassically in the absence of quantization effects. Other TCAD-based bandgap widening techniques can also be found in the literature [202–204].

5.3.3 Description of simulated devices and results

5.3.3.1 Tunnel Field-Effect Transistor (TFET)

The device herein analyzed is a ultrascaled silicon-based n–type TFETs. Its structure differs essentially from that of the MOSFET in the nature of the dopants used in the source and the drain as mentioned in Section 2.2. The conventional MOSFETs have the same type of dopants whereas the source and the drain have opposite types in TFETs. In general, the device has to be reverse biased and a voltage applied to the gate. For example, the source, channel, and drain are p^+, n^-, and n^+, receptively, in a n–type TFET and so a positive voltage is required in the gate and drain. Then, the BTBT phenomenon appears between the two opposite regions and its tunneling direction is mainly perpendicular. This type of BTBT is known as point tunneling because the induced electric fields and the tunneling path are opposite. In a p–type TFET, the dopings and voltages are the opposite but the operating principles are the same.

The simulated n–type TFET is schematically depicted in Figure 5.20 along with its doping concentrations and dimensions. The center of the device in the x direction corresponds to $X = 0$nm. It features an n–doped drain with $N_{\mathrm{D}} = 10^{19}\mathrm{cm}^{-3}$, a p–type doped source with $N_{\mathrm{S}} = 10^{20}\mathrm{cm}^{-3}$, and an n–doped channel with $10^{15}\mathrm{cm}^{-3}$, a gate oxide with EOT = 1nm, gate workfunctions of 4.05eV, and the body thickness has been varied from $T_{\mathrm{Si}} = 5$nm to $T_{\mathrm{Si}} = 7$nm. In addition, a set of simulations including low bias condition and saturation regime has been performed to determine the importance of the drain bias on the BTBT.

5.3.3.2 Results

As described in Subsection 5.3.2, two different tunneling trajectories have been considered to mimic in a semiclassical way the quantum mechanical phenomenon

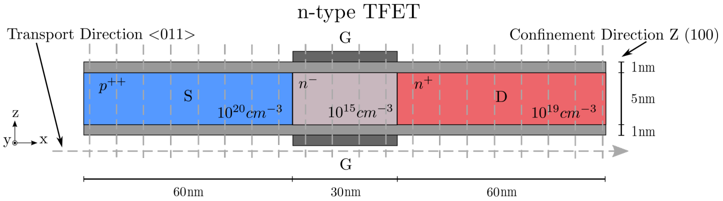


Figure 5.20: Si n-type TFET analyzed in this work. 1D Schrödinger equation is solved for each grid point in the transport direction and BTE is solved by the MC method in the transport plane.

of inter-band tunneling. Both tunneling paths (the one following the maximum gradient and the one accounting for the minimum tunneling length) are an acceptable bet for including in the MS-EMC simulator a feasible BTBT model. Accepting this premise, the different spatial distributions of the BTBT generation rates obtained from each method (F_{max} and L_{min}) and depicted in Figure5.21(a) and (b) should not be interpreted as the evidence of an inconsistency in the simulation approach, but as the proof of the conceptual difference between two independent ways of modeling the mentioned semiclassical trajectories. On the one side, it is worth noticing that the L_{min} criterion tends to distribute the electron generation uniformly inside the channel along the z-direction, whereas F_{max} provides a generation profile more concentrated towards the gate dielectrics and nearer to the source-to-channel junction. On the other side, there was almost no difference in $G_{BTBT,h}$ between both tunneling paths, L_{min} and F_{max} due to the fact that the differences between both methods as it accounts for the ending points of the paths (see Figure 5.19). For the sake of comparison, Figure 5.21(c) shows the mapping of G_{BTBT} obtained for the same biasing and from TCAD-based simulations.

One important and pertinent question arises in light of the different G_{BTBT} profiles obtained in each case: does this difference in the generation rate distributions entail a difference in the total number of carriers injected by BTBT? The answer to this question is given in Figure 5.22 where we show the total number of electrons generated in the channel by means of BTBT for the MS-EMC (considering both F_{max} and L_{min}) and for the TCAD approach. One of the main advantages of the method implemented in our MS-EMC code is that it allows to inject in the most probable slice (from Eq. 5.9) the number of generated carriers corresponding to that slice (from Eq. 5.10), and not the total number of carriers

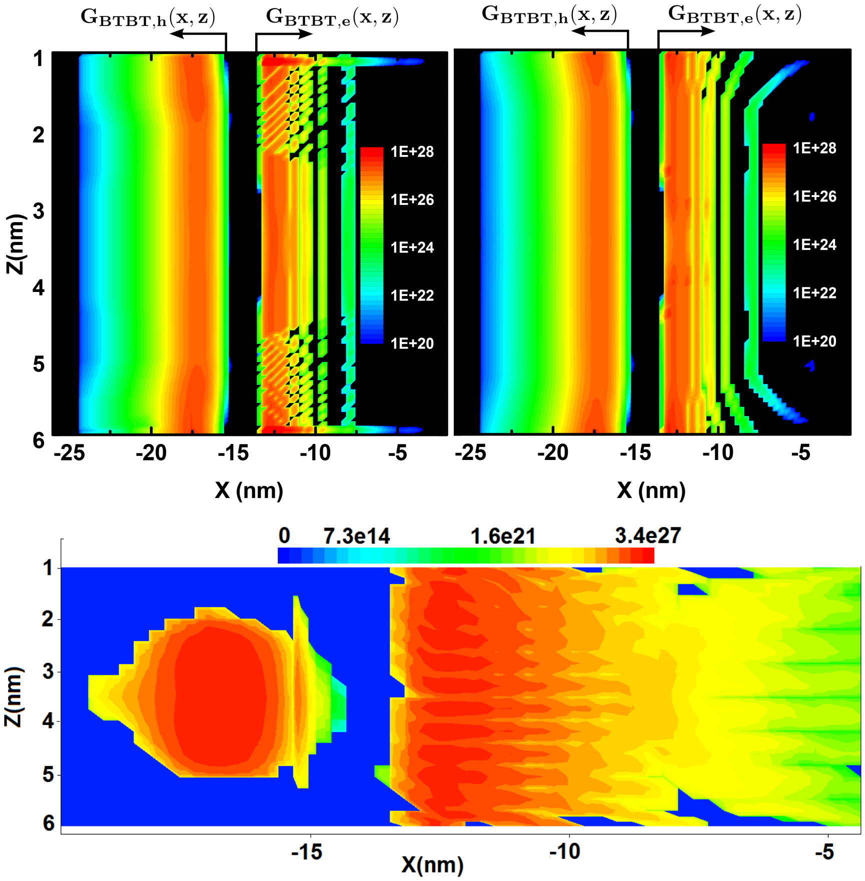


Figure 5.21: Generation rate distributions for holes (right) and electrons (left): MS-EMC with tunneling path following the maximum gradient trajectory (top), MS-EMC with minimum length tunneling path (middle), TCAD-based approach with quantum corrections included (bottom). All figures correspond to $T_{\mathrm{Si}} = 5nm$, $V_{\mathrm{GS}} = 0.8$V, $V_{\mathrm{DS}} = 1$V. The center of the device is taken as reference position with $X = 0$nm. The color scale is in cm^{-3}s^{-1}.

(as done in [181]). As a result, Figure 5.22 is not the BTBT rate multiplied by a certain time, but the average of the generated electrons. Observe that overall there is a good matching between the displayed curves suggesting that our BTBT treatment in the MS-EMC provides accurate results for this type of devices.

In general, the number of particles when F_{max} is considered is slightly higher than for the L_{min} assumption as depicted in Figure 5.22. How electrons tend to follow the maximum gradient trajectory in F_{max} and they are therefore pushed

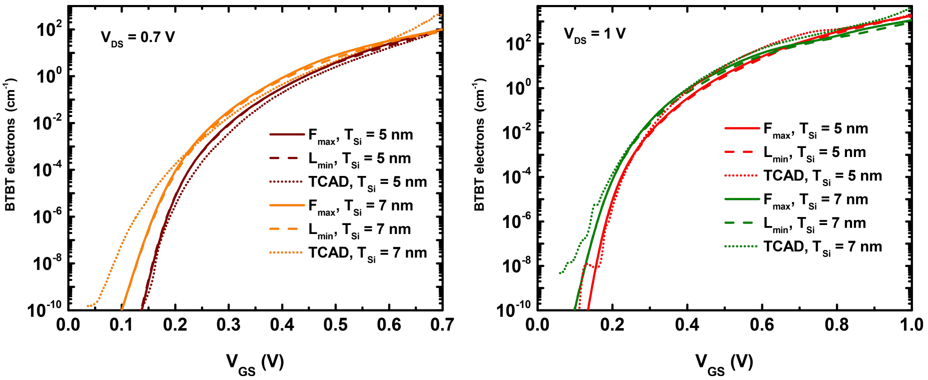


Figure 5.22: Number of electrons generated by BTBT inside the channel for both tunneling path assumptions in MS-EMC ($F_{\max}$ and $L_{\min}$) compared to the result obtained from the TCAD-based simulation approachat $V_{\mathrm{DS}} = 0.7\mathrm{V}$ (top) and $V_{\mathrm{DS}} = 1\mathrm{V}$ (bottom) conditions with $T_{\mathrm{Si}} = 5nm$ and $T_{\mathrm{Si}} = 7nm$.

into particular region, as shown in Figures 5.19 and 5.21. Accordingly, the most probability slice includes this region increasing the number of electrons. Nonetheless, electrons emerge uniformly in $L_{\min}$ and so they are more distributed along the z direction.

Moreover, as the total number of carriers injected by BTBT in this TFET device turns out to be mostly negligible in comparison with the total carrier distribution inside the channel, the tunneling path choice does not have a noticeable impact on the energy profile of the lower subband, E_{e1}. This can be observed for $V_{\mathrm{DS}} = 1\mathrm{V}$ and $V_{\mathrm{GS}} = 0.8\mathrm{V}$ in Figure 5.23 along with the corresponding subband profile obtained from TCAD simulations.

The transfer characteristics calculated from the integration of the BTBT generation rate along the selected (most probable) slice in the MS-EMC simulations are shown in Figure 5.24. Again, the comparison with the results arising from the TCAD-based approach suggests a good performance of our MS-EMC even in the low subthreshold region. This confirms the fact that the Monte Carlo method is also suitable for assessing this type of devices at very reduced current levels, which traditionally was known to be a problematic issue. Of course, the extremely reduced ON currents featured by the analyzed device are due to the fact that we are dealing with silicon. Once that the MS-EMC simulator has proven to be reliable for handling BTBT phenomena, next steps would be oriented to the

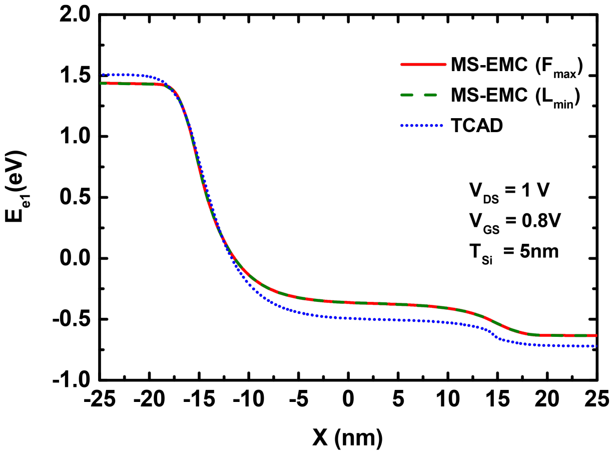


Figure 5.23: Behavior of the first subband for electrons following the x direction. As expected, the two different tunneling path criteria inside the MS-EMC provide almost identical profiles.

consideration of direct materials with lower bandgaps and alternative geometries like that of the heterogate EHBTFET or the Fin EHBTFET. Accordingly, this issue is included as a future work.

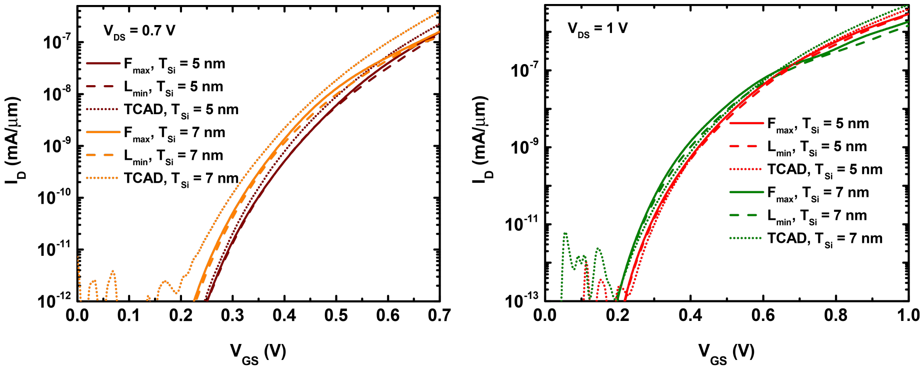


Figure 5.24: $I_{\rm D} - V_{\rm GS}$ curves obtained with the MS-EMC (for both $F_{\rm max}$ and $L_{\rm min}$ methods) compared to the one resulting from the TCAD-based approach at $V_{\rm DS} = 0.7$V (top) and $V_{\rm DS} = 1$V (bottom) with $T_{\rm Si} = 5nm$ and $T_{\rm Si} = 7nm$. Reduced ON current levels are due to the utilization of silicon as channel material.

Finally, once the device performance has been analyzed and compared to TCAD, it can be highlighted the impact on the drain current resulting from the choice of the placement in the most likely slice the whole electron charge calculated for the entire device and not the fraction of it that would correspond to that slice. Figures 5.25 and 5.26 show the number of generated electrons as well as the equivalent drain current, respectively. Observe how, as expected in the methodology, the curves for the integration of the total charge overestimate the drain current as they inject in the selected slice that total BTBT charge corresponding to the whole device and not only the fraction associated to that slice.

