Quantum transport simulation in III-V semiconductor transistors

Corentin Grillet

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : NANO ELECTRONIQUE ET NANO TECHNOLOGIES

Arrêté ministériel : 25 mai 2016

Présentée par

Corentin GRILLET

Thèse dirigée par **Marco PALA**, , CNRS, et
codirigée par **FRANCOIS TRIOZON** cea
préparée au sein du **Laboratoire Institut de Microélectronique, Electromagnétisme et Photonique - Laboratoire d'hyperfréquences et de caractérisation**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique, Traitement du Signal (EEATS)**

Simulation du transport quantique dans les transistors en semi-conducteurs III-V

Quantum transport simulation in III-V semiconductor transistors

Thèse soutenue publiquement le **7 avril 2017**,
devant le jury composé de :

Monsieur Arnaud BOURNEL
Professeur, Université Paris-Sud, Président
Monsieur Massimo MACUCCI
Professeur, Université de Pise, Examinateur
Monsieur Raphael CLERC
Professeur, Université Jean Monnet, Rapporteur
Monsieur Alessandro CRESTI
Chargé de recherche, CNRS Délégation alpes, CoDirecteur de these
Monsieur François TRIOZON
Ingénieur de recherche, CEA, CoDirecteur de these
Monsieur Marco PALA
Chargé de recherche, CNRS Ile-de-France Sud, Directeur de these

Nomenclature

Green's functions formalism

$\Sigma^{\lessgtr}$	Lesser/greater self-energy
$\Sigma^{\rightleftarrows}$	Right/left-connected self-energy
g	Left connected Green's function
D_{ac}	Acoustic phonons deformation potential
D_{opt}	Optical phonons deformation potential
$G^{\pm}$	(anti)chronological Green's function
$G^{\lessgtr}$	Lesser/greater Green's function
$G^{(0)}$	Unperturbed Green's function
$G^{R/A}$	Retarded/advanced Green's function
Σ_{ph}	Phonon self-energy
Σ^{R}	Retarded self-energy
ω_{opt}	Optical phonons frequency
NEGF	Non-equilibrium Green's function
SCBA	Self-consistent Born approximation

Physics and second quantization

$\lvert\psi\rangle$	Wave function
$\lvert u\rangle$	Bloch function
ϕ	Electric potential
f	Fermi distribution
J	Electrical current
m^*	Effective mass
$\hat{c}^\dagger/\hat{c}$	Creation/annihilation operators
E_{F}	Fermi level energy
E_{G}	Gap energy
$\hat{\mathcal{H}}$	Hamiltonian operator
Θ	Heaviside function

I Identity matrix

$\hat{\rho}$ Density matrix

$\mathcal{T}$ Time-ordering operator

$\hat{U}$ Time-evolution operator

$\uparrow/\downarrow$ Up/down spin

$\hat{\mathcal{W}}$ Perturbation operator

CB/VB Conduction/valence band

e Elementary charge

HV/LC Highest valence/lowest conduction (subband)

LDOS Local density of states

n/p Electron/hole density

Device-related terms

E_{sw} Switching energy

I_{DS} Source-drain current

$I_{\mathrm{on}}/I_{\mathrm{off}}$ *on*- and *off*-state currents

L_{dop} Doped region length

L_{ext} Extension region length

L_{G} Gate length

L_{ov} Overlap length

L_{sp} Spacer length

N_{SD} Source/drain doping

t_{B} Tunnel barrier thickness

t_{ch} Channel thickness

t_{ox} Oxide thickness

$t_{\mathrm{S}}/t_{\mathrm{D}}$ Source/drain layer thickness

T_{sw} Intrinsic switching time

V_{DD} Supply voltage

V_{DS} Source-drain voltage

V_{GS} Gate voltage

$V_{\mathrm{on}}/V_{\mathrm{off}}$ *on*- and *off*-state voltages

BTBT Band-to-band tunneling

DIBL Drain-induced barrier lowering

FET Field effect transistor

HIBL Hole-induced barrier lowering

MOS Metal-Oxide-Semiconductor

NW Nanowire

SCE Short channel effect

SR Surface roughness

SS Subthreshold swing

STDT Source-to-drain tunneling

TFET Tunnel-FET

UTB Ultra thin body

Contents

Acknowledgments

I would like to express my sincere gratitude to my advisor Dr. Marco Pala for his continuous support, patience, guidance, and of course kindness.
I would also like to show my greatest appreciation to my co-advisors Dr. Alessandro Cresti and Dr. François Triozon, for sharing many pearls of wisdom with me.
I am immensely grateful to my colleagues Dr. Jiang Cao and Dr. Demetrio Logoteta, who provided continuous insight and help.
I would also like to thank Thomas Moncond'Huy, Timothée Allenet and my other fellow doctoral students for their feedback, cooperation and of course friendship.
Finally, I cannot say thank you enough to my parents and everyone who supported me during these years.

CHAPTER 1

Introduction

1.1 Context

Nearly 70 years ago, W.B. Shockley, J. Bardeen and W.H. Brattain made a discovery that would become one of the most notable technological break-through of the last century. They were subsequently awarded the 1956 Nobel Prize in Physics *"for their researches on semiconductors and their discovery of the transistor effect"*. The invention of the bipolar transistor in the Bell Laboratories rapidly led the scientific community to enrich the field of semiconductor physics. Building over the knowledge brought by previous solid-state physicists such as L. Brioullin or F. Bloch, great progress in the understanding of the energy band structure was achieved.
In 1959, the famous speech of yet-to-be Nobel laureate R.P. Feynman – *There's Plenty of Room at the Bottom* – was a prelude to the development of more advanced microfabrication techniques. The next major leap in semiconductor physics came with the development of the field effect transistor, and notably the MOSFET. As a side effect, a wide range of new phenomena could be studied in the emerging devices. Some were related to the small size of the structures, like the ballistic transport. A handful was observed at low temperature, like the Kondo effect (1964). Some others required the presence of high magnetic fields, like the quantum Hall effect (1980), which was first observed in the two-dimensional inversion layer of a MOSFET [1]. In the following decades, the fabrication techniques kept improving, and electronic devices got more efficient and powerful as their size decreased. In 1986, the term "nanotechnology" was introduced for the first time in the science-fiction literature [2].
With several billions of transistors in every modern computer and phone, semiconductor devices have become the most elementary building block of today's information technology. Even though quantum theory proved itself vital to understand the fundamentals of solid-state physics, classical and semi-classical models could still be elaborated on top of the quantum formalism. For a long time, semiconductor engineering has been taking place in the *mesoscopic* domain (*i.e.* between the macroscopic and the atomic scales). In the 1990s, the 1 μm limit was broken. In the 2000s, we reached the nanometric scale by manufacturing 100 nm long devices. Today, the transistor channel length is getting close to the 10 nm mark. Henceforth, taking quantum phenomena into account is not only important for understanding the underlying theory: it is also vital to accurately predict the behavior of the future nanoelectronic devices.

Indeed, size reduction has triggered many undesired effects, which prevent transistors from fulfilling their essential tasks. Before describing these effects, let us remind that one expects a transistor to:

- switch between its *on* and *off* states as fast as possible (*i.e.* billions of times per second);
- keep these states sufficiently distinguishable (*i.e.* distant of several orders of magnitude in terms of current);
- perform the aforementioned tasks with as little energy dissipation as possible.

In practice, at the nanoscale, we can be confronted to electron flow reduction due to quantum confinement effects, electrostatic control deterioration over the channel, and tunneling or leakage current in the *off* state. These effects degrade the performances of the transistors by hindering their ability to switch efficiently (subthreshold swing degradation), and by bringing their *on* and *off* states closer to one another (*on*-current lowering and *off*-current increase).
Modern transistor technology is mainly based on silicon. Improving the performance of the nanoelectronic devices is likely to require alternative channel semiconductors, like III-V materials. These semiconductors, which belong to the 13^{th} and 15^{th} columns of the periodic table, can be arranged into various compounds such as indium-arsenide (InAs), gallium-arsenide (GaAs) or aluminum-antimonide (AlSb), to mention just a few. Some of these materials exhibit remarkable transport properties due to their high electron mobility and low effective mass [3]. Another benefit of the III-V semiconductors lies in the fact that they can be arranged in a broad variety of heterostructures. We can, for example, take advantage of their different bandgaps to form quantum wells (useful in ultra-thin devices) or even tunnel barriers. However, compared to silicon, these materials also suffer from increased short-channel effects [4], which can be detrimental for the design or efficient nanotransistors. For this reason, new device architectures shall also be investigated. In this work, we will show that we can either try to counteract the drawbacks of size reduction through the use of multi-gate architectures, or take advantage of the quantum behavior by designing tunnel devices. Among the important issues that have to be addressed, we can cite the variability of III-V devices induced by surface fluctuations, or the impact of quantum effects, strain and device geometry on carrier transport.

1.2 Full-quantum simulations

In this work, we will simulate nanoelectronic devices with custom-made algorithms. Computing the behavior and the properties of transistors (or any other physical system, for that matter) can usually be done *via* two different approaches: the analytical method or the numerical method. The devices we will consider in this work are too small to be correctly described by the bulk, macroscopic models, but are large enough to contain thousands of atoms, and there is no simple way to describe the system accurately. That is the reason why we perform numerical simulations, using a full-quantum approach.
The band structure of the materials is computed with an eight-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian [5] that we have discretized in the real space. Such an Hamiltonian is well suited

for the study of III-V semiconductors since it accounts for direct-gap band structures and includes conduction, light-hole, heavy-hole and spin-orbit bands, that all contribute to transport in this type of materials. This approach is also compatible with other models developed with the $\mathbf{k} \cdot \mathbf{p}$ perturbation theory, like the Pikus-Bir Hamilonian [6], that we use to account for the effects of strain.
The electron transport is computed with the non-equilibrium Green's functions (NEGF) approach [7–9], that is a very versatile quantum framework able to consider the effects relevant for the study of nanoscale transistors. Since these devices are subjected to an electric potential, they are indeed brought in a state of non-equilibrium. The NEGF allows us to solve the subsequent electron flow while accounting for the impact of confinement, tunneling, or interferences. The electron-phonon interactions are also included with the self-consistent Born approximation (SCBA).
Since the simulations are rather resource demanding, they are performed on a dedicated computer cluster. On a technical point of view, the implementation of the code and the choice of a proper balance between the complexity of the model and the feasibility of the simulations are the main hurdles to overcome.

1.3 Outline

As explained above, in order to understand which is the best couple of material and device architecture, we will run predictive simulations based on rigorous models. We will perform 2D-3D self-consistent simulations of III-V nanodevices based on advanced physical models within the NEGF formalism, using eight-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonians.
The first half of this dissertation is focused on detailing the simulation approach and its implementation. In Chap.2, we explain how to model the band structure of an isolated device. We give some basics of solid state physics and quantum mechanics, which are then used to introduce the perturbation $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian. In Chap.3, we explain how to connect the system to external contacts and then compute the resulting electron transport. To that end, we introduce the NEGF formalism and the SCBA scheme and illustrate their computational implementation.
The second half of the manuscript aims to apply the quantum transport algorithm to practical cases of nanotransistors. In Chap.4, we simulate an ultra-thin MOSFET made of either InAs or InGaAs. We study the bias voltage scalability, the effect of the gate length and the spacer, or the influence of the channel semiconductor, as well as its thickness. This chapter serves us as an introductory case study, before moving to more elaborate architectures. Chap.5 presents a comparison between InAs and Si nanowire-FETs and shows how the gate-all-around architectures can be helpful in the quest for size reduction. The investigation resembles to that of Chap.4, but we also demonstrate the impact of interface fluctuations on the quality of the transport, by means of a statistical study led on hundreds of nanowire simulations. Finally, in Chap.6, we address the case of a tunnel-FET device made of stacked III-V compounds. We provide details on its operation by reviewing the effect of various geometrical parameters, such as the barrier thickness, the size of the overlap region, or the gate extensions. In addition, we assess the effect of phonon scattering in this type of device. We then formulate our general conclusions about this work in Chap.7 and propose possible perspectives.

Bibliography

[1] J. Wakabayashi and S. Kawaji. Hall effect in silicon MOS inversion layers under strong magnetic fields. *Journal of the Physical Society of Japan*, 44(6):1839–1849, jun 1978.

[2] E. Drexler. *Engines of Creation: The Coming Era of Nanotechnology.* Anchor Library of Science, 1987.

[3] J.A. del Alamo. Nanometre-scale electronics with III-V compound semiconductors. *Nature*, 479(7373):317–323, nov 2011.

[4] T. Rollo D. Esseni, M.G. Pala. Essential physics of the OFF-state current in nanoscale MOSFETs and tunnel FETs. *IEEE Transactions on Electron Devices*, 62(9):3084–3091, sep 2015.

[5] T. B. Bahder. Eight-band k·p model of strained zinc-blende crystals. *Physical Review B*, 41(17):11992–12001, jun 1990.

[6] J. Singh. *Physics of semiconductors and their heterostructures.* MaGraw-Hill, 1993.

[7] M. Pourfath. *Non-Equilibrium Green's Function Method for Nanoscale Device Simulation.* Springer, 2014.

[8] S. Datta. *Quantum Transport: Atom to Transistor.* Cambridge University Press, 2005.

[9] R. van Leeuwen G. Stefanucci. *Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction.* Cambridge University Press, 2013.

CHAPTER 2

Band structure modeling

In which we give an introduction on perturbation theory and then detail the method from which the eight-band $\boldsymbol{k}\cdot\boldsymbol{p}$ Hamiltonian is derived and used in our simulations.

The study of nanoelectronic devices requires precise predictions of the electron behavior at the nanoscale. To this aim, a realistic description of the electronic structure is of central importance. A good understanding of the band structure is indeed essential for the design of field effect transistors, since it dictates the way electrons spread and flow through the device.

2.1 Approximations

There are different ways to model a semiconductor's band structure. However, all these methods tend to use a similar set of approximations. An intuitive way of modeling a solid is to consider a general electron Hamiltonian of the form

$$\hat{\mathcal{H}} = \sum_i -\frac{\hbar^2}{2m_i}\nabla_i^2 + \sum_i V(r_i) + \sum_{i\neq j} U(r_i, r_j) = \hat{T} + \hat{V} + \hat{U}, \tag{2.1}$$

where $\hat{T}$ is the kinetic energy, and $\hat{U}$ and $\hat{V}$ are the electron-electron and electron-nucleus interaction potentials:

$$\begin{aligned} \hat{U} &= \sum_{j\neq i} \frac{e^2}{4\pi\epsilon_0\ |\mathbf{r}_j - \mathbf{r}_i|} \\ \hat{V} &= -\sum_{j\neq i} \frac{Z_j\ e^2}{4\pi\epsilon_0\ |\mathbf{R}_j - \mathbf{r}_i|} \end{aligned} \tag{2.2}$$

with the elementary charge e, the electron positions $\{\mathbf{r_i}\}$, the atomic nuclei positions $\{\mathbf{R_i}\}$, and the protons number $\{Z_i\}$. Yet, such a Hamiltonian would require substantial computation power to deal with. Therefore, some approximations are usually made:

- First, we forget about the core electrons, and only consider the outermost orbitals. As explained in Sec.2.4, the electrical properties of a semiconductor are mainly influenced by the outer shell, and we combine together the lower energy orbitals and the nuclei into ions.
- Second, we assume that the motion of the electrons is much faster than that of the heavier ions. Therefore, the wavefunction can be factorized in its electronic and ionic components (Born-Oppenheimer approximation). When calculating the electronic structure, we consider the ions as static.

- Third, the electron-electron interactions are not taken into account exactly. Instead, we introduce the mean field theory, which states that one can approximate a many-body problem by averaging a one-body problem. This leads to the one-electron approximation.

The resulting Hamiltonian can be expressed in terms of an atomistic basis set (like the tight binding method), a plane wave basis set (like the empirical pseudopotential approach), or a Bloch-state approach (like the $\mathbf{k}\cdot\mathbf{p}$ theory). Even though the accuracy and the complexity of these models vary, they are all fundamentally meant to solve one-electron problems in the one-electron formalism. The $\mathbf{k}\cdot\mathbf{p}$ model offers a good compromise between accuracy and computational burden, and is the framework used in this work.
In practice, most of the transport properties of semiconductors can be understood by focusing on a reduced portion of the band structure. This zone is centered on the minimum of the conduction band, and the maximum of the valence band. When the bottom of the conduction band is aligned with the top of the valence band in the k-space, the material is categorized as "direct gap". All the III-V compounds treated is this work will belong to this category, and have the region of interest around $\mathbf{k} = 0$ (also called the Γ-point).

2.2 Perturbation theory

The perturbation theory [1, 2] allows one to describe complex quantum systems by adding a small perturbation to an initial configuration. The unperturbed Hamiltonian generally corresponds to an ideal, simplified case. The perturbation plays the role of other additional phenomena affecting the system, which refine the model and lead to more precise and realistic results.
Let us first consider a simple unperturbed system whose eigenstates $|\psi_n\rangle$ are known

$$\hat{\mathcal{H}}^{(0)}\,|\psi_n\rangle = E_n^{(0)}\,|\psi_n\rangle\,. \tag{2.3}$$

We now introduce a perturbation potential $\hat{\mathcal{W}}$ in the Hamiltonian, such that $\hat{\mathcal{H}} = \hat{\mathcal{H}}^{(0)} + \lambda\hat{\mathcal{W}}$, where λ is a small scalar parameter. The Schrödinger equation can now be expanded as

$$\begin{aligned}&\Big(\hat{\mathcal{H}}^{(0)} + \lambda\hat{\mathcal{W}}\Big)\Big(\,|\psi_n^{(0)}\rangle + \lambda\,|\psi_n^{(1)}\rangle + \lambda^2\,|\psi_n^{(2)}\rangle + \mathcal{O}(\lambda^3)\Big)\\ &\quad= \Big(E_n^{(0)} + \lambda E_n^{(1)} + \lambda^2 E_n^{(2)} + \mathcal{O}(\lambda^3)\Big)\Big(\,|\psi_n^{(0)}\rangle + \lambda\,|\psi_n^{(1)}\rangle + \lambda^2\,|\psi_n^{(2)}\rangle + \mathcal{O}(\lambda^3)\Big).\end{aligned} \tag{2.4}$$

The exponent in brackets indicates the order of each term in the series. The zeroth-order of this equation corresponds to the unperturbed case. At the order λ (*i.e.* first order), this equation simplifies as

$$\lambda\Big(\hat{\mathcal{H}}^{(0)} - E_n^{(0)}\Big)\,|\psi_n^{(1)}\rangle + \lambda\Big(\hat{\mathcal{W}} - E_n^{(1)}\Big)\,|\psi_n^{(0)}\rangle = 0. \tag{2.5}$$

Taking the scalar product of this expression with $\psi_n^{(0)}$ leads to

$$\langle\psi_n^{(0)}|\,\hat{\mathcal{H}}^{(0)}\,|\psi_n^{(1)}\rangle + \langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle = E_n^{(0)}\,\langle\psi_n^{(0)}|\psi_n^{(1)}\rangle + E_n^{(1)}\,\langle\psi_n^{(0)}|\psi_n^{(0)}\rangle\,, \tag{2.6}$$

which yields to the first-order energy correction

$$E_n^{(1)} = \frac{\langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle}{\langle\psi_n^{(0)}|\psi_n^{(0)}\rangle} = \langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle\,. \tag{2.7}$$

As we can see, the first-order energy shift can be obtained by assuming that the system stays in the unperturbed state under the application of a perturbation $\hat{\mathcal{W}}$. In other words, we can compute the first-order eigenvalues of the perturbation Hamiltonian without considering any change in the eigenvectors (which makes sense, since $\hat{\mathcal{W}}$ has been chosen to affect lightly the system compared to $\hat{\mathcal{H}}^{(0)}$). However, to completely account for the first-order perturbation, the next step is to update the value of the eigenstate. To do so, we consider a state $\psi_{m\neq n}^{(0)}$ orthogonal to $\psi_n^{(0)}$, such as $\langle\psi_m^{(0)}|\psi_n^{(0)}\rangle = 0$, and we compute again its scalar product with Eq.(2.5) . We deduce the following expression for the first-order eigenstate correction

$$\langle\psi_m^{(0)}|\psi_n^{(1)}\rangle = \frac{\langle\psi_m^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle}{E_m^{(0)} - E_n^{(0)}} \implies |\psi_n^{(1)}\rangle = \sum_{m\neq n} \frac{\langle\psi_m^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle}{E_m^{(0)} - E_n^{(0)}}\,|\psi_m^{(0)}\rangle\,, \tag{2.8}$$

which shows that the perturbed wave function can also be deduced from the knowledge of the unperturbed system. We can keep iterating through the next perturbation orders while applying the same process. For the second order, one will find

$$\begin{cases} E_n^{(2)} = \langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(1)}\rangle = \displaystyle\sum_{m\neq n} \frac{|\,\langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_m^{(0)}\rangle\,|^2}{E_n^{(0)} - E_m^{(0)}} \\[2ex] |\psi_n^{(2)}\rangle = \displaystyle\sum_{m\neq n}\left[-\frac{\langle\psi_n^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle\,\langle\psi_m^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle}{\left(E_n^{(0)} - E_m^{(0)}\right)^2} + \sum_{k\neq n} \frac{\langle\psi_m^{(0)}|\,\hat{\mathcal{W}}\,|\psi_k^{(0)}\rangle\,\langle\psi_k^{(0)}|\,\hat{\mathcal{W}}\,|\psi_n^{(0)}\rangle}{\left(E_n^{(0)} - E_m^{(0)}\right)\left(E_m^{(0)} - E_k^{(0)}\right)}\right] |\psi_m^{(0)}\rangle \end{cases} \tag{2.9}$$

The perturbation corrections are thus fully described by the matrix elements $\hat{\mathcal{W}}_{ij} \equiv \langle\psi_i^{(0)}|\,\hat{\mathcal{W}}\,|\psi_j^{(0)}\rangle$. The smaller $\left|E_i^{(0)} - E_j^{(0)}\right|$, the more $\hat{\mathcal{W}}_{ij}$ shifts the energy of the system. Thus, the second-order energy perturbation is due to the interaction between different states. Whether two eigenvalues interact or not is determined by the elements of the perturbation Hamiltonian. The $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian described in the next section is obtained through a similar second-order expansion.

2.3 2-band Hamiltonian: conduction and valence bands

A crystal is a periodic system of ordered matter, and essentially consists of a unit cell which can contain several atoms. This cell is repeated in space by a set of discrete

translation operations, defined by linear combinations of the *primitive vectors*. In this work, we assume that we are dealing with perfectly crystalline materials. In such solids, electrons are subjected to a periodic potential

$$V(\mathbf{r}) = V(\mathbf{r} + \mathbf{R}), \tag{2.10}$$

where $\mathbf{R}$ can be any of the linear combinations mentioned above. The same periodicity holds for the electrons' wave functions, since the Hamiltonian is invariant under a translation by $\mathbf{R}$. The Bloch theorem takes advantage of the periodicity of the system, and allows us to express the wave function as

$$\psi_k(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_k(\mathbf{r}), \tag{2.11}$$

where u_k is a Bloch lattice function. Put together, Bloch functions will form the band structure of the material. To predict the electrical properties of a material, one will focus on the lowest partially filled band, called the conduction band (CB). The valence band (VB), located energetically below the CB, also has to be considered if it contributes to the transport, as this is the case for III-V semiconductors [3].

In the Schrödinger's equation, the kinetic term $-\hbar^2\nabla^2$ corresponds to the square of the momentum operator $\hat{p} = -i\hbar\nabla$. Applying this squared operator to the Bloch wave function $e^{i\mathbf{k}.\mathbf{r}} u_{n,k}(\mathbf{r})$ yields

$$-\hbar^2\nabla^2 e^{i\mathbf{k}.\mathbf{r}} u_{n,k}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}}(-i\hbar\nabla)^2 u_{n,k}(\mathbf{r}) + \hbar^2 k^2 e^{i\mathbf{k}.\mathbf{r}} u_{n,k}(\mathbf{r}) + 2\hbar\mathbf{k}\cdot e^{i\mathbf{k}.\mathbf{r}} \underbrace{(-i\hbar\nabla)}_{\mathbf{p}} u_{n,k}(\mathbf{r}). \tag{2.12}$$

Therefore, if one injects the Bloch wave function into the Schrödinger equation, one obtains:

$$\Big(\underbrace{-\frac{\hbar^2}{2m_0}\nabla^2 + V(\mathbf{r})}_{\hat{\mathcal{H}}^{(0)}} + \underbrace{\frac{\hbar^2 k^2}{2m_0} + \frac{\hbar}{2m_0}\mathbf{k}\cdot\mathbf{p}}_{\hat{\mathcal{W}}} \Big) u_{n,k}(\mathbf{r}) = E(k) u_{n,k}(\mathbf{r}). \tag{2.13}$$

This corresponds to the system without spin-orbit interaction. As explained before, the quantum perturbation theory can be applied to any system whose Hamiltonian can be split into two parts: an easy to solve unperturbed part and an additional perturbation. In Eq.(2.13), we have highlighted these two components. Here, we will use this approach around $\mathbf{k} = 0$, by assuming that the eigenfunctions $u_{n,0}$ and their eigenvalues $E_n(0)$ are known. The dispersion relation developed at the second order writes [4, 5]

$$E_n(\mathbf{k}) = E_n(0) + \frac{\hbar^2 k^2}{2m_0} + \underbrace{\frac{\hbar}{m_0}\mathbf{k}\cdot\langle u_{n,0}|\mathbf{p}|u_{n,0}\rangle}_{\mathbf{k}=0} + \frac{\hbar^2}{m_0^2}\sum_{m\neq n} \frac{|\langle u_{n,0}|\mathbf{k}\cdot\mathbf{p}|u_{m,0}\rangle|^2}{E_n(0) - E_m(0)} + \mathcal{O}(\lambda^3). \tag{2.14}$$

Around $\mathbf{k} = 0$, the third term from the left vanishes. By summing on $\alpha, \beta = x, y, z$, this equation becomes

$$E_n(\mathbf{k}) = E_n(0) + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar^2}{m_0^2} \sum_{\alpha,\beta} k_\alpha k_\beta \sum_{m \neq n} \frac{\langle u_{n,0}|p_\alpha|u_{m,0}\rangle \langle u_{m,0}|p_\beta|u_{n,0}\rangle}{E_n(0) - E_m(0)} + \mathcal{O}(\lambda^3). \quad (2.15)$$

This can also be written in a more compact form

$$E_n(\mathbf{k}) \approx E_n(0) + \frac{\hbar^2}{2} \sum_{\alpha,\beta} k_\alpha k_\beta \left[\frac{1}{m_n^*}\right]_{\alpha,\beta}, \quad (2.16)$$

where we introduced the effective mass m^*, whose tensor is defined as

$$\left[\frac{1}{m_n^*}\right]_{\alpha,\beta} = \frac{\delta_{\alpha,\beta}}{m_0} + \frac{2}{m_0^2} \sum_{m \neq n} \frac{\langle u_{n,0}|p_\alpha|u_{m,0}\rangle \langle u_{m,0}|p_\beta|u_{n,0}\rangle}{E_n(0) - E_m(0)}. \quad (2.17)$$

Eq.(2.16) corresponds to the simplest form of the effective mass approximation. In this model, electrons are considered to have a mass $m^* \neq m_0$ (that can also be negative in the valence band). In a simple 2-band model, we can represent the conduction and valence bands by two Bloch functions $|u_c\rangle$ and $|u_v\rangle$. In this case, the system obeys

$$\begin{aligned} \hat{\mathcal{H}}^{(0)} |u_{c,0}\rangle &= E_c(0) |u_{c,0}\rangle \\ \hat{\mathcal{H}}^{(0)} |u_{v,0}\rangle &= E_v(0) |u_{v,0}\rangle \end{aligned}, \quad (2.18)$$

where E_c and E_v are the conduction and valence energy levels. The momentum interaction term between those bands is

$$\begin{aligned} \langle u_c|\mathbf{p}|u_v\rangle &= P \\ \langle u_v|\mathbf{p}|u_c\rangle &= P^* \end{aligned} \quad (2.19)$$

P is the coupling term between the conduction band (CB) and the valence band (VB). In this basis, the 2-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian for Eq.(2.13) is

$$\hat{\mathcal{H}}_{k.p,2} = \begin{array}{c} \\ |u_c\rangle \\ |u_v\rangle \end{array} \begin{array}{c} \begin{array}{cc} |u_c\rangle & |u_v\rangle \end{array} \\ \begin{pmatrix} E_c(0) + \frac{\hbar^2 k^2}{2m_0} & \frac{\hbar}{m_0} kP \\ \frac{\hbar}{m_0} kP^* & E_v(0) + \frac{\hbar^2 k^2}{2m_0} \end{pmatrix} \end{array}, \quad (2.20)$$

where the first row/column corresponds to the CB and the second row/column refers to the VB. To lighten the notation, we define the gap energy as $E_G \equiv E_c(0) - E_v(0)$ and we take the energy reference at $E_v(0) = 0eV$. The eigenvalues of Eq.(2.20) are

$$E(k) = \frac{E_G}{2} + \frac{\hbar^2 k^2}{2m_0} \pm \frac{1}{2} \sqrt{E_G^2 + \left(\frac{2\hbar}{m_0} k|P|\right)^2}. \quad (2.21)$$

For small values of k, these eigenvalues are close to

$$E(k) \simeq \begin{cases} E_G + \dfrac{\hbar^2 k^2}{2m_0} + \dfrac{\hbar^2}{E_G m_0^2}|P|^2 k^2 \ , & \text{(CB)} \\ \dfrac{\hbar^2 k^2}{2m_0} - \dfrac{\hbar^2}{E_G m_0^2}|P|^2 k^2 \ , & \text{(VB)} \end{cases} \tag{2.22}$$

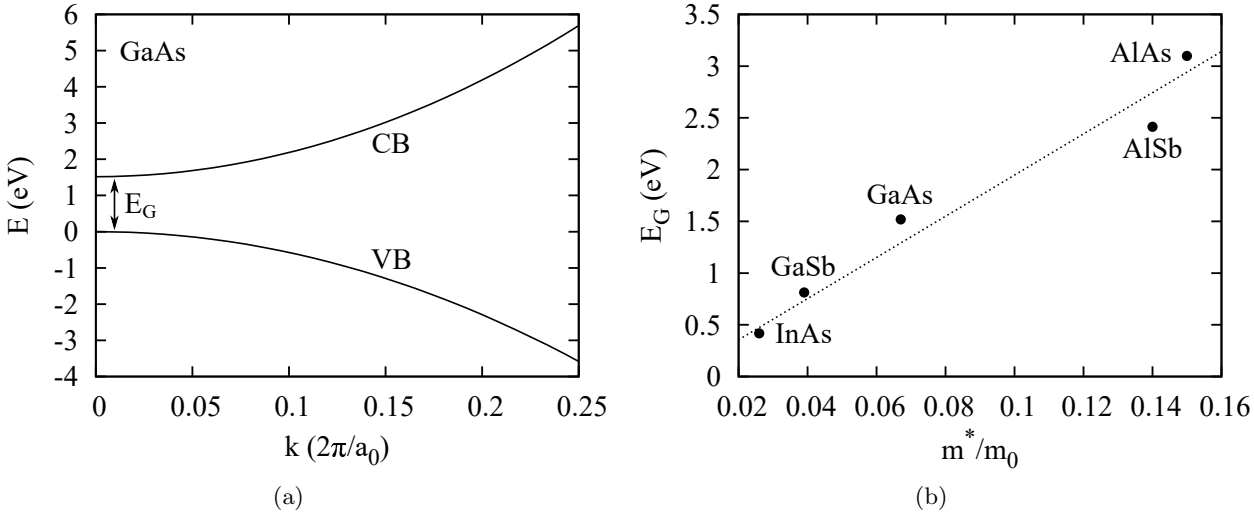


Figure 2.1: (a) GaAs band structure plotted with the 2-band $\boldsymbol{k}\cdot\boldsymbol{p}$ Hamiltonian. (b) E_G versus m^* in the III-V compounds used in this work (data from [6]). The semiconductors with the smallest gaps tend to have lower effective masses.