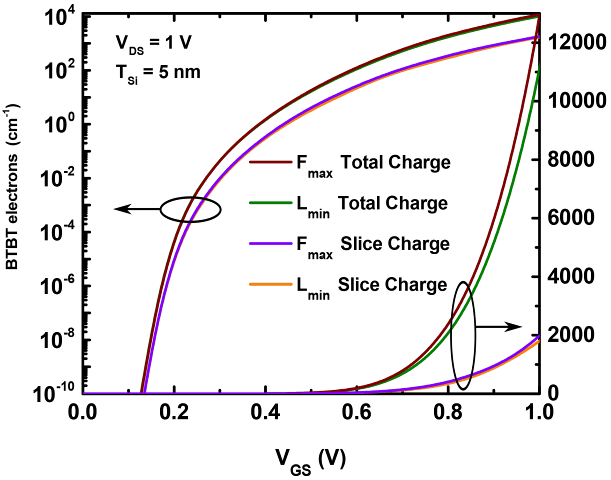


Figure 5.25: Number of electrons generated by BTBT inside the channel for both tunneling path assumptions in MS-EMC (F_{max} and L_{min}) where a assumption that considers the total integration of the generation rate (labeled as “Total Charge”) is compared to the other one that integrates the generation rate only across the selected slice (labelled as “Slice Charge”).

5.3.4 Conclusions

This section presents the successfully implementation of a non-local BTBT model inside an existing MS-EMC tool. This phenomenon has been herein considered as the working principle of some attractive devices instead of an unwanted parasitic effect in short-channel devices. In particular, TFETs are in the way to become an alternative to conventional MOSFETs due to the possibility of achieving low subthreshold swing (SS) combined with small OFF current levels which allows operation at low V_{DD}. Accordingly, we have focused on the study of ultra-

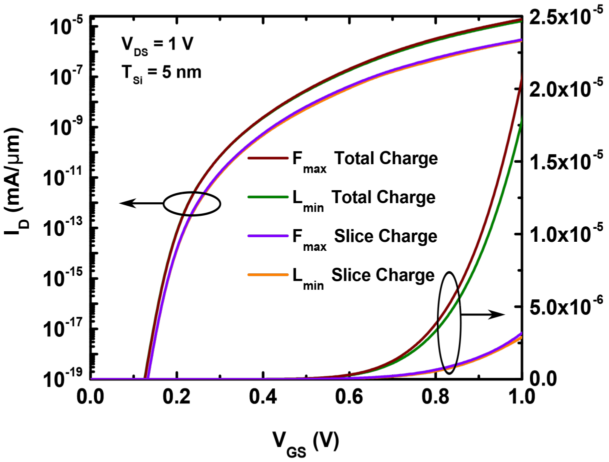


Figure 5.26: $I_{\text{DS}} - V_{\text{GS}}$ curves obtained with the MS-EMC (for both F_{max} and L_{min} methods) where a assumption that considers the total integration of the generation rate (labeled as "Total Charge") is compared to the other one that integrates the generation rate only across the selected slice (labelled as "Slice Charge").

scaled silicon-based n-type TFETs allowing a detailed scattering description and a moderate computational cost. The necessary semiclassical adaptation of the quantum mechanical process of interband tunneling led us to develop two alternative methods for defining the concept of tunneling path inside the forbidden barrier. Quantum corrections associated to subband discretization phenomena have been considered for both electrons and holes. Results from TCAD simulations including quantization effects have been used for comparison with the proposed simulation approach. In other words, we use the TCAD results in order to validate the ones given by the novel BTBT block in the MS-EMC. Moreover, one of the main results is that both L_{min} and F_{max} tunneling path assumptions are admissible in ultrascaled silicon-based n-type TFETs. In consequence, it is desirable in this specific work that the three approaches are equivalent. Despite Monte Carlo could be very time consuming when compared with a simple TCAD, TCAD must be calibrated before hand by other more physical methods such as Monte Carlo.

5.4 Gate Leakage

5.4.1 Introduction

Reducing the gate oxide thickness involves an increase in the field across the oxide. The high electric field coupled with thin oxides leads to the possibility that charge carriers can overcome the barrier for transport set up by the dielectric layer, resulting in the tunneling of carriers from substrate to gate and also from gate to substrate through the gate oxide [21]. This tunnel is known as gate oxide tunneling current or the gate leakage mechanism (GLM) and it can be divided into two categories: intrinsic and extrinsic mechanisms.

In the following section, we will focus the discussion on electron transport, since holes travel mostly in the low field source region and so its transport in the relevant material is described by a drift-diffusion approach. Moreover, if we compare both the electron tunneling from the conduction band and the hole tunneling from the valence band in the injection from Si to SiO_2, due to the higher potential barrier for holes (4.5eV) than for electrons (3.1eV), the tunneling current associated with holes is much less than the one with electrons [21,205,206].

The intrinsic mechanisms are always present, even if a dielectric film of perfect quality is assumed. The most typical one is the direct tunneling in which electrons tunnel to the gate through the energetically forbidden band gap of the dielectric material. An energy-band diagram of a MOS structure with metal gate and silicon substrate is shown in Figure 5.27 when a positive bias is applied to the gate. In case that a very thin oxide layer (less than $3-4nm$) is considered, ensuing in a small width of the potential barrier, electrons forming the inversion layer can tunnel into or through the dielectric layer and thus give rise to a gate current. Similarly, if a negative gate bias is applied, electrons from the n- metal can tunnel into or through the oxide layer (not shown in Figure 5.27).

Additionally, direct tunneling merges into a Fowler-Nordheim tunneling for high voltages because electrons tunnel into the conduction band of the oxide layer instead of tunneling directly to the gate. In general, this FN tunneling is negligible for normal device operation whereas the direct tunneling current is significant, especially for low oxide thicknesses. Another intrinsic mechanism is the Schottky emission, which is usually considered to be a thermionic emission of electrons over a reduced work-function barrier, resulting from the combined effects of image potential and an applied electric fields.

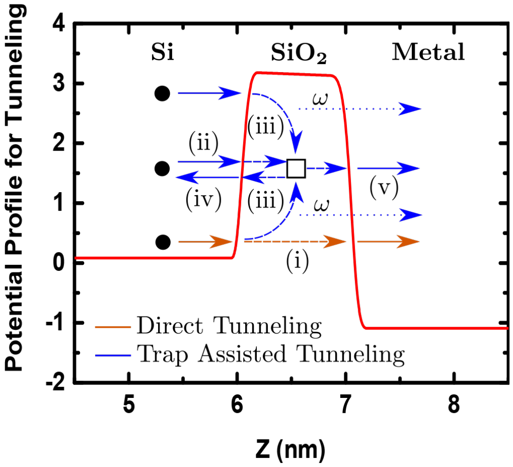


Figure 5.27: Schematic band diagram of a MOS structure with metal gate and silicon substrate where transport mechanisms implemented in the MS-EMC simulator are described: *i*) direct tunneling, *ii*) elastic and *iii*) inelastic tunneling into a trap emitting or capturing a phonon with energy ω, *iv*) detrapping to the substrate, and *v*) tunneling from the trap to the gate.

The extrinsic mechanisms are related to the existence of defect states such as elastic/inelastic tunneling of electrons into and out of defects, tunneling between defects, and field-enhanced thermal emission of electrons from defects, so called PF emission. These mechanisms are modeled as the direct tunneling but a defect state must be near the electron position and the tunneling path is calculated according to the trap position.

For the sake of simplicity, only four tunneling directions are taken into account when it comes to the implementation of the GLM in the simulation tool as sketched in Figure 5.27: the direct tunneling from the substrate to the gate, tunnel into a trap, from the trap to the substrate again, and finally from the trap to the gate. We consider elastic direct tunneling through a constant barrier of infinite extent in the two lateral dimensions that varies only in confinement direction. However, for the trap assisted tunneling both the elastic and inelastic cases are considered. In addition, the noisy nature of this mechanism due to the random number of electrons affected by the gate leakage is included. For this reason, the selection of the tunneling effect is obtained in accordance with a uniformly distributed random number and the electron position, the same way as the random position of the traps along the oxide is calculated. Thus, in the remainder of this section, a detailed discussion of this transport mechanism is given together with its sensitivity towards the device structure. This is intented to form the basis for forthcoming investigations, such as the study of the Random Telegraph Noise (RTN) and its comparison to experimental data [207, 208]. Finally, a meticulous comparative of how GLM modifies the performance of FDSOI, DGSOI, and FinFET is presented.

5.4.2 Simulation set-up

Before starting the Monte Carlo iterations, it is necessary to define the input of both the physical and simulation parameters as mentioned in Section 3.2. In consequence, some initial characteristics must be fixed in order to include the gate leakage mechanism in the simulation.

Initially, the number of traps and its location in the oxide are chosen bearing in mind its random nature. Firstly, the number of traps is deterministically calculated according to the oxide dimensions and the trap density, which depends on the material and the wafer orientation. For example, a trap density for a good quality gate oxide is between $10^{11} cm^{-2}$ and $10^{12} cm^{-2}$ when the dielectric is SiO_2 and the wafer orientation is (100).

Secondly, a uniform distributed random sequence of numbers is reckoned in order to calculate the trap location in the x and y directions, and its energy level that is usually between 2.9eV and 3.9eV below the conduction band of the SiO_2. This energy level is chosen in accordance with a random number ($r_{E_{trap}} \epsilon [2.9 - 3.9]$eV) calculated with the initial conditions and the shift of conduction band during the whole simulation. Furthermore, if a double gate device is considered, the number of traps and its random parameters are calculated independently for each gate.

Thirdly, it is indispensable to keep in mind that the MS-EMC code makes use of a 2D description, whereas an electron can be trapped only when it is located near a trap location in the oxide with 3D coordinates. Subsequently, the dimension of the trap is defined as a cube with the same sizes in the three directions and the percentage of charge that can be located in that cube is estimated. This percentage n_{perz} will be compared to a random number in the MC iterations, so that it is possible to determine the probability of finding an electron located near the trap in the x and y directions that can be set near the trap in the z direction.

Then, when the traps are totally defined, the number of particles near the dielectric is required in order to improve the computational cost, because the distance between its location and the interface modifies the tunnel probability. Due to the quantization effect, the carrier density is different from the classical prediction, since it peaks at a small distance away from the surface and not at the surface, as predicted by classical physics. This fact can be considered as an effective increase in the oxide thickness. Thus, quantization effect modulates the gate leakage current [209]. Moreover, the 2D Boltzmann equation (BTE) used in MS-EMC code characterizes the semiclassical motion of the particles in the transport direction even though its location in the confinement direction is unknown. The simulation particles are distributed along the whole device and hence the percentage of the ones near the interface (n_{intf}) is measured with respect to the total number of particles ($n(x,z)$):

$$n_{intf} = \frac{\sum_{intf}\sum_x n(x,z)}{\sum_z\sum_x n(x,z)}, \tag{5.13}$$

where $intf$ represents the region near the interface in the z direction. In this study, $intf$ is considered as the 10% of the T_{Si} for each gate. This percentage will be also compared to a random number in order to estimate if the particle is located near the interface.

The last step required before starting the Monte Carlo iterations is the calculation of the initial tunneling probabilities of each mechanism: *i*) the direct tunneling probability (T_{DT}), *ii*) the probability of electrons tunnel into a trap (T_{ST}) or *iii*) out of a trap to the substrate again (T_{TS}), and *iv*) the tunneling probability of electrons out of a trap to the gate electrode (T_{TG}). The WKB approximation is considered in order to calculate the tunneling probability of an electron as in S/D tunneling (Section 5.2). For the GLM, this transmission coefficient does not only depend on the barrier thickness, height, and structure, which, in our case, is set up by the band gap of the dielectric material [32, 33, 178, 179], but also on some specific factors of each mechanism.

i) Direct tunneling probability (T_{DT}): this assessment only considers the tunneling path starting in the interface between the substrate and the dielectric (z_{ox}), and it finishes in the interface between the dielectric and the gate contact (z_g). Then, the T_{DT} is given directly by the WKB approximation:

$$T_{DT}(E) = \exp\left\{-\frac{2}{\hbar}\int_{z_{ox}}^{z_g} \sqrt{2m_z^*(E_{CB}(x,z)-E)}\,\mathrm{d}z\right\} \tag{5.14}$$

where E and m_z^* are the energy and the confinement effective mass of the electron respectively, and $E_{CB}(x,z)$ corresponds to the energy of the conduction band in the point x, z. Additionally, a Fermi-Dirac distribution of the electrons and available states at any given energy in the gate electrodes are assumed considering that, after tunneling, the electron is going to thermalize.