The CB and the VB are defined by parabolic dispersion curves, which is a basic, yet efficient way to treat any direct-gap semiconductor around the Γ point. Fig.2.1-a shows the resulting band structure for GaAs, E_G=1.519eV, a_0=5.65Å and E_P=20eV (extracted from [6]), where E_p controls the mixing between the CB and the VB, and follows the definition introduced by Kane [7]

$$E_P = \frac{2}{m_0} P^2. \tag{2.23}$$

$\mathbf{E_G}$ and $\mathbf{E_P}$ are taken as input parameters in the simulation code. From Eq.(2.17), we can also derive the value of the conduction and valence band effective masses in the 2-band case

$$\frac{1}{m^*_{c/v}} = \frac{1}{m_0} \pm \frac{2P^2}{E_G m_0^2} = \frac{1}{m_0}\left(1 \pm \frac{E_P}{E_G}\right). \tag{2.24}$$

In III-V semiconductors, we have generally $E_G \ll E_P$ [6]. For that reason, the energy gap E_G and the effective mass m^* are roughly proportional, as illustrated in Fig.2.1-b. Indeed, we are dealing with two coupled energy bands that tend to strongly repel each other when the gap is small. This increases the curvature of the bands, which in turn means that the effective mass is reduced. On the contrary, in large gap compounds, the bands are weakly coupled and the effective mass is closer to the free electron mass.

2.4 4-band Hamiltonian: electronic orbitals

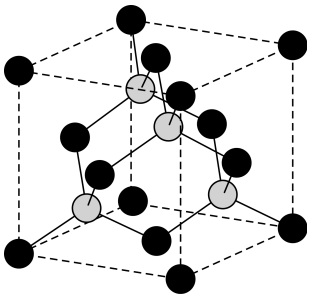


Figure 2.2: Zinc blende crystal structure of the III-V semiconductors. The dark atoms form a FCC lattice.They correspond to elements of group III and the light ones are group-V materials (or conversely).

Some insights about the electronic orbitals of semiconductors will be helpful for the rest of this discussion. The III-V compounds we want to model are made of two interpenetrating face cubic centered (FCC) lattices. They form a so called "zinc-blende" structure (Fig.2.2). This structure is related to their tetrahedral bonding which can, in turn, be explained by their electronic orbitals. Table 2.1 shows the electronic configuration of the targeted semiconductors. Group-III materials have a s^2p^1 outer shell configuration, whereas group-V materials have a s^2p^3 type configuration. Since the outermost electrons tend to delocalize in the material, the overall electronic configuration of a III-V crystal becomes s^2p^2. However, once the atoms are brought together, the orbitals will also hybridize to reach a more stable collective state. In the case of III-V semiconductors, this process leads to the formation of four sp^3 orbitals, as shown in Fig.2.3. This corresponds to the diamond geometry.

Group	Element	Electronic configuration
III	Al	$(1s^22s^22p^6)$ $\mathbf{3s^23p^1}$
V	P	$(1s^22s^22p^6)$ $\mathbf{3s^23p^3}$
III	Ga	$(1s^22s^22p^63s^23p^63d^{10})$ $\mathbf{4s^24p^1}$
V	As	$(1s^22s^22p^63s^23p^63d^{10})$ $\mathbf{4s^24p^3}$
III	In	$(1s^22s^22p^63s^23p^63d^{10}4s^24p^64d^{10})$ $\mathbf{5s^25p^1}$
V	Sb	$(1s^22s^22p^63s^23p^63d^{10}4s^24p^64d^{10})$ $\mathbf{5s^25p^3}$

Table 2.1: Electronic configuration of some III-V semiconductors. s and p orbitals are the main components of the band structure, while the core electrons can be neglected.

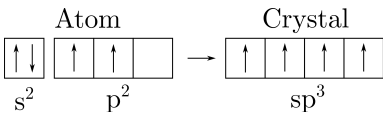


Figure 2.3: Outermost electron shell of a III-V semiconductor in its atomic form (left) and after sp^3 hybridization in the crystal (right). The direction of the spins is arbitrary.

The upmost orbitals are those that contribute the most to the band structure. Thus, the conduction and valence bands of III-V semiconductors present s- and p-like characteristics. Even if bands result from a mixing between those two types of orbitals, one

can make the approximation that the conduction band is mostly s-like around $k = 0$; while the valence band corresponds to a linear combination of p-like orbitals [8].
The Bloch lattice functions $u_{n,k}(\mathbf{r})$ retain most of the symmetries of those orbitals around $\mathbf{k} = 0$. Therefore, let us replace the $\{|u_c\rangle, |u_v\rangle\}$ basis by a different set of Bloch functions. We denote $|u_s\rangle$ the s-like conduction state and $|u_x\rangle$, $|u_y\rangle$, $|u_z\rangle$, the p_x-, p_y- and p_z-like valence states. These Bloch functions act as lattice-periodic repetitions of the orbitals. They can be written

$$|u_{n,k}\rangle = \sum_{i}^{\{s,x,y,z\}} c_i(\mathbf{k}) \, |u_{i,0}\rangle \, , \tag{2.25}$$

with c_i some scalar coefficients. By generalizing the ideas introduced for the 2-band Hamiltoninan, the momentum operator between the conduction and the valence bands will obey

$$\langle u_s|p_i|u_j\rangle = P\delta_{ij} \quad i,j \in \{x,y,z\}. \tag{2.26}$$

We can use this new basis to write an improved version of $\hat{\mathcal{H}}_{k.p,2}$. First, we consider a single s-like conduction band and a 3-fold degenerate p-like valence band (one light-hole and two heavy-hole states). With the orbital basis, we can expand the 2-band Hamiltonian from Eq.(2.20) to a 4-band expression

$$\hat{\mathcal{H}}_{k.p,4} = \begin{array}{c} \begin{matrix} |u_s\rangle & |u_x\rangle & |u_y\rangle & |u_z\rangle \end{matrix} \\ \begin{pmatrix} E_c(0) + \frac{\hbar^2 k^2}{2m_0} & \frac{\hbar}{m_0} k_x P & \frac{\hbar}{m_0} k_y P & \frac{\hbar}{m_0} k_z P \\ \frac{\hbar}{m_0} k_x P^* & E_v(0) + \frac{\hbar^2 k^2}{2m_0} & 0 & 0 \\ \frac{\hbar}{m_0} k_y P^* & 0 & E_v(0) + \frac{\hbar^2 k^2}{2m_0} & 0 \\ \frac{\hbar}{m_0} k_z P^* & 0 & 0 & E_v(0) + \frac{\hbar^2 k^2}{2m_0} \end{pmatrix} \end{array} \tag{2.27}$$

In the present case, we only consider CB-VB coupling (represented by the $\hat{\mathcal{H}}_{1,i}$ and $\hat{\mathcal{H}}_{i,1}$ non-diagonal terms). To go beyond the scope of the effective mass model and account for the effects of other bands, we then introduce the Löwdin perturbation theory [9]. Indeed, distant bands (above the CB and below the VBs) also have a non-zero impact on the dispersion profile and can be indirectly included in the Hamiltonian, without being explicitly modeled. We first recall that the Hamiltonian of the system can be written

$$\sum_n \hat{\mathcal{H}}_{m,n} c_n = E_m c_m \;\longrightarrow \sum_{n \neq m} \hat{\mathcal{H}}_{m,n} c_n = (E_m - \hat{\mathcal{H}}_{m,m}) c_m. \tag{2.28}$$

Before introducing Kane's 4-band Hamiltonian, we focus on the valence bands only. Löwdin renormalization consists in splitting the bands into two categories [5]. Class A denotes the valence bands formed by the p-orbital states $|u_x\rangle$, $|u_y\rangle$ and $|u_z\rangle$, whereas all the other bands are from class B. The coefficients c can be obtained from Eq.(2.28) and divided according the classes that have just been defined. Thus, we can write

$$c_m = \sum_{i \neq m}^{A} \frac{\hat{\mathcal{H}}_{m,i}}{E_m - \hat{\mathcal{H}}_{m,m}} c_i + \sum_{j \neq m}^{B} \frac{\hat{\mathcal{H}}_{m,j}}{E_m - \hat{\mathcal{H}}_{m,m}} c_j \tag{2.29}$$

We want to eliminate the states from class B (*i.e.* the bands distant from the degenerate p-like valence bands) by a process of iterations, which is limited to the first order, such that

$$\begin{aligned} c_m &= \sum_{i\neq m}^{A} \frac{\hat{\mathcal{H}}_{m,i}}{E_m - \hat{\mathcal{H}}_{m,m}} c_i + \sum_{j\neq m}^{B} \frac{\hat{\mathcal{H}}_{m,j}}{E_m - \hat{\mathcal{H}}_{m,m}} \overbrace{\Big(\sum_{i\neq j}^{A} \frac{\hat{\mathcal{H}}_{j,i}}{E_j - \hat{\mathcal{H}}_{j,j}} c_i + \underbrace{\sum_{k\neq j}^{B} \frac{\hat{\mathcal{H}}_{j,k}}{E_j - \hat{\mathcal{H}}_{j,j}} c_k}_{=0 \text{ (first order)}} \Big)}^{c_j} \\ &= \sum_{i\neq m}^{A} \frac{c_i}{E_m - \hat{\mathcal{H}}_{m,m}} \underbrace{\Big(\hat{\mathcal{H}}_{m,i} + \sum_{j\neq m}^{B} \frac{\hat{\mathcal{H}}_{m,j}\hat{\mathcal{H}}_{j,i}}{E_j - \hat{\mathcal{H}}_{j,j}} \Big)}_{U_{m,i}} \end{aligned} \tag{2.30}$$

From this expression, we can extract the renormalized Hamiltonian U, which contains the effect of the remote bands on the valence bands. Thanks to this reformulation, we only have to solve the following eigenvalue problem, that is now limited to the bands from class A:

$$\sum_{n}^{A} (U^A_{m,n} - E_n \delta_{m,n}) c_m = 0 \tag{2.31}$$

The renormalized Hamiltonian can be fragmented as $U^A = \hat{\mathcal{H}} + \hat{\mathcal{H}}^{int}$. The left-hand contribution contains the diagonal valence-band terms already included in the $\hat{\mathcal{H}}_{k.p,4}$ Hamiltonian (Eq.(2.27)), that are

$$\hat{\mathcal{H}}_{m,n} = \langle u_{m,0} | \, E_m(0) + \frac{\hbar^2 k^2}{2m_0} \, | u_{n,0} \rangle = \Big(E_m(0) + \frac{\hbar^2 k^2}{2m_0} \Big) \delta_{m,n} \quad \text{(m,n} \in \text{A)}. \tag{2.32}$$

The left-hand Hamiltonian contains additional interaction terms and reads

$$\hat{\mathcal{H}}^{int}_{m,n} = \langle u_{m,0} | \frac{\hbar}{m_0} \mathbf{k} \cdot \mathbf{p} | u_{n,0} \rangle = \sum_{i}^{\{x,y,z\}} \frac{\hbar k_i}{m_0} \langle u_m | p_i | u_n \rangle \quad \text{(m} \in \text{A, n} \in \text{B)} \tag{2.33}$$

We are ultimately interested in including the supplementary effects generated by this Hamiltonian, to improve the precision of the model. To summarize, by combining these contributions, the expression of U^A is

$$\begin{aligned} U^A_{m,n} &= \Big(E_m(0) + \frac{\hbar^2 k^2}{2m_0} \Big) \delta_{m,n} + \sum_{\alpha,\beta}^{\{x,y,z\}} k_\alpha k_\beta \frac{\hbar^2}{m_0^2} \sum_{i\neq m,n}^{B} \frac{\langle u_m | p_\alpha | u_i \rangle \langle u_i | p_\beta | u_n \rangle}{E_0 - E_i} \\ &= E_m(0)\delta_{m,n} + \sum_{\alpha,\beta} k_\alpha k_\beta D^{\alpha,\beta}_{m,n} \end{aligned}, \tag{2.34}$$

with

$$D_{m,n}^{\alpha,\beta} = \frac{\hbar^2}{2m_0}\Big(\delta_{m,n}\delta_{\alpha,\beta} + \frac{1}{m_0}\sum_{i}^{B}\frac{\langle u_m|p_\alpha|u_i\rangle\langle u_i|p_\beta|u_n\rangle + \langle u_m|p_\beta|u_i\rangle\langle u_i|p_\alpha|u_n\rangle}{E_0 - E_n}\Big). \tag{2.35}$$

Kohn and Luttinger argue that this tensor plays the same role in the theory of degenerate bands (*i.e.* when multiple valence bands are present) that the effective mass in the non-degenerate case [10]. They propose to take advantage of the symmetries of the diamond lattice to express this matrix in the $|u_x\rangle$, $|u_y\rangle$, $|u_z\rangle$ basis as

$$D = k_\alpha k_\beta D^{\alpha,\beta} = \begin{array}{c} \begin{matrix} |u_x\rangle & |u_y\rangle & |u_z\rangle \end{matrix} \\ \begin{pmatrix} Lk_x^2 + M(k_y^2+k_z^2) & Nk_xk_y & Nk_xk_z \\ Nk_xk_y & Lk_y^2 + M(k_x^2+k_z^2) & Nk_yk_z \\ Nk_xk_z & Nk_yk_z & Lk_z^2 + M(k_x^2+k_y^2) \end{pmatrix} \end{array}. \tag{2.36}$$

The definition of the terms L, M and N, also given by Dresselhaus, Kip and Kittel [11], is

$$\begin{aligned} L \equiv D_{xx}^{xx} &= \frac{\hbar^2}{2m_0} + \underbrace{\frac{\hbar^2}{m_0^2}\sum_{n}^{\{x,y,z\}}\frac{|\langle u_x|p_x|u_n\rangle|^2}{E_v(0) - E_n}}_{L'}, \\ M \equiv D_{xx}^{yy} &= \frac{\hbar^2}{2m_0} + \underbrace{\frac{\hbar^2}{m_0^2}\sum_{n}^{\{x,y,z\}}\frac{|\langle u_x|p_y|u_n\rangle|^2}{E_v(0) - E_n}}_{M'}, \\ N \equiv D_{x,y}^{x,y} &= 0 + \frac{\hbar^2}{m_0^2}\sum_{n}^{\{x,y,z\}}\frac{\langle u_x|p_x|u_n\rangle\langle u_n|p_y|u_y\rangle + \langle u_x|p_y|u_n\rangle\langle u_n|p_x|u_y\rangle}{E_v(0) - E_n}. \end{aligned} \tag{2.37}$$

In practice, all these parameters can be experimentally calibrated by means of cyclotron resonance. Note that we are only interested in the right-hand element of L and M (denoted L' and M') since the left-hand term is already included in $\hat{\mathcal{H}}_{k.p,4}$.
Similarly, for the conduction band, Kane adds another set of terms [7, 12]

$$\begin{aligned} A &= \frac{\hbar^2}{m_0^2}\sum_{n}^{\{x,y,z\}}\frac{|\langle u_s|p_x|u_n\rangle|^2}{E_c(0) - E_n}, \\ B &= \frac{2\hbar^2}{m_0^2}\sum_{n}^{\{x,y,z\}}\frac{\langle u_s|p_x|u_n\rangle\langle u_n|p_y|u_z\rangle}{\big[E_c(0) + E_v(0)\big]/2 - E_n}, \end{aligned} \tag{2.38}$$

where B couples the conduction and the valence bands. Combining the VB terms L',

M', N, the CB term A and the CB-VB term B , the final renormalized interaction Hamiltonian in the $|u_s\rangle$, $|u_x\rangle$, $|u_y\rangle$, $|u_z\rangle$ basis reads

$$\hat{\mathcal{H}}^R_{k.p,4} = \begin{array}{c} \begin{matrix} |u_s\rangle & |u_x\rangle & |u_y\rangle & |u_z\rangle \end{matrix} \\ \begin{pmatrix} Ak^2 & Bk_yk_z & Bk_xk_z & Bk_xk_y \\ & L'k_x^2 + M'(k_y^2+k_z^2) & Nk_xk_y & Nk_xk_z \\ & & L'k_y^2 + M'(k_x^2+k_z^2) & Nk_yk_z \\ & & & L'k_z^2 + M'(k_x^2+k_y^2) \end{pmatrix} \end{array} \tag{2.39}$$

In the next section, we show how to combine $\hat{\mathcal{H}}_{k.p,4}$, $\hat{\mathcal{H}}^R_{k.p,4}$ and another spin-orbit matrix to form a 8-band Hamiltonian.

2.5 8-band Hamiltonian: spin-orbit interaction

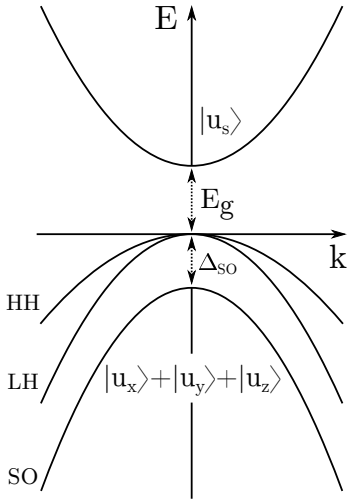


Figure 2.4: Schematic band structure of a direct-gap semiconductor with spin orbit splitting. The light-hole, heavy-hole and split-off valence bands result from linear combinations of the $|u_x\rangle$,$|u_y\rangle$ and $|u_z\rangle$ Bloch states.

As a consequence of relativistic effects, an electron moving in a potential V feels an effective magnetic field, that acts on its spin. The spin-orbit coupling increases with the atomic number of the atoms. Thus, the III-V compounds that contain heavy elements such as In, Ga, As, or Sb, are expected to be especially sensitive to this effect. The spin-orbit (SO) interaction can be included in the unperturbed (*i.e.* k independent) Hamiltonian with the following expression [13, 14]

$$\hat{\mathcal{H}}^{(0)} = \frac{-\hbar^2}{2m_0}\nabla^2 + \underbrace{\frac{\hbar}{4m_0^2c^2}(\sigma \times \nabla V)\cdot \mathbf{p}}_{SO}, \tag{2.40}$$

where σ corresponds to the Pauli matrices

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_y = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.41}$$

With this new contribution, Eq.(2.13) becomes:

$$\left(\hat{\mathcal{H}}^{(0)} + \frac{\hbar^2 k^2}{2m_0} + \frac{\hbar}{m_0}\mathbf{k}\cdot\pi\right)u_{n,k}(\mathbf{r}) = E(k)u_{n,k}(\mathbf{r}), \tag{2.42}$$

with

$$\pi = \mathbf{p} + \frac{\hbar}{4m_0c^2}(\sigma \times \nabla V). \tag{2.43}$$

Expanding the SO component in Eq.(2.42) yields

$$\hat{\mathcal{H}}_{SO} = \frac{\hbar}{4m_0c^2}(\sigma \times \nabla V)\cdot\mathbf{p} + \underbrace{\frac{\hbar^2}{4m_0c^2}(\sigma \times \nabla V)\cdot\mathbf{k}}_{neglected} \tag{2.44}$$

The right hand side term can be neglected, as we are working around $\mathbf{k} = 0$. To take the SO interaction into account in the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian, the basis has to be extended to consider the spin, and the Hamiltonian has to include not only four, but actually eight bands. In this new basis, the spin-orbit Hamiltonian can be built from the Pauli matrices [4]:

$$\hat{\mathcal{H}}_{\mathrm{SO},8} = \begin{array}{c} \begin{array}{cccccccc} |u_s\uparrow\rangle & |u_x\uparrow\rangle & |u_y\uparrow\rangle & |u_z\uparrow\rangle & |u_s\downarrow\rangle & |u_x\downarrow\rangle & |u_y\downarrow\rangle & |u_z\downarrow\rangle \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & i \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & -i & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -i & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & i & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \frac{\Delta_{\mathrm{SO}}}{3i} \tag{2.45}$$

where the spin-orbit splitting term can be expressed by different ways, using the symmetry properties of the orbitals

$$\begin{aligned} \frac{4m_0^2c^2}{3i\hbar}\Delta_{\mathrm{SO}} &= \langle u_x|\frac{\partial V}{\partial x}p_y - \frac{\partial V}{\partial y}p_x|u_y\rangle \\ &= \langle u_y|\frac{\partial V}{\partial y}p_z - \frac{\partial V}{\partial z}p_y|u_z\rangle \\ &= \langle u_z|\frac{\partial V}{\partial z}p_x - \frac{\partial V}{\partial x}p_z|u_x\rangle \end{aligned} \tag{2.46}$$

Here, we assume that the states at the valence band maximum are formed by electrons from the p-orbitals [8]. In the expression of $\hat{\mathcal{H}}_{\mathrm{SO},8}$, it can be observed that the spin-orbit contribution for the s-like states (corresponding to the CB) is zero. Indeed, the orbital

angular momentum $\mathbf{L}$ is equal to 0 in the s-like states and the spin-orbit interaction is proportional to $\mathbf{L}\cdot\mathbf{S}$ (with $\mathbf{S}$ the spin angular momentum). Finally, the CB-VB coupled, renormalized and SO $\mathbf{k}\cdot\mathbf{p}$ Hamiltonians can be assembled in the orbital basis with spin as

$$\hat{\mathcal{H}}^{\text{orb}}_{k.p,8} = \begin{pmatrix} \hat{\mathcal{H}}_{k.p,4} & 0 \\ 0 & \hat{\mathcal{H}}_{k.p,4} \end{pmatrix} + \begin{pmatrix} \hat{\mathcal{H}}^{R}_{k.p,4} & 0 \\ 0 & \hat{\mathcal{H}}^{R}_{k.p,4} \end{pmatrix} + \hat{\mathcal{H}}_{\text{SO},8} \tag{2.47}$$

$\lvert J, J_z\rangle$ state	Orbital combination	Energy at Γ
$\lvert u_{C\uparrow}\rangle = \lvert\frac{1}{2},\frac{1}{2}\rangle$	$i\lvert u_s\uparrow\rangle$	$E_g(E_c)$
$\lvert u_{C\downarrow}\rangle = \lvert\frac{1}{2},-\frac{1}{2}\rangle$	$i\lvert u_s\downarrow\rangle$	$E_g(E_c)$
$\lvert u_{HH\uparrow}\rangle = \lvert\frac{3}{2},\frac{3}{2}\rangle$	$-\frac{1}{\sqrt{2}}\big(\lvert u_x\uparrow\rangle + i\lvert u_y\uparrow\rangle\big)$	0 (E_v)
$\lvert u_{HH\downarrow}\rangle = \lvert\frac{3}{2},-\frac{3}{2}\rangle$	$-\frac{1}{\sqrt{2}}\big(\lvert u_x\downarrow\rangle - i\lvert u_y\downarrow\rangle\big)$	0 (E_v)
$\lvert u_{LH\uparrow}\rangle = \lvert\frac{3}{2},\frac{1}{2}\rangle$	$-\frac{1}{\sqrt{6}}\big(\lvert u_x\downarrow\rangle + i\lvert u_y\downarrow\rangle + 2\lvert u_z\uparrow\rangle\big)$	0 (E_v)
$\lvert u_{LH\downarrow}\rangle = \lvert\frac{3}{2},-\frac{1}{2}\rangle$	$\frac{1}{\sqrt{6}}\big(\lvert u_x\uparrow\rangle - i\lvert u_y\uparrow\rangle + 2\lvert u_z\downarrow\rangle\big)$	0 (E_v)
$\lvert u_{SO\uparrow}\rangle = \lvert\frac{1}{2},\frac{1}{2}\rangle$	$\frac{1}{\sqrt{3}}\big(\lvert u_x\downarrow\rangle + i\lvert u_y\downarrow\rangle + \lvert u_z\uparrow\rangle\big)$	$-\Delta_{SO}$
$\lvert u_{SO\downarrow}\rangle = \lvert\frac{1}{2},-\frac{1}{2}\rangle$	$-\frac{1}{\sqrt{3}}\big(\lvert u_x\uparrow\rangle - i\lvert u_y\uparrow\rangle + \lvert u_z\downarrow\rangle\big)$	$-\Delta_{SO}$

Table 2.2: Table of correspondence between the orbital basis $\lvert u_{\{s,x,y,z\}}\uparrow/\downarrow\rangle$ and the angular momentum basis $\lvert J, J_z\rangle$, quantized in the z direction [13, 15–17]. In this new basis, the states can be associated to specific bands, which simplifies the physical interpretation of the model.