ii) Probability of tunneling into a trap (T_{ST}): two important factors must be included in this calculation. In the first place, the trap occupation makes sure that the Pauli exclusion principle is considered and T_{ST} is therefore multiplied by $(1-f)$, being $f=0$, $f=0.5$, $f=1$ if the trap is empty, or it has one or two particles, respectively. Secondly, if the tunneling is inelastic, it must emit or absorb a phonon. When the particle energy is higher than the trap energy, a phonon is emitted and the probability is multiplied by $(1+n(\omega))$, being ω the difference between the particle and the trap energy, and $n(\omega)$ the Bose–Einstein factor:

$$n(\omega) = \frac{1}{\exp^{\frac{\omega}{k_\mathrm{B}T}} - 1} \tag{5.15}$$

Conversely, if the particle energy is lower than for the trap energy, a phonon is absorbed and so it is multiplied by $n(\omega)$. In summary, if the other parameters are considered in the same way as the T_{DT} and the tunneling path starts in the interface between the substrate and the dielectric (z_{ox}) and finishes in the trap location (z_{trap}), T_{ST} is defined for the emitted and absorption cases as follows:

$$T_{ST,em}(E) = (1 - f)\,(1 + n(\omega)) \exp\left\{ -\frac{2}{\hbar} \int_{z_{ox}}^{z_{trap}} \sqrt{2m_z^*(E_{CB}(x,z) - E)}\,\mathrm{d}z \right\} \tag{5.16}$$

$$T_{ST,abs}(E) = (1 - f)\,n(\omega) \exp\left\{ -\frac{2}{\hbar} \int_{z_{ox}}^{z_{trap}} \sqrt{2m_z^*(E_{CB}(x,z) - E)}\,\mathrm{d}z \right\} \tag{5.17}$$

iii) Probability of tunneling out of a trap to the substrate (T_{TS}): the WKB approximation is only multiplied by the trap occupation f. In addition, energy is transferred to a phonon via inelastic collisions excess and the electron loses memory of its previous state. For this reason, if this trapped electron tunnels again to the substrate, a new energy level must be chosen. As the carriers tend to be at the subband with lower kinetic energy, the approximation used in this mechanism calculates the subband in which the electron has lower kinetic energy. Furthermore, if the energy trap state is lower than the minimum subband, this tunnel is forbidden turning this probability into zero. Keeping in mind these hypotheses and being the rest of the parameters the same as for the direct tunneling, this probability is given by:

$$T_{TS}(E) = f \exp\left\{ -\frac{2}{\hbar} \int_{z_{trap}}^{z_{ox}} \sqrt{2m_z^*(E_{CB}(x,z) - E)}\,\mathrm{d}z \right\} \tag{5.18}$$

iv) Probability of tunneling out of a trap to the gate electrode (T_{TG}): the same assumptions as for T_{TS} are made with the difference that the particle has available states at any given energy in the gate electrode (as for the direct tunneling), the start point is the trap location (z_{trap}), and the final point is the interface between the dielectric and the gate contact (z_g).

$$T_{TG}(E) = f \exp\left\{ -\frac{2}{\hbar} \int_{z_{trap}}^{z_{g}} \sqrt{2m_z^*(E_{CB}(x,z) - E)}\,\mathrm{d}z \right\} \tag{5.19}$$

As a conclusion of the tunneling probabilities calculation, it is imperative to emphasize that the trap assisted probabilities must be calculated for each trap in the dielectric (or both dielectrics in a double gate), due to the different location of each one. In the same way, all probabilities must be recalculated when the conduction band changes, or in case that an electron is trapped or detrapped in each Monte Carlo iteration. However, GLM has a very low frequency and hence a different period that the used for MS-EMC self-consistency (Δt_{GLM}) in which the probabilities of undergoing a direct or trap assisted tunneling are checked, is defined. In that way, the particles can only undergo this mechanism according to an accurate period of occurrence instead of after each integration step. For this reason, this assumption has been considered to reduce the computational effort. Moreover, another advantage of the MS-EMC simulator is the ability to switch on and off the tunneling process, as it is included in a separate routine after each iteration.

When all the initial parameters of the system are deeply introduced, the Monte Carlo iterations begin as described in section 3.2. The fundamentals of the free-flight technique of an electron used in Monte Carlo algorithms are based on stochastic and ergodicity processes. It calculates the positions of each electron in the transport direction after a random flight time, which finishes because of the random choice of a scattering event. In particular, it is computed in the Monte Carlo Transport for electrons block shown in Figure 4.3. After each flight, the new position and transport properties of the electrons are calculated. Depending on the carrier location, its energy and the event period Δt_{GLM}, our algorithm estimates the probability of undergoing a tunnel process. Two different scenarios determined by the particle location can be considered as depicted in Figure 5.28: when it is in the channel or in the event it is trapped.

- Particle in the channel: the first step is to decide if the particle is located near the substrate-dielectric interface using a uniform distributed random number r_{ch1}. Only if $r_{ch1} < n_{intf}$, the particle can undergo direct or trap assisted tunneling. Otherwise, the particle can continue with its normal motion. Due to the comparison between the location of the particle and all the traps in the dielectric, always bearing in mind its 3D dimension, this scenario can be divided into two different sub-scenarios thanks to another

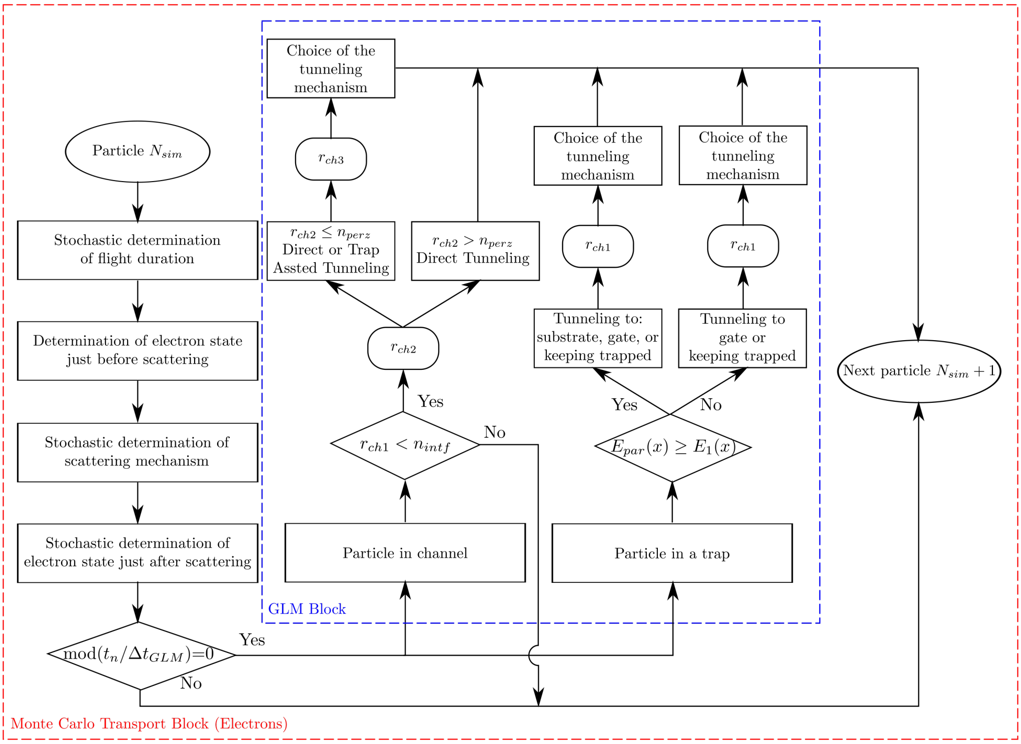

Particle N_{sim}
Stochastic determination of flight duration
Determination of electron state just before scattering
Stochastic determination of scattering mechanism
Stochastic determination of electron state just after scattering
mod($t_n/\Delta t_{GLM}$)=0
Yes
No
Choice of the tunneling mechanism
r_{ch3}
$r_{ch2} \leq n_{perz}$ Direct or Trap Assted Tunneling
$r_{ch2} > n_{perz}$ Direct Tunneling
r_{ch2}
Yes
$r_{ch1} < n_{intf}$
No
Particle in channel
Choice of the tunneling mechanism
r_{ch1}
Tunneling to: substrate, gate, or keeping trapped
Choice of the tunneling mechanism
r_{ch1}
Tunneling to gate or keeping trapped
Yes
No
$E_{par}(x) \geq E_1(x)$
Particle in a trap
GLM Block
Next particle $N_{sim}+1$
Monte Carlo Transport Block (Electrons)

Figure 5.28: Flowchart of the Monte Carlo Transport Block for electrons included in the MS-EMC simulator shown in Figure 4.3 with the additional blocks of the gate leakage mechanism (GLM) code where N_{sim} is the considered particle, subscript n stands for the iteration number, Δt_{GLM} is the time step where GLM is calculated, r_{ch1}, r_{ch2}, or r_{ch3} are uniformly distributed random numbers, n_{intf} is the percentage of particles near the interface between the substrate and the oxide, n_{perz} is the percentage of charge that can be located near a trap taking into account the 3D direction, $E_{par}(x)$ and $E_1(x)$ are the particle and the lowest subband energies in the transport direction (x), respectively.

uniform distributed random number r_{ch2}. On the one side, if $r_{ch2} \leq n_{perz}$, the particle can undergo both direct and trap assisted tunneling choosing its mechanism according to the comparison between the tunneling probabilities and a uniform distributed random number r_{ch3}. If the particle is trapped and its energy is different from the trap energy, the particle goes to the trap state emitting or absorbing a phonon. On the other side, if $r_{ch2} > n_{perz}$, it can only undergo direct tunneling through the dioxide.

- Particle in a trap: the first step is to determine whether the particle can go back to the substrate, whether it leaves the device going to the gate contact, or whether it remains in the trap. Like the above mentioned case, this choice is made by comparing tunneling probabilities and a uniform distributed random number r_{trap1}. As a result of the high substrate doping level and the large electric field at the $Si - SiO_2$ surface, the quantization of carrier energy occurs within the Si substrate. This leads to less occupied energy states from which electrons can tunnel. Subsequently, a particle with the same energy as the trap one and higher than the lower subband energy can only tunnel from the trap to the substrate. If this state is available, the particle can go back to the channel and it will continue with its normal motion. Contrarily, there are no available states and the particle will be trapped in the oxide or it will tunnel to the gate contact.

In that point, the tunneling path for all the mechanisms must be described. Due to the negligible tunneling time caused by the thin dielectric involved in the gate leakage mechanism and the low frequency of these events, it is a realistic assumption to consider the instantaneous tunnel. The electron goes directly from the starting point to the ending point on the same time step. In other words, the electron is not going to fly to the potential barrier.

It is important to highlight again that the tunneling probabilities are recalculated considering the same assumptions as for the initial calculation, both when the conduction band changes and when an electron is trapped or detrapped in each event period, Δt_{GLM}, because a change in the trap occupation modifies the rates. Apart from that, the charge trapped is dynamically keeping in mind in the 2D Poisson solution in order to preserve the self-consistency during the time simulation.

The last step in the gate leakage methodology is the calculation of the gate leakage current in proportion to the event period Δt_{GLM}, the tunneling probabilities, and the average number of particles affected by each mechanism. It can be divided into four major components, namely: direct tunneling current (I_{DT}), tunneling current of electrons into a trap (I_{ST}), tunneling of electrons out of a trap to the substrate again (I_{TS}), and tunneling current of electrons out of a trap to the gate electrode (I_{TG}). Finally, the steady-state net current through the surface resulting from the presence of a trap implies that the overall gate leakage current is calculated as the net difference of all tunneling currents passing through the substrate-dielectric interface: $I_{GL} = I_{DT} + I_{ST} - I_{TS}$.

5.4.3 Description of simulated devices and results

5.4.3.1 Device Structures

The same devices than for the S/D tunneling (FDSOI, DGSOI, and FinFET) are herein analyzed in order to compare their performance when the gate leakage mechanism is included. These devices are described in Figure 5.5 and in Tables 5.1 and 5.2. As a summary, the considered confinement direction of these devices on standard wafers changes from (100) for both planar FDSOI and DGSOI to $(0\bar{1}1)$ for FinFET, whereas the transport direction remains constant <011>. In addition, the channel thickness ranges from $T_{Si} = 3nm$ to $T_{Si} = 5nm$, the gate oxide with Equivalent Oxide Thickness is $EOT = 1nm$, and the gate work function is 4.385eV. For the FDSOI device, a Back-Plane with a $UTBOX = 10nm$, Back-Bias polarization of $V_{BB} = 0V$, and Back-Plane work function of 5.17eV have been chosen.

In this study, the gate length has remained constant as $L_G = 15nm$ and the effective mass involved in GLM is the effective confinement mass (m_z) being the higher populated valley the Δ_2 in both the FDSOI and the DGSOI and Δ_4 in the FinFET. Finally, the number of traps is calculated considering a trap density of $10^{12}cm^{-2}$ in the SiO_2 oxide and the device dimensions. However, the traps position and its energy for this particular work are set equal in the three devices, so that instead of calculating them randomly and study the variability of this mechanism, we compare how such phenomenon affects the aforementioned devices.

5.4.3.2 Results

If we consider the study of this phenomenon according to the tunnel probability for any gate leakage mechanism calculated by the WKB approximation (Equations 5.14, 5.16, 5.17, 5.18, and 5.19), it can be reduced by the longer tunneling path, the higher potential barrier, or the bigger m_z. As this phenomenon undergoes in the confinement direction, all the devices have the same potential barrier between the substrate and the gate contact caused by the dielectric. For this reason, the difference in the effective confinement mass only modifies the tunnel probability among FDSOI, DGSOI, and FinFET. Conclusively, the higher effective confinement mass for both the FDSOI and the DGSOI ($m_z = m_l = 0.916m_0$) than for the FinFET orientation ($m_z = \frac{2m_l m_t}{m_l + m_t} = 0.326m_0$) involves that the tunnel probability is lower for the planar devices than for the vertical one. However, the effective confinement mass has the same value $m_z = m_t = 0.198m_0$ for the less populated valleys; Δ_2 in FinFET and Δ_4 in FDSOI and DGSOI.

The higher tunnel probability for the FinFET than for FDSOI and DGSOI can be distinguished in Figure 5.29 where the electron distribution as a function of the transport and confinement directions taking into account the charge trapped is shown. The higher tunnel probability increases the number of particles that can undergo any gate leakage mechanism and so the charge trapped in FinFET is higher. The random nature of the trap location is also depicted in this Figure.

Moreover, the different electron distribution conforming to its location in each device must be highlighted. On one side, if both double gate devices are compared, the total charge near the interface between the substrate and the oxide in Figure 5.29 is lower for the FinFET than for the DGSOI. On the other side, if both planar devices are analyzed, the inclusion of an additional gate increases the electron confinement.