The drawback of such an expression is that it does not allow distinguishing the different valence bands. Indeed, as shown in Fig.2.4, the light-hole, heavy-hole and spin-orbit VBs are formed by linear combinations of the p_x, p_y and p_z states. Another possibility is to express the Hamiltonian with the eigenstates of the total angular momentum $\lvert J, J_z\rangle$, with $\mathbf{J} = \mathbf{L} + \mathbf{S}$, in order to diagonalize the spin-orbit term proportional to $\mathbf{L}\cdot\mathbf{S}$ [13, 15, 16]. In this basis, the orbital angular momentum L = 0 for the s-like states, whereas it is equal to 1 in the case of p-like states. Since the spin angular momentum S=1/2 or -1/2, the total angular momentum will take the value 1/2 in the CB and 1/2 or 3/2 in the VBs. Tab.2.2 shows the linear combinations of the (orbital) Bloch states that diagonalize the spin-orbit interaction at $\mathbf{k} = 0$, in the $\lvert J, J_z\rangle$ basis. Bahder [17] applies such a change of basis to the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian. This

forms the Hamiltonian used in this work

$$\hat{\mathcal{H}}_{k.p,8} = \begin{array}{c} \begin{array}{cccccccc} |u_C\uparrow\rangle & |u_C\downarrow\rangle & |u_{HH}\uparrow\rangle & |u_{LH}\uparrow\rangle & |u_{HH}\downarrow\rangle & |u_{LH}\downarrow\rangle & |u_{SO}\uparrow\rangle & |u_{SO}\downarrow\rangle \end{array} \\ \begin{pmatrix} A & 0 & V^\dagger & 0 & \sqrt{3}V & -\sqrt{2}U & -U & \sqrt{2}V^\dagger \\ & A & -\sqrt{2}U & -\sqrt{3}V^\dagger & 0 & -V & \sqrt{2}V & U \\ & & -P+Q & -S^\dagger & R & 0 & \sqrt{\frac{3}{2}}S & -\sqrt{2}Q \\ & & & -P-Q & 0 & R & -\sqrt{2}R & \frac{1}{\sqrt{2}}S \\ & & & & -P-Q & S^\dagger & \frac{1}{\sqrt{2}}S^\dagger & \sqrt{2}R^\dagger \\ & & & & & -P+Q & \sqrt{2}Q & \sqrt{\frac{3}{2}}S^\dagger \\ & & & & & & -P-\Delta_{\mathrm{SO}} & 0 \\ & & & & & & & -P-\Delta_{\mathrm{SO}} \end{pmatrix} \end{array} \tag{2.48}$$

with the k-dependent parameters:

$$\begin{aligned} A &= E_c + \gamma_c \frac{\hbar^2}{2m_0}(k_x^2+k_y^2+k_z^2) \\ U &= \frac{1}{\sqrt{3}}P_0 k_z \\ V &= \frac{1}{\sqrt{6}}P_0(k_x - ik_y) \\ P &= -E_v + \gamma_1 \frac{\hbar^2}{2m_0}(k_x^2+k_y^2+k_z^2) \\ Q &= \gamma_2 \frac{\hbar^2}{2m_0}(k_x^2+k_y^2-2k_z^2) \\ R &= -\sqrt{3}\frac{\hbar^2}{2m_0}\Big(\gamma_2(k_x^2-k_y^2) - i\gamma_3 k_x k_y\Big) \\ S &= \sqrt{3}\ \gamma_3 \frac{\hbar^2}{m_0} k_z(k_x - ik_y) \end{aligned} \tag{2.49}$$

With this expression, it is possible to associate each line/column of the Hamiltonian with a specific band. The term A is related to the conduction band, while P, Q, R and S correspond to the valence bands. In addition, U and V are the CB-VB coupling terms and include the mixing parameter P_0 (similar to the P term in 2.19).
The expressions of the modified Luttinger parameters are

$$\gamma_1 = \gamma_1^L - \frac{E_P}{3E_G},\ \gamma_2 = \gamma_2^L - \frac{E_P}{6E_G},\ \gamma_3 = \gamma_3^L - \frac{E_P}{6E_G}, \tag{2.50}$$

and

$$\gamma_c = \frac{1}{m_c^*} - \frac{E_p}{3}\Big(\frac{2}{E_g} + \frac{1}{E_g+\Delta_{SO}}\Big). \tag{2.51}$$

The γ_n are derived from the Luttinger parameters γ_n^L that are listed, for III-V binary

and ternary compounds, in Ref. [6].
Fig.2.5 shows a comparison of 8-band and tight-binding Hamiltonians band structures for bulk InAs. Both methods give similar results around the Γ-point. A calibration of the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian with the empirical pseudopotential method can also be found in Ref. [18].

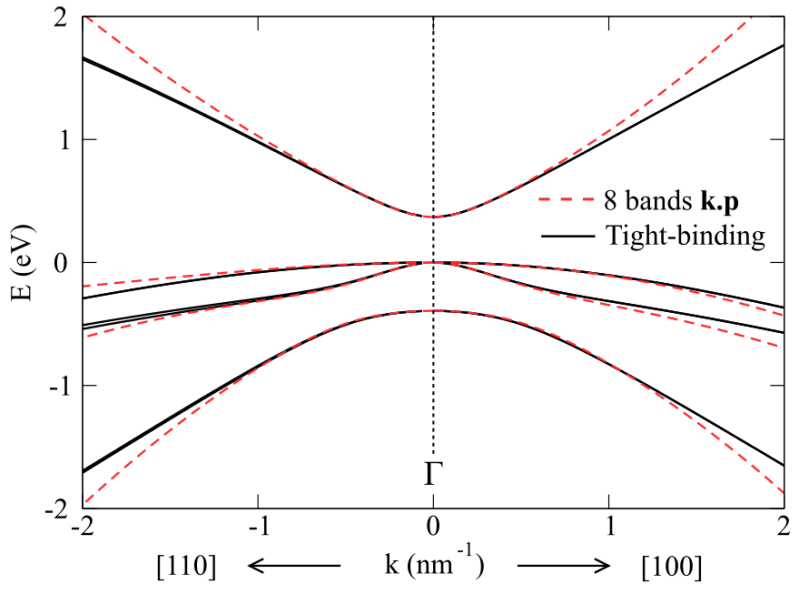


Figure 2.5: 8-band $\boldsymbol{k}\cdot\boldsymbol{p}$/tight-binding band structure comparison for bulk InAs along the [110] and [100] directions.

2.6 Numerical implementation for a 2D system

In order to use the $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian to treat transport problems, it is convenient to work in the real space representation. To that end, we have to assume a local basis that, in our case, will be given by the nodes of a discrete mesh.
To illustrate how this Hamiltonian is implemented in the code, we treat the case of a 2D system. The 8 band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian presented in Eq.2.48 is written in the k-space. In order to transfer it into the real space, we adopt the usual prescription of quantum mechanics (shown below) and then discretize the real space operator with a finite difference method. Hence, the transition from the momentum- to the position-space can be done *via*

$$\begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \longrightarrow -i \begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix}. \tag{2.52}$$

This replacement of the $\mathbf{k}$ wave vector is only done in the non-periodic directions of the system (here, x and z). Along the periodic dimension (y-axis) the k_y component is meant to be used as a parameter. The solutions for each k_y are computed independently

(in a selected range) and are later summed up to form the total solution. For example, the element A of $\hat{\mathcal{H}}_{8,k.p}$ (see Eq.(2.49)) becomes

$$A\psi = E_C\ \psi + \frac{\hbar^2}{2m}\Big(-\frac{\partial}{\partial x}\gamma_c\frac{\partial}{\partial x} - \gamma_c\ k_y^2 - \frac{\partial}{\partial z}\gamma_c\frac{\partial}{\partial z}\Big)\psi. \tag{2.53}$$

The system is now ready to be discretized with the finite difference method (Fig.2.6).

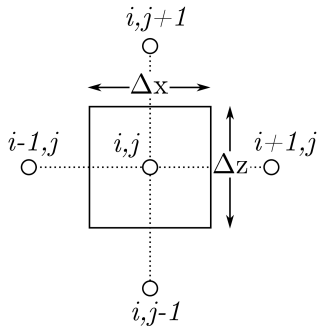


Figure 2.6: Finite differences discretization of a 2D grid. Each site is contained in a subdomain of area $\Delta x\Delta z$

In practice, it means that we can apply the following discretization procedure for the first and second derivatives of ψ around the node (i,j)

$$\begin{aligned} -i\frac{\partial}{\partial x}\psi\Big|_{i,j} &\approx -i\frac{\psi_{i+1,j} - \psi_{i-1,j}}{2\Delta x}, \\ -\frac{\partial^2}{\partial x^2}\psi\Big|_{i,j} &\approx -\frac{\psi_{i+1,j} + \psi_{i-1,j} - 2\psi_{i,j}}{\Delta x^2}. \end{aligned} \tag{2.54}$$

The same goes for k_z (and k_y in the case of a 3D problem). In our nanodevice simulations, we typically resort to a step size Δ of 0.2 nm. Thus, in our example, the discretization of A using the box integration method yields

$$\begin{aligned} A\psi_{i,j}\Delta x\Delta z = &(E_{Ci,j} + \gamma_{i,j}\frac{\hbar^2}{2m}k_y^2)\ \psi_{i,j}\Delta x\Delta z \\ &- \frac{\hbar^2}{2m}\Big(\Delta z\frac{\psi_{i+1,j} - \psi_{i,j}}{\Delta x}\ \frac{\gamma_{i+1,j} + \gamma_{i,j}}{2} + \Delta z\frac{\psi_{i-1,j} - \psi_{i,j}}{\Delta x}\ \frac{\gamma_{i-1,j} + \gamma_{i,j}}{2} \\ &+ \Delta x\frac{\psi_{i,j+1} - \psi_{i,j}}{\Delta z}\ \frac{\gamma_{i,j+1} + \gamma_{i,j}}{2} + \Delta x\frac{\psi_{i,j-1} - \psi_{i,j}}{\Delta z}\ \frac{\gamma_{i,j-1} + \gamma_{i,j}}{2}\Big). \end{aligned} \tag{2.55}$$

The other elements of $\hat{\mathcal{H}}_{k.p,8}$ can be discretized the same way. Let us divide the system in a series of vertical slices, arranged along the x-axis, in accordance with the procedure shown in Fig.2.7.

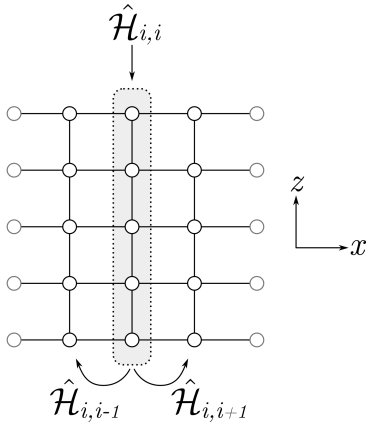


Figure 2.7: Schematic view of a 2D discretized system. Each dot corresponds to a site in the lattice and is connected to its nearest neighbors. At position $x = x_i$ The on-site term and the top/bottom interaction terms are contained in the $\hat{\mathcal{H}}_{i,i}$ submatrix, while the right/left couplings are accounted for in $\hat{\mathcal{H}}_{i,i\pm1}$.

In such a discretization, A can be written as matrices

$$A_{i,i} = \overbrace{\begin{pmatrix} \ddots & \ddots & & \\ \ddots & \mathcal{E}_j & \alpha_j & \\ & \alpha_j^\dagger & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}}^{N_z}, \quad A_{i,i+1} = \overbrace{\begin{pmatrix} \ddots & & \\ & \beta_j & \\ & & \ddots \end{pmatrix}}^{N_z}. \tag{2.56}$$

where N_z is the number of sites in the vertical direction. $A_{i,i}$ corresponds to the i-th slice and $A_{i,i+1}$ is responsible for the coupling between the i-th slice and its neighbor. The index j still denotes the vertical position.
The on-site terms $\mathcal{E}$ correspond to the (i,j) part of $A\psi_{i,j}\Delta x\Delta z$ and read

$$\begin{aligned}\mathcal{E}_j = &(E_{Ci,j} + \gamma_{i,j}k_y^2)\ \Delta x\Delta z \\ &+ \frac{\hbar^2}{2m}\Big[(\frac{\gamma_{i-1,j}+\gamma_{i,j}}{2} + \frac{\gamma_{i+1,j}+\gamma_{i,j}}{2})\frac{\Delta z}{\Delta x} + (\frac{\gamma_{i,j+1}+\gamma_{i,j}}{2} + \frac{\gamma_{i,j-1}+\gamma_{i,j}}{2})\frac{\Delta x}{\Delta z}\Big],\end{aligned} \tag{2.57}$$

while the hopping terms α and β correspond respectively to the $(i, j+1)$ and $(i+1, j)$ parts of $A\psi_{i,j}\Delta x\Delta z$

$$\begin{aligned}\alpha_j &= -\frac{\hbar^2}{2m}\frac{\Delta x}{\Delta z}\ \frac{\gamma_{i,j+1}+\gamma_{i,j}}{2}, \\ \beta_j &= -\frac{\hbar^2}{2m}\frac{\Delta z}{\Delta x}\ \frac{\gamma_{i+1,j}+\gamma_{i,j}}{2}.\end{aligned} \tag{2.58}$$

The same derivation can naturally be applied to the other terms of $\hat{\mathcal{H}}_{k.p,8}$. We can

then write the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian for the i-th slice as

$$\hat{\mathcal{H}}_{i,i'} = \overbrace{\begin{pmatrix} A_{i,i'} & 0 & \dots & & \\ 0 & A_{i,i'} & & & \\ \vdots & & \ddots & & \vdots \\ & & & -P_{i,i'} - \Delta^{SO}_{i,i'} & 0 \\ & & \dots & 0 & -P_{i,i'} - \Delta^{SO}_{i,i'} \end{pmatrix}}^{N_z \times N_b}, \tag{2.59}$$

where i' can take the values i, $i+1$ or $i-1$, and N_b is the number of bands. Here, $N_b = 8$ since this matrix is the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian (Eq.(2.48)) projected on a system of N_z points. The elements of this Hamiltonian are now themselves square (diagonal or tridiagonal) submatrices of rank N_z (as shown in Eq.(2.56)). As illustrated in Fig.2.8, these terms are defined in such a way that the in-slice $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian $\hat{\mathcal{H}}_{i,i}$ remains Hermitian, while the slice-slice hopping Hamiltonians only have to verify $\hat{\mathcal{H}}_{i+1,i} = \hat{\mathcal{H}}^{\dagger}_{i,i+1}$.
Once all the in-slice and coupling Hamiltonians $\hat{\mathcal{H}}_{i,i'}$ have been defined The **total** Hamiltonian takes the form of a block tridiagonal matrix

$$\hat{\mathcal{H}}_{tot} = \overbrace{\begin{pmatrix} \hat{\mathcal{H}}_{1,1} & \hat{\mathcal{H}}_{1,2} & & & \\ & \dots & & & \\ & \hat{\mathcal{H}}_{i,i-1} & \hat{\mathcal{H}}_{i,i} & \hat{\mathcal{H}}_{i,i+1} & \\ & & & \dots & \\ & & & \hat{\mathcal{H}}_{N_x-1,N_x} & \hat{\mathcal{H}}_{N_x,N_x} \end{pmatrix}}^{N_x \times N_z \times N_b}, \tag{2.60}$$

where N_x is the number of sites in the x direction. The total rank of the problem is thus $N_x \times N_z \times N_b$.

2.7 Coupled mode-space approach

Due to the size of the devices treated in this work, the computational cost of the simulations can rapidly increase (especially in the case of 3D simulations). In order to reduce the size of the problem, we resort to the coupled mode-space technique [19, 20]. In the real-space, the problem is defined by a block tridiagonal Hamiltonian, similar to that of Eq.(2.60). Thus, we have

$$(\hat{\mathcal{H}}_{i,i-1} + \hat{\mathcal{H}}_{i,i} + \hat{\mathcal{H}}_{i,i+i})\chi^m_i = E^m_i \chi^m_i, \tag{2.61}$$

where the m-th eigenfunction χ^m_i represents the so called m-th *mode* of the i-th slice. These $N_z N_b$ eigenfunctions can be arranged as column vectors in a matrix X, that takes the form

$$X_i = \left[\chi^1_i, \chi^2_i, ..., \chi^{NzNb}_i\right]. \tag{2.62}$$

These N_x rectangular matrices can, in turn, be used to generate the transformation

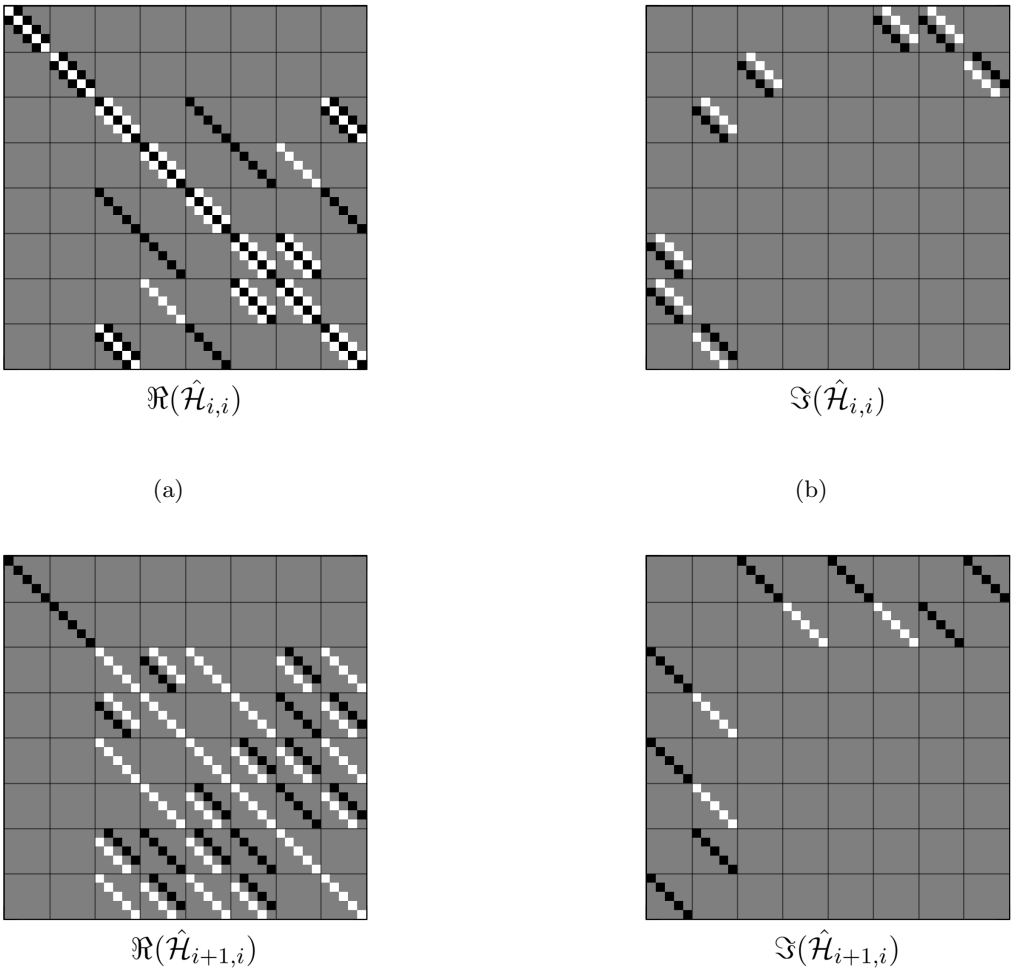


Figure 2.8: Graphical representation of the real and imaginary parts of the $\hat{\mathcal{H}}_{i,i}$ and $\hat{\mathcal{H}}_{i+1,i}$ 8×8 $\boldsymbol{k}\cdot\boldsymbol{p}$ Hamiltonians. This example corresponds to a 2D system with N_z=5, discretized along k_x and k_z (at k_y=0). The grid delimits the $N_z \times N_z$ submatrices. The white elements are positive numbers and the black ones are negative. For the total Hamiltonian $\hat{\mathcal{H}}_{tot}$ to be Hermitian, the diagonal Hamiltonian $\hat{\mathcal{H}}_{i,i}$ (and all its submatrices) must be Hermitian as well. However, $\hat{\mathcal{H}}_{i+1,i}$ can be non-Hermitian since it is an off-diagonal element of $\hat{\mathcal{H}}_{tot}$.

matrix $\mathcal{U}$, that reads

$$\mathcal{U} = \overbrace{\begin{pmatrix} X_1 & & & \\ & X_2 & & \\ & & \ddots & \\ & & & X_{Nx} \end{pmatrix}}^{N_x \times N_z \times N_b}. \tag{2.63}$$

This unitary matrix can be applied to the original real-space (RS) Hamiltonian in order to form the mode-space (MS) Hamiltonian, while preserving the tridiagonal form of

this operator:

$$\hat{\mathcal{H}}_{k.p}^{\mathrm{MS}} = \mathcal{U}^{\dagger}\, \hat{\mathcal{H}}_{k.p}^{\mathrm{RS}}\, \mathcal{U}. \tag{2.64}$$

In this equivalent reformulation of the 2D problem, $\hat{\mathcal{H}}_{k.p}^{\mathrm{MS}}$ still has a rank $N_x N_z N_b$ and does not present actual benefits with respect to $\hat{\mathcal{H}}_{k.p}^{\mathrm{RS}}$ yet. However, we can use a truncated version of X_i, by selecting only M modes. By doing so, the rank of the MS Hamiltonian becomes MN_x, which can indeed reduce the computational burden. In the case of 3D systems, the rank is even reduced from $N_x N_y N_z N_b$ to MN_x. For example, if we consider 20 modes in an 5×5 nm^2 cross-section nanowire, with a step size $\Delta z = \Delta y = 0.2$ nm, the rank of the Hamiltonian will be decreased 250-fold. Note that the mode-space approximation is only valid if we keep the lowest modes, since they are the most relevant for electron transport.

2.8 Appendix: Strain Hamiltonian

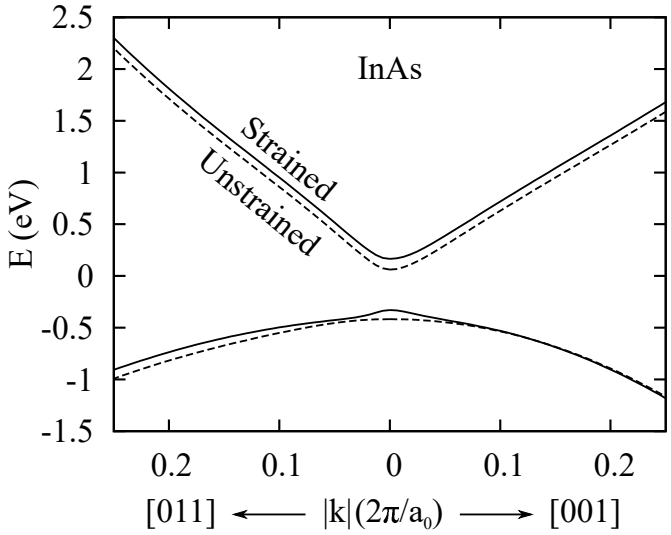


Figure 2.9: Effect of compressive strain on the band structure of bulk InAs

Following the work of Pikus and Bir [21], it is possible to include the effect of strain within the perturbation theory framework. The first order perturbation elements to add to the $\hat{\mathcal{H}}_{8,k.p}$ Hamiltonian are given by

$$\hat{\mathcal{H}}^{\text{strain}}_{m,n} = \sum_{i,j}^{\{x,y,z\}} -\mathbf{k}_i \epsilon_{i,j} \langle u_m | \mathbf{p}_j | u_n \rangle + \epsilon_{i,j} \langle u_m | \mathcal{D}_{i,j} | u_n \rangle \,, \tag{2.65}$$

where $\epsilon_{i,j}$ is the strain tensor and $\mathcal{D}$ is the deformation potential, which describes the effect of strain on the potential and kinetic energy of the electrons. The resulting eight-band $\mathbf{k} \cdot \mathbf{p}$ strain Hamiltonian is [17]

$$\hat{\mathcal{H}}_{\text{strain},8} = \begin{pmatrix} p' & 0 & -v^* & 0 & -\sqrt{3}v & \sqrt{2}u & u & -\sqrt{2}v^* \\ & p' & \sqrt{2}u & -\sqrt{3}v^* & 0 & v & -\sqrt{2}v & -u \\ & & -p+q & -s^* & r & 0 & \sqrt{3/2}s & -\sqrt{2}q \\ & & & -p-q & 0 & r & -\sqrt{2}r & s/\sqrt{2} \\ & & & & -p-q & s^* & s^*/\sqrt{2} & \sqrt{2}r^* \\ & & & & & -p+q & \sqrt{2}q & \sqrt{3/2}s \\ & & & & & & -p & 0 \\ & & & & & & & -p \end{pmatrix}, \tag{2.66}$$

where

$$
\begin{aligned}
p' &= a_c\ (\epsilon_{x,x} + \epsilon_{y,y} + \epsilon_{z,z}),\\
p &= a_v\ (\epsilon_{x,x} + \epsilon_{y,y} + \epsilon_{z,z}),\\
q &= b\ \big(\epsilon_{z,z} - (\epsilon_{x,x} + \epsilon_{y,y})/2\big)\\
u &= \frac{P_0}{\sqrt{3}}\ (\epsilon_{x,z}k_x + \epsilon_{y,z}k_y + \epsilon_{z,z}k_z),\\
v &= \frac{P_0}{\sqrt{6}}\ \big((\epsilon_{x,x} - i\epsilon_{x,y})k_x + (\epsilon_{x,y} - i\epsilon_{y,y})k_x + (\epsilon_{x,z} - i\epsilon_{y,z})k_z\big),\\
r &= \frac{\sqrt{3}}{2}b\ (\epsilon_{x,x} - \epsilon_{y,y}) - id\epsilon_{x,y},\\
s &= -d\ (\epsilon_{x,z} - i\epsilon_{y,z}),
\end{aligned}
\tag{2.67}
$$

where the Pikus-Bir deformation-potential constants a_v, b and d describe the VB-strain coupling, while a_c describes the CB-strain coupling. The deformation potentials of different III-V compounds can be found in [6]).

Bibliography

[1] L.D. Landau and E.M. Lifshitz. *Quantum Mechanics: Non-relativistic Theory, chap. 6*, volume 3. Pergamon Press, 3 edition, 1965.

[2] F. Laloe C. Cohen-Tannoudji, B. Diu. *Quantum Mechanics, chap. 11*, volume 2. Wiley, 1965.

[3] H. Ehrenreich. Band structure and transport properties of some 3–5 compounds. *Journal of Applied Physics*, 32(10):2155–2166, oct 1961.

[4] C. Galeriu. *k.p theory of semiconductor nanostructures.* PhD thesis, Worcester Polytechnic Institute, 2005.

[5] P. Marconcini and M. Macucci. The k.p method and its application to graphene, carbon nanotubes and graphene nanoribbons: the dirac equation. *Rivisita del Nuovo Cimento*, 34(8), 2011.

[6] L.R. Ram-Mohan I. Vurgaftman, J.R. Meyer. Band parameters for III–v compound semiconductors and their alloys. *Journal of Applied Physics*, 89(11):5815–5875, jun 2001.

[7] E.O. Kane. Energy band structure in p-type germanium and silicon. *Journal of Physics and Chemistry of Solids*, 1(1-2):82–99, sep 1956.

[8] J. Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures.* Cambridge Univ. Press, 2003.

[9] P. Lowdin. A note on the quantum-mechanical perturbation theory. *The Journal of Chemical Physics*, 19(11):1396–1401, nov 1951.

[10] W. Kohn J. M. Luttinger. Motion of electrons and holes in perturbed periodic fields. *Physical Review*, 97(4):869–883, feb 1955.

[11] G. Dresselhaus, A.F. Kip, and C. Kittel. Cyclotron resonance of electrons and holes in silicon and germanium crystals. *Physical Review*, 98(2):368–384, apr 1955.

[12] E. O. Kane. *Handbook on Semiconductors vol.1: Energy band theory.* W. Paul, 1982.

[13] S. L. Chuang. *Physics of Optoelectronic Devices.* Wiley, 1995.

[14] H. Kroemer. *Quantum mechanics: for engineering, materials science, and applied physics.* Englewood Cliffs, 1994.

[15] A. Zakharova, S. Yen, and K. Chao. Hybridization of electron, light-hole, and heavy-hole states in InAs/GaSb quantum wells. *Physical Review B*, 64(23), nov 2001.

[16] C. Pryor. Eight-band calculations of strained InAs/GaAs quantum dots compared with one-, four-, and six-band approximations. *Physical Review B*, 57(12):7190–7195, mar 1998.

[17] T. B. Bahder. Eight-band k·p model of strained zinc-blende crystals. *Physical Review B*, 41(17):11992–12001, jun 1990.

[18] F. Conzatti. *Numerical Simulation of Advanced CMOS and Beyond CMOS Devices*. PhD thesis, Univeristy of Udine, 2011.

[19] M.Luisier, A.ndreas Schenk, and W. Fichtner. Quantum transport in two- and three-dimensional nanoscale transistors: Coupled mode effects in the nonequilibrium green's function formalism. *Journal of Applied Physics*, 100(4):043713, aug 2006.

[20] M. Shin. Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the k·p method. *Journal of Applied Physics*, 106(5):054505, sep 2009.

[21] G. L. Bir and G. E. Pikus. *Symmetry and Strain- Induced sects in Semiconductors*. Wiley, 1974.

CHAPTER 3

Electron transport

In which we show that we can compute the non-equilibrium Green's functions to extract meaningful physical properties, essential to simulate quantum transport – and then give extensive details about the computational implementation of the methods.

The ultimate aim of our approach is to compute the current density inside the nanodevices. This quantity depends on many phenomena occurring inside the system, like phonon scattering, wave interferences, tunneling or quantum confinement. Up to this point, we have only shown how to build an isolated quantum system, by describing a physical system by its Hamiltonian and calculating its energy structure. We know the available energy levels, but we still have to determine how the electrons flow through the channel. Simulating electron transport at the nanoscale entails solving a non-equilibrium statistical problem. Different approaches exist, but we choose to use the non-equilibrium Green's functions (NEGF) formalism [1–4], motivated by two main reasons:

- It not feasible to solve the eigenvalue problem for the whole system. Indeed, the Schrödinger equation would include not only the device itself, but also the electron reservoirs. The dimension of the Hilbert space describing a quantum system grows with the number of particles. So does the complexity of the algorithm and the computation time. Green's functions offer a decent compromise to estimate the observable quantities we are interested in, without having to calculate the wave functions of the entire system.
- Even though we are dealing with nanometric devices, these systems are far from being completely ballistic and scattering processes must be properly taken into account. In particular, the electron-phonon coupling is an unavoidable source of scattering at room temperature. The NEGF is a versatile framework, which enables us to implement such a many-body interaction in a computationally convenient way.

In practice, the studied devices are connected to their environment *via* the contacts (or leads), which can be modeled by infinite electron reservoirs. By allowing new carriers to enter and leave the device, these leads bring the system out of equilibrium. NEGF outputs the energy availability and occupancy of the quantum states (in a given energy range), in order to evaluate the non-equilibrium current. Since the complexity of the problem does not allow for an analytical solution, the computation is numerical (see Sec.3.5). For the charge, the energy and the current to be conserved, one will have to resort to self-consistent calculations.
Even though the NEGF formalism can, in principle, handle time-dependent problems, **we will eventually move to the energy domain, where we will assume the**

existence of stationary solutions. Those steady states will correspond to successive steps in the device operation, that can be computed independently from one another.

3.1 Preliminary concepts

On a mathematical point of view, Green's functions can be applied to problems of the form

$$\hat{\mathcal{O}}\ \psi(x) = f(x) = \int dx'\ \delta(x - x')f(x'), \tag{3.1}$$

where $\hat{\mathcal{O}}$ is a linear differential operator, ψ is analogous to the system's wavefunction and f plays the role of an external contribution – for example, a force. If f is a continuous function, then it can be written as a Dirac distribution, as shown on the right-hand side of Eq.(3.1). This ordinary differential equation can be solved by the knowledge of a Green's function G, defined as

$$\hat{\mathcal{O}}\ G(x - x')\ =\ \delta(x - x'). \tag{3.2}$$

Indeed, by combining Eq.(3.2) and Eq.(3.1), one can solve the problem for ψ

$$\begin{aligned} \hat{\mathcal{O}}\ \psi(x) &= \int \mathrm{d}x'\ \hat{\mathcal{O}}\ G(x - x')f(x'), \\ \psi(x) &= \int \mathrm{d}x'\ G(x - x')f(x'). \end{aligned} \tag{3.3}$$

Physically, the Green's function can be seen as the relation which links a cause f to its effect ψ. The x variable can refer to a position in space, time, or to a space-time coordinate (in which case $\hat{\mathcal{O}}G = \delta(t)\delta(\mathbf{r})$). To improve readability, we now resort to the following notation

$$G(x', x) \equiv G(x' - x). \tag{3.4}$$

3.1.1 Green's functions in quantum mechanics

Since the time-dependent Schrödinger equation complies with the above definition, one can apply Eq.(3.2) to a one-particle Hamiltonian

$$\left(i\hbar\frac{\partial}{\partial t} - \hat{\mathcal{H}}\right)G(\mathbf{r}', t'; \mathbf{r}, t) = \delta(\mathbf{r}' - \mathbf{r})\delta(t' - t). \tag{3.5}$$

In this context, the Green's function is often called the *propagator*. It can be expressed as

$$G(\mathbf{r}', t'; \mathbf{r}, t) = -\frac{i}{\hbar}\ \Theta(t' - t)\ \langle \mathbf{r}'|\hat{U}(t', t)|\mathbf{r}\rangle\ , \tag{3.6}$$

where Θ is the *Heaviside* step function, which ensures that $t' > t$. The matrix element on the right contains the probability amplitude of a transition from state $\{\mathbf{r}, t\}$ to state

$\{\mathbf{r}', t'\}$, performed in two steps

$$\{\mathbf{r}, t\} \xmapsto{\hat{U}(t',t)} \{\mathbf{r}, t'\} \xmapsto{\langle \mathbf{r}'|} \{\mathbf{r}', t'\}$$

$\hat{U}$ is the *evolution operator*, transforming the quantum state at time t into a quantum state at time t'

$$\boxed{|\Psi(t')\rangle = \hat{U}(t', t)\,|\Psi(t)\rangle} \tag{3.7}$$

In the context of a time-independent Hamiltonian, $\hat{U}$ simply reads

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\hat{U}(t', t) &= \hat{\mathcal{H}}\hat{U}(t', t) \\ \hat{U}(t', t) &= e^{-i(t'-t)\hat{\mathcal{H}}/\hbar} \end{aligned}. \tag{3.8}$$

By Stone's theorem, this operator is unitary, since it can be expressed as the exponential of $\hat{\mathcal{H}}$, which is a self-adjoint operator

$$\hat{U}^\dagger(t', t)\hat{U}(t', t) = \hat{U}(t', t)\hat{U}(t, t') = I. \tag{3.9}$$

Eq.(3.7) corresponds to the so called *Schrödinger picture* of quantum mechanics, where the states are the only time-dependent parts of the equation. However, in this work, it will be convenient to introduce another quantum mechanics scheme, detailed hereafter.

3.1.2 Heisenberg picture

In the *Heisenberg picture*, the quantum state is time invariant, while the time dependence is shifted to the operators. This representation is connected to the Schrödinger picture by the expressions

$$\begin{cases} |\psi_H\rangle = \hat{U}(0, t)\,|\psi_S(t)\rangle = |\psi_S(0)\rangle \\ \hat{\mathcal{O}}_H(t) = \hat{U}(0, t)\ \hat{\mathcal{O}}_S\ \hat{U}(t, 0) \end{cases} \tag{3.10}$$

Additionally, in the Heisenberg picture, the time evolution of an operator is given by

$$i\hbar\frac{d}{dt}\hat{\mathcal{O}}_H(t) = [\hat{\mathcal{O}}_H(t), \hat{\mathcal{H}}], \tag{3.11}$$

where the square brackets denote the commutator.
In the rest of this work, any operator associated with a time variable will belong to the Heisenberg scheme. The subscript "H" will often be omitted for readability purposes.

In the case of a time-dependent Hamiltonian, the expression of $\hat{U}$ becomes [5]

$$\hat{U}(t', t) = 1 - \frac{i}{\hbar}\int_t^{t'} d\tau\ \hat{\mathcal{H}}(\tau)\hat{U}(\tau, t), \tag{3.12}$$

where the evolution operator is calling itself recursively. This can be developed as

$$\hat{U}(t',t) = 1 + \sum_{n=1}^{\infty} \Big(-\frac{i}{\hbar} \Big)^n \int_t^{t'} dt_0 \cdots \int_t^{t_n} dt_n \hat{\mathcal{H}}(t_0) \cdots \hat{\mathcal{H}}(t_n). \tag{3.13}$$

The order in which the Hamiltonians are written is important and we can resort to the following *time-ordering* procedure to rearrange them

$$\begin{aligned} \mathcal{T}[\hat{\mathcal{H}}(t')\hat{\mathcal{H}}(t)] &= \Theta(t'-t)\hat{\mathcal{H}}(t')\hat{\mathcal{H}}(t) + \Theta(t-t')\hat{\mathcal{H}}(t)\hat{\mathcal{H}}(t') \\ &= \begin{cases} \hat{\mathcal{H}}(t)\hat{\mathcal{H}}(t') & t' > t \\ \hat{\mathcal{H}}(t')\hat{\mathcal{H}}(t) & t' < t \end{cases} \end{aligned} \tag{3.14}$$

The expression for $\hat{U}(t',t)$ becomes

$$\hat{U}(t',t) = 1 + \sum_{n=1}^{\infty} \frac{(-\frac{i}{\hbar})^n}{n!} \int_t^{t'} dt_0 \cdots \int_t^{t_n} dt_n \; \mathcal{T}[\hat{\mathcal{H}}(t_0) \cdots \hat{\mathcal{H}}(t_n)], \tag{3.15}$$

which is the series expansion of the exponential function.
To account for the adjoint of the evolution operator (which corresponds to a backwards propagation in time), let us split the time-ordering procedure into $\mathcal{T}^+$ and $\mathcal{T}^-$, corresponding to its *chronological* ($t' > t$) and *antichronological* ($t' < t$) version. $\hat{U}$ can finally be written

$$\begin{aligned} \hat{U}(t',t) &= \mathcal{T}^+ \exp\Big(-\frac{i}{\hbar} \int_t^{t'} d\tau \; \hat{\mathcal{H}}(\tau) \Big) \\ \hat{U}^\dagger(t',t) &= \mathcal{T}^- \exp\Big(\frac{i}{\hbar} \int_t^{t'} d\tau \; \hat{\mathcal{H}}(\tau) \Big) \\ &= \hat{U}(t,t') \end{aligned} \tag{3.16}$$

3.2 Second quantization in many-fermion systems

The second quantization formalism allows us to describe a many-particle system in a very convenient way, able to automatically take into account the Pauli principle for identical particles.

3.2.1 Creation and annihilation operators

The *creation* operator $\hat{c}^\dagger$ is a procedure which turns an empty state $|0\rangle$ into an occupied state $|n\rangle$

$$\hat{c}_n^\dagger |0\rangle = |n\rangle \,. \tag{3.17}$$

In the case of electrons (which are fermions), the Pauli exclusion principle allows only one particle per state. For that reason, in this section, any state denoted by anything

else than "0" corresponds to an occupied state, containing one particle. When dealing with multiple states, one can use the *Fock state* notation

$$\begin{aligned} |\{n\}\rangle &\equiv |n_1, n_2, ..., n_N\rangle\,, \\ |n_i\rangle &\equiv |0, 0, .., n_i, ..., 0\rangle\,, \\ |0\rangle &\equiv |0, 0, ..., 0\rangle\,. \end{aligned} \tag{3.18}$$

Using this notation, Eq.(3.17) can then be generalized as follows:

$$\begin{aligned} |n, m\rangle &= \hat{c}_m^\dagger\,|n\rangle = \hat{c}_m^\dagger \hat{c}_n^\dagger\,|0\rangle\,, \\ |n_1, ..., n_N\rangle &= \hat{c}_N^\dagger\,|n_1, ..., n_{N-1}\rangle = \hat{c}_N^\dagger ... \hat{c}_1^\dagger\,|0\rangle\,. \end{aligned} \tag{3.19}$$

Additionally, with fermions, exchanging any two particles will result in a sign change of the state ket, such that

$$|..., n_i, ..., n_j, ...\rangle = -\,|..., n_j, ..., n_i, ...\rangle\,. \tag{3.20}$$

Applying this rule to Eq.(3.19) yields

$$\hat{c}_m^\dagger \hat{c}_n^\dagger\,|0\rangle = |n, m\rangle = -\,|m, n\rangle = -\hat{c}_n^\dagger \hat{c}_m^\dagger\,|0\rangle\,, \tag{3.21}$$

from which one can deduce an important property of the creation operator

$$\hat{c}_m^\dagger \hat{c}_n^\dagger + \hat{c}_n^\dagger \hat{c}_m^\dagger \equiv \left[\hat{c}_m^\dagger, \hat{c}_n^\dagger\right]_+ = 0, \tag{3.22}$$

where we have introduced the *(anti)commutator* notation

$$\left[\hat{A}, \hat{B}\right]_\pm = \hat{A}\hat{B} \pm \hat{B}\hat{A}. \tag{3.23}$$

Eq.(3.22) is a consequence of the Pauli exclusion principle, since it implies that $\hat{c}_n^\dagger \hat{c}_n^\dagger = 0$, which prevents a fermion from being created twice in the same state.

Similarly, the *annihilation* operator $\hat{c}_n$ acts on a filled state $|n\rangle$ to turn it into an empty state $|0\rangle$

$$\hat{c}_m\,|n\rangle = \delta_{m,n}\,|0\rangle\,, \tag{3.24}$$

where the Kronecker delta function δ prevents the state n from being annihilated by $\hat{c}_{m\neq n}$ (and more specifically, protects the empty state from undergoing any further annihilation). This operator anticommutes the same way than the creation operator (Eq.(3.22))

$$\left[\hat{c}_m, \hat{c}_n\right]_+ = \left[\hat{c}_m^\dagger, \hat{c}_n^\dagger\right]_+ = 0. \tag{3.25}$$

Last but not least, the most useful commutation relation for the rest of this work is

$$\boxed{\left[\hat{c}_m, \hat{c}_n^\dagger\right]_+ = \delta_{m,n} = \langle m|n\rangle} \tag{3.26}$$

Note that creating an electron equates to annihilating a hole and conversely. For example, $\hat{c}_m^\dagger \hat{c}_n \left|0,n\right\rangle = \left|m,0\right\rangle$ first removes an electron from state n and places it in state m. Symmetrically, $\hat{c}_m \hat{c}_n^\dagger \left|m,0\right\rangle = \left|0,n\right\rangle$ "fills" the hole at state n and then creates a new hole at state m.

3.2.2 Statistical ensemble

The density operator is a statistical tool used to describe a quantum system in a mixture of states

$$\hat{\rho} = \sum_i p_i \left|\psi_i\right\rangle \left\langle\psi_i\right|, \tag{3.27}$$

with $p_i = \left\langle\psi_i\right| \hat{\rho} \left|\psi_i\right\rangle$ the probability for the system to be in the state $\left|\psi_i\right\rangle$. The off-diagonal elements $\hat{\rho}_{i,j}$ correspond to a transition probability. The trace of $\hat{\rho}$ can be easily calculated by using the basis for which it is diagonal

$$\mathrm{Tr}(\hat{\rho}) = \sum_n \hat{\rho}_{n,n} = \sum_{i,n} p_i \left\langle\psi_n|\psi_i\right\rangle \left\langle\psi_i|\psi_n\right\rangle = \sum_i p_i \left\langle\psi_i\right| I \left|\psi_i\right\rangle = \sum_i p_i \overset{\text{def}}{=} 1. \tag{3.28}$$

At thermal equilibrium, the operator reads

$$\hat{\rho} = \frac{e^{-(\hat{\mathcal{H}}-\mu N)/k_\mathrm{B}T}}{\mathrm{Tr}(e^{-(\hat{\mathcal{H}}-\mu N)/k_\mathrm{B}T})}, \tag{3.29}$$

where N is the number operator, and the denominator serves as a normalization factor (or *partition function*), so that $\sum p_i = 1$
The density matrix is useful to describe the *expectation value* of an operator. This corresponds to the average eigenvalue of an operator and is denoted

$$\langle\hat{\mathcal{O}}\rangle \equiv \left\langle\psi\right| \hat{\mathcal{O}} \left|\psi\right\rangle, \tag{3.30}$$

where $\left|\phi\right\rangle$ is a pure state. Using $\langle m|n\rangle = \mathrm{Tr}(\left|n\right\rangle \left\langle m\right|)$, one can derive

$$\langle\hat{\mathcal{O}}\rangle = \sum_i p_i \left\langle\psi_i\right| \hat{\mathcal{O}} \left|\psi_i\right\rangle = \sum_i p_i \ \mathrm{Tr}\big(\left|\psi_i\right\rangle \left\langle\psi_i\right| \hat{\mathcal{O}}\big) = \boxed{\mathrm{Tr}(\hat{\rho}\hat{\mathcal{O}})}. \tag{3.31}$$

Let us get back to second quantization and define the *number operator* $\hat{n}$ which returns 0 or 1 when applied to an empty or a filled fermion state

$$\hat{n}_i = \hat{c}_i^\dagger \ \hat{c}_i. \tag{3.32}$$

While this operator only has two possible eigenvalues, its expectation value can still be anywhere between 0 and 1, and corresponds to the average occupation number n of the targeted state

$$\boxed{n_i = \langle \hat{n}_i \rangle = \langle \hat{c}_i^\dagger \hat{c}_i \rangle = \mathrm{Tr}(\hat{\rho}\ \hat{c}_i^\dagger \hat{c}_i) = f_i} \tag{3.33}$$

Where we can see that, at equilibrium, the states occupancy follows a Fermi-Dirac distribution

$$f_i = \frac{1}{e^{(E_i - \mu)/k_\mathrm{B}T} + 1}. \tag{3.34}$$

3.3 Equilibrium Green's functions

In order to make the connection with the computer simulations, we shall now employ discrete space variables for the Green's functions and use the matrix notation

$$G_{i,j}(t', t) \equiv G(\mathbf{r}', t'; \mathbf{r}, t).$$

The time indexes are kept as continuous variables, as they are useful for the forthcoming derivations, but are not part of the final simulations.

Before computing the electrical current and the many-body phenomena that result from it, we first consider a non-interacting Hamiltonian. In the second quantization formalism, this Hamiltonian can be written as

$$\hat{\mathcal{H}} = \sum_i E_i\ \hat{c}_i^\dagger \hat{c}_i + \sum_{i \neq j} t_{i,j}\ \hat{c}_i^\dagger \hat{c}_j, \tag{3.35}$$

where the diagonal terms $\hat{\mathcal{H}}_{i,i} = E_i$ are the energies of each state and the off-diagonal terms $\hat{\mathcal{H}}_{i,j} = t_{i,j}$ are the hopping energies between two given states.
Following the idea expressed in Sec.3.1.1 and using the second quantization tools introduced in Sec.3.2, one defines the *chronological* and *antichronological* Green's functions G^+ and G^-

$$G^\pm_{i,j}(t', t) = -\frac{i}{\hbar} \frac{\langle \psi_0 |\ \mathcal{T}^\pm[\hat{c}_i(t') \hat{c}_j^\dagger(t)]\ |\psi_0\rangle}{\langle \psi_0 | \psi_0 \rangle}, \tag{3.36}$$

where $|\psi_0\rangle$ is the ground state of the system. Using the expectation value notation of Eq.(3.30), this expression can be compacted as

$$\boxed{G^\pm_{i,j}(t', t) = -\frac{i}{\hbar} \langle \mathcal{T}^\pm[\hat{c}_i(t') \hat{c}_j^\dagger(t)] \rangle} \tag{3.37}$$

Let us analyze the physical meaning of this expression before moving any further.

Depending on the relative position of t and t' along the real axis, G^+ can take two forms

$$G^+_{i,j}(t',t) = \begin{cases} -\frac{i}{\hbar}\,\langle \hat{c}_i(t')\hat{c}^\dagger_j(t)\rangle & t' > t \\ -\frac{i}{\hbar}\,\langle \hat{c}^\dagger_j(t)\hat{c}_i(t')\rangle & t' < t \end{cases}$$

The top expression gives the probability amplitude for a hole to move from $\{j,t\}$ to $\{i,t'\}$ whereas the bottom one corresponds to an electron going from $\{i,t'\}$ to $\{j,t\}$. The time-ordering operator has ensured that the operators act forward in time. Hence the name "chronological" for this Green's function. Similarly, G^- gives the probability amplitude for an electron or a hole to travel between two sites, backward in time.
If we drop the time ordering operator, the (anti)chronological Green's function can be split as follows

$$G^\pm_{i,j}(t',t) = \Theta(t'-t)\; G^{\gtrless}_{i,j}(t',t) + \Theta(t-t')\; G^{\lessgtr}_{i,j}(t',t), \tag{3.38}$$

where we define the *greater* and *lesser* Green's functions $G^>$ and $G^<$ as

$$\boxed{\begin{aligned} G^>_{i,j}(t',t) &\equiv -\frac{i}{\hbar}\,\langle \hat{c}_i(t')\hat{c}^\dagger_j(t)\rangle \\ G^<_{i,j}(t',t) &\equiv \;\; \frac{i}{\hbar}\,\langle \hat{c}^\dagger_j(t')\hat{c}_i(t)\rangle \end{aligned}} \tag{3.39}$$

Which are related to G^+ and G^- by

$$G^+ + G^- = G^> + G^<. \tag{3.40}$$

$G^<$ and $G^>$ are only connected to one type of particle (respectively an electron or a hole). However, the direction in time can be either forward or backward (depending on the sign of $t'-t$).
These Green's functions can be combined to form another set of functions: the so called *retarded* and *advanced* Green's functions $G^R = (G^A)^\dagger$

$$\begin{aligned} G^R_{i,j}(t',t) &\equiv \Theta(t'-t)\,\Big(G^>_{i,j}(t',t) - G^<_{i,j}(t',t)\Big), \\ G^A_{i,j}(t',t) &\equiv \Theta(t-t')\,\Big(G^<_{i,j}(t',t) - G^>_{i,j}(t',t)\Big). \end{aligned} \tag{3.41}$$

Hence

$$\boxed{\begin{aligned} G^R_{i,j}(t',t) &= -\frac{i}{\hbar}\,\langle \big[\hat{c}_i(t'),\hat{c}^\dagger_j(t)\big]_+\rangle\,\Theta(t'-t) \\ G^A_{i,j}(t',t) &= \;\; \frac{i}{\hbar}\,\langle \big[\hat{c}_i(t'),\hat{c}^\dagger_j(t)\big]_+\rangle\,\Theta(t-t') \end{aligned}} \tag{3.42}$$

also connected to $G^<$ and $G^>$ by

$$G^R - G^A = G^> - G^<. \tag{3.43}$$

Finding a physical meaning for these retarded and advanced quantities is less straightforward, since they contain both electron and hole contributions (which arise when one develops the anticommutator). G^R and G^A will mainly be used to determine the density of states. In the following sections, we will show that we can ultimately discard G^A since it does not provide any supplementary information. Indeed, the stationary state time invariance implies that $G^A = (G^R)^\dagger$.

3.3.1 Energy domain

In this work, we are not actually interested in solving the time-dependent Green's functions. As explained in this chapter's introduction, we are looking for the steady-state solution, within a given energy window. If we consider a diagonal Hamiltonian, the elements of the evolution operator are

$$\hat{U}_{n,n}(t',t) = e^{-iE_n(t'-t)/\hbar}. \tag{3.44}$$

Using the anticommutation relation of Eq.(3.26) and the fact that $\hat{c}_i^{(\dagger)}(t) = \hat{U}^{(\dagger)}(t,0)\hat{c}_i^{(\dagger)}(0)$, we find

$$\big[\hat{c}_i(t'), \hat{c}_j^\dagger(t)\big]_+ = \hat{U}_{i,i}(t',t)\big[\hat{c}_i(0), \hat{c}_j^\dagger(0)\big]_+ = \hat{U}_{i,i}(t',t)\ \delta_{i,j}, \tag{3.45}$$

which allows us to re-express the retarded and advanced Green's functions

$$G_{i,j}^{R/A}(t',t) = \mp\frac{i}{\hbar}\hat{U}_{i,j}(t',t)\ \Theta(\pm t \mp t'). \tag{3.46}$$

Moreover, based on Eq.(3.33), the lesser/greater equilibrium Green's functions become

$$\begin{aligned} G_{i,j}^{<}(t',t) &= \ \ \frac{i}{\hbar}\hat{U}_{i,j}(t',t)\ f_i\ \delta_{i,j}, \\ G_{i,j}^{>}(t',t) &= -\frac{i}{\hbar}\hat{U}_{i,j}(t',t)\big(1-f_i\big)\delta_{i,j}. \end{aligned} \tag{3.47}$$

We can now move from the time-domain to the energy-domain by resorting to a Fourier transform

$$G^{R/A}(E) = \int dt\ G^{R/A}(t)e^{iEt/\hbar} \longrightarrow\ G_{i,i}^{R/A}(E) = \mp\frac{i}{\hbar}\int_{0/-\infty}^{\infty/0} dt\ e^{i(E-E_i)t/\hbar}, \tag{3.48}$$

which ultimately yields

$$\boxed{G_{i,i}^{R/A}(E) = \frac{1}{E \pm i\epsilon - E_i}} \tag{3.49}$$

where $i\epsilon$ is a small imaginary part introduced to help convergence.

Let us now introduce the *spectral function*, which provides information about the density of states (also useful in Sec.3.3.3) and reads

$$A(E) = i\big(G^R(E) - G^A(E)\big) = -2\ \Im(G^R(E)). \tag{3.50}$$

From Eqs.(3.49) and (3.50), and from the Cauchy identity $\Im[1/(x \pm i\epsilon)] = \mp i\pi\delta(x)$, it follows

$$A_{i,i}(E) = 2\pi\ \delta(E - E_i). \tag{3.51}$$

As for the lesser and greater Green's functions, they are related to the density of occupied states, since

$$\begin{aligned} G^<_{i,i}(E) &= -f_i(G^R_{i,i}(E) - G^A_{i,i}(E)) = i f_i\ A_{i,i}(E), \\ G^>_{i,i}(E) &= (1 - f_i)(G^R_{i,i}(E) - G^A_{i,i}(E)) = -i(1 - f_i)A_{i,i}(E), \end{aligned} \tag{3.52}$$

$$\boxed{\begin{aligned} G^<_{i,i}(E) &= \ \ 2\pi i\ f_i\ \delta(E - E_i) \\ G^>_{i,i}(E) &= -2\pi i\ (1 - f_i)\ \delta(E - E_i) \end{aligned}} \tag{3.53}$$

We have obtained the equilibrium steady-state Green's functions expressed in terms of energy and position. The next step is to show how the retarded, greater and lesser Green's functions can actually be used to compute physical quantities of interest.

3.3.2 Electron and hole densities

The *charge distribution* is one of the most basic quantities that one can observe in order to understand the behavior of a nanodevice. From the definition of the number operator of Eq.(3.33), we can obtain the *electron and hole occupation probability* for the i-th state

$$\begin{aligned} n_i(t) &= \langle \hat{c}_i^\dagger(t)\hat{c}_i(t)\rangle = -i\hbar\ G^<_{i,i}(t), \\ p_i(t) &= \langle \hat{c}_i(t)\hat{c}_i^\dagger(t)\rangle = \ \ i\hbar\ G^>_{i,i}(t). \end{aligned} \tag{3.54}$$

As we just explained, we are interested in the steady-state solutions and may therefore translate these equations in the energy domain

$$\boxed{\begin{aligned} n_i &= -\frac{i}{2\pi}\int \mathrm{d}E\ G^<_{i,i}(E) \\ p_i &= \ \ \frac{i}{2\pi}\int \mathrm{d}E\ G^>_{i,i}(E) \end{aligned}} \tag{3.55}$$

We can then readily obtain the electron and hole charges by multiplying these carrier concentrations by the electronic charge.

3.3.3 Density of states

The *density of states* is another essential quantum transport quantity. It is related to the spectral function and thus (see Eq.(3.50)) to the retarded Green's function by

$$\boxed{D_i(E) = \frac{1}{2\pi}\, A_{i,i}(E) = -\frac{1}{\pi}\Im\big(G^R_{i,i}(E)\big)} \tag{3.56}$$

D corresponds to the total number of states in a specific site. It is therefore called the *local density-of-states* (LDOS). The total density-of-states is then simply the trace of the spectral function.