A meticulous study of how the different electron distribution modifies the probability of undergoing GLM is depicted in Figures 5.30 and 5.31 where different observations must be highlighted. Firstly, it is necessary to emphasize that the direct tunneling through the oxide is the dominant phenomenon in the three devices due to the small oxide thickness. Secondly, the reduction of the total charge near the interface decreases the probability of undergoing direct or trap assisted tunneling through the oxide. It therefore reduces the number of electrons affected in FinFET in comparison with the DGSOI. The same remark is also shown in FDSOI, specially at high gate bias, because of the lower electron confinement. Thirdly, the probability of a trapped electron to return to the substrate should generally be lower than the probability of tunneling to the gate contact because the electrons need available energy states in order to be detrapped to the substrate. Nonetheless, the difference in the electron confinement must modify these energy states increasing substantially the probability of

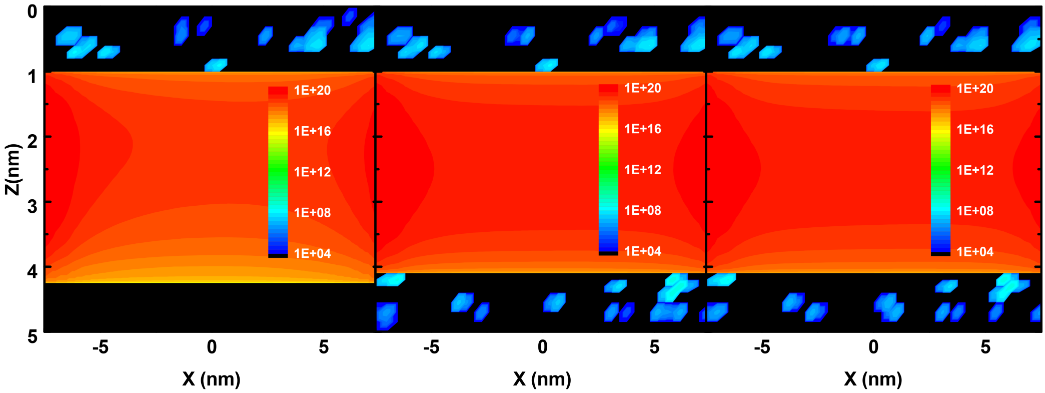


Figure 5.29: Electron distribution in cm^{-2} as a function of the transport and confinement directions, X and Z, respectively, in the 15nm device including gate leakage mechanism for FDSOI (left), DGSOI (middle), and FinFET (right) with $T_{Si} = 3nm$, $V_{GS} = 0.6V$ and $V_{DS} = 100mV$. Recall that $X = 0$nm corresponds to the center of the device.

tunneling to the substrate in FDSOI in contrast with the general behavior of the DGSOI and FinFET. Furthermore, the available states decrease as the gate bias increases until this tunnel becomes forbidden. Due to the lower charge near the interface in the FinFET, this fact starts at lower gate bias. Eventually, the slope in the total number of electrons affected by GLM as well as in the direct tunneling is higher for the FinFET. In other words, the number of particles that undergoes direct tunneling tends to be similar for the FDSOI and FinFET as the gate bias increases.

These statements can be extended for the current generated by the total gate leakage mechanism and each one individually (direct tunneling, tunnel into a trap, and dettraping to the substrate and the gate contact) as a function of V_{GS} (Figure 5.31). In addition, the trapped charge can be perfectly observable in Figure 5.31 because the flowing current from the substrate to all the traps is not the same as the sum of the currents of particles that leave the traps.

When the gate leakage is included, an important number of particles that should contribute to the thermionic current undergoes direct or trap assisted tunneling through the oxide. For this reason, the loss of charge reduces the potential barrier between the source and the drain as depicted in Figure 5.32. This

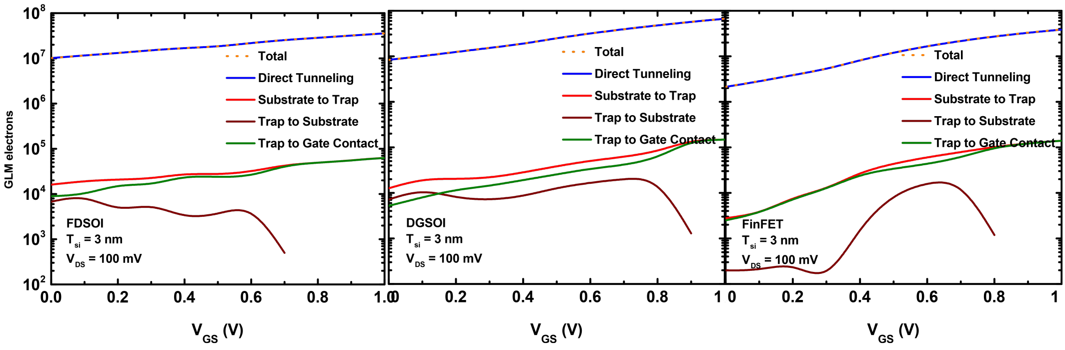


Figure 5.30: Average number of electrons in arbitrary units affected by the total gate leakage mechanism and each one individually (direct tunneling, tunnel into a trap, and dettraping to the substrate and the gate contact) as a function of V_{GS} in the 15nm device and $T_{Si} = 3nm$ at low drain bias for FDSOI (left), DGSOI (middle), and FinFET (right).

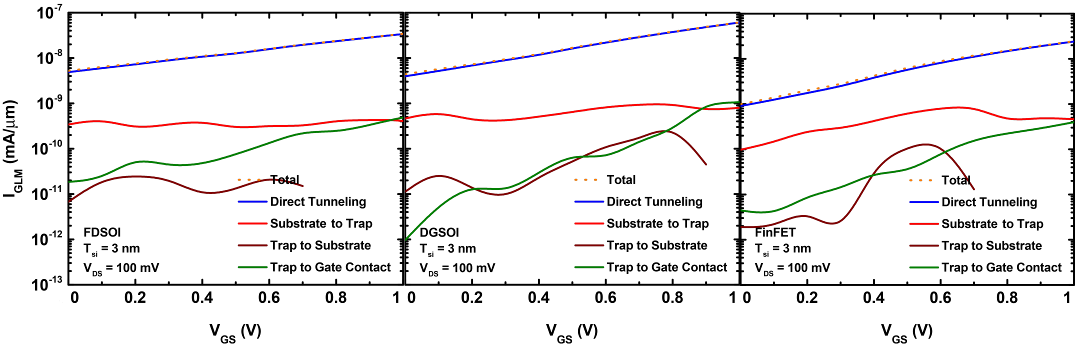


Figure 5.31: Current generated by the total gate leakage mechanism and each one individually (direct tunneling, tunnel into a trap, and dettraping to the substrate and the gate contact) as a function of V_{GS} in the 15nm device and $T_{Si} = 3nm$ at low drain bias for FDSOI (left), DGSOI (middle), and FinFET (right).

reduction is not very severe because the total lost charge is small in comparison with the total charge (Figure 5.29).

Finally, Figure 5.33 shows the impact of the gate leakage mechanism on the drain current variation ($\Delta I_D/I_{D,w/o}$) of a simulation considering GLM and another simulation without taking GLM into account. By way of explanation, the parameter $\Delta I_D/I_{D,w/o}$ represents the variation in current when this phenomenon

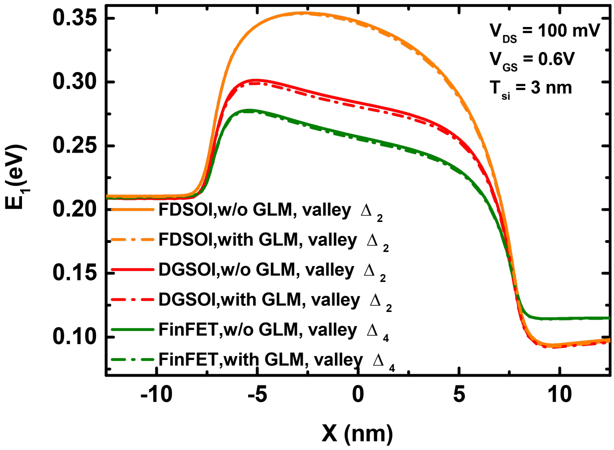


Figure 5.32: Energy profile of the lower energy subband in the 15nm device for FDSOI (valley Δ_2), DGSOI (valley Δ_2), and FinFET (valley Δ_4) with gate leakage mechanism and w/o considering it with $T_{Si} = 3nm$, $V_{GS} = 0.6V$ and $V_{DS} = 100mV$. Recall that $X = 0$nm corresponds to the center of the device.

is included in comparison with the drain current $I_{D,w/o}$ and, hence, it will be more remarkable when the variation is in the order of $I_{D,w/o}$. Two different scenarios should be highlighted. The first one is when the gate bias is lower than the threshold voltage. In this region, ΔI_D is negative because the reduction of the number of particles owing to GLM lower the current. The second one is when the gate bias is higher than the threshold voltage being ΔI_D slightly negative because the reduction in the potential barrier increases the thermionic current and so the impact of GLM in comparison with $I_{D,w/o}$ is very small. The big onset shift between both scenarios is due to the quantum confinement of the electrons near the interface. The reduction of the I_{OFF} and the low effectiveness in the I_{ON} due to this phenomenon can be advantageous if a device with the highest I_{ON}/I_{OFF} ratio is demanded.

When the channel thickness increases, the percentage of the region near the interface decreases. Therefore, the number of particles affected and the $\Delta I_D/I_{D,w/o}$ decreases. Nevertheless, the contribution of a single gate in FDSOI reduces the drain bias and so the impact of GLM on the drain current is larger in this device. In addition, $\Delta I_D/I_{D,w/o}$ is more negative for the FinFET than for the DGSOI as the gate bias increases, as predicted in Figure 5.30. When the electron confine-

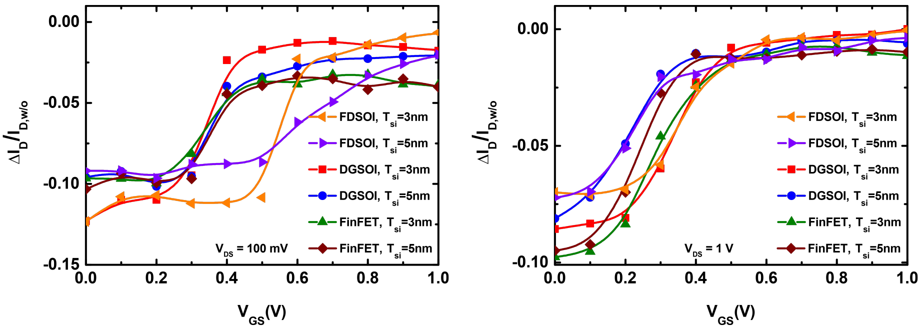


Figure 5.33: Difference between the drain current (ΔI_D) of a simulation considering gate leakage mechanism and a simulation w/o taking it into account as a function of V_{GS} in the 15nm device for FDSOI, DGSOI, and FinFET at low drain bias condition (left) and saturation (right) conditions.

ment is strong, the number of particles that can undergo GLM is similar between both devices. For this reason, the bigger m_z for the DGSOI reduces the tunnel probability.

This effect gets even less relevance when the drain bias is increased due to the higher carrier confinement. but the influence of this quantum effect is lower for $T_{Si} = 5nm$. In addition, FDSOI shows less degradation at saturation regime.

As a conclusion, if focusing on the variation of the drain current, the DGSOI is more tolerant at any drain bias. On the one hand, the current of a double gate is higher than in a single one and any variation in the drain bias is therefore less observable. On the other hand, the number of particles affected in both DGSOI and FinFET tends to be similar in regions of strong carrier confinement. Accordingly, the higher tunneling probability of the FinFET involves higher particles that undergo this phenomenon. As a result, the variation in the drain current is larger in FinFETs.

5.4.4 Conclusions

This section presents the implementation of gate leakage mechanism (GLM) including direct and trap assisted tunneling in a MS-EMC tool for the comparison of its effectiveness in ultrascaled FDSOI, DGSOI and FinFET devices. Three different assumptions of the motion of a particle when it undergoes trap assisted tunneling are also described: tunnel into a trap or out of a trap to the substrate again, and tunneling of electrons out of a trap to the gate electrode. Our calculations show that the main difference in our simulations among the three devices when this quantum effect is bearing in mind is the electron confinement near the interface. If the charge located near the interface in comparison to the total charge is larger, the probability of undergoing this phenomenon is higher. The loss of charge decreases I_{OFF} whereas it increases I_{ON} due to the reduction of the potential barrier between the source and the drain. In conclusion, the single gate FDSOI and the vertical FinFET are the devices that show less electron confinement in contrast with DGSOI, and hence the number of electrons that tunnels through the oxide is lower. However, the DGSOI shows more tolerance in terms of drain current variations at any bias regime due to the higher drain current in comparison with a single gate and the lower tunneling probability.

Chapter 6

Conclusions and Future work

6.1 Conclusions

The main aim of this PhD Thesis is the development of a Multi-subband Ensemble Monte Carlo simulator which includes different quantum effects and the assessment of different nanodevices using this tool.

In this context, the main contributions of this work are listed below:

- A broad description of the MS-EMC tool has been provided which is able to deal with transient and non-homogeneous phenomena, arbitrary orientations, different valleys and subbands, multiple gates (MuG) architectures and SOI technology. Moreover, the phonon and the surface roughness scatterings, as well as the ionized impurities have been taken into account.

- As the addition of new models are required due to the device scaling, the computational resources have been improved. The MS-EMC code has been parallelized using the OpenMP standard in order to reduce the simulation time. It has been demonstrated that, despite the transport simulation shows a dependency among the simulation parts, if the most computational cost blocks are divided into different threads for a multi core processors, the total simulation time can be substantially reduced. Furthermore, the computational cost reduction can be beneficial to the code developers and, hence, to the elaboration of novel models.