3.4 Non-equilibrium Green's functions

When a current actually starts flowing through the device, the system is brought away from its thermodynamical equilibrium. The Hamiltonian of such a perturbed system can be split into two parts

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{W}}. \tag{3.57}$$

Where $\hat{\mathcal{H}}_0$ is the one-body non-interacting Hamiltonian whose solutions are known and $\hat{\mathcal{W}}$ is the interacting (or perturbed) Hamiltonian, which accounts for additional terms that drive the system out of equilibrium. The corresponding interacting stationary state reads

$$|\psi_{int}\rangle = \hat{U}_I(0, \mp\infty)\,|\psi_0(\mp\infty)\rangle\,, \tag{3.58}$$

where the system is prepared in the ground state $|\psi_0\rangle$ at time $t = -\infty$ and is brought to the present ($t = 0$), while the perturbation $\hat{\mathcal{W}}$ is **adiabatically** switched on. In Sec.3.1.2, we have introduced the Schrödinger and the Heisenberg representations. They can be combined to form the *interaction picture*, where both the state kets and the operators depend on time. Here above, the evolution operator has been written in the interaction picture (hence the subscript "I") since it is function of a time-dependent Hamiltonian and takes the form shown in Eq.(3.16).
Under non-equilibrium conditions, the (anti)chronological Green's function from Eq.(3.40) becomes

$$\begin{aligned} G^{\pm}_{i,j}(t',t) &= \frac{-\frac{i}{\hbar}\,\langle\psi_0(\infty)|\;\hat{U}_I(\infty,0)\;\mathcal{T}^{\pm}[\hat{c}_i(t')\hat{c}^{\dagger}_j(t)]\;\hat{U}_I(0,-\infty)\;|\psi_0(-\infty)\rangle}{\langle\psi_0(\infty)|\hat{U}_I(\infty,0)\hat{U}_I(0,-\infty)|\psi_0(-\infty)\rangle} \\ &= \frac{-\frac{i}{\hbar}\,\langle\psi_0(\infty)|\;\mathcal{T}^{\pm}[\;\hat{U}_I(\infty,-\infty)\hat{c}_i(t')\hat{c}^{\dagger}_j(t)]\;|\psi_0(-\infty)\rangle}{\langle\psi_0(\infty)|\hat{U}_I(\infty,-\infty)|\psi_0(-\infty)\rangle}. \end{aligned} \tag{3.59}$$

Scanning the equation from right to left: The system is brought adiabatically from the remote past to the present, where a particle is created and annihilated (electron or hole, depending on the sign of $t' - t$) and the system finally returns to equilibrium

in the distant future. Out of equilibrium, however, the remote past and distant future ground states differ by a phase factor

$$|\psi_0(\infty)\rangle = e^{iL}\,|\psi_0(-\infty)\rangle\,. \tag{3.60}$$

To avoid using $|\psi_0(\infty)\rangle$ int the expression of G, we can take advantage of the fact that $|\psi_0(\infty)\rangle = \hat{U}(\infty,-\infty)\ \ |\psi_0(-\infty)\rangle$, to obtain

$$G^{\pm}_{i,j}(t',t) = -\frac{i}{\hbar}\,\langle\psi_0(-\infty)|\ \hat{U}_I(-\infty,\infty)\mathcal{T}^{\pm}[\ \hat{U}_I(\infty,-\infty)\hat{c}_i(t')\hat{c}^{\dagger}_j(t)]\ |\psi_0(-\infty)\rangle\,, \tag{3.61}$$

where the denominator vanishes since $\langle\psi_0(\infty)|\hat{U}_I(-\infty,-\infty)|\psi_0(-\infty)\rangle = I$. Yet, the numerator's evolution operators cannot be combined because one belongs to the time-ordering $\mathcal{T}^{\pm}$ and the other does not.
In order to solve the NEGF, we will show that one has to expand it perturbatively. However, since $|\psi_0(\infty)\rangle$ is not connected to $|\psi_0(-\infty)\rangle$ in an obvious way, the solution cannot be computed easily. Indeed, the evolution operators shown in the above expression contain series expansions (see Eq.(3.15)), which are too intricate to be solved this way.

Note: In this section, for the sake of simplicity, we have resorted to the fundamental state $|\psi_0\rangle$ that describes the system at $T = 0$. However, in the general case, we can use the density matrix, as discussed in Sec.3.2.2.

3.4.1 Keldysh contour

The *Keldysh formalism* [1–3] proposes to discard any explicit reference the distant future $t = \infty$ by using a specific integration path, composed of two time branches (Fig.3.1). The upper branch $C^{\uparrow}$ extends from $t = -\infty$ to $t = \infty$, whereas the lower branch $C^{\downarrow}$ does the opposite. If the upper/lower branch times are denoted "$t_{\uparrow/\downarrow}$", one gets a total contour $C = C^{\uparrow} \cup C^{\downarrow}$ which goes from $t = -\infty_{\uparrow}$ to $t = -\infty_{\downarrow}$. In that case, the contour-ordered Green's function reads

$$G_{i,j}(t',t) = -\frac{i}{\hbar}\,\langle\ \mathcal{T}^{C}[\hat{c}_i(t')\hat{c}^{\dagger}_j(t)\hat{U}_I(-\infty_{\downarrow},-\infty_{\uparrow})]\ \rangle\,, \tag{3.62}$$

where we have dropped the equilibrium time-ordering operator and introduced the contour-ordering operator $\mathcal{T}^{C}$, which orders the elements according to their position on the Keldysh contour. The time-evolution operator takes the form shown in Eq.(3.16), integrated along C

$$\hat{U}(-\infty_{\downarrow},-\infty_{\uparrow}) = \mathcal{T}^{C}\exp\Big(-\frac{i}{\hbar}\oint_{C} d\tau\ \hat{\mathcal{W}}(\tau)\Big). \tag{3.63}$$

Fig.3.1 shows the time contour C and represents graphically where are located the time boundaries t and t' of the various Green's functions. In order to help the interpretation of their physical meaning, one can also make the connection with the explanations given

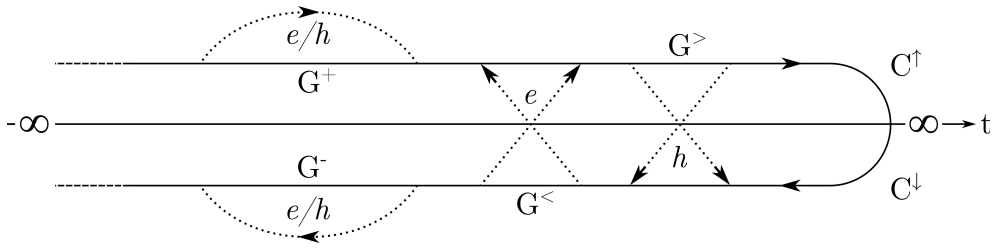


*Figure 3.1: Plain line: The Keldysh time-contour and its two branches $C^{\uparrow}$ and $C^{\downarrow}$. Dotted lines: graphical representations of the Green's functions on the time contour. The arrows always go from t to t′. The chronological and antichronological functions G^{+} and G^{-} correspond to the probability amplitude for a particle (electron **or** hole) to travel respectively forward in time (upper branch) or backward in time (lower branch). The lesser and greater functions $G^{<}$ and $G^{>}$ are each associated to a specific type of particle, but can either correspond to a forward **or** a backward time propagation (represented by the crossing arrows). Figure partially inspired from [6].*

after Eq.(3.43).
Mathematically, they can be expressed as follows:

$$\boxed{\begin{aligned} G^{+}_{i,j}(t',t) &= -\frac{i}{\hbar}\,\langle\, \mathcal{T}^{C}[\hat{c}_i(t'_{\uparrow})\hat{c}^{\dagger}_j(t_{\uparrow})]\,\rangle \\ G^{-}_{i,j}(t',t) &= -\frac{i}{\hbar}\,\langle\, \mathcal{T}^{C}[\hat{c}_i(t'_{\downarrow})\hat{c}^{\dagger}_j(t_{\downarrow})]\,\rangle \\ G^{>}_{i,j}(t',t) &= -\frac{i}{\hbar}\,\langle \hat{c}_i(t'_{\downarrow})\hat{c}^{\dagger}_j(t_{\uparrow})\rangle \\ G^{<}_{i,j}(t',t) &= \;\;\frac{i}{\hbar}\,\langle \hat{c}^{\dagger}_i(t'_{\uparrow})\hat{c}_j(t_{\downarrow})\,\rangle \end{aligned}} \tag{3.64}$$

The equilibrium and non-equilibrium Green's function formalisms are, otherwise, structurally equivalent (as long as the adiabatic assumption holds). The retarded and advanced non-equilibrium Green's functions can be obtained and the definitions introduced in the equilibrium case in Sec.3.3 (charge, density of states, ...) still hold at non-equilibrium. In other words, the NEGF formalism can be used to treat quantum transport as a stationary problem, while a current is nonetheless flowing through the device. The aim of the next sections is to show how this current is actually computed.

3.4.2 Dyson equation and self-energy

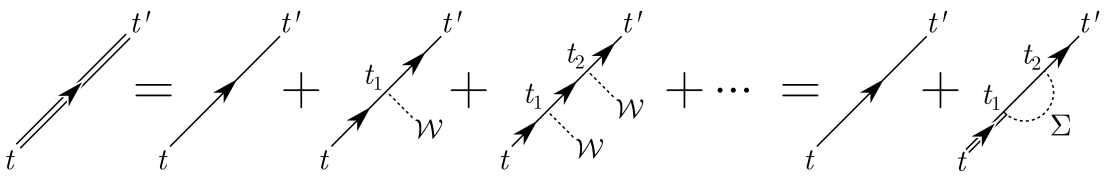


Figure 3.2: Diagram representation of the interacting Green's function (left hand side), perturbatively expanded into a unperturbed term (simple line) and an infinite series of n-th order scattering processes. The terms of this series can be brought together in a single quantity: the self-energy Σ (right hand side).

In nanodevices, the electrons are subjected to various interactions as they travel from one contact to the other. In order to include these interactions one can exploit the recursive structure of the Green's functions. If we consider the Green's function for an unperturbed system $G^{(0)}$ and a perturbation $\hat{\mathcal{W}}$, the total Green's function can be expanded as [5, 7]

$$\begin{aligned} G_{i,j}(t',t) = G^{(0)}_{i,j}(t',t) &+ \int_t^{t'} \mathrm{d}t_1 \; G^{(0)}_{i,m}(t',t_1) \; \hat{\mathcal{W}}_{m,n} \; G^{(0)}_{n,j}(t_1,t) \\ &+ \int_t^{t'} \mathrm{d}t_2 \int_t^{t_2} \mathrm{d}t_1 \; G^{(0)}_{i,\alpha}(t',t_2) \; \hat{\mathcal{W}}_{\alpha,\beta} \; G^{(0)}_{\beta,m}(t_2,t_1) \; \hat{\mathcal{W}}_{m,n} \; G^{(0)}_{n,j}(t_1,t) \\ &+ ... \end{aligned} \tag{3.65}$$

This series expansion can be interpreted as a succession of scattering events, represented in the form of diagrams in Fig.3.2. The most basic event corresponds to the free propagation from time t to time t' with no scattering. It can be represented as a single line. One then adds a first order scattering event, occurring between times t and t'. We then keep adding intermediate interactions as we iterate through this process. Ultimately, the overall scattering perturbation can be merged in a single term, called the *self-energy* Σ. The Green's function now reads

$$G_{i,j}(t',t) = G^{(0)}_{i,j}(t',t) + \int_t^{t'} \mathrm{d}t_2 \int_t^{t_2} \mathrm{d}t_1 \; G^{(0)}_{i,m}(t',t_2) \; \Sigma_{m,n}(t_2,t_1) \; G_{n,j}(t_1,t). \tag{3.66}$$

In other words, the self-energy allows us to obtain the perturbed (out-of-equilibrium) Green's functions in terms of the unperturbed (equilibrium) ones. In a more compact form, for the retarded quantity, we have:

$$\boxed{G^R = G^R_{(0)} + G^R_{(0)} \; \Sigma^R \; G^R} \tag{3.67}$$

This equation is called the Dyson equation and Σ^R is the retarded self-energy. Deriving the expression for the greater and lesser Green's functions is less straightforward and requires to perform the integration on the Keldysh contour described in Sec.3.4.1 and to use the Langreth' rules [1], which state that a function D defined by

$$D(t'_\uparrow, t_\downarrow) = \int_C \mathrm{d}t_1 \; A(t'_\uparrow, t_1) B(t_1, t_\downarrow), \tag{3.68}$$

corresponds to a lesser quantity of

$$D^<(t',t) = \int \mathrm{d}t_1 \; \big(A^R(t',t_1) B^<(t_1,t) + A^<(t',t_1) B^A(t_1,t) \big). \tag{3.69}$$

As a result, it can be proven that [2]

$$\boxed{G^< = G^R \; \Sigma^< \; G^A} \tag{3.70}$$

Resorting to a self-energy term is a very versatile approach that allows one to include different types of interactions. This high adaptability is one of the main strengths of the NEGF method. In practice, Σ can actually represent all the interactions of an electron with the other particles in the system. However, even though the general form of the Dyson equation is rather compact and straightforward, the main challenge comes from the derivation and the computation of the various self-energies. We note that many of the expressions relating the Green's functions to one another also hold true for the self-energies. For example

$$\Sigma^R - \Sigma^A = \Sigma^> - \Sigma^<, \tag{3.71}$$

is similar to Eq.(3.43) and we can also write

$$\Sigma^R = (\Sigma^A)^\dagger. \tag{3.72}$$

3.4.3 Electron-electron interaction: Poisson's equation

Even though we treat one-electron problems, the electron-electron interaction can be partially accounted for, by considering that each electron moves in a potential that comes from its average interaction with all of the other carriers [1]. In second quantization, a many-body system can be described by the following Hamiltonian

$$\hat{\mathcal{H}} = \underbrace{\int \mathrm{d}\mathbf{r}\ \hat{c}^\dagger(\mathbf{r})\big[-\frac{1}{2}\nabla^2 + V(\mathbf{r})\big]\hat{c}(\mathbf{r})}_{\text{one-body}} + \underbrace{\frac{1}{2}\int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}'\ \hat{c}^\dagger(\mathbf{r})\hat{c}^\dagger(\mathbf{r}')\ v(\mathbf{r}-\mathbf{r}')\ \hat{c}(\mathbf{r}')\hat{c}(\mathbf{r})}_{\text{two-body}}. \tag{3.73}$$

The one-body part (left hand side) is the usual one-body Hamiltonian and the two-body part (right hand side) contains a Coulomb interaction term v.

$$v(\mathbf{r}-\mathbf{r}') = \frac{e^2}{4\pi\epsilon\ |\mathbf{r}-\mathbf{r}'|}, \tag{3.74}$$

where ϵ is the dielectric constant of the medium. Similarly to the perturbative expansion detailed in Sec.3.4.2, the contribution of the two-body part of the Hamiltonian can be approximated (at the lowest order) by the Hartree self-energy, as [8]

$$\Sigma_{\text{Hartree}}(\mathbf{r},t) = -i\hbar \int \mathrm{d}\mathbf{r}'\ v(\mathbf{r}-\mathbf{r}')\ G^<(\mathbf{r}',t;\mathbf{r}',t). \tag{3.75}$$

Since $n(\mathbf{r},t) = -i\hbar\ G^<(\mathbf{r},t;\mathbf{r},t)$ (see Eq.(3.54)), we obtain

$$\begin{aligned}\Sigma_{\text{Hartree}}(\mathbf{r},t) &= \int \mathrm{d}\mathbf{r}'\ v(\mathbf{r}-\mathbf{r}')\ n(\mathbf{r}',t)\\ &= \int \mathrm{d}\mathbf{r}' \frac{\rho(r',t)}{4\pi\epsilon\ |\mathbf{r}-\mathbf{r}'|}\\ &= -e\phi(\mathbf{r}),\end{aligned} \tag{3.76}$$

where ρ is the charge density and ϕ is the electric potential. Finally, we obtain the Poisson's equation

$$\boxed{\nabla \cdot \big(\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})\big) = -\rho(\mathbf{r})} \tag{3.77}$$

In our simulations, the charge density due to the free carrier concentration can be obtained by solving the retarded Green's function (Eq.(3.54)). The charge will also be affected by the introduction of dopants, or by the electrostatic potential at the electrodes (due to the application of voltages). For that reason, the Poisson's equation and the NEGF are computed iteratively in a so called *Poisson-Schrödinger* cycle. Once the Poisson's equation has been solved by accounting for the charge density obtained by the NEGF, ϕ is used as an input for the potential energy of the one-body Hamiltonian, to recompute the NEGF. The process is repeated until a desired level of convergence is reached.

3.4.4 Electron-phonon interaction: SCBA

Scattering mechanisms can be relevant when considering transport in nanodevices. Indeed, vibrations in the lattice (*i.e.* phonons) can cause changes in the crystal potential, which, in turn, can affect the flow of the electrons. Phonons are implemented within the deformation potential method and the electron-phonon interaction is approximated as being local. The lesser/greater phonon self-energy reads

$$\Sigma^{\lessgtr}_{\text{ph}} = D_0^{\lessgtr} G^{\lessgtr}, \tag{3.78}$$

where $D_0^{\lessgtr}$ is the Green's function of the unperturbed phonon bath. In the expression adopted in this work, acoustic phonons are treated within the elastic approximation, whereas polar optical phonons are assumed to be dispersionless.
For acoustic phonons, the corresponding lesser/greater quantity at the i-th slice and for the n-th mode reads [1, 9]

$$\Sigma^{\lessgtr}_{\text{ac},i}(n,n,E) = \frac{D_{\text{ac}}^2 k_B T}{\rho \nu_S^2} \sum_m \mathcal{I}_i(m,n)\; G_i^{\lessgtr}(m,m,E)\,, \tag{3.79}$$

where D_{ac} stands for the acoustic deformation potential, ρ is the material density, ν_S is the sound velocity and $\mathcal{I}$ is the form factor, that is used to transform the expression of the self-energy from the real-space to the mode-space (see Sec.2.7)

$$\mathcal{I}_i(m,n) = \int dz\; |\chi_i(m,z)|^2\; |\chi_i(n,z)|^2. \tag{3.80}$$

Likewise, the lesser self-energy term for the optical phonons is written

$$\Sigma^{\lessgtr}_{\text{opt},i}(n,n,E) = \frac{\hbar D_{\text{opt}}^2}{2\rho\omega_{\text{opt}}} \sum_m \mathcal{I}_i(m,n) G_i^{\lessgtr}(m,m,E \pm \hbar\omega_{\text{opt}}) \left[N_{\text{BE}}(\hbar\omega_{\text{opt}}) + \frac{1}{2} \pm \frac{1}{2} \right], \tag{3.81}$$

where D_{opt} is the optical deformation potential, ω_{opt} is the optical phonon frequency and N_{BE} is the Bose-Einstein distribution function, corresponding to the average phonon density at the energy $\hbar\omega_{\text{opt}}$

$$N_{\text{BE}} = \frac{1}{e^{(\hbar\omega_{\text{opt}}/k_B T)} - 1}. \tag{3.82}$$

Once these self-energies have been computed, the total lesser/greater self-energy for phonons is simply given by

$$\Sigma^{\lessgtr}_{\text{ph}} = \Sigma^{\lessgtr}_{\text{ac}} + \Sigma^{\lessgtr}_{\text{opt}}. \tag{3.83}$$

Finally, from Eq.(3.71), the retarded self-energy for phonons is approximated by

$$\Sigma^{R}_{\text{ph},i} = \frac{1}{2}\left(\Sigma^{>}_{\text{ph},i} - \Sigma^{<}_{\text{ph},i}\right). \tag{3.84}$$

As discussed in Sec.3.4.2, computing the self-energies is necessary to implement interaction phenomena in the NEGF. However, as shown above, the Green's functions are also required to solve the phonons' self-energies. Consequently, in a similar way than the Poisson-Schrödinger cycle detailed in Sec.3.4.3, we perform a *self-consistent Born approximation* (SCBA) cycle, that consists in solving Σ_{ph} and $G^{\lessgtr}$ iteratively, until a desired degree of convergence is achieved. More details on the connection between the SCBA and the Poisson-Schrödinger cycles are given in Sec.3.5.

3.4.5 Electrical current

In Sec.3.3 and 3.4.1, we have shown that the NEGF allowed one to compute the carrier concentration and the density of states in a system brought out of equilibrium (that can typically correspond to a transistor, connected to its environment *via* two leads). The last quantity essential for electron transport is evidently the electrical current. Let us consider a Hamiltonian of the form

$$\hat{\mathcal{H}} = \sum_i E_i \ \hat{c}_i^\dagger \hat{c}_i + \sum_{i \neq j} t_{i,j} \ \hat{c}_i^\dagger \hat{c}_j, \tag{3.85}$$

where E is the on-site energy and t is a hopping term. The charge on a given site i is

$$\hat{q}_i = -e \ \hat{c}_i^\dagger \hat{c}_i. \tag{3.86}$$

The total current that leaves site i towards the coupled sites can be obtained with the continuity equation [2]

$$\hat{J}_i = -\frac{\partial q_i}{\partial t} = -\frac{1}{i\hbar}\left[q_i, \hat{\mathcal{H}}\right]_-. \tag{3.87}$$

Hence, we obtain

$$\hat{J}_i = \frac{-ie}{\hbar} \sum_j (t_{i,j}\hat{c}_i^\dagger \hat{c}_j - t_{j,i}\hat{c}_j^\dagger \hat{c}_i). \tag{3.88}$$

The current flowing from site i to a specific neighbor j is thus restricted to

$$\hat{J}_{i\to j} = \frac{-ie}{\hbar}(t_{i,j}\hat{c}_i^\dagger \hat{c}_j - t_{j,i}\hat{c}_j^\dagger \hat{c}_i). \tag{3.89}$$

From Eq.(3.64), this current also reads

$$J_{i\to j} = -e\,\big(t_{i,j}G^<_{i,j}(t=0) - t_{j,i}G^<_{j,i}(t=0)\big). \tag{3.90}$$

In this work, we are interested in the solutions in the energy domain. The conversion can be done accordingly to the procedure presented in Sec.3.3.1

$$\begin{aligned} J_{i\to j} &= -\frac{e}{h}\int \mathrm{d}E\,\big(t_{i,j}G^<_{i,j}(E) - t_{j,i}G^<_{j,i}(E)\big) \\ &= -\frac{e}{h}\int \mathrm{d}E\; 2\Re\big[t_{j,i}G^<_{i,j}(E)\big]. \end{aligned} \tag{3.91}$$

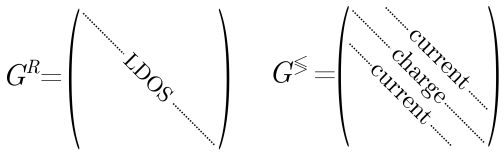


Figure 3.3: Simplified view of the NEGF that highlights the relevant parts of the NEGF matrices that are used to compute the LDOS, the current and the carrier density (and consequently, the charge). The LDOS and the current are obtained with $G^<$ via Eqs.(3.55) and (3.91), whereas the LDOS comes from G^R, as shown in Eq.(3.56).

To conclude, we have shown in Eq.(3.54) that the electron and hole densities could be obtained with the diagonal elements of the $G^<$ matrix. Moreover, we have also explained of the local density of states could be computed with the diagonal part of G^R (Eq.(3.56)). In the present section, we consider a third physical quantity that is connected to the off-diagonal terms of $G^<$, namely the electrical current. As illustrated in Fig.3.3, we will thus only be interested in a limited portion of the G^R and $G^<$ matrices in the simulations.

3.5 Computational implementation

Even though finding the eigenvalues of the Hamiltonian and solving Poisson's equation is within the reach of a standard desktop computer, the computation of the SCBA and the NEGF are very resource demanding. In order to optimize the performances of the code, the algorithm is implemented in *Fortran*. This compiled language is well suited for high performance scientific computing and gives access to optimized linear algebra libraries that are particularly convenient for the purpose of this work. As detailed in Sec.3.5.2, we also resort to parallel programming to speed up the computation of the SCBA cycles. Fig.3.4 presents a general view of the simulation code, in the form of a block diagram. The *Schrödinger* block corresponds to the ideas developed in Chap.2, whereas the *SCBA* and *Poisson* blocks contain parts of the formalism derived in the

present chapter. The inputs required in the code are a finite elements mesh file that details the structure (geometry, materials, regions), a list of the materials' properties and a file that contains the physical (voltages, phonons, doping) and computational (cycles precision and cutoff, parallelization) parameters. The outputs contain a wide range of physical quantities related to transport, among which is the current, the carrier distribution, the electrical potential, the density of states, the bands, or the wave functions.