- The boundary condition has been optimized for the ohmic contacts considering that when a superparticle reaches the source or drain terminals, it exits the simulation domain being able to be reused without any matrix reorganization process. By doing this, the neutrality condition of the carriers is fulfilled because the superparticles that exit the source or drain are reused to create the new ones to be injected. This procedure also allows the use of parallel code and the optimization of the computational cost. Regarding the injection distribution, the equilibrium Maxwell-Boltzmann and the Fermi-Dirac statistics have been assumed. We demonstrated that the second one gives a more accurate picture of the distribution function because the superparticle re-injection depends on the relationship between the angle and the injection energy. As a result, the drain current is higher in the Fermi-Dirac statistics than in the Maxwell-Boltzmann ones owing to the initial motion parameters.

- A energy-dependent model for calculating the superparticle weight (the number of electrons per superparticle) has been developed in order to reduce the stochastic noise that the high energy superparticles include in the device performance. This model has been motivated because this kind of superparticles are very unlikely and a high weight can modify the dominant statistic. It has been shown that a proper weight choice according to the energy decreases the noise in the subthreshold region and allows the use of MS-EMC codes in this particular bias condition.

- The S/D tunneling has been introduced as a scaling limit owing to the threshold voltage modification. Regarding the WKB approximation, the tunnel probability of a superparticle with lower energy than the potential barrier near this position to go through it depends mainly on: the tunneling path, the potential barrier and the transport effective mass. An assessment of its impact on FDSOI, DGSOI, and FinFET has been provided being their energy profile and confinement orientation which determine the physical degradation. It has been concluded that the higher transport effective mass in the FinFET makes this architecture more undamaged facing this phenomenon, specially in the subthreshold and saturation regimes. In addition, our simulations have shown that the tunneling time can not be neglected because an instantaneous tunnel overestimates the number of particles that undergoes this effect by increasing their tunnel probability.

- The implementation of the BTBT effect has been considered in order to study a device which injection current is this mechanisms, instead of studying it as a leakage mechanism in conventional architectures. This device has been a TFET which principal advantage compared to conventional MOSFETs is the low subthreshold swing and small I_{OFF}. In particular, a non-local BTBT model has been developed combining the local equation for the generation rates, the choice of a non-local tunneling path, and the modification of the conduction and valence band according to their first available energy state. Moreover, the assessment of different approaches for the tunneling path, the tunneling path follows the maximum gradient trajectory or the minimum length, provides differences in both the current levels and the spatial distribution of the generated carriers. Finally, a comparison between these results and other results obtained from commercial TCAD simulations has shown that the model herein developed provides very accurate agreement and so it validates our assumptions.

- A model for evaluating the gate leakage mechanism (GLM) including direct and trap assisted tunneling has been implemented for the MSB-EMC tool. Three different assumptions for trap assisted tunneling have been considered: tunnel into a trap or out of a trap to the substrate again, and tunneling of electrons out of a trap to the gate electrode. In the case of direct tunneling, the tunneling probability has been calculated by using the WKB approximation as in the S/D tunneling. On the contrary, for the trap assisted tunneling, this transmission coefficient does not only depend on the dielectric barrier structure but also on some specific factors of each mechanism such as the trap occupancy or the emission or absorption of a phonon. However, our simulations have shown that the direct tunneling is the dominant phenomenon due to the small oxide thickness. Eventually, a comparison of its effectiveness in ultrascaled FDSOI, DGSOI and FinFET devices has been studied. We have demonstrated that the main difference in our simulations among the three devices when this quantum effect is bearing in mind is the electron confinement near the interface being the

DGSOI which shows more tolerance in terms of drain current variations at any bias regime.

6.2 Future Work

Some preliminary results have been reported in this document, nonetheless, the full potential of the developed simulator has to be still exploited in future research works. Accordingly, we propose the following issues as future steps to continue the study of the enhancement effects and the quantum mechanisms in the MS-EMC tool:

1. The study of the electron mobility may be completed by the inclusion of other scattering mechanism, such as for example the Remote Coulomb Scattering, the Impact Ionization, or the Heat Transport.

2. Concerning the silicon-based 2D MS-EMC simulator, three improvements can be considered: firstly, the introduction of the Monte Carlo transport for holes; secondly, the extrapolation from de 2D code to a 3D one in order to simulate 3D device architectures such as nanowires; and thirdly, the adaptation of the electrostatic and transport numerical solvers in order to include high–κ oxides and new materials.

3. Both the S/D tunneling and the strain technology modify the device characteristics, such as the potential barrier, and so the device performance. A thorough study of this effect on strained devices will be addressed.

4. We have approximated the non-local BTBT phenomenon by using local equations and non-local tunneling paths for the generation rate calculations. The WKB approximation would be useful to be included in the generation rate in order to describe the particle movement in the band gap according to the imaginary wave vector. This task should be linked with a robust estimator for the drain current calculated by the spatial average of the superparticles generated by BTBT along the channel. In that way, the current will consider the initial free-flight conditions of the BTBT superparticles.

5. Once that the MS-EMC simulator has proven to be reliable for handling BTBT phenomena in TFETs, the next steps would be oriented to the consideration of alternative geometries such as the heterogate electron–hole bilayer TFET (HG-EHBTFET). The main difference between both devices is that TFETs are based on point tunneling (gate induced electric fields and tunneling directions mainly perpendicular), whereas HG-EHBTFETs are based on line tunneling (electric fields and tunneling directions mostly aligned).

6. A motion assumption where the superparticle tunnel through the gate oxide in the GLM should be developed in order to not disregard the tunneling time inside the gate oxide.

7. The study of the quantum effects herein developed has been independently carried out. However, real devices are affected by a large number of phenomena as the device dimensions decrease. For this reason, an analysis of the simultaneous impact of the major number of quantum effects can be considered to get a more accurate solution.

List of publications

Journal papers

- Padilla, J. L., Alper, C., **Medina-Bailón, C.**, Gámiz, F., and Ionescu, A. M. . Assessment of pseudo-bilayer structures in the heterogate germanium electron-hole bilayer tunnel field-effect transistor. *Applied Physics Letters*, 106(26):262102, June 2015.

- **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Impact of non uniform strain configuration on transport properties for FD14+ devices. *Solid-State Electronics*, 115(B):232-236, January 2016.

- **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Confinement orientation effects in S/D tunneling. *Solid-State Electronics*, Available Online.

- **Medina-Bailón, C.**, Padilla, J. L., Sampedro, C., Alper, C., Gámiz, F., and Ionescu, A.M. Implementation of Band-to-Band Tunneling Phenomena in Multi-Subband-Ensemble Monte Carlo simulator: Application to Silicon TFETs. *Transactions on Electron Devices*, Submitted.

Conference contributions

- **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Impact of non uniform strain configuration on transport properties for FD14+ devices. *2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS).*

- Diaz-Llorente, C., **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Sub-22nm scaling of UTB2SOI devices for Multi-V_T applications. *2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS).*

- **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Impact of S/D tunneling in ultrascaled devices, a Multi-Subband Ensemble Monte Carlo study. *2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD).*

- **Medina-Bailón, C.**, Sampedro, C., Gámiz, F., Godoy, A., and Donetti, L. Confinement orientation effects in S/D tunneling. *2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS).*

- **Medina-Bailón, C.**, Sampedro, C., Padilla, J. L., Gámiz, F., Godoy, A., and Donetti, L. Multi-subband ensemble Monte Carlo study of band-to-band tunneling in silicon-based TFETs. *2016 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD).*

Bibliography

[1] "The international Technology Roadmap For Semiconductors (ITRS)," http://www.itrs.net/, 2013.

[2] R. Dennard, F. Gaensslen, L. Kuhn, and H. Yu, "Desing of micron MOS switching devices," vol. 18. 1972 International Electron Device Meeting, 1972, pp. 168–170.

[3] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, pp. 256–268, October 1974.

[4] G. Baccarani, M. Wordeman, and R. Dennard, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Transactions on Electron Devices*, vol. 31, pp. 452–462, April 1984.

[5] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, L. Shih-Hsien, G. Sai-Halasz, R. Viswanathan, H.-J. Wann, S. Wind, and H. Wong, "CMOS scaling into the nanometer regime," *Proceedingd of the IEEE*, vol. 85, pp. 486–504, April 1997.

[6] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 μm MOSFET's: A 3-D ?atomistic? simulation study," *IEEE Transactions on Electron Devices*, vol. 45, pp. 2505–2513, December 1998.

[7] D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H. Wong, "Device scaling limits of Si MOSFETs and their application dependences," *Proceedingd of the IEEE*, vol. 89, pp. 259–288, March 2001.

[8] M. Fischetti, "Scaling MOSFETs to the Limit: A Physicist's Perspective," *Journal of Computational Electronics*, vol. 2, pp. 73–79, December 2003.

[9] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1837–1852, September 2003.

[10] A. Asenov, S. Kaya, and A. Brown, "Generalized scaling theory and its application to a 1/4 micrometer MOSFET design," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1254–1260, May 2003.

[11] A. Asenov, "Simulation of Statistical Variability in Nano MOSFETs." Symposium on VLSI Technology Digest of Technical Papers, June 2007, pp. 86–87.

[12] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, April 1965.

[13] H.-S. Wong, "Beyond the conventional transistor," *IBM Journal of Research and Development*, vol. 46, pp. 133–168, March 2002.

[14] K. Roy and S. Prasad, *Low-Power CMOS VLSI Circuit Design.* New York: Wiley, 2000.

[15] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, pp. 305–327, February 2003.

[16] T. Skotnicki, J. Hutchby, T.-J. King, H.-S. Wong, and F. Boeuf, "The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance," *IEEE Circuits and Devices Magazine*, vol. 21, pp. 16–26, January 2005.

[17] M. Bohr, "The Evolution of Scaling from the Homogeneous Era to the Heterogeneous Era." 2011 IEEE International Electron Devices Meeting (IEDM), 2011, pp. 1–6.

[18] K. J. Kuhn, “Considerations for ultimate CMOS scaling,” *IEEE Transactions on Electron Devices*, vol. 59, no. 7, pp. 1813–1828, 2012.

[19] A. Ionescu and H. Riel, “Tunnel field-effect transistors as energy-efficient electronic switches,” *Nature*, vol. 479, pp. 329–337, November 2011.

[20] R. Pierret, *Semiconductor Device Fundamentals.* Reading, Massachusetts: Addision-Wesley, 1996.

[21] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI devices.* New York: Cambridge University Press, 2009.

[22] C. Sampedro, F. Gámiz, A. Godoy, R. Valín, A. García-Loureiro, and F. G. Ruiz, “Multi-Subband Monte Carlo study of device orientation effects in ultra-short channel DGSOI,” *Solid-State Electronics*, vol. 54, no. 2, pp. 131–136, 2010.

[23] C. Sampedro, F. Gámiz, A. Godoy, R. Valín, A. García-Loureiro, N. Rodríguez, I. M. Tienda-Luna, F. Martinez-Carricondo, and B. Biel, “Multi-Subband Ensemble Monte Carlo simulation of bulk MOSFETs for the 32 nm-node and beyond,” *Solid-State Electronics*, vol. 65-66, no. 1, pp. 88–93, 2011.

[24] C. Sampedro, F. Gámiz, L. Donetti, and A. Godoy, “Reaching sub-32 nm nodes: ET-FDSOI and BOX optimization,” *Solid-State Electronics*, vol. 70, pp. 101–105, 2012.

[25] C. Sampedro, F. Gámiz, and A. Godoy, “On the extension of ET-FDSOI roadmap for 22 nm node and beyond,” *Solid-State Electronics*, vol. 90, pp. 23–27, 2013.

[26] M. V. Fischetti and S. Laux, “Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects,” *Physical Review B*, vol. 38, no. 14, pp. 9721 – 9745, 1988.

[27] C. Medina-Bailon, C. Sampedro, F. Gámiz, A. Godoy, and L. Donetti, “Impact of S / D Tunneling in Ultrascaled Devices, a Multi-Subband Ensemble Monte Carlo Study.” 2015 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2015, pp. 21–24.

[28] ——, “Confinement orientation effects in S / D tunneling.” 2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), 2016, pp. 100–103.

[29] ——, “Confinement orientation effects in S / D tunneling,” p. Available Online, 2016.

[30] J. W. J. Wang and M. Lundstrom, “Does source-to-drain tunneling limit the ultimate scaling of MOSFETs?” *Digest. International Electron Devices Meeting,*, pp. 707–710, 2002.

[31] I. Hiroshi, “Future of nano CMOS technology,” *Solid-State Electronics*, vol. 112, pp. 56–67, March 2015.

[32] D. Bohm, *Quantum Theory.* New Jersey: Prentice-Hall, 1951.

[33] D. J. Griffiths, “The WKB approximation,” in *Introduction to Quantum Mechanics.* New Jersey: Prentice Hall, 1995, ch. 8, pp. 274–297.

[34] L. De Michielis, M. Iellina, P. Palestri, A. M. Ionescu, and L. Selmi, “Effect of the choice of the tunnelling path on semi-classical numerical simulations of TFET devices,” *Solid-State Electronics*, vol. 71, pp. 7–12, May 2012.

[35] R. Valín, “Paralelización y optimización de un simulador 2d monte carlo sobre arquitecturas grid y cluster: Estudio de fluctuaciones en transistores mosfet basados en soi,” Ph.D. dissertation, Universidad de Santiago de Compostela, Santiago de Compostela, September 2011.

[36] J. Colinge, J. Alderman, W. Xiong, and C. Cleavelin, “Quantum-mechanical effects in trigate SOI MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 53, pp. 1131–1136, May 2006.