3.5.1 Recursive scheme

Similarly to the method employed to discretize the Hamiltonian (Sec.2.6), we adopt a finite difference grid to compute the NEGF. We recall from Eq.(3.49) that the complete retarded Green's function reads [4]

$$G^R = (EI - \hat{\mathcal{H}}_{k.p} - \Sigma)^{-1}. \tag{3.92}$$

As we notice, this matrix has the same dimension than the total eight-band $\mathbf{k} \cdot \mathbf{p}$ Hamiltonian, which represents as much as $(N_b \times N_x \times N_y \times N_z)^2$ elements in the case of a 3D system. Directly solving this equation represents a very heavy task and would have to be done multiple times on a large energy grid. However, only a limited portion of the Green's function is actually required to compute the transport (Fig.3.3). This is one of the reasons why, in this manuscript, we have focused on expressing the NEGF as matrix elements, rather than using the full matrix notation. If the total Hamiltonian is written as a tri-diagonal block matrix, the NEGF can be computed step by step by slicing the problem along the transport direction and by considering only first-neighbor slice-slice interactions, as shown in Fig.3.5. In this chapter, the notation "$G_{i,i}$" refers to the Green's function of the i-th slice, and no explicit reference to the direction transverse to transport is made.
Additionally, Caroli *et al.* have proposed to split the system into three distinct parts [10]. The *left* and *right* regions surround a *central* region, where all the interactions occur. Conversely, the right and left contact regions are viewed as inert and do not host any interaction. In the case of nanostructures, this assumption is justified since most of the interaction processes happen in the structure itself (due to its high susceptibility to its environment) rather than in the contacts. Once an electron exits the system, we assume that it does not affect the state of the leads, whose Fermi energies remain perfectly stable. As explained in the introduction of this chapter, we simply consider that the device is connected to semi-infinite electron reservoirs whose behavior is perfectly under control. In the case of a transistor, these reservoirs' Fermi levels are directly related to an applied voltage, that can be freely chosen. The position of the right, central and left regions also results from an arbitrary choice. In practice, this means that the self-energy can be split into three components:

$$\Sigma = \underbrace{\Sigma^{\leftarrow}}_{\text{Left}} + \underbrace{\Sigma_{\text{ph}}}_{\text{Center}} + \underbrace{\Sigma^{\rightarrow}}_{\text{Right}}, \tag{3.93}$$

where the $\Sigma^{\leftarrow}$ and $\Sigma^{\rightarrow}$ the left- and right-connected self-energies are related to the contacts and the phonon self-energy comes from the central region (since we consider

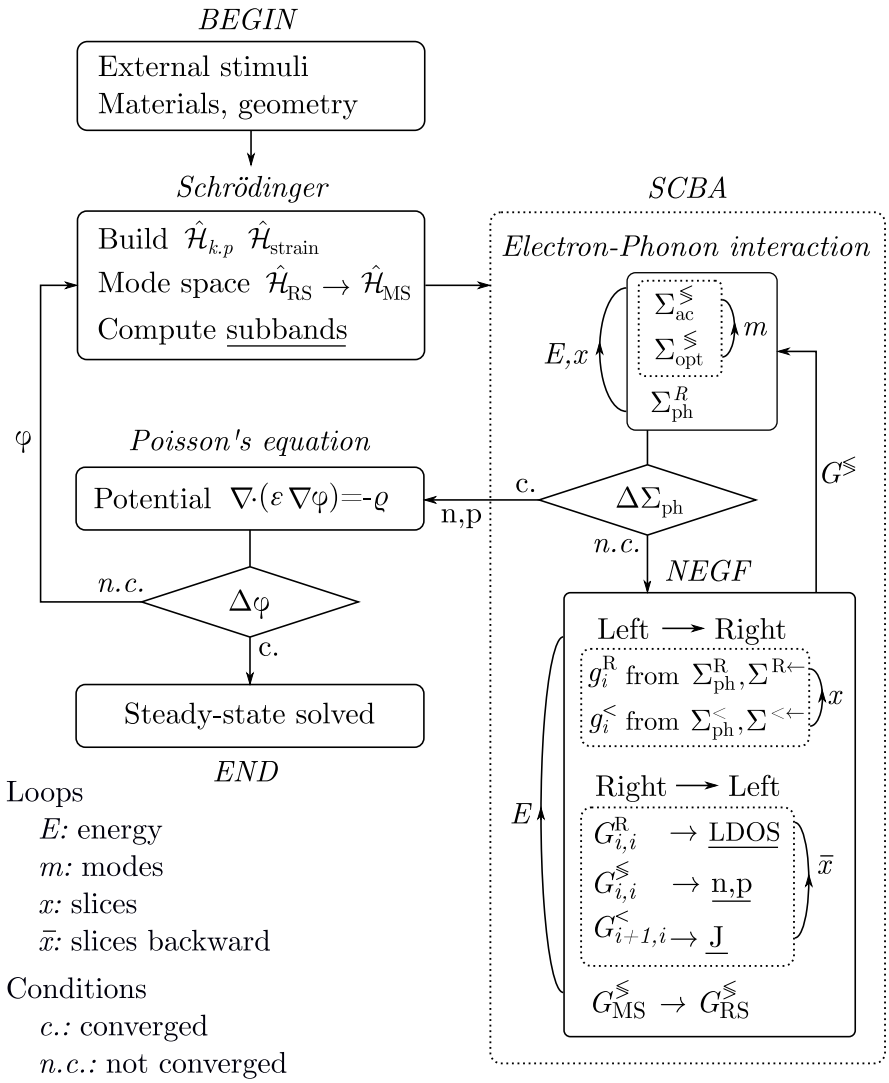


Figure 3.4: Structure of the code. The algorithm is composed of two main loops: the outer Poisson-Schrödinger loop and the inner SCBA loop, which encompasses the computation of the NEGF (using the recursive scheme detailed in Sec.3.5.1) and the phonons. The underlined terms are the main outputs of the code, since they correspond to tangible physical quantities and contain many information about the transport. More pragmatically, in the case of a transistor, this piece of program allows us to obtain the (steady-state) current for specific gate and supply voltages (that are included as input parameters via the "external stimuli" initialisation block). Note that once the convergence is reached by the Poisson's equation, a last SCBA cycle is actually performed before extracting the final physical quantities.

that no scattering occurs outside of the device). This implies that Eqs.(3.92) and (3.70) now read

$$\begin{aligned} G^R &= g^R + g^R(\Sigma^R_{\leftarrow} + \Sigma^R_{\text{ph}} + \Sigma^R_{\rightarrow})G^R, \\ G^< &= G^R(\Sigma^<_{\leftarrow} + \Sigma^<_{\text{ph}} + \Sigma^<_{\rightarrow})G^A, \end{aligned} \tag{3.94}$$

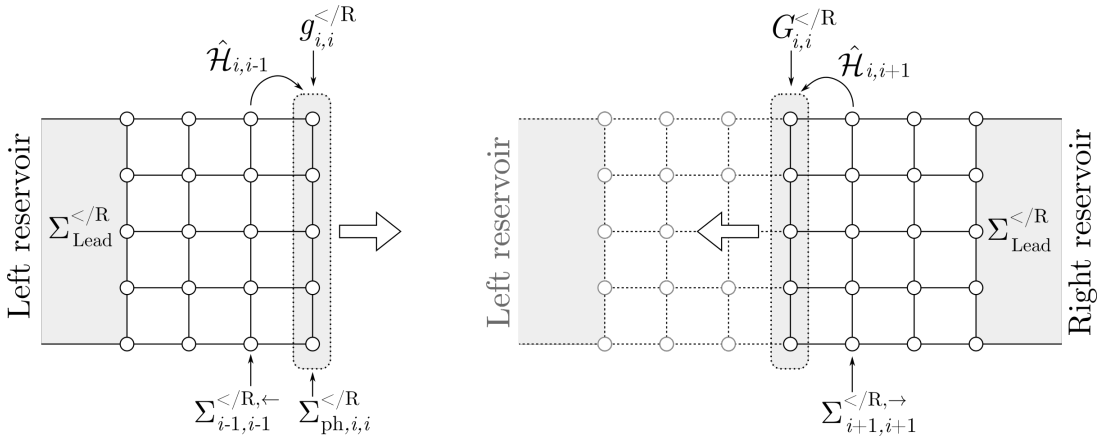


Figure 3.5: Recursive approach corresponding to the NEGF block of Fig.3.4. In the first stage (left), we start from the left contact and compute the left-connected Green's functions $g^<$ and g^R. The self-energies of the previous slice $\Sigma^{<,\leftarrow}$ and $\Sigma^{R,\leftarrow}$ are included through Eq.(3.95) and the phonon self-energies $\Sigma^<_{\text{ph}}$ and Σ^R_{ph} are also added. The resulting Green's functions are incomplete, since the right self-energies are missing in the computation. Thus, in the second stage (right), we repeat the process from right to left and the complete Green's functions $G^<$ and G^R are finally obtained. At the contacts, since no previous step exists to determine the right/left-connected self-energy, $\Sigma^{</R}_{Lead}$ is used.

where the lower case Green's functions correspond to a right- or left-connected system. Using the notation of Eq.(3.85), the left-connected components can be written

$$\begin{aligned} \Sigma^{R,\leftarrow}_{i,j} &= t_{i,m}\ g^R_{m,n}\ t_{n,j}, \\ \Sigma^{<,\leftarrow}_{i,j} &= t_{i,m}\ g^<_{m,n}\ t_{n,j}. \end{aligned} \tag{3.95}$$

where the terms t are the Hamiltonian's hopping terms.
The recursive method [11] consists in solving the NEGF on each slice iteratively by considering its interactions with its previous neighbor. Each computation step can be seen as a small lead-device-lead system (where each slice plays successively the role of the device). A slice is affected by the self-energy $\Sigma^{\rightarrow}$ of its right neighbor, by the self-energy $\Sigma^{\leftarrow}$ of its left neighbor and by a phonon self-energy Σ_{ph}. If we start to solve the problem from the left contact, the right side terms will be unknown, since $\Sigma^{\rightarrow}$ is related to $\hat{\mathcal{H}}_{i+1,i}$, which has not been computed yet. For that reason, the recursive approach comprises two stages:

- First, the problem is solved from left to right and allows one to obtain the left-connected lesser and retarded Green's functions, $g^<$ and g^R.
- Second, once we reach the right contact, the problem is solved backwards, from right to left. This time, both the previous (right) and the next (left) neighbors are known and we can derive the complete Green's functions G^R and $G^<$.

At the beginning of each stage, it is necessary to compute the initial self-energy, coming from the contact. This is done with the Sancho-Rubio iterative procedure, that is designed to determine the surface Green's function at an interface with a semi-infinite

homogeneous reservoir (more details can be found in [12]). Once we know the lead's self-energy Σ_{lead}, the first left-connected terms can be obtained from Eq.(3.94):

$$g^R_{1,1} = (EI_{1,1} - \hat{\mathcal{H}}_{1,1} - \Sigma^R_{\text{ph},1,1} - \Sigma^R_{\text{lead}})^{-1}, \tag{3.96}$$

$$g^<_{1,1} = g^R_{1,1}\ (\Sigma^<_{\text{ph},1,1} + \Sigma^<_{\text{lead}})\ g^A_{1,1}. \tag{3.97}$$

where the phonon self-energies Σ_{ph} have been calculated with the procedure detailed in Sec.3.4.4. We can then solve the diagonal elements iteratively, by using the previous slice's results to determine the left-connected self-energy *via* Eq.(3.95)

$$g^R_{i,i} = (EI_{i,i} - \hat{\mathcal{H}}_{i,i} - \Sigma^R_{\text{ph},i,i} - \underbrace{\hat{\mathcal{H}}_{i,i-1}\ g^R_{i-1,i-1}\ \hat{\mathcal{H}}_{i-1,i}}_{\Sigma^{R,\leftarrow}_{i-1,i-1}})^{-1}, \tag{3.98}$$

$$g^<_{i,i} = g^R_{i,i}\ (\Sigma^<_{\text{ph},i,i} + \underbrace{\hat{\mathcal{H}}_{i,i-1}\ g^<_{i-1,i-1}\ \hat{\mathcal{H}}_{i-1,i}}_{\Sigma^{<,\leftarrow}_{i-1,i-1}})\ g^A_{i,i}. \tag{3.99}$$

Once the right contact has been reached, the same procedure can be repeated backwards. We obtain the complete retarded Green's function

$$G^R_{i,i} = G^R_{i,i} + g^R_{i,i}\ \underbrace{(\hat{\mathcal{H}}_{i,i+1}\ G^R_{i+1,i+1}\ \hat{\mathcal{H}}_{i+1,i})}_{\Sigma^{R,\rightarrow}_{i+1,i+1}}\ g^R_{i,i}. \tag{3.100}$$

The resulting trace $G^R(i,E)$ allows us to compute the **local density of states (LDOS)** along the transport direction, as a function of the energy, *via* Eq.(3.56).
In order to solve $G^<$, we introduce an additional expression

$$G^< = g^< + g^R\Sigma^< g^A + g^R\Sigma^R g^< + g^<\Sigma^A g^A. \tag{3.101}$$

Thus, the diagonal part of the lesser Green's function can be computed as

$$\begin{aligned} G^<_{i,i} = g^<_i &+ g^R_{i,i}\ (\hat{\mathcal{H}}_{i,i+1}\ G^<_{i+1,i+1}\ \hat{\mathcal{H}}_{i+1,i})\ g^A_{i,i} \\ &+ g^R_{i,i}\ (\hat{\mathcal{H}}_{i,i+1}\ G^R_{i+1,i+1}\ \hat{\mathcal{H}}_{i+1,i})\ g^<_i \\ &+ g^<_{i,i}\ (\hat{\mathcal{H}}_{i,i+1}\ G^A_{i+1,i+1}\ \hat{\mathcal{H}}_{i,i+1})\ g^A_{i,i}, \end{aligned} \tag{3.102}$$

The same procedure can be applied to $G^>$ and the **electron and hole densities** can be extracted with Eq.(3.54).
Finally, the off-diagonal values are obtained with

$$\begin{aligned} G^<_{i,i+1} &= G^<_{i,i}\ \hat{\mathcal{H}}_{i,i+1}\ g^A_{i+1,i+1} + G^R_{i,i}\ \hat{\mathcal{H}}_{i,i+1}\ g^<_{i+1,i+1}, \\ G^<_{i+1,i} &= G^<_{i+1,i+1}\ \hat{\mathcal{H}}_{i+1,i}\ g^A_{i,i} + G^R_{i+1,i+1}\ \hat{\mathcal{H}}_{i+1,i}\ g^<_{i,i}. \end{aligned} \tag{3.103}$$

These values are used to compute the **spectral current** $J(i,E)$ with the expression developed in Eq.(3.91)

$$J_{i\rightarrow\ i+1} = -\frac{e}{h}\int \mathrm{d}E\ \left(\hat{\mathcal{H}}_{i,i+1}G^{<}_{i,i+1}(E) - \hat{\mathcal{H}}_{i+1,i}G^{<}_{i+1,i}(E)\right). \tag{3.104}$$

Once these quantities have been calculated on all the energy grid, the SCBA approach consists in using $G^{\lessgtr}$ to solve the phonon self-energies $\Sigma^{<}_{\mathrm{ph}}$ and Σ^{R}_{ph}, with the approach discussed in Sec.3.4.4. These self-energies will be used to recompute $G^{\lessgtr}$ and G^{R}, and the process will be repeated until the convergence of the algorithm is reached.
As illustrated in Fig.3.4, when the convergence of the SCBA has been achieved, the resulting carrier density is transferred to the Poisson's equation, where the electric potential can be computed. Note that since we are working in the mode-space, it is necessary to transform $G^{\lessgtr}$ back into the real-space to obtain the spatial carriers distribution. The transformation is done with

$$G^{\lessgtr}_{\mathrm{RS},i,j} = \sum_{m,n}^{\mathrm{modes}} \chi_i^{m*}\ G^{\lessgtr}_{\mathrm{MS},m,n}\ \chi_j^{n}, \tag{3.105}$$

where we used the notation introduced in Sec.2.7. Finally, the electric potential is used as an input to recompute the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian and to start another Schrödinger-NEGF-Poisson cycle.
Ultimately, the non-equilibrium steady-state transport – with a given set of external conditions – is solved. In practice, this corresponds to the state of the nanoelectronic device when its electrodes are connected to a given voltage.

3.5.2 Convergence

The accuracy and the speed of the computation is related to many parameters. Both the SCBA and the Poisson-Schrödinger loops have their own tolerances and cut-off limits. The size of the device and the number of modes also play an important role in the simulation time.
To make the computation faster, we resort to parallelization of the code on several CPUs. The parallel resolution is performed on the energy grid during the SCBA loop (where the simulation spends a large amount of time). The total energy range is split into fragments of size ΔE (defined as an input) and each of these fragments is shared by N_{sub} subprocesses (that correspond to individual CPU cores). The energy resolution is thus given by $R = \Delta E/N_{\mathrm{sub}}$, which is close to 1 meV in our case. If N_{sub} is increased, the resolution is improved; and if ΔE is increased, the simulation gets faster. Thereby, if both quantities are increased so that R remains unchanged, the computation is sped up with no loss of precision. However, ΔE must not be chosen arbitrarily, since another constraint comes into play: we also have to consider an integer parameter N_{op} such that $\hbar\omega_{\mathrm{opt}} = N_{\mathrm{op}} \times \Delta E$, with ω_{opt} the optical phonons frequency (see Sec.3.4.4). In practice, we want $\hbar\omega_{\mathrm{opt}} = 30$ meV, which means that ΔE cannot exceed 30 meV either (for which we would set $N_{\mathrm{op}} = 1$).
Finally, on a larger scale, the whole computation can also be divided into multiple individual steps. Indeed, the operation of the device is split into separate stages, each corresponding to a steady-state solution under a given set of external stimuli

(source-drain voltage, gate potential, etc.). These steps being independent, they can be computed in parallel. However, for a given number of available cores, this kind of parallelization does not necessarily lead to a faster simulation. When several states of the device are computed in series, the solution for the potential ϕ_n is kept as an input "guess" for step $n+1$. There is a significant time difference between a simulation step performed with and without such an initial guess. When all these steps are solved in parallel, one does not benefit from this boosting effect anymore.
We shall also note that only a fraction of the code is parallelized. Thus, we cannot establish a linear relation between the number of subprocesses and the expected duration of the simulation. As a perspective, the convergence of the algorithm could be accelerated by performing the coupled mode-space, form factor or Poisson computation in parallel. We may also consider keeping the solution of the last Green's functions as a starting guess for the following iteration.

Bibliography

[1] M. Pourfath. *Non-Equilibrium Green's Function Method for Nanoscale Device Simulation*. Springer, 2014.

[2] A. Cresti. *Theoretical Imaging of Currents in Nanostructures*. PhD thesis, Univeristy of Pisa, 2005.

[3] R. van Leeuwen G. Stefanucci. *Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction*. Cambridge University Press, 2013.

[4] S. Datta. *Quantum Transport: Atom to Transistor*. Cambridge University Press, 2005.

[5] M. Di Ventra. *Electrical Transport in Nanoscale Systems*. Cambridge University Press, 2008.

[6] P. Kakashvili and C. J. Bolech. Time-loop formalism for irreversible quantum problems: Steady-state transport in junctions with asymmetric dynamics. *Physical Review B*, 78(3), jul 2008.

[7] J. D. Walecka A. L. Fetter. *Quantum Theory of Many-Particle Systems*. McGraw-Hill, 1980.

[8] W. Schäfer and M. Wegener. *Semiconductor Optics and Transport Phenomena*. Springer, 2002.

[9] K. Rogdakis, S. Poli, E. Bano, K. Zekentes, and M.G. Pala. Phonon- and surface-roughness-limited mobility of gate-all-around 3c-SiC and Si nanowire FETs. *Nanotechnology*, 20(29):295202, jul 2009.

[10] C. Caroli, R. Combescot, D. Lederer, P. Nozieres, and D. Saint-James. A direct calculation of the tunnelling current. II. free electron description. *Journal of Physics C: Solid State Physics*, 4(16):2598–2610, nov 1971.

[11] M.P. Anantram, M.S. Lundstrom, and D.E. Nikonov. Modeling of nanoscale devices. *Proceedings of the IEEE*, 96(9):1511–1550, sep 2008.

[12] M. P. Lopez Sancho, J. M. Lopez Sancho, J. M. L. Sancho, and J. Rubio. Highly convergent schemes for the calculation of bulk and surface green functions. *Journal of Physics F: Metal Physics*, 15(4):851–858, apr 1985.

CHAPTER 4

In(Ga)As planar MOSFET

In which we introduce the basics of FET design, apply these principles to the case of III-V semiconductor nanotransistors and understand the role of the various geometrical and physical parameters.

As discussed in Chap.1, the goal of this work is to identify possible alternatives to silicon-based field effect transistors (FETs). For this assessment to be quantitative, we have to define what criteria should be evaluated. In practice, the aim is to find a device that - when compared to silicon FETs - would:

- show an improved *subthreshold swing* (SS),
- and/or show a larger *on-current* (I_{on}).

These two properties are related to the fundamental goals of a transistor, formulated in the introduction (*i.e.* "switch between its *on* and *off* states as fast as possible" and "keep these states sufficiently distinguishable"). The third task ("generate as little energy dissipation as possible") will not be explicitly treated in this chapter. However, different operating regimes, which correspond to various fields of application (and therefore different power consumptions), will be described later.

4.1 Quantities and definitions

Despite not being identical, all the devices presented in this work share some common properties. Geometrically, they are essentially made of three distinct parts, whose name allude to the field of hydraulics: the *source*, the *channel* and the *drain*. The electrons enter the device from the source, travel through the channel, and exit from the drain. The channel is the active part of the transistor, since it is (partially or totally) covered by a *gate* electrode, which generates a potential barrier when a gate voltage V_{GS} is applied. This metallic electrode is separated from the channel by an oxide layer, hence the acronym "MOS" (Metal-Oxide-Semiconductor). The voltage V_{GS} modulates the shape of the barrier (usually its height - but also its width in the case of a Tunnel-FET), which can allow (or prevent) a current I_{DS} to flow from the source to the drain. This current also increases with the energy difference between source and drain electrons, which is connected to the source-drain voltage V_{DS}.
The performance of a transistor is evaluated for a given *off* current I_{off} and a fixed value of V_{DS} (corresponding to the supply voltage V_{DD}). The initial *off*-state gate voltage is denoted V_{off}. It is then increased until $V_{\mathrm{GS}} = V_{\mathrm{off}} + V_{\mathrm{DS}}$, where one can finally extract the *on*-current I_{on}

$$I_{\mathrm{on}} \equiv I_{\mathrm{DS}}(V_{\mathrm{off}} + V_{\mathrm{DS}}) \tag{4.1}$$

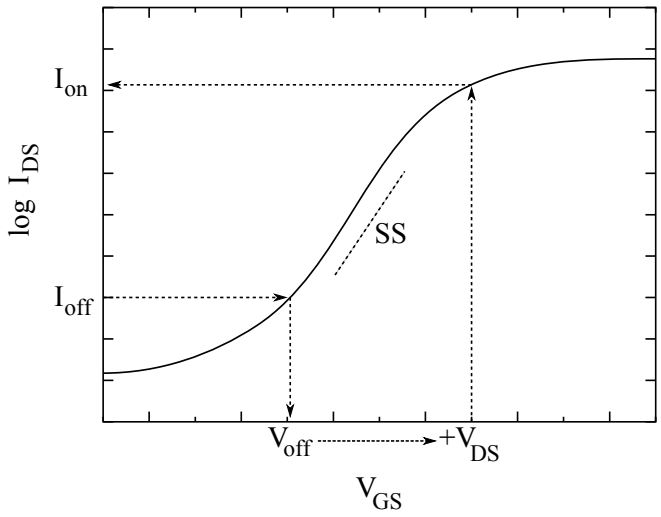


Figure 4.1: Typical transfer characteristic of a transistor, showing the evolution of the current I_{DS} as a function of the gate voltage V_{DS}. The arrows show how, from a given set of I_{off} and V_{DS}, one can obtain the corresponding on current I_{on} and subthreshold swing SS.

The subthreshold-swing (SS) is proportional to the increase rate of V_{GS} for a given change in I_{DS}. It is conventionally expressed in mV/dec and corresponds to the variation of V_{GS} required to increase the source-drain current by one order of magnitude. It is usually measured between V_{off} and a so called *threshold voltage* V_{th}, which denotes the minimum gate voltage needed to create a conducting channel between the source and the drain. It is evaluated with the formula

$$\mathrm{SS} = \frac{V_{\mathrm{th}} - V_{\mathrm{off}}}{\log_{10}\left(I_{\mathrm{th}}/I_{\mathrm{off}}\right)} \tag{4.2}$$

Where I_{th} is the *threshold current*, defined here as I_{th}=1 A/m. Moreover, in a MOSFET, the SS can also be modeled by the following expression [1]

$$\mathrm{SS} = \ln(10)\frac{k_{\mathrm{B}}T}{e}\left(1 + \frac{C_d}{C_{ox}}\right) \tag{4.3}$$

Where C_{ox} is the capacitance of the oxide and C_d is the capacitance of the depletion region; *i.e.* the domain from which the carriers have been forced away by the gate potential (which roughly corresponds to the channel). In an ideal MOSFET device, the ratio C_d/C_{ox} should tend towards zero. At room temperature (300K), the best possible subthreshold swing is thus close to 60 mV/dec. For this reason, we shall not expect our MOSFET device to go beyond this limit. We may nonetheless try to get as close to it as possible.
To summarize, depending on the range of applications targeted, one chooses an appropriate set of I_{off} and V_{DS} and evaluates the performance of the device by extracting the values of I_{on} and SS, as shown in Fig.4.1. The specifications to meet are listed in the *International Technology Road-map for Semiconductors* (ITRS) [2] and will serve us as a reference for the rest of this work.

4.2 Channel material

Before focusing on novel device architectures, we shall study a ultra-thin-body (UTB) MOSFET in which the silicon channel has been replaced by a III-V compound semiconductor. Indeed, by combining materials from groups III and V, it is possible to obtain various compounds that present better transport properties than silicon, due to their higher mobility and lower effective mass [3]. These semiconductors promote the establishment of a ballistic transport regime [4], which is beneficial for the device performance. On the other hand, III-V semiconductors are also expected to be more sensitive to *short channel effects* (SCE) [5]. This term refers to the detrimental phenomena that arise when the channel length is decreased. Indeed, handling the flow of electrons through a nanoscale channel requires a good electrostatic control, which is more easily achieved in silicon. The class of materials studied in this work suffers from noticeable electron tunneling and leakage. Given this bittersweet information, one cannot readily conclude whether III-V could outperform silicon technology. The aim of this chapter is to carry an introductory investigation on this issue, by focusing on the specific case of an InAs n-type MOSFET.

4.3 Description of the device

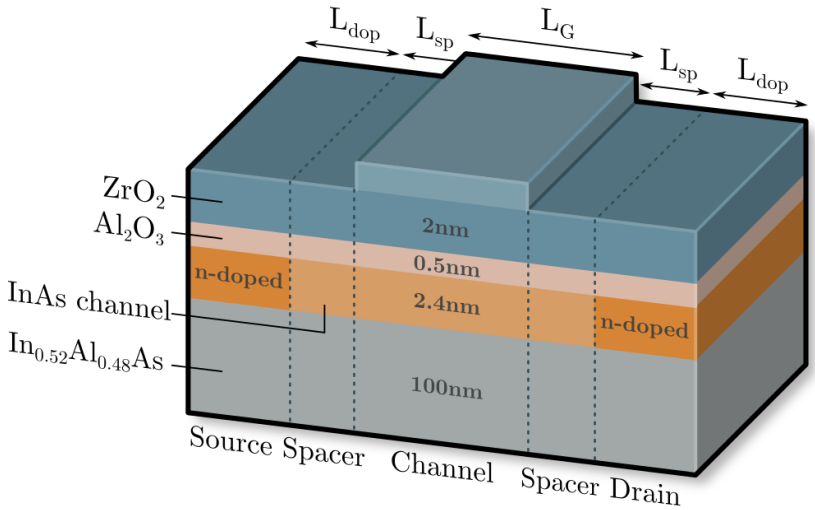


Figure 4.2: Scheme of the InAs UTB MOSFET. The thickness is indicated on each layer, while the lengths of the doped regions, spacers and gate (respectively L_{dop}, L_{sp} and L_G) are detailed in Tab.4.1. The block drawn on top of the oxide layer corresponds to the position of the gate contact. The doping of the InAs regions marked with "n-doped" is $3 \times 10^{19} cm^{-3}$, while the rest of the channel remains undoped.

4.3.1 Channel

We design an InAs MOSFET with a 2.4 nm thick channel, inspired from the experimental work of [6]. This InAs layer is grown on a $In_{0.52}Al_{0.48}As$ surface, where the

Thicknesses (nm)				Lengths (nm)		
t_{InAs}	t_{InAlAs}	$t_{Al_2O_3}$	t_{ZrO_2}	L_{G}	L_{dop}	L_{sp}
2.4	100	0.5	2	8, 15, 25, 40	20	17

Table 4.1: Geometrical parameters of the studied device. The corresponding lengths are shown on Fig.4.2

subscripts denote the respective molar fractions of each element. In practice, this aluminum/indium molar ratio allows the ternary compound to match the InP substrate lattice constant (not modeled here), thus avoiding dislocations. However, there is still a lattice mismatch between InAs and InAlAs ($a_{InAs} > a_{InAlAs}$). For that reason, the channel undergoes compressive strain, while the InAlAs is assumed unstrained in our simulations. As explained in Sec.2.8, the effect of strain in the transport is modeled by adding a strain interaction matrix to the 8×8 $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian. In order to compute the value of the axial strain $\epsilon_{\parallel}$, we compare the lattice parameters of the stacked materials. The transverse strain $\epsilon_{\perp}$ is then simply obtained with a macroscopic approach, using the material's Poisson's ratio ($\nu = \epsilon_{\perp}/\epsilon_{\parallel}$). Consequently, in the InAs channel, we obtain $\epsilon_{\parallel} = -0.0311$ and $\epsilon_{\perp} = 0.0338$. The total Hamiltonian then writes

$$\hat{\mathcal{H}}_{tot} = \hat{\mathcal{H}}_{8,k.p} + \hat{\mathcal{H}}_{8,strain}, \tag{4.4}$$

where the $\hat{\mathcal{H}}_{8,k.p}$ and $\hat{\mathcal{H}}_{8,strain}$ parameters are shown in Tab.4.2. Additional parameters, such as the number of conduction and valence modes and the phonon parameters are listed in Tab.4.3.

4.3.2 Oxide

To reproduce the experimental results as precisely as possible, the gate oxide is composed of 0.5 nm of Al_2O_3 on top of 2 nm of ZrO_2. To evaluate the properties of these gate layers, a comparison can be made with the conventional SiO_2 oxide by computing the equivalent oxide thickness (EOT)

$$\mathrm{EOT} = t_{ox}\frac{\epsilon_{SiO_2}}{\epsilon_{ox}}, \tag{4.5}$$

where t_{ox} corresponds to the actual oxide thickness and ϵ denotes a dielectric constant. Here, the EOT of the oxide layers add up to a total of 0.52 nm, which is an aggressive, but still achievable value.

4.3.3 Regions

The transistor is composed of three different parts, as shown on Fig.4.2. The corresponding dimensions are listed in Tab.4.1.
First, the source and the drain regions are 20 nm long. This shall allow the subbands to stabilize before reaching the right and left contacts. These regions are doped (*i.e.* charge impurities are introduced in the semiconductor) in order to make available mobile charge for the injection into the channel. In the simulations, the dopant concentration is set to $N_{\mathrm{SD}} = 3\times10^{19} cm^{-3}$ in both the source and the drain.