[37] C. Lee, A. Nazarov, I. Ferain, N. D. Akhavan, R. Yan, P. Razavi, R. Yu, R. Doria, and J. Colinge, “Low subthreshold slope in junctionless multigate transistors,” *Journal of Applied Physics*, vol. 96, no. 102106, March 2010.

[38] I. Ferain, C. Colinge, and J. Colinge, “Multigate transistors as the future of classical metal-oxide-semiconductor field-effect transistors,” *Nature*, vol. 479, pp. 310–316, November 2011.

[39] C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, and B. Ricco, "Scaling the MOS Transistor Below 0.1 μ m: Methodology, Device Structures, and Technology Requirements," *IEEE Transactions on Electron Devices*, vol. 41, pp. 941–951, June 1994.

[40] P. Packan, S. Akbar, M. Armstrong, D. Bergstrom, M. Brazier, H. Deshpande, and et al., "High performance 32nm logic technology featuring 2nd generation high–κ + metal gate transistors." 2009 IEEE International Electron Devices Meeting (IEDM), 2009, pp. 1 – 4.

[41] G. Ye, P.D.and Wilk, J. Kwo, B. Yang, H.-J. Gossmann, M. Frei, S. Chu, J. Mannaerts, M. Sergent, M. Hong, K. Ng, and J. Bude, "GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition," *Electron Device Letters, IEEE*, vol. 24, pp. 209–211, April 2003.

[42] S. Takagi and A. Toriumi, "Quantitative understanding of inversion-layer capacitance in Si MOSFETs," *IEEE Transactions on Electron Devices*, vol. 42, pp. 2125–2130, December 1995.

[43] M. Lundstrom and Z. Ren, "Essential physics of carrier transport in nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, pp. 133–141, January 2002.

[44] R. Kotlyar, M. D. Giles, S. Cea, T. Linton, L. Shifren, C. Weber, and M. Stettler, "Modeling the effects of applied stress and wafer orientation in silicon devices: From long channel mobility physics to short channel performance," *Journal of Computational Electronics*, vol. 8, pp. 110–123, June 2009.

[45] S. Takagi and M. Takenaka, "III-V/Ge CMOS technologies on Si platform." 2010 Symposium on VLSI Technology (VLSIT), June 2010, pp. 147–148.

[46] H. Shang, J. Chu, S. Bedell, E. Gusev, P. Jamison, Y. Zhang, J. Ott, M. Copel, D. Sadana, K. Guarini, and M. Ieong, "Selectively formed high mobiity strained Ge PMOSFETs for high performance CMOS." 2004 IEEE International Electron Devices Meeting (IEDM), December 2004, pp. 157–160.

[47] J. del Álamo, "Nanometre-scale electronics with III-V compound semiconductors," *Nature*, vol. 479, pp. 317–323, November 2011.

[48] S. Datta, "III-V field-effect transistors for low power digital logic applications," *Microelectronic Engineering*, vol. 84, pp. 2133–2137, September–October 2007.

[49] Y. Sun, S. Koester, E. Kiewra, J. de Souza, N. Ruiz, J. Bucchignano, A. Callegari, K. Fogel, J. Fompeyrine, K. Sadana, D. Webb, P. Locquet, M. Sousa, and R. Germann, "Post-Si CMOS: III-V n-MOSFETs with high–κ gate dielectrics." Proceedings of the 2007 Compound Semiconductor Integrated Circuit Symposium, May 2002, pp. 231–234.

[50] S. Deleonibus, F. Andrieu, P. Batude, X. Jehl, F. Martin, F. Milesi, S. Morvan, F. Nemouchi, M. Sanquer, and M. Vinet, "Future micro/nanoelectronics: Towards full 3D and zero variability." 2013 13th International Workshop on Junction Technology (IWJT), 2013, pp. 1 – 5.

[51] L. Grenouillet, Q. Liu, R. Wacquez, P. Morin, N. Loubet, and D. Cooper, "UTBB FDSOI scaling enablers for the 10 nm node." 2013 IEEE SOI-3D-subthreshold microelectronics technology unified conference (S3S), 2013, pp. 1 – 2.

[52] O. Faynot, F. Andrieu, C. Fenouillet-Beranger, O. Weber, P. Perreau, and L. Tosti, "Planar FDSOI technology for sub 22 nm nodes." 2010 International symposium on VLSI technology systems and applications (VLSI-TSA, 2010, pp. 26 – 27.

[53] H. Hall, J. Bardeen, and G. Pearson, "The effects of pressure and temperature on the resistance of pn junctions in germanium," *Physical Review*, vol. 84, no. 1, pp. 129 – 132, 1951.

[54] C. Smith, "Piezoresistance effect in germanium and silicon," *Physical Review*, vol. 94, no. 1, pp. 42 – 49, 1954.

[55] S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, and M. Alavi, "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 /spl mu/m/sup 2/

SRAM cell." 2002 IEEE International Electron Devices Meeting (IEDM), 2002.

[56] Y. Sun, S. E. Thompson, and T. Nishida, "Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors," *Journal of Applied Physics*, vol. 101, no. 10, p. 104503, 2007.

[57] ——, *Strain Effect in Semiconductors*. Boston, MA: Springer US, 2010.

[58] F. Gámiz, L. Donetti, N. Rodriguez, C. Sampedro, O. Faynot, J. C. Barbe, and A. Biaxial, "Combined effect of mechanical stressors and channel orientation on mobility in FDSOI n and p MOSFETs." 2012 IEEE International SOI Conference (SOI), 2012, pp. 31–32.

[59] C. Medina-Bailon, C. Sampedro, F. Gámiz, A. Godoy, and L. Donetti, "Impact of non uniform strain configuration on transport properties for FD14 + devices," *Solid-State Electronics*, 2015.

[60] G. C. F. Yeap, S. Krishnan, and M.-R. Lin, "Fringing-induced barrier lowering (FIBL) in sub-100 nm MOSFETs with high–κ gate dielectrics," *Electronics Letters*, vol. 34, pp. 1150 – 1152, 2007.

[61] M. M. Frank, "High-κ/metal gate innovations enabling continued CMOS scaling." Proceedings of the ESSCIRC 2011, 2011, pp. 50–58.

[62] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, and et al., "A 45nm logic technology with high–κ + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193nm dry patterning, and 100 % Pb-free packaging." 2007 IEEE International Electron Devices Meeting (IEDM), 2007, pp. 247 – 250.

[63] T. Koyanagi, K. Tachi, K. Okamoto, K. Kakushima, P. Ahmet, K. Tsutsui, N. Sugii, T. Hattori, and H. Iwai, "Electrical Characterization of La_2O_3-Gated Metal Oxide Semiconductor Field Effect Transistor with Mg Incorporation," *Japanese Journal of Applied Physics*, vol. 48, no. 5S1, 2009.

[64] M. Alles, "Thin film SOI emerges," *IEEE Spectrum*, vol. 34, pp. 37 – 45, 1997.

[65] S. Cristoloveanu, D. Munteanu, and M. S. T. Liu, “A review of the pseudo-MOS transistor in SOI wafers: operation, parameter extraction, and applications,” *IEEE Transactions on Electron Devices*, vol. 47, pp. 1018 – 1027, 2000.

[66] G. K. Celler and S. Cristoloveanu, “Frontiers of silicon-on-insulator,” *Journal of Applied Physics*, vol. 93, no. 9, pp. 4955 – 4978, 2003.

[67] G. G. Shahidi, “Soi technology for the GHz era.” 2001 International Symposium on VLSI Technology, Systems, and Applications, 2002, pp. 11 – 14.

[68] K. Oshimaa, S. Cristoloveanua, B. Guillaumotd, H. Iwaic, and S. Deleonibusb, “Advanced SOI MOSFETs with buried alumina and ground plane: self-heating and short-channel effects,” *Solid-State Electronics*, vol. 48, pp. 907 – 917, 2004.

[69] “SOITEC: revolutionary semiconductor materials for energy and electronics,” http://www.soitec.com/en/index.php, 2014.

[70] F. Gámiz, “Electron mobility in extremely thin single-gate silicon-on-insulator inversion layers,” *Journal of Applied Physics*, vol. 86, no. 6269, 1999.

[71] J. P. Colinge, “Multiple-gate SOI MOSFETs,” *Solid-State Electronics*, vol. 48, pp. 897 –? 905, 2004.

[72] P. Ramm, A. Klumpp, J. Weber, and M. M. V. Taklo, “3D Systems-on-Chip Technology for More than Moore systems,” *Microsystem Technologies*, vol. 16, pp. 1051 – 1055, 2010.

[73] D. Sun and J. G. Fossum, “Dynamic floating-body instabilities in partially depleted SOI CMOS circuits.” 1994 IEEE International Electron Devices Meeting (IEDM), 1994, pp. 661 – 664.

[74] J. P. Colinge, *Silicon-On-Insulator Technology: Materials to VLSI*, 3rd ed. Kluwer Academic Press, 2004.

[75] S. Thompson, P. Packman, and M. Bohr, “MOS scaling: Transistor challanges for the 21st century,” *Intel Technology Journal*, vol. Q-3, pp. 1 – 19, 1998.

[76] H. K. Lim and J. G. Fossum, “Threshold voltage of thin-film silicon-on-insulator (SOI) MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 30, pp. 1244 – 1251, 1983.

[77] J. Colinge, “Transconductance of silicon-on-insulator MOSFETs,” *IEEE Electron Device Letters*, vol. 6, no. 11, pp. 573–574, 1985.

[78] O. Faynot, F. Andrieu, O. Weber, C. Fenouillet-Beranger, P. Perreau, J. Manzurier, T. Benoist, L. Tosti, and et al., “Planar fully depleted SOI technology: A powerful architecture for sub 20 nm node and beyond.” 2010 IEEE International Electron Devices Meeting (IEDM), 2010, pp. 50 – 53.

[79] S. Borkar, “Circuit techniques for subthreshold leakage avoidance, control and tolerance.” 2004 IEEE International Electron Devices Meeting (IEDM), 2004, pp. 421–424.

[80] J. Manzurier, O. Weber, F. Allain, L. Tosti, L. Brevard, O. Faynot, and et al., “Drain current variability and MOSFET parameters correlations in planar FDSOI technology.” 2011 IEEE International Electron Devices Meeting (IEDM), 2011, pp. 366 – 369.

[81] T. J. K. Liu and L. Chang, *Into the Nano Era*. New York: Springer-Verlag, 2009.

[82] S. Cristoloveanu, “How many gates do we need in a transistor?” vol. 1. 2007 International Semiconductor Conference, 2007, pp. 3–10.

[83] M. Jurczak, N. Collaert, A. Veloso, T. Hoffmann, and S. Biesemans, “Review of FinFET technology.” 2009 IEEE International SOI Conference, 2009, pp. 1 – 4.

[84] H. Kawasaki, V. S. Basker, T. Yamashita, C. H. Lin, Y. Zhu, J. Faltermeier, S. Schmitz, J. Cummings, S. Kanakasabapathy, H. Adhikari, H. Jagannathan, A. Kumar, K. Maitra, J. Wang, C. C. Yeh, C. Wang, and et al.,

"Challenges and solutions of FinFET integration in a SRAM cell and a logic circuit for 22 nm node and beyond." 2009 IEEE International Electron Devices Meeting (IEDM), 2009, pp. 289 – 292.

[85] D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, "A fully depleted lean-channel transistor (DELTA)-a novel vertical ultra thin SOI MOSFET." 1989 IEEE International Electron Devices Meeting (IEDM), 1989, pp. 833 – 836.

[86] J. P. Colinge, "Multi-gate SOI MOSFETs," *Microelectronic Engineering*, vol. 84, pp. 2071 – 2076, 2007.

[87] K. Tachi, T. Ernst, C. Dupré, A. Hubert, S. Bécu, H. Iwai, S. Cristoloveanu, and O. Fay, "Transport optimization with width dependence of 3D-stacked GAA silicon nanowire FET with high–κ/metal gate stack." 2009 Silicon Nanoelectronics Workshop, 2009, pp. 13 – 14.

[88] S. Bangsaruntip, G. M. Cohen, A. Majumdar, Y. Zhang, S. U. Engelmann, N. C. M. Fuller, L. M. Gignac, S. Mittal, J. S. Newbury, M. Guillorn, T. Barwicz, L. Sekaric, M. M. Frank, and J. W. Sleight, "High performance and highly uniform gate-all-around silicon nanowire MOSFETs with wire size dependent scaling." 2009 IEEE International Electron Devices Meeting (IEDM), 2009, pp. 1 – 4.

[89] K. Kuhn, M. D. Giles, D. Becher, P. Kolar, K. A., R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai, "Process technology variation," *IEEE Transactions on Electron Devices*, vol. 58, pp. 2197 – 2208, 2011.

[90] K. Kuhn, "CMOS scaling for the 22nm node and beyond: device physics and technology." 2011 international symposium on VLSI technology, systems and applications (VLSI-TSA), April 2011.

[91] D. Hisamoto, W.-C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, T.-J. King, J. Bokor, and C. Hu, "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm," *IEEE Transactions on Electron Devices*, vol. 47, pp. 2320 – 2325, 2000.

[92] B. Yu, L. Chang, A. S., H. Wang, S. Bell, C. Yang, C. Tabery, C. Ho, Q. Xiang, T. King, J. Bokor, C. Hu, M. Lin, and D. Kyser, “FinFET scaling to 10 nm gate length.” 2002 IEEE International Electron Devices Meeting (IEDM), 2002, pp. 251 – 254.

[93] M. J. H. van Dal, N. Collaert, G. Doornbos, G. Vellianitis, G. Curatola, B. J. Pawlak, R. Duffy, C. Jonville, and B. Degroote, “Highly manufacturable FinFETs with sub-10nm fin width and high aspect ratio fabricated with immersion lithography.” 2007 IEEE Symposium on VLSI Technology, 2007, pp. 110 – 111.

[94] J. Kavalieros, B. Doyle, S. Datta, G. Dewey, M. Doczy, B. Jin, D. Lionberger, M. Metz, W. Rachmady, M. Radosavljevic, U. Shah, N. Zelick, and R. Chau, “Tri-Gate Transistor Architecture with High-k Gate Dielectrics, Metal Gates and Strain Engineering.” 2006 Symposium on VLSI Technology. Digest of Technical Papers, 2006, pp. 50 – 51.

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

[96] F. Balestra, S. Cristoloveanu, M. Benachir, J. Brini, and T. Elewa, “Double-gate silicon-on-insulator transistor with volume inversion: A new device with greatly enhanced performance,” *Solid-State Electronics*, vol. 8, pp. 410 – 412, 1987.

[97] S. Cristoloveanu, T. Ernst, D. Munteanu, and T. Ouisse, “Ultimate MOSFETs on SOI: Ultra thin, single gate, double gate, or ground plane,” *International Journal of High Speed Electronics and Systems*, vol. 10, pp. 217 – 230, 2000.

[98] T. Ernst, S. Cristoloveanu, G. Ghibaudo, T. Ouisse, S. Horiguchi, Y. Ono, Y. Takahashi, and K. Murase, “Ultimately thin double-gate SOI MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 50, pp. 830 – 838, 2003.

[99] B. Doyle, B. Boyanov, S. Datta, M. Doczy, S. Hareland, B. Jin, T. Kavalieros, T. Linton, R. Rios, and R. Chau, “Tri-gate fully-depleted CMOS

transistors: Fabrication, design and layout." 2003 Symposium on VLSI Technology. Digest of Technical Papers, 2003, pp. 133 – 134.

[100] B. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, T. Kavalieros, T. Linton, R. Murthy, R. Rios, and R. Chau, "High performance fully-depleted tri-gate CMOS transistors," *IEEE Electron Device Letters*, vol. 24, pp. 263 – 265, 2003.

[101] B. Lu, E. Matioli, and T. Palacions, "Tri-gate normally-off GaN power MISFET," *IEEE Electron Device Letters*, vol. 33, pp. 360 – 362, 2012.

[102] J.-T. Park, J. P. Colinge, and C. H. Diaz, "Pi-Gate SOI MOSFET," *IEEE Electron Device Letters*, vol. 22, pp. 405 – 406, 2001.

[103] F.-L. Yang, H.-Y. Chen, F.-C. Chen, C.-C. Huang, C.-Y. Chang, H.-K. Chiu, and et al., "25 nm CMOS Omega FETs." 2002 IEEE International Electron Devices Meeting (IEDM), 2002, pp. 255–258.

[104] C. Auth, C. Allen, A. Blattner, D. Bergstrom, M. Brazier, M. Bost, M. Buehler, V. Chikarmane, T. Ghani, and et al., "A 22 nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors." 2012 Symposium on VLSI Technology, 2012, pp. 131 – 132.

[105] S. Natarajan, M. Agostinelli, S. Akbar, M. Bost, and et al., "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size." 2014 IEEE International Electron Devices Meeting (IEDM), 2014, pp. 3.7.1 – 3.7.3.

[106] B. Iniguez, D. Jimenez, J. Roig, H. A. Hamid, L. F. Marsal, and J. Pallares, "Explicit continuous model for long-channel undoped surrounding gate MOSFETs," *IEEE Transactions on Electron Devices*, vol. 52, pp. 1868–1873, 2005.

[107] V. Scu.hmidt, H. Riel, S. Senz, S. Karg, W. Riess, and U. Gösele, "Realization of a Silicon Nanowire Vertical Surround?Gate Field?Effect Transistor," *Small*, vol. 2, pp. 85 – 88, 2006.

[108] S. Suk, K. H. Yeo, K. H. Cho, M. Li, Y. Yeoh, S.-Y. Lee, and et al., "High-performance twin silicon nanowire MOSFET (TSNWFET) on bulk Si wafer," *IEEE Transactions on Nanotechnology*, vol. 7, pp. 181 – 184, 2008.

[109] M. Jagadesh Kumar, M. A. Reed, G. A. J. Amaratunga, G. M. Cohen, D. B. Janes, C. M. Lieber, M. Meyyappan, L.-E. Wernersson, K. L. Wang, R. S. Chau, T. I. Kamins, M. Lundstrom, B. Yu, and C. Zhou, "Special Issue on Nanowire Transistors: Modeling, Device Design, and Technology," *IEEE Transactions on Electron Devices*, vol. 55, pp. 2813–2819, 2008.

[110] J. M. Hergenrother, D. Monroe, F. P. Klemens, A. Komblit, G. R. Weber, W. M. Mansfield, M. R. Baker, and et al., "The vertical replacement-gate (VRG) MOSFET: A 50-nm vertical MOSFET with lithography-independent gate length." 1999 IEEE International Electron Devices Meeting (IEDM), 1999, pp. 75 – 78.

[111] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, and et al., "A 90 nm high volume manufacturing logic technology featuring novel 45 nm gate length strained silicon CMOS transistors." 2003 IEEE International Electron Devices Meeting (IEDM), 2003, pp. 978 – 980.

[112] S. Suk, S.-Y. Lee, S.-M. Kim, E.-J. Yoon, M.-S. Kim, M. Li, C. W. Oh, K. H. Yeo, and et al., "High performance 5 nm radius twin silicon nanowire MOSFET (TSNWFET): Fabrication on bulk Si wafer, characteristics, and reliability." 2005 IEEE International Electron Devices Meeting (IEDM), 2005, pp. 717–720.

[113] K. H. Yeo, S. Suk, M. Li, Y.-Y. Yeoh, K. H. Cho, K.-H. Hong, S. Yun, M. S. Lee, N. Cho, and et al., "Gate-all-around (GAA) twin silicon nanowireMOSFET (TSNWFET) with 15 nm length gate and 4 nm radius nanowires." 2006 IEEE International Electron Devices Meeting (IEDM), 2006, pp. 539–542.

[114] C. Dupré, A. Hubert, S. Bécu, M. Jublot, V. Maffini-Alvaro, C. Vizioz, F. Aussenac, and et al., "15 nm-diameter 3D stacked nanowireswith in-

dependent gates operation: Φ FET." 2008 IEEE International Electron Devices Meeting (IEDM), 2008, pp. 749 – 752.

[115] M. Li, K. H. Yeo, S. D. Suk, Y. Y. Yeoh, D.-W. Kim, T. Y. Chung, K. S. Oh, and W.-S. Lee, "Sub-10 nm gate-all-around CMOS nanowire transistors on bulk Si substrate." Symposium on VLSI Technology Digest of Technical Papers, June 2009, pp. 94 – 95.

[116] S. Bangsaruntip, A. Majumdar, G. M. Cohen, S. U. Engelmann, Y. Zhang, M. Guillorn, L. M. Gignac, and et al., "Sub-10 nm gate-all-around CMOS nanowire transistors on bulk Si substrate." Symposium on VLSI Technology Digest of Technical Papers, June 2010, pp. 21 – 22.

[117] U. Ravaioli, "Hierarchy of simulation approaches for hot carrier transport in deep submicron devices," *Semiconductor Science and Technology*, vol. 13, no. 1, pp. 1 – 10, 1998.

[118] S. Selberherr, *Analysis and Simulation of Semiconductor Devices.* New York: Springer-Verlag, 1984.

[119] Y. Zhao, J. Watling, S. Kaya, A. Asenov, and J. Barker, "Drift diffusion and hydronynamic simulations of Si/SiGe p-MOSFETs," *Materials Science and Engineering: B*, vol. 72, pp. 180 – 183, 2000.

[120] A. García-Loureiro, K. Kalna, and A. Asenov, "Efficent three-dimensional parallel simulations of PHEMTs," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 15, pp. 327 – 340, 2005.

[121] A. García-Loureiro, M. Aldegunde, N. Seoane, K. Kalna, and A. Asenov, "3D Drift-Diffusion Simulation with Quantum-Corrections of Tri-Gate MOSFETs." Spanish Conference on Electron Devices, 2009. CDE 2009., February 2009, pp. 200 – 203.

[122] W. Fawcett, A. D. Boardman, and S. Swain, "Monte Carlo determination of electron transport properties in gallium arsenide," *Journal of Physics and Chemistry of Solids*, vol. 31, no. 9, pp. 1963 – 1990, 1963.

[123] R. W. Hockney and J. W. Eastwood, *Numerical simulations using particles.* New York: McGraw – Hill, 1981.

[124] C. Jacoboni and L. Reggiani, “The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials,” *Reviews of Modern Physics*, vol. 55, no. 3, pp. 645 – 705, 1983.

[125] P. L. C. Jacoboni, *The Monte Carlo method for semiconductor device simulation.* Vienna: Springer-Verlag, 1989.

[126] K. Hess, *Monte Carlo Device Simulation: Full Band and Beyond.* Norwell, MA: Kluwer, 1991.

[127] M. V. Fischetti and S. Laux, “Monte Carlo study of electron transport in silicon inversion layers,” *Physical Review B*, vol. 48, no. 4, pp. 2244 – 2274, 1993.

[128] K. Tomizawa, *Numerical Simulation of Submicron Semiconductor devices.* Boston and London: Artech House, 1993.

[129] F. Gámiz, “Monte Carlo simulations of electron transport properties in extremely thin SOI MOSFETs,” *IEEE Transactions on Electron Devices*, vol. ED-45, no. 1122, 1998.

[130] F. Gámiz and M.V.Fischetti, “Monte Carlo simulation of double gate silicon on insulator inversion layers: The role of volume inversion,” *Journal of Applied Physics*, vol. 89, no. 5478, 2001.

[131] G. Formicone, M. Saraniti, D. Vasileska, and D. Ferry, “Study of a 50 nm nMOSFET by ensemble Monte Carlo simulation including a new approach to surface roughness and impurity scattering in the Si inversion layer,” *IEEE Transactions on Electron Devices*, vol. 49, no. 1, pp. 125–132, 2002.

[132] F. Bufler, Y. Asahi, H. Yoshimura, C. Zechner, A. Schenk, and W. Fichtner, “Monte Carlo simulation and measurement of a nanoscale n-MOSFETs,” *IEEE Transactions on Electron Devices*, vol. 50, no. 2, pp. 418–424, 2003.

[133] e. a. C. Sampedro, F. Gámiz, “Monte Carlo simulation of double gate silicon on insulator devices operated as velocity modulation transistors,” *Appl.Phys.Lett.*, vol. 86, no. 202115, 2005.

[134] ——, “The multivalley effective conduction band-edge method for Monte Carlo simulation of nanoscale structures,” *Electron Dev*, vol. 53, no. 2703, 2006.

[135] S. Datta, “Nanoscale device modelling: the Green’s function method,” *Superlattices and Microstructures*, vol. 24, pp. 253 – 278, 2000.

[136] A. Martinez, A. Svizhenko, M. Anantram, J. Barker, A. Brown, and A. Asenov, “A study of the effect of the interface roughness on a DG–MOSEFT using a full 2D NEGF technique.” 2005 IEEE International Electron Devices Meeting (IEDM), December 2005, pp. 613 – 616.

[137] A. Martinez, N. Seoane, A. Brown, J. Barker, and A. Asenov, “Variability in Si nanowire MOSFETs due to the combined effect of interface roughness and random dopants: A fully three-dimensional NEGF simulation study,” *IEEE Transactions on Electron Devices*, vol. 57, no. 7, pp. 1626–1635, 2010.

[138] M. G. Ancona and G. J. Iafrate, “Quantum correction to the equation of state of an electron gas in a semiconductor,” *IEEE Transactions on Electron Devices*, vol. 39, no. 13, pp. 9536–9540, 1989.

[139] A. Wettstein, A. Schenk, and W. Fichtner, “Quantum device-simulation with the density-gradient model on unstructured grids,” *IEEE Transactions on Electron Devices*, vol. 48, no. 2, pp. 279–289, 2001.

[140] D. K. Ferry, R. Akis, and D. Vasileska, “Quantum effects in MOSFETs: Use of an effective potential in 3D Monte Carlo simulation of ultra-short channel devices.” 2000 IEEE International Electron Devices Meeting (IEDM), December 2000, pp. 287 – 290.

[141] C. Cercignani, *Theory and application of the Boltzmann equation.* Scottish Academic Press, 1975.

[142] L. E. Reichl and I. Prigogine, *A modern course in statistical physics.* Austin: University of Texas press, 1980, vol. 71.

[143] H. F. Budd, “Hot carriers and the path variable method,” vol. 21, no. 1966. Proceedings of International Conference on the Physics of Semiconductors, J. Phys. Soc. Japan, Kyoto, January 1966, pp. 420 – 423.

[144] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods.* London: Springer, 1964.

[145] N. P. Buslenko, D. I. Golenko, Y. A. Shreider, I. M. Sobol, and V. G. Sragovich, *The Monte Carlo method: the method of statistical trials.* Pergamon, 1966, vol. 87.

[146] L. I. Schiff, *Quantum Mechanics.* New York: McGraw-Hill, 1968.

[147] L. S. D. Esseni, P. Palestri, *Nanoscale MOS Transistors.* New York: Cambridge University Press, 2011.

[148] N. W. Ashcroft and N. D. Mermin, *Introduction to Solid State Physics.* Philadelphia: Saunders College, 1976.

[149] C. Kittel, *Introduction to Solid State Physics*, 1995th ed. Pennsylvania: John Wiley $\wedge$ Sons, 1995.

[150] H. Brooks, "Scattering by ionized impurities in semiconductors," *Physical Review*, vol. 83, no. 4, pp. 83 – 879, 1951.