	InAs	$In_{0.52}Al_{0.48}As$	
E_G	0.417	1.879	(eV)
E_P	18.0	18.29	(eV)
γ_C	2.25	1.04	
γ_1^L	20.00	12.20	
γ_2^L	8.50	4.81	
γ_3^L	9.20	5.47	
m^*	0.026	0.098	(m_0)
Δ_{SO}	0.39	0.37	(eV)
ϵ	15.5	12.88	
b	-1.80	-2.04	(eV)
d	-3.60	-3.50	(eV)
a_v	-1.00	-1.71	(eV)
a_c	-5.35	-5.35	(eV)

Table 4.2: $\boldsymbol{k}\cdot\boldsymbol{p}$ (top) and strain (bottom) parameters used in the simulation [7]. The InAlAs parameters are obtained using Bowing's interpolation between InAs and AlAs.

Phonons			Modes		Transport	Temp.
D_{ac}(eV)	$D_{\text{opt}}(10^8\text{eV/cm})$	$\hbar\omega_{op}$(meV)	CB	VB	direction	(K)
5.8	2	30	4	4	[100]	300

Table 4.3: Simulation parameters used for the MOSFET device. The phonon deformation potentials are taken from [8]

Second, *spacer* regions are added before and after the gate. These undoped channel extensions are often introduced to reduce SCEs. Indeed, even though they are not located directly under the gate contact, they will still contribute to increase the width of the potential barrier, which shall reduce *off*-state tunneling phenomena [5]. They also help to decouple the doped regions and the gate, thus improving the electrostatic control of the gate.
Third, the gate length L_G will range from 8 to 40 nm. Size reduction is indeed a keystone in the design of novel devices, but can lead to various detrimental phenomena that have been described earlier. For that reason, we will study the effect of gate length reduction on the performance of the device in Sec.4.7

4.4 Simulations

Let us define x as the direction of transport and z as the vertical direction (the direction in which the layers are stacked). The system is periodic in the third direction y and the total wave function can be written

$$\psi(x,y,z) = \psi_W(x,z)e^{ik_y y}, \tag{4.6}$$

where ψ_W is the wave function of a device of width W along the y direction. Thus,

the discretization of the Hamiltonian is performed in the real-space along x and z and in the k-space along y. Put another way, we have

$$\hat{\mathcal{H}} \equiv \hat{\mathcal{H}}(-i\frac{\partial}{\partial x}, k_y, -i\frac{\partial}{\partial z}). \tag{4.7}$$

Here, we have set $W = 10$ nm after observing that the results remained unchanged if we kept increasing the width beyond this value. A more detailed derivation of such a 2D Hamiltonian is done in Sec.2.6. This choice of discretization allows us to account for the heterostructures present in the device. Indeed, the materials have been chosen in such a way that the band structure gives rise to a two-dimensional electron gas (2DEG) in the channel. The small-gap InAs is sandwiched between two large-gap materials, which create a potential well and confine the electrons along the z direction. If the whole bulk of the device was made of InAs, the electron flow could occur outside of the 2.4 nm channel layer and this confinement would not arise. Such an architecture is similar to the *silicon on insulator* (SOI) MOSFETs, where the Silicon channel in grown on a buried oxide. Fig.4.3 shows a vertical section of the energy bands and the way they are modified when a gate voltage is applied.

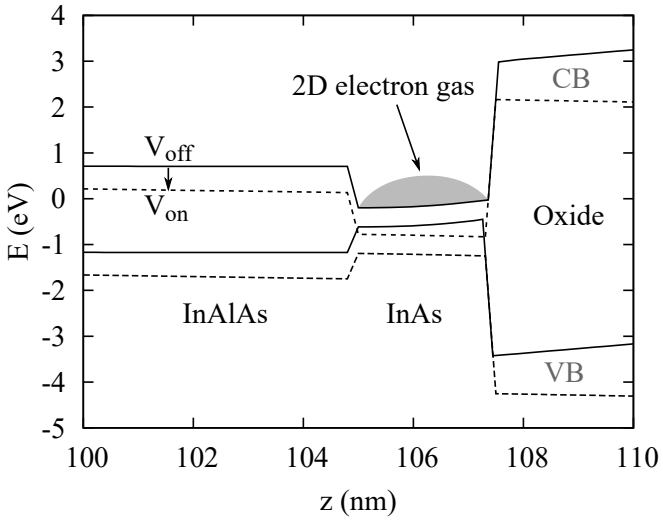


Figure 4.3: Conduction (CB) and valence (VB) bands of the off-state (plain lines) and on-state (dotted lines) transistor, along the z direction. The cut has been done in the middle of the channel, at position x=50 nm, for V_{DS}=0.7V. The electron gas (sketched in gray) is confined in the InAs channel, due to the larger gap of the surrounding materials. Apart from a slight band deformation near the oxide, the whole band structure is shifted towards lower energies when the gate is switched on.

The figure also illustrates the effect of V_{GS} on the band structure. The fact that the bands are slightly bent shows that the coupling with the gate electrode increases as we get closer to the oxide. However, this bend is quite weak when compared to the overall downshift of the bands, created by the application of a gate voltage. This means that, despite this slight curvature, the gate still controls the top and the bottom parts of the channel nearly equally. Fig.4.4 confirms this statement, by showing that the electron density in the channel is rather homogeneous in the vertical direction.

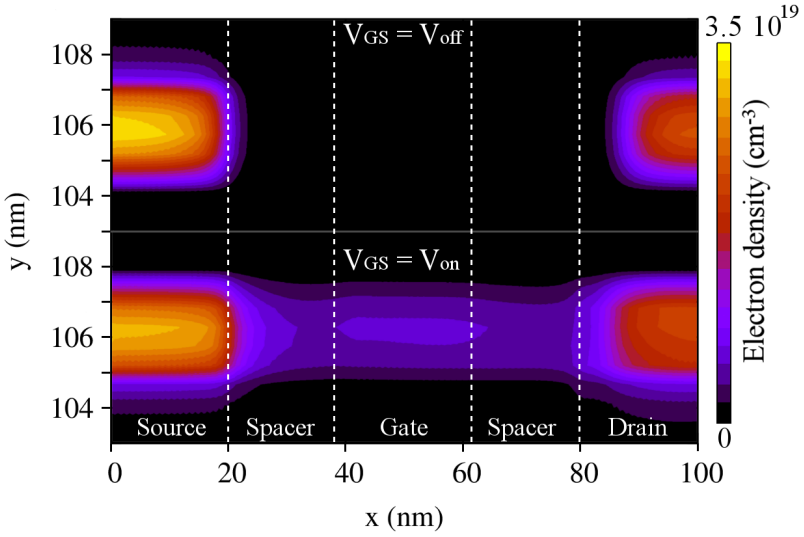


Figure 4.4: Electron density in the InAs layer at the off state (top) and on state (bottom). The source and the drain appear clearly more populated than the channel, due to their doping. In the drain, the density is lower than $N_{SD} = 3 \times 10^{19} cm^{-3}$ due to the presence of a source-drain voltage V_{DS}=0.7V.

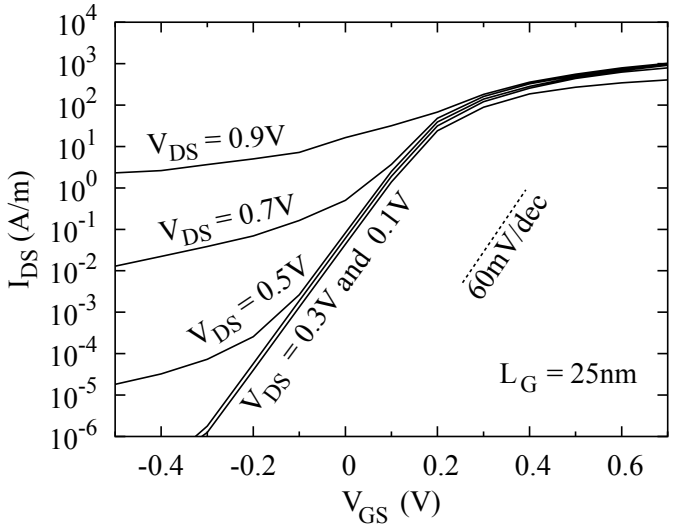


Figure 4.5: Transfer characteristics of the L_G=25 nm device, under various bias voltages. The SS degrades and I_{off} increases as V_{DS} gets larger. To understand this behavior, we shall first focus on the V_{DS}=0.7V device at the off state (V_{GS} <0V).

4.5 Off state behavior

This MOSFET device is mainly suitable for high power (HP) applications. It means that it shall be operated with $V_{DS} \geq$ 0.5V and I_{off}=0.1 A/m. In this first section, we focus on the L_G=25 nm device. I may also be noted that, from now on, the gate voltage V_{GS} is defined in such a way that $V_{off} = 0V$. In Fig.4.5, we plot the device's transfer characteristics for different V_{DS} values varying from 0.1 to 0.9V. This plot is

characterized by a significant increase of the off current and SS degradation as V_{DS} increases. To gain physical insight on this behavior, we first focus on the off state of the V_{DS}=0.7V device (whose SS is strongly degraded) and we plot, in Fig.4.6, the lowest conduction (LC) and the highest valence (HV) subbands of this device, at low V_{GS}. We note that the effective gap is larger than in bulk InAs due to quantum-confinement and strain effects [9]. A first observation which can be made from this figure is that the LC subband is energetically lower than the HV subband near the channel/drain junction area. The occupied states are determined by the position of the source and drain Fermi levels E_F^S and E_F^D. This energy configuration, paired with the shape of the bands, gives rise to *band-to-band tunneling* (BTBT) from the valence band in the channel to the conduction band in the drain. However, as shown in the spectral current distribution (Fig.4.7) such a tunneling does not directly contribute to the overall I_{DS} current: the electrons tunneling towards the valence band balance those tunneling towards the conduction band. Moreover, the HV subband at the source remains lower that the LC subband at the drain. The higher *off* current at high V_{DS} is predominantly due to thermionic current above the conduction-band potential barrier, which is amplified as the source-drain voltage is increased.

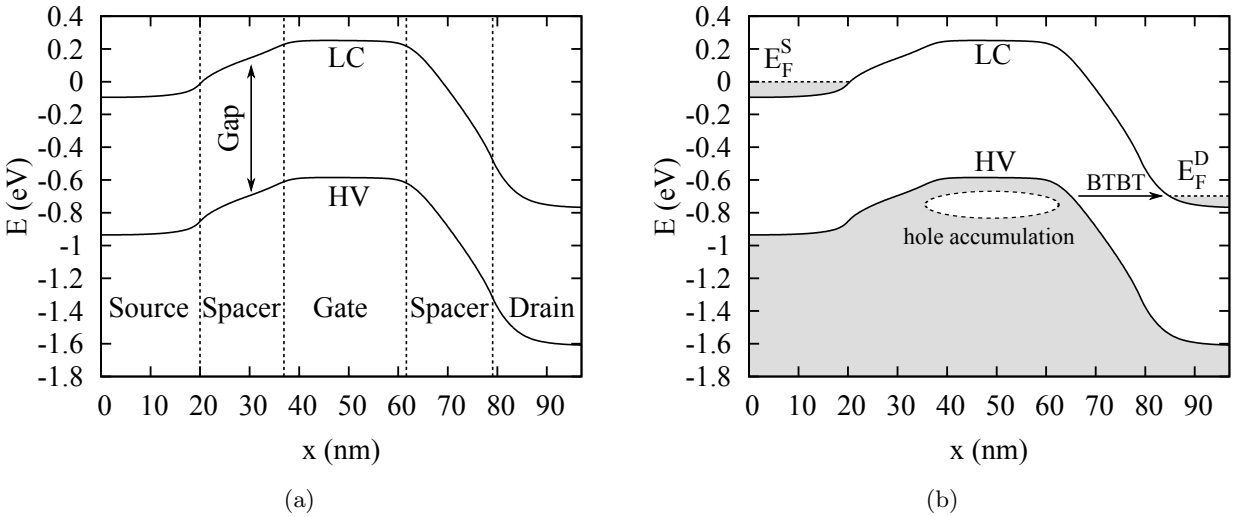


Figure 4.6: (a) Highest valence (HV) and lowest conduction (LC) subbands of the V_{DS}=0.7V device (at V_{GS}=-0.5V), plotted in the channel direction and overlayed with the regions' labels. It also illustrates the role of the spacers as transition regions. (b) Sketch of the electronic occupation of these subbands (shown in grey). E_F^S and E_F^D are the source and drain Fermi levels. The small gap of the InAs allows electrons from the HV subband to tunnel to the drain's LC subband, which generates a hole accumulation under the gate.

To summarize, electrons can tunnel from the channel to the drain via a local BTBT effect, without causing any direct increase of the *off* current. To explain why this BTBT is still detrimental for the InAs MOSFET, we shall investigate the behavior the device when V_{GS} increases. Even though it does not contribute to the net current, the BTBT actually induces a positive charge in the channel (or *hole accumulation*, as shown in Fig.4.6-b), that acts against the gate potential and therefore decreases the gate control on the channel barrier. This phenomenon, called *hole-induced barrier lowering* (HIBL) [10] is especially present at low values of V_{GS}. This can be explained

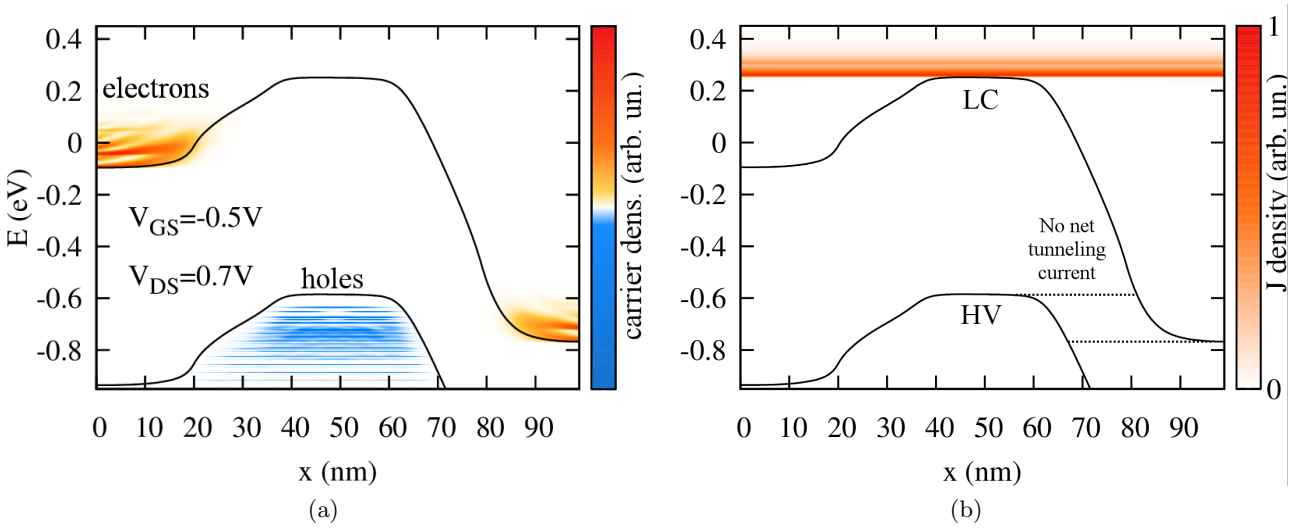


Figure 4.7: (a) Electron and hole densities in the subbands already described in Fig.4.6. This corresponds to the carrier occupation in the HV and LC states, and confirms the presence of holes in channel (HV) states energetically higher than drain (LC) states. (b) Spectral current at the off state (V_{GS}=-0.5V). The off current is actually due to thermionic effects above the top of the LC subband, while the BTBT solely generate local charge accumulation without contributing to the net current.

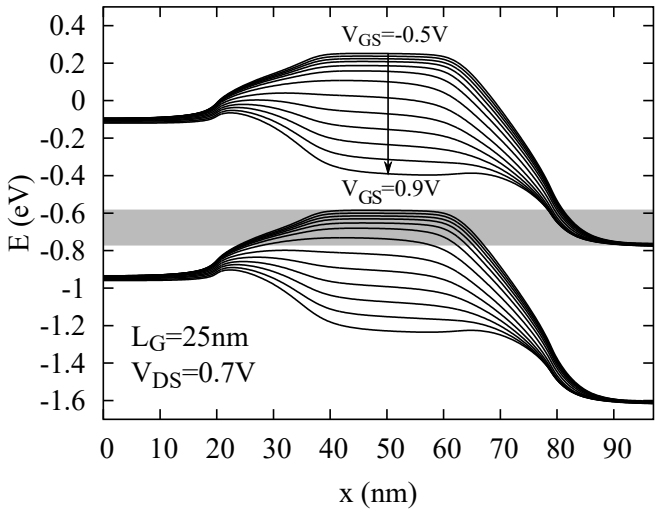


Figure 4.8: Spatial profile of the highest valence subband and of the lowest conduction subband at different V_{GS}. The BTBT window (grey) only encompasses the lowest values of V_{GS}. The gate is thus more effective at larger gate voltages, since the positive charge accumulation in the channel vanishes.

by looking at the evolution of the subbands as V_{GS} increases: the BTBT window only exists for $V_{GS} \lesssim 0V$, but vanishes as the transistor is being switched on (and as the HV subband, responsible of the HIBL, is lowered). This is more clearly illustrated by Fig.4.8.

4.6 Bias voltage scaling

We have just explained how BTBT and HIBL can be detrimental to the *off* state of the device and thus to the overall performance of the InAs MOSFET. Let us come back to our initial observation, which was that the *off* state was more degraded at high values of V_{DS} (Fig. 4.5). When comparing the subbands of low and high bias voltage devices (Fig. 4.9), one notices that the BTBT window is greatly reduced when V_{DS} decreases. As depicted in Fig.4.10-a, values of V_{DS} lower than 0.6V see their SS substantially improved, due to the better control of the gate on the channel. This result confirms what we have been saying up to now. Moreover, while we would normally expect an *on* current increase with V_{DS}, there is actually a breaking point after V_{DS}=0.6V where I_{on} starts to decline, due to the relevance of BTBT.

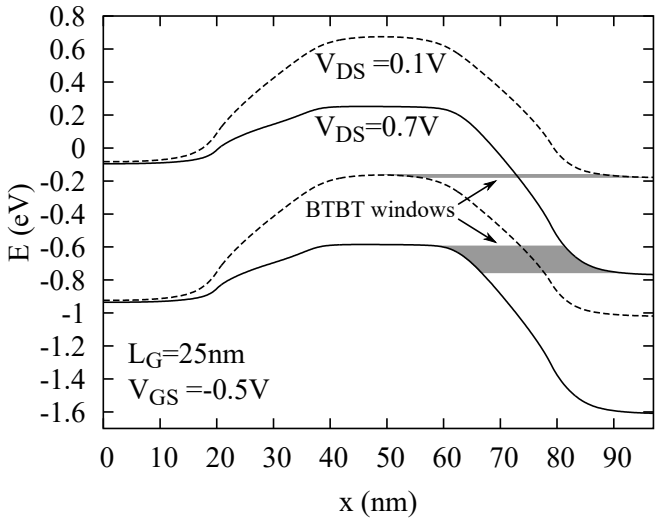


Figure 4.9: Off state subbands of low and high V_{DS} devices. The BTBT window almost vanishes in the V_{DS} = 0.1 V transistor, since the Fermi level at the drain is now above the top of the HV subband.

Now that the influence of the bias voltage has been clarified, we shall select a more optimal V_{DS} value for the rest of this study. To benefit from an improved *off* state while still responding to the HP specifications, we will now focus V_{DS} = 0.5V, which shall not present the HIBL responsible of the SS degradation shown in Sec.4.5. In the next steps, we will review some aspects which are more intrinsic to the device, like its geometry of the materials it is made of.

4.7 Gate length scaling

The gate length is one of the key metrics of a transistor. By reducing the length of the channel, one can integrate a greater quantity of devices in a given chip. However this usually come with various side effects, which we will now evaluate in the case of the V_{DS}=0.5V device. After running simulations for different values of L_G, ranging from 8 to 40 nm, we observe that the shortening of the gate leads to a worse electrostatic integrity. This is revealed by Fig. 4.11-a, which shows that the SS increases as L_G is

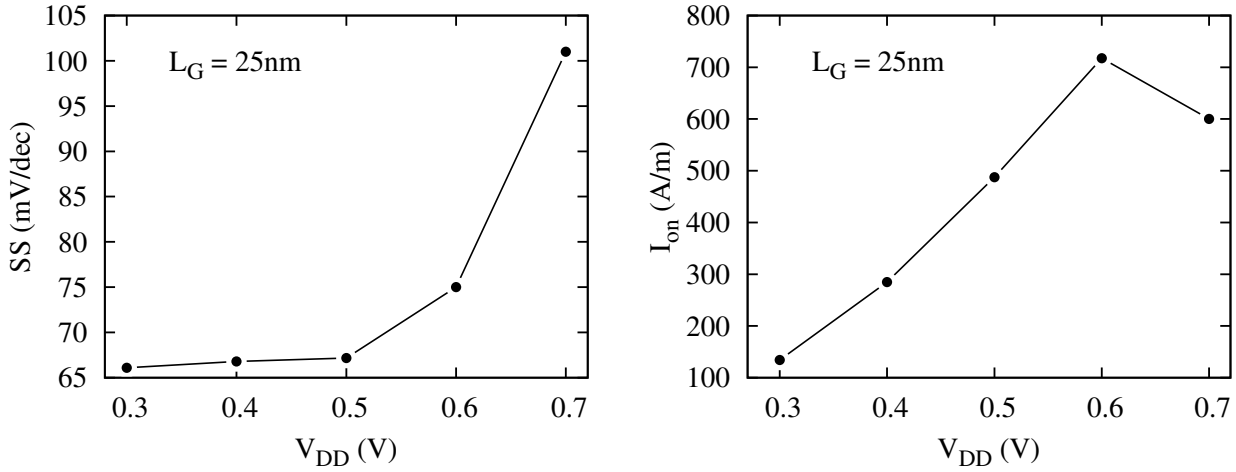


Figure 4.10: (a) SS evolution for different values of supply voltages $V_{DD} = V_{DS} = V_{GS}$. The slope shows decent values up to $V_{DD} > 0.6V$, beyond which it is quickly deteriorated due to the HIBL described previously. This effect ultimately results in a drop of I_{on} at high V_{DD} (b), from which we can conclude that there is not real interest of using this device beyond $V_{DD} = 0.6V$

reduced. On a positive note, we notice that the devices with a gate longer than 25 nm present a subthreshold swing close to the theoretical limit of 60 mV/dec.

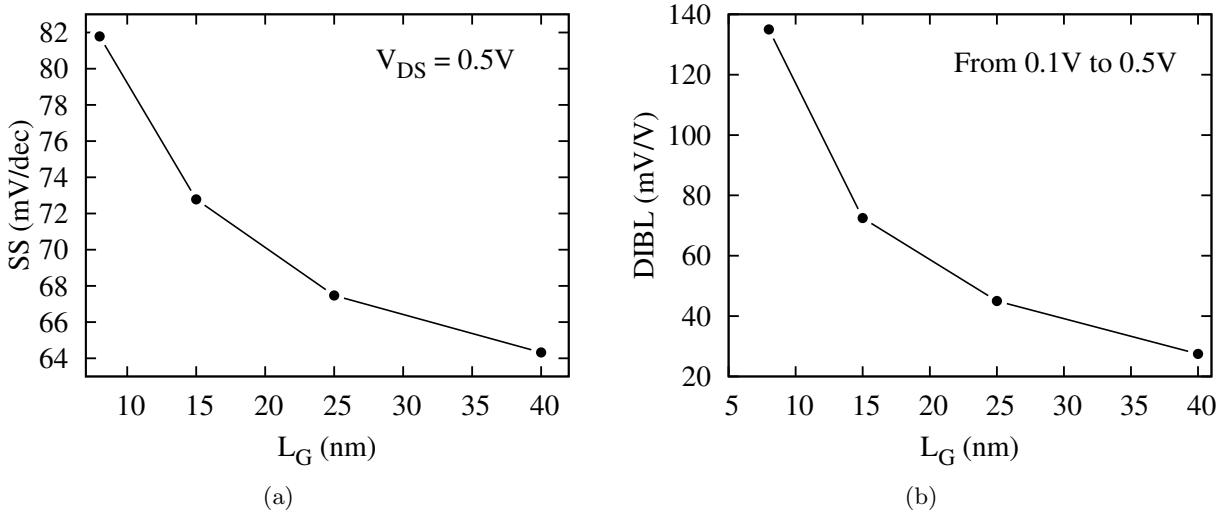


Figure 4.11: (a) Subthreshold swing evolution as a function of L_G in the V_{DS}=0.5V device. A reduction of the gate length deteriorates its electrostatic control on the channel, thus hindering the ability of the device to switch quickly from an off to an on state. (b) This behavior is confirmed by an increase of the DIBL as L_G is shortened. Indeed, when the gate size is reduced, it starts to be affected by the potential of the surrounding low-energy regions, especially the drain.

Another phenomenon resulting from a reduced electrostatic control is the so called *drain-induced barrier lowering* (DIBL). The gate electrode has for sole purpose the

creation (and the removal) of a potential barrier in the channel. When L_G is reduced, the source and the drain tend to be more strongly coupled to the channel region. Due to the voltage difference between the source and the drain, the latter naturally tends to reduce the barrier height in the channel. This effect is amplified when the region on which the gate potential is applied - *i.e.* the channel - is reduced. Such a lowering of the potential barrier facilitates the flow of thermionic electrons at the *off* state, which increases the *off* current. In practice, this effect is evaluated by comparing the change of the *threshold voltage* V_{th} for a given V_{DS} difference:

$$\text{DIBL} = \frac{V_{th}^{high} - V_{th}^{low}}{V_{DS}^{high} - V_{DS}^{low}}. \tag{4.8}$$

In the present case, V_{th} is extracted a $I_{th} = 1$ A/m. The DIBL is then computed by comparing the devices supplied at $V_{DS} = 0.1$V and 0.5V. The higher the DIBL, the more difficult it is for the gate to properly control the current. In accordance with the previous observations, Fig.4.11-b confirms that the short-gate transistors suffer from a decreased electrostatic integrity.

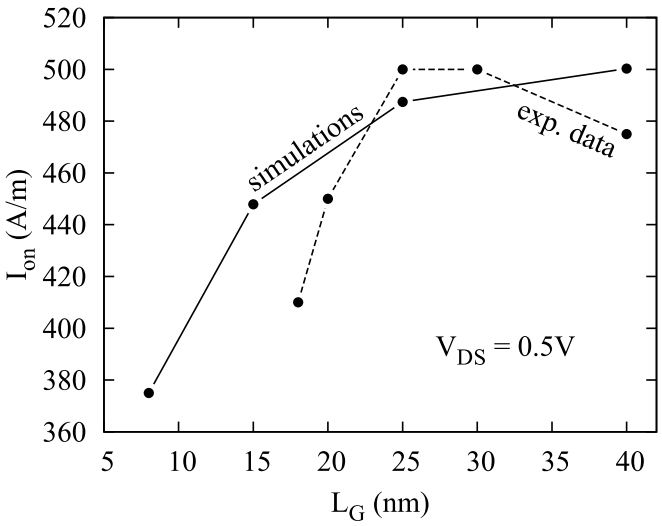


Figure 4.12: Evolution of the simulated I_{on} (extracted for I_{off}=0.1 A/m) as a function of L_G, compared with the experimental values obtained in a similar 2.4 nm thick InAs MOSFET [6]. In both cases, the current is reduced in the shorter devices, due to the worse electrostatic integrity described earlier. However, the simulations do not account properly for the on current degradation observed in actual long-gate devices.