[151] C. Sampedro, F. Gámiz, A. Godoy, M. Prunnila, and J. Ahopelto, "A comprehensive study of carrier velocity modulation in DGSOI transistors," *Solid–State Electronics*, vol. 49, pp. 1504–1509, September 2005.

[152] K. Fuchs, "The conductivity of thin metallic films according to the electron theory of metals," vol. 34, no. 1. Mathematical Proceedings of the Cambridge Philosophical Society, January 1938, pp. 100 – 108.

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

[154] F. Gamiz, J. B. Roldan, J. A. Lopez-Villanueva, P. Cartujo-Cassinello, and J. E. Carceller, "Surface roughness at the Si-SiO_2 interfaces in fully depleted silicon-on-insulator inversion layers," *Journal of Applied Physics*, vol. 86, no. 12, pp. 6854 – 6863, 1999.

[155] G. Baccarani, C. Jacoboni, and A. Mazzone, "Current transport in narrow-base transistors," *Solid–State Electronics*, vol. 20, pp. 5–10, January 1977.

[156] J. Zimmermann and E. Constant, "Application of Monte Carlo techniques to hot carrier diffusion noise calculation in unipolar semiconducting components," *Solid–State Electronics*, vol. 23, pp. 915–925, September 1980.

[157] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, "Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches," *Journal of Applied Physics*, vol. 92, no. 7, pp. 3730–3739, 2002.

[158] I. P. Christiansen, "Numerical simulation of hydrodynamics by the method of point vortices," *Journal of Computational Physics*, vol. 13, no. 3, pp. 363 – 379, 1973.

[159] G. R. Baker, "The ?cloud in cell? technique applied to the roll up of vortex sheets," *Journal of Computational Physics*, vol. 31, no. 1, pp. 76 – 95, 1979.

[160] D. Geer, "Chip makers turn to multicore processors," *Computer*, vol. 38, no. 5, pp. 11 – 13, 2005.

[161] L. Dagum and R. Menon, "OpenMP: an industry standard API for shared-memory programming," *IEEE computational science and engineering*, vol. 5, no. 1, pp. 46 – 55, 1998.

[162] R. Chandra, *Parallel programming in OpenMP.* Morgan Kaufmann, 2001.

[163] B. Chapman, G. Jost, and R. Van Der Pas, *Using OpenMP: portable shared memory parallel programming.* MIT press, 2008, vol. 10.

[164] D. L. Woolard, H. Tian, a. Littlejohn, K. V. Kim, and S. Member, "Efficient Ohmic Boundary Conditions for the Monte Carlo Simulation of Electron Transport," *IEEE Transactions on Electron Devices*, vol. 41, no. 4, 1994.

[165] T. González and D. Pardo, "Physical models of ohmic contact for Monte Carlo device simulation," *Solid-State Electronics*, vol. 39, no. 4, pp. 555–562, 1996.

[166] M. V. Fischetti, L. Wang, B. Yu, C. Sachs, P. M. Asbeck, Y. Taur, and M. Rodwell, "Simulation of electron transport in high-mobility MOSFETs:

Density of states bottleneck and source starvation," *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 109–112, 2007.

[167] X. Oriols, E. Fernàndez-Díaz, a. Alvarez, and a. Alarcón, "An electron injection model for time-dependent simulators of nanoscale devices with electron confinement: Application to the comparison of the intrinsic noise of 3D-, 2D- and 1D-ballistic transistors," *Solid-State Electronics*, vol. 51, no. 2, pp. 306–319, 2007.

[168] I. Riolino, M. Braccioli, L. Lucci, P. Palestri, D. Esseni, C. Fiegna, and L. Selmi, "Monte-Carlo simulation of decananometric nMOSFETs: Multi-subband vs. 3D-electron gas with quantum corrections," *Solid-State Electronics*, vol. 51, no. 11-12, pp. 1558–1564, 2007.

[169] C. Jungemann, S. Decker, R. Thoma, W. L. Eng, and H. Goto, " Phase space multiple refresh: A versatile statistical enhancement method for Monte Carlo device simulation." In 1996 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 1996, pp. 65–66.

[170] A. Duncan, U. Ravaioli, and J. Jakumeit, "Full-band Monte Carlo investigation of hot carrier trends in the scaling of metal-oxide-semiconductor field-effect transistors," *IEEE Transactions on Electron Devices*, vol. 45, no. 4, pp. 867–876, 1998.

[171] J. Kim, H. Shin, C. Lee, Y. J. Park, and H. S. Min, "A new weight redistribution technique for electron-electron scattering in the MC simulation," *IEEE Transactions on Electron Devices*, vol. 51, no. 9, pp. 1448–1454, 2004.

[172] V. De and S. Borkar, " Technology and design challenges for low power and high performance." In Proceedings of the 1999 international symposium on Low power electronics and design, 1999, pp. 163 – 168.

[173] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI devices.* New York: Cambridge University Press, 2009.

[174] R. Pierret, *Semiconductor Device Fundamentals.* Reading, Massachusetts: Addision-Wesley, 1996.

[175] J. A. Mandelman and J. Alsmeier, "Anomalous narrow channel effect in trench-isolated buried-channel p-MOSFET's ," *IEEE Electron Device Letters*, vol. 15, no. 12, pp. 496 – 498, 1994.

[176] S. Chung and C.-T. Li, "An analytical threshold-voltage model of trench-isolated MOS devices with nonuniformly doped substrates ," *IEEE Transactions on Electron Devices*, vol. 39, pp. 614–622, 1992.

[177] V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, S. Borkar, and et al., "Techniques for leakage power reduction," in *Design of high-performance microprocessor circuits.* Wiley-IEEE press, 2001, pp. 48 – 52.

[178] R. Tsu and L. Esaki, "Tunneling in a finite superlattice," *Applied Physics Letters*, vol. 22, pp. 562 – 564, 1973.

[179] D. Ferry and S. M. Goodnick, *Transport in Nanostructures.* New York: Cambridge University Press, 1997.

[180] G. D. L. Sun, X. Y. Liu and R. Q. Han, "Monte Carlo Simulation of Schottky contact with direct tunneling model," *Semiconductor Science and Technology*, vol. 18, pp. 576–581, 2003.

[181] A. Revelant, P. Palestri, and L. Selmi, "Multi-subband semi-classical simulation of n-type Tunnel-FETs," *2012 13th International Conference on Ultimate Integration on Silicon (ULIS)*, pp. 187–190, Mar. 2012.

[182] Z. Huang, T. E. Feuchtwang, P. H. Cutler, and E. Kazes, "Wentzel-Kramers-Brillouin method in multidimensional tunneling," *Physical Review A*, vol. 41, no. 1, pp. 32–41, 1990.

[183] C. Shen, L.-T. Yang, G. Samudra, and Y.-C. Yeo, "A new robust non-local algorithm for band-to-band tunneling simulation and its application to Tunnel-FET," *Solid-State Electronics*, vol. 57, no. 1, pp. 23–30, Mar. 2011.

[184] S. O'uchi, T. Matsukawa, T. Nakagawa, K. Endo, Y. X. Liu, T. Sekigawa, and et al., "Characterization of metal-gate FinFET variability based on

measurements and compact model analyses." 2008 IEEE International Electron Devices Meeting, December 2008, pp. 1 – 4.

[185] T. Matsukawa, S. O'uchi, K. Endo, Y. Ishikawa, H. Yamauchi, Y. X. Liu, and et al., "Comprehensive analysis of variability sources of FinFET characteristics." Symposium on VLSI Technology Digest of Technical Papers, June 2009, pp. 118 – 119.

[186] X. Wang, A. R. Brown, B. Cheng, and A. Asenov, "Statistical variability and reliability in nanoscale FinFETs." 2011 IEEE International Electron Devices Meeting (IEDM), December 2011, pp. 5.4.1–5.4.4.

[187] C. Sampedro, L. Donetti, F. Gámiz, and A. Godoy, "3D Multi-Subband Ensemble Monte Carlo Simulator of FinFETs and Nanowire Transistors." 2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2014, pp. 21–24.

[188] A. Rahman, M. S. Lundstrom, A. W. Ghosh, A. Rahman, M. S. Lundstrom, and A. W. Ghosh, "Generalized effective-mass approach for n-type metal-oxide-semiconductor field-effect transistors on arbitrarily oriented wafers," *Journal of Applied Physics*, vol. 97, no. 053702, 2005.

[189] C. Diaz-Llorente, C. Medina-Bailon, C. Sampedro, F. Gámiz, A. Godoy, and L. Donetti, "Sub-22nm scaling of UTB2SOI devices for Multi-V_T applications." 2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), 2015, pp. 281–284.

[190] J. L. Padilla, C. Alper, F. Gamiz, and A. M. Ionescu, "Assessment of field–induced quantum confinement in heterogate germanium electron–hole bilayer tunnel field–effect transistor," *Applied Physics Letters*, vol. 105, no. 8, pp. 082 108–1–082 108–4, 2014.

[191] J. L. Padilla, C. Alper, A. Godoy, F. Gamiz, and A. M. Ionescu, "Impact of Asymmetric Configurations on the Heterogate Germanium Electron–Hole Bilayer Tunnel Field–Effect Transistor Including Quantum Confinement," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3560–3566, 2015.

[192] J. L. Padilla, C. Alper, C. Medina-Bailon, F. Gamiz, and A. M. Ionescu, "Assessment of pseudo–bilayer structures in the heterogate germanium electron–hole bilayer tunnel field–effect transistor," *Applied Physics Letters*, vol. 106, no. 26, pp. 262 102–1–262 102–4, 2015.

[193] C. Medina-Bailon, C. Sampedro, J. L. Padilla, F. Gámiz, A. Godoy, and L. Donetti, "Multi-subband ensemble Monte Carlo study of band-to-band tunneling in silicon-based TFETs." 2016 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2016, pp. 253–256.

[194] O. Manasreh, "Parabolic potential well," in *Introduction to nanomaterials and devices.* New Jersey: Willey, 2011, pp. 437 – 441.

[195] E. O. Kane, "Zener tunneling in semiconductors," *Journal of Physics and Chemistry of Solids*, vol. 12, no. 2, pp. 181 – 188, 1960.

[196] ——, "Theory of Tunneling," *Journal of Applied Physics*, vol. 32, no. 1, p. 83, 1961.

[197] W. Vandenberghe, B. Sorée, W. Magnus, and M. V. Fischetti, "Generalized phonon-assisted Zener tunneling in indirect semiconductors with non-uniform electric fields: A rigorous approach," *Journal of Applied Physics*, vol. 109, no. 12, p. 124503, 2011.

[198] Synopsys, "Sentaurus Device User Version 2014.09," no. September, 2014.

[199] K.-H. Kao, A. S. Verhulst, W. G. Vandenberghe, B. Soree, G. Groeseneken, and K. De Meyer, "Direct and Indirect Band-to-Band Tunneling in Germanium-Based TFETs," *IEEE Transactions on Electron Devices*, vol. 59, no. 2, pp. 292–301, 2012.

[200] J. L. Padilla, C. Alper, F. Gamiz, and A. M. Ionescu, "Quantum Mechanical Confinement in the Fin Electron?Hole Bilayer Tunnel Field-Effect Transistor," *IEEE Transactions on Electron Devices*, vol. 63, no. 8, pp. 3320–3326, 2016.

[201] S. Sant and A. Schenk, "Methods to Enhance the Performance of InGaAs/InP Heterojunction Tunnel FETs," *IEEE Transactions on Electron Devices*, vol. 63, no. 5, pp. 2169 – 2175, 2016.

[202] A. M. Walke, A. S. Verhulst, A. Vandooren, D. Verreck, E. Simoen, V. R. Rao, G. Groeseneken, N. Collaert, and A. V. Y. Thean, "Part I: Impact of Field-Induced Quantum Confinement on the Subthreshold Swing Behavior of Line TFETs," *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4057–4064, 2013.

[203] J. L. Padilla, F. Gamiz, and A. Godoy, "A Simple Approach to Quantum Confinement in Tunneling Field–Effect Transistors," *IEEE Electron Device Letters*, vol. 33, no. 10, pp. 1342–1344, 2012.

[204] G. B. Beneventi, E. Gnani, A. Gnudi, S. Reggiani, and G. Baccarani, "Optimization of a Pocketed Dual–Metal–Gate TFET by Means of TCAD Simulations Accounting for Quantization-Induced Bandgap Widening," *IEEE Transactions on Electron Devices*, vol. 62, no. 1, pp. 44–51, 2015.

[205] K. Cao, W.-C. Lee, W. Liu, X. Jin, P. Su, S. Fung, J. An, B. Yu, and C. Hu, "BSIM4 gate leakage model including source-drain partition." 2000 IEEE International Electron Devices Meeting (IEDM). Technical Digest., December 2000, pp. 815–818.

[206] F. Hamzaoglu and M. R. Stan, "Circuit-level techniques to control gate leakage for sub-100nm CMOS." Proceedings of the 2002 international symposium on Low power electronics and design (ISLPED), 2002, pp. 60–63.

[207] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, "Discrete Resistance Switching in Submicrometer Silicon Inversion Layers: Individual Interface Traps and Low-Frequency (1/f) Noise," *Applied Physics Letters*, vol. 52, pp. 228 – 231, January 1984.

[208] K. R. Farmer, C. T. Rogers, and R. A. Buhrman, "Localized-State Interactions in Metal-Oxide-Semiconductor Tunnel Diodes," *Applied Physics Letters*, vol. 58, pp. 2255 – 2258, May 1987.

[209] N. Yang, W. Henson, J. Hauser, and J. Wortman, "Modeling study of ultrathin gate oxides using tunneling current and capacitance-voltage measurement in MOS devices," *IEEE Transactions on Electron Devices*, vol. 46, pp. 1464 – 1471, July 1999.