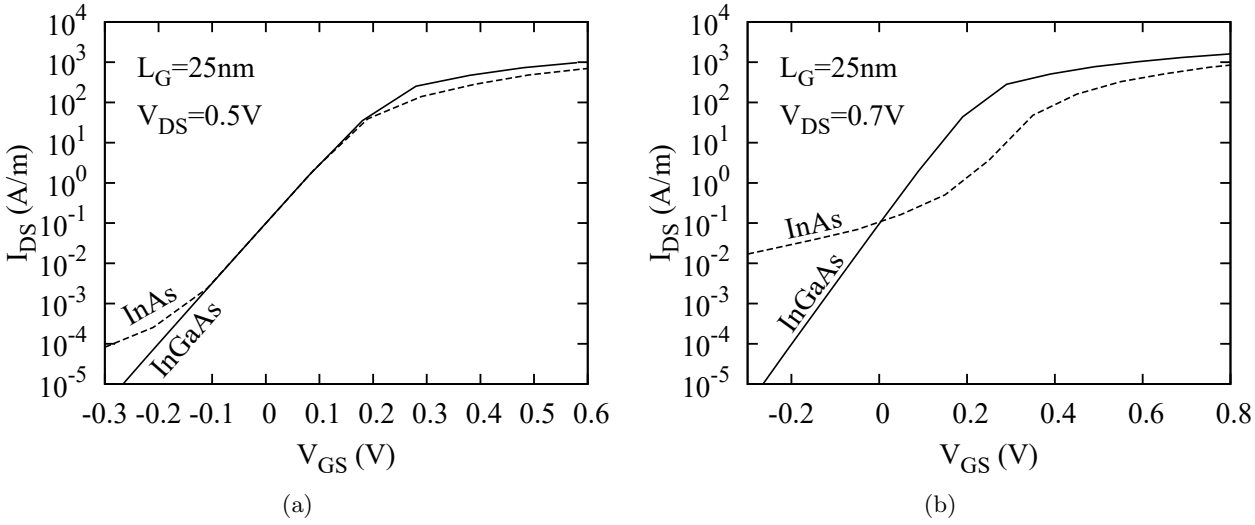
As a consequence, the *on* current is also reduced in shorter devices. In Fig.4.12, we plot the evolution of I_{on} with L_G and our simulations are compared with the experimental data of Ref. [6]. We observe that the simulations, which are closer to ideal devices, are less sensitive to L_G reduction. However, and quite surprisingly, the experimental devices reach slightly higher *on* currents that the simulated ones at moderate gate lengths. Finally, for $L_G \gtrsim 30$ nm, the experimental case exhibits a I_{on} drop, whereas the latter is not visible in our simulations. This I_{on} decrease may be caused by some scattering mechanisms that cause an increased channel resistance in longer devices, or by the presence of other sources of disorder. In the simulations, neglecting these effects

	$In_{0.75}Ga_{0.25}As$	
E_G	0.78	(eV)
E_P	19.60	(eV)
γ_C	4.15	
γ_1^L	17.23	
γ_2^L	6.89	
γ_3^L	7.63	
m^*	0.038	(m_0)
Δ_{SO}	0.41	(eV)
ϵ	14.85	
b	-1.85	
d	-3.70	
a_v	-1.04	
a_c	-5.60	

Table 4.4: $\boldsymbol{k}\cdot\boldsymbol{p}$ (top) and strain (bottom) parameters used for the InGaAs channel MOSFET.

may allow the transport to remain ballistic in the largest devices.

4.8 Comparison with InGaAs



(a) (b)

Figure 4.13: Transfer caracteristics of the InAs and InGaAs devices, at V_{DS} = 0.5V (a) and V_{DS} = 0.7V (b). The gate voltage is chosen so that $V_{GS}(I_{off})$ = 0V with I_{off} = 0.1 A/m in both devices. The InAs transistor is strongly degraded as V_{DS} is increased (both in terms of SS and I_{on}) whereas the InGaAs device remains unperturbed.

We now propose to study a similar device, where the channel material has been replaced with $In_{0.75}Ga_{0.25}As$ (denoted InGaAs in the rest of this section), as this compounds is also a potential candidate for nanotransistor design [11]. The $\mathbf{k}\cdot\mathbf{p}$ parameters of this material are shown in Tab.4.4. Due to the different lattice parameter of this

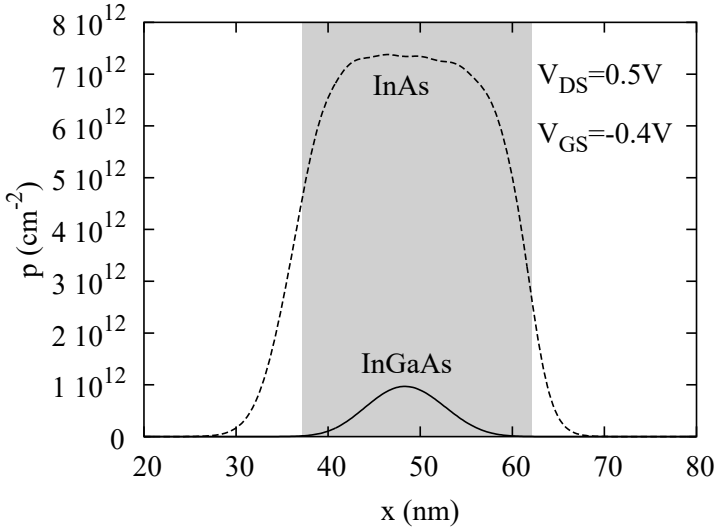


Figure 4.14: Off state hole densities in InAs and InGaAs, integrated along the vertical direction and plotted along the transport direction. The grey area represents the position of the gate.

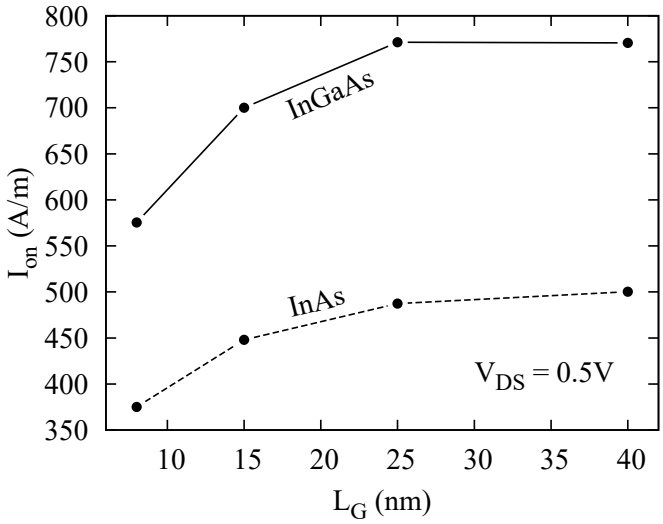


Figure 4.15: On current evolution with L_G in InAs and InGaAs devices.

material, the longitudinal and transverse strain components in the channel are now $\epsilon_{\parallel} = -0.0140$ and $\epsilon_{\perp} = 0.0048$.

In Fig.4.13-a, we plot the transfer characteristic of both the InAs and the InGaAs devices for V_{DS}=0.5V. We note that the InGaAs MOSFET does not suffer from the *off*-state degradation previously observed in the InAs device. However, for V_{DS}=0.5V, by the time the current has reached I_{off}=0.1 A/m, the SS of both devices stabilizes to the same value. Similar simulations done for L_G ranging from 8 to 40 nm confirm that the introduction of InGaAs does not change the SS of the transistor in the operating regime (*i.e.* above I_{DS} = 0.1 A/m) or the ratio of tunneling current at V_{GS} = 0 V. However, as stated above, the remote *off* state ($V_{GS} < V_{off}$) is greatly improved with this new channel material – *i.e.* I_{DS} keeps getting smaller as V_{GS} is

decreased. This can be explained by the larger gap of the InGaAs, which breaks the BTBT window, thus preventing the HIBL. Fig.4.14 confirms that the hole accumulation is significantly weaker in the InGaAs device, even at really small V_{GS}. Even though this property is not an actual advantage at $V_{DS} = 0.5$ V, it can allow the InGaAs device to perform decently at higher bias voltages, where the InAs transistor suffers from a degraded I_{off}. The V_{DD} scalability of the InGaAs transistor will not be thoroughly detailed, as we have chosen to focus on the $V_{DS} = 0.5$ V case. Nevertheless, we can briefly mention Fig.4.13-b, which represents the transfer characteristic of III-V MOSFETs for $V_{DS} = 0.7$ V. Here, the InGaAs device shows an excellent performance, with SS = 68 mV/dec and $I_{on} = 1340$ A/m (which clearly outperforms its InAs counterpart, with SS = 192 mV/dec and $I_{on} = 600$ A/m).
Both materials present a similar SS at $V_{DS} = 0.5$ V, but another strong point of the InGaAs device is its improved I_{on}. Fig.4.15 shows that the *on* current is affected the same way by L_G in both devices (which confirms that the electrostatic control is similar in both cases) but the InGaAs MOSFET is shifted up of more than 200 A/m. This semiconductor has a higher effective mass than InAs (due to the presence of GaAs), which means that it should present a lower electron mobility. However, this also allows this material to benefit from a larger density of states. This can be seen more intuitively in the effective mass model, where the dispersion relation gets wider as m^* is increased. In low effective mass semiconductors such as InAs, the charge in channel may thus be reduced. In turn, this generates a drop of the current, which can, in some cases, cancel out the *on* current enhancement resulting from the increased electron velocity. Ultimately, InGaAs devices show an improved performance in such a n-type thin-body MOSFET by benefiting from a better I_{on} without suffering from a SS degradation.

4.9 Channel thickness

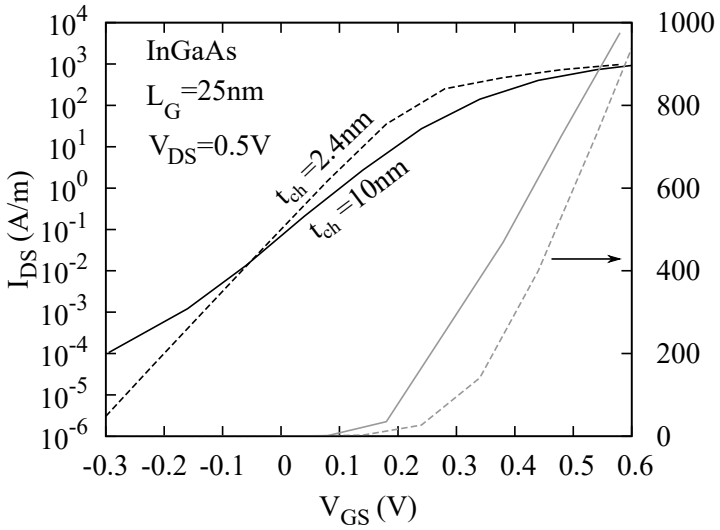
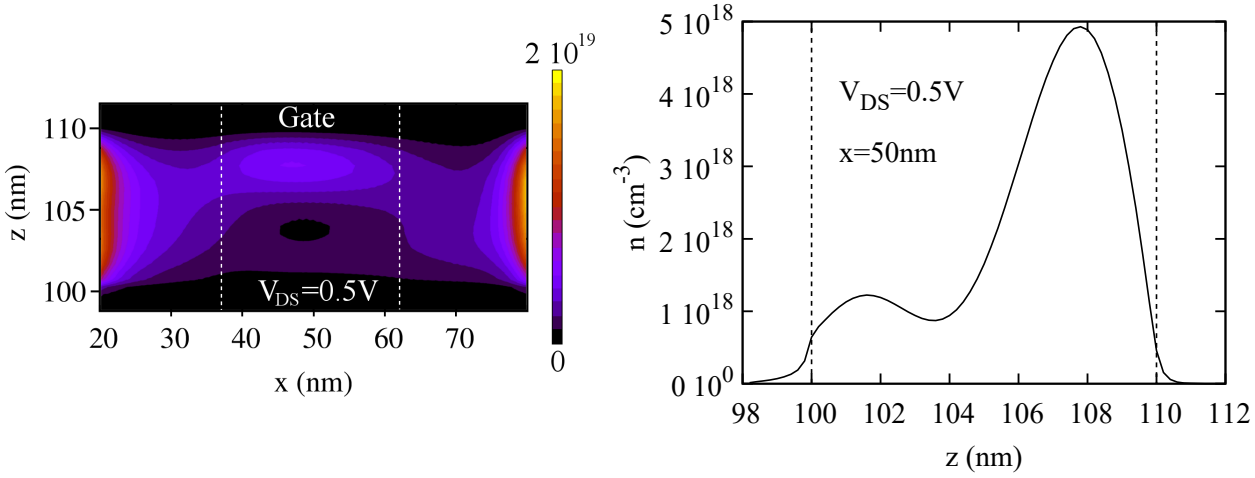


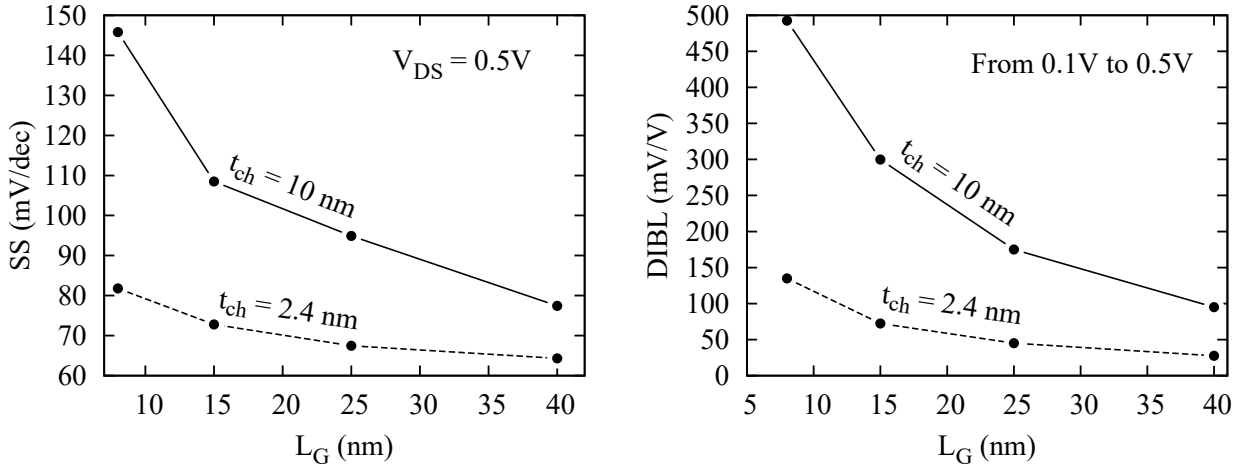
Figure 4.16: Transfer characteristics of 2.4 nm and 10 nm thick InGaAs MOSFETS

We now investigate the effect of the channel thickness on the transport. In practice,



(a) (b)

Figure 4.17: (a) On-state electron density in the 10 nm thick InGaAs channel. The carrier distribution shows a lack of homogeneity under the gate. (b) Vertical cross section of Fig.(a) in the middle of the channel.



(a) (b)

Figure 4.18: Subthreshold swing (a) and DIBL (b) evolution as a function of L_G in the 2.4 nm and 10 nm thick devices. The thicker transistor shows an overall worse SS and DIBL (especially at small L_G) and a greater sensitivity to gate length scaling.

growing thick InAs layers can result in a loss of material quality, by inducing dislocations or other types of defects. For that reason, we simulate a device similar to that of the last section, with an InGaAs channel of thickness t_{ch}=10 nm. The I(V) curves of the thick and thin devices are compared in Fig.4.16. At L_G=25 nm, the 10 nm thick transistor appears to be degraded, both in terms of SS and I_{on}. After investigating the *off* state, we conclude that the larger thickness is not responsible for any increase in *off*-tunneling. However, the electrostatic control is strongly impacted by the new geometry of the channel. As depicted in Figs.4.17-a and 4.17-b, the *on*-state charge

distribution is far less homogeneous in the thick device than it was in the original 2.4 nm MOSFET (see Fig.4.4 for comparison). The carriers now accumulate near the oxide, while the bottom of the channel is less populated. We can postulate that this effect is due to the gate potential having a greater influence on the uppermost part of the channel. However, this is only a part of the story, since the charge density also appears more directed toward the bottom of the channel in some other V_{GS} steps (not shown here). We can thus conclude that this behavior is mainly related to quantum confinement effects. This irregular carrier density participates to the degradation of the electrostatic control. Indeed, as shown in Fig.4.18-a, the SS of the thick device is increased compared to the thin-body case. As a result, the DIBL is also far more pronounced (Fig.4.18-b). On top of showing degraded performance, this thick-channel InGaAs transistor is also more sensitive to gate length scaling. This is clearly brought to light by Fig.4.19, which shows that I_{on} drops quickly as V_{GS} is decreased. Only for $L_G \gtrsim 35$ nm, this device starts to benefit from an *on* current better than that of the thin MOSFET. This corresponds to the point where the device does not undergo significant electrostatic degradation anymore and can start to take advantage of its larger channel to conduct more current.

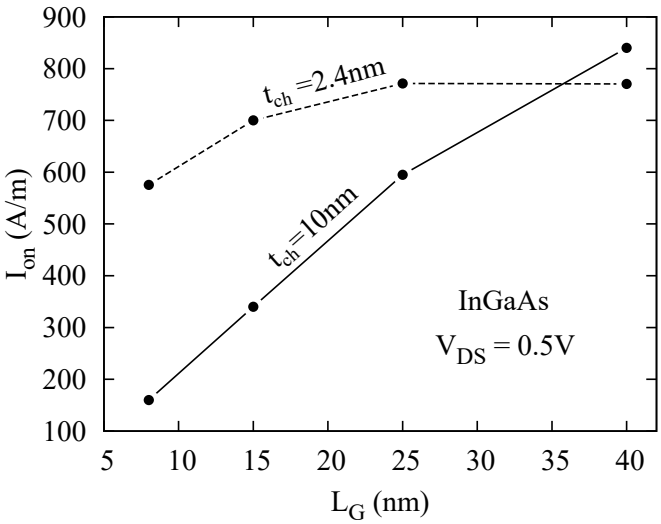
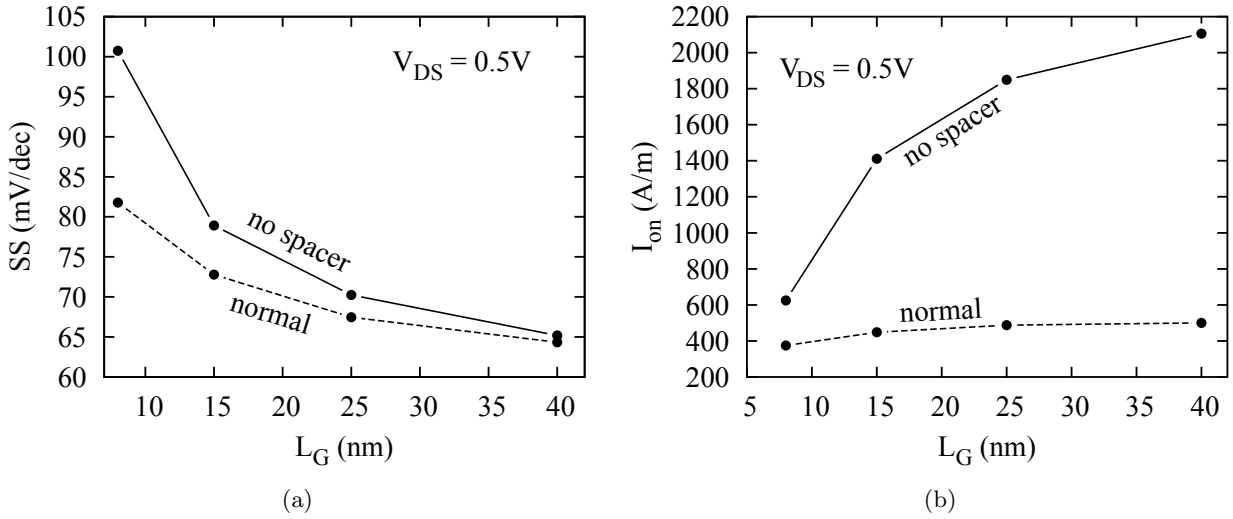


Figure 4.19: I_{on} evolution as a function of L_G in the 2.4 nm and 10 nm thick devices. Due to the its degraded electrostatics, the 10 nm thick device cannot benefit from its wider channel, except at large gate length where it starts to outperform the 2.4 nm thick channel.

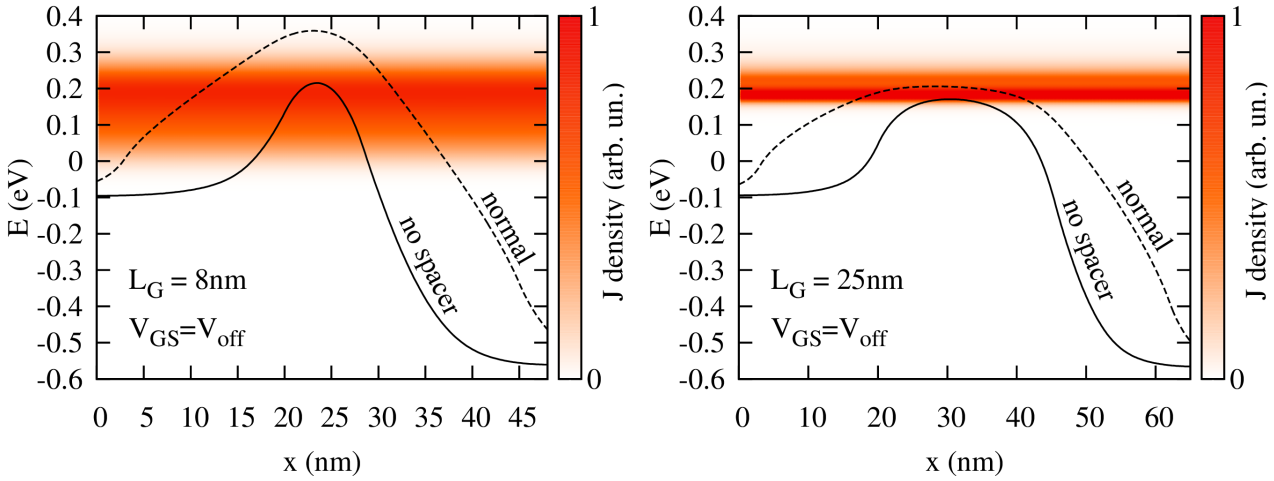
4.10 Spacer effect

We come back to the original 2.4 nm InAs device, to study the effect of the spacer. As explained earlier, these undoped regions located on both sides of the gate help to improve the electrostatic control by putting the center of the channel away from the influence of the source and the drain. This induces a reduced electrostatic screening under the gate, which enhances the modulation of the gate potential barrier. Nevertheless, we propose to study a device where the spacer has been removed.
Fig.4.20-a confirms the predictions that have just been made, by showing a SS degradation in the device without spacers. However, when L_G is increased, the SS of both



(a) (b)

Figure 4.20: (a) SS and (b) I_{on} evolution with L_G of an InAs device with and without spacer (respectively denoted "normal" and "no spacer"). The SS reduction is only significant for the smallest gate lengths, suggesting that the spacers could be omitted is the channel is long enough. Moreover, the removal of the spacers positively impacts the on current, by facilitating the flow of electrons.



(a) (b)

*Figure 4.21: LC subband and current density of the device without spacer, for a L_G=8 nm (a) and 25 nm (b). The dotted line corresponds to the subband of the device with spacer, at the same V_{GS}. V_{off} corresponds to the gate voltage extracted at I_{off}=0.1 A/m. The barrier is decreased both in terms of height and width: the first generates thermionic current **above** the barrier, while the second contributes to STDT **through** the barrier.*

devices becomes very close. Actually, the spacers seem to only be useful for the short gate lengths, where the gate screening is stronger. To gain physical insight on this behavior, we plot the *off*-state LC subband and the current density of long and short devices in Figs.4.21-a and -b. We first note that the barrier height is lower in the

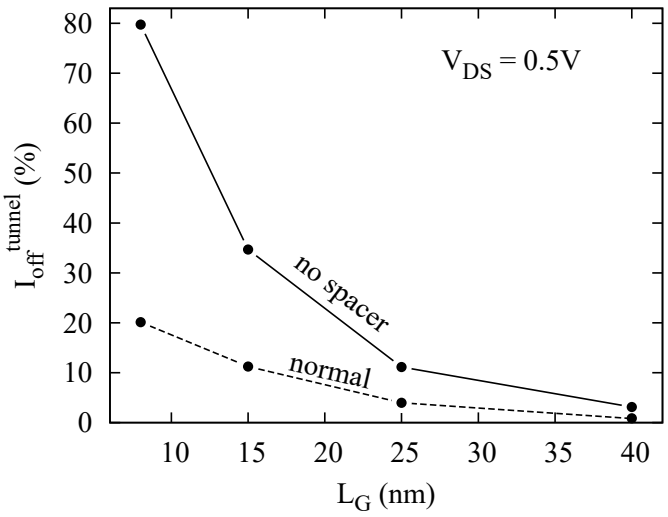


Figure 4.22: Tunneling current ratio as a function of L_G, extracted at I_{DS}=I_{off}=0.1 A/m. The contribution of STDT to the off current is significant in the short devices without spacer. In longer devices (with or without spacer), the off current is essentially due to thermionic effects.

devices without spacer. This barrier lowering, due to the coupling of the gate with the contacts, is even more noticeable at low gate length. The decline of the potential barrier promotes the flow of thermionic current above the top of the LC, from which a parasitic *off* current originates. Indeed, even though the Fermi level is below the top of the barrier, electrons excited by thermal fluctuations can still jump above it. As it could be expected, the barrier also appears thinner when the spacers are removed, since the transition between the contacts and the gate gets sharper. This creates an additional undesired effect at the *off* state, namely source to drain tunneling (STDT). To evaluate the respective contributions of thermionic and STDT currents, we plot the tunneling ratio as a function of L_G in Fig.4.22 (more details about the tunneling current extraction can be found in the next chapter, where we use this method more extensively). In both cases, the tunneling component is more significant at low gate lengths. But the most striking result is that STDT in the transistor without spacer prevails overs thermionic current for $L_G \lesssim 15$ nm. When the gate length is increased, however, both devices show a similar *off* current composition (which is essentially due to thermionic effects). Again, this confirms that when L_G is large enough, the *off* state is not strongly impacted by the removal of the spacers. Further investigations show no significant change in HIBL when the spacers are removed, which allows us to conclude that most of the impact of the spacers takes place in the higher energy range, near the top of the LC.

Finally, the removal of the spacers causes a dramatic increase of I_{on}, as shown in Fig.4.20-b. The shorter channel is indeed in favor of a larger flow of electron due to ballistic effects. Overall, we conclude that the spacers are only useful at small gate lengths, where they improve the SS, with practically no I_{on} deterioration. However, when $L_G \gtrsim 20$ nm, removing the spacers has a substantial positive effect on the *on* current, without strong impact on the *off* state and the slope. In the case of InGaAs devices, the removal of the spacers leads to similar results.

Depending on the targeted gate length for future transistor devices, further investiga-

tion would be required to determine the most optimized choice of spacer length.

4.11 Conclusions and perspectives

As a first application for our NEGF code, we performed 2D full-quantum simulations of InAs and InGaAs nMOSFETs. We started by conducting supply-voltage scaling investigation to evaluate the performance of such devices at different channel lengths. It appeared that the InAs transistor could not work properly for $V_{DS} > 0.6$ V due to its small gap, enabling strong HIBL. For that reason, it seems that InGaAs devices are the most suitable for high V_{DD} applications.
Even though both InAs and InGaAs are sensitive to gate-length scaling we have shown that, for a given L_G, the latter presented an overall better I_{on}, while the SS remains identical in both cases (and compatible with the ITRS requirements for HP applications).
The advantage of ultra-thin-body architectures for III-V MOSFETs has also been demonstrated. The degradation of the electrostatic control in a 10 nm-thick channel deteriorates both the slope and the *on* current. While we have only carried out this investigation for an InGaAs channel (since thick InAs layers are not practical to grow experimentally), we can mention a similar simulation performed for a 10 nm-thick InAs MOSFET, which reveals that the I_{DS} current could not even reach 0.1 A/m (*i.e.* I_{off}) in the selected V_{GS} range.
We also demonstrated that the spacers were not really relevant if the gate-length was large enough. For example, for $L_G = 25$ nm and $V_{DS} = 0.5$ V, the spacers only improve the SS by 2 or 3 mV/dec in both InAs and InGaAs devices. In addition, they can even be quite detrimental for I_{on}, since they introduce additional length to the channel, which prevents to current to be strictly ballistic. In the examples cited above, removing the spacer takes the *on* current from 500 A/m (InAs with spacers) or 800 A/m (InGaAs with spacers) to 1800 A/m. We finally remind that getting rid of the spacers improves the overall size of the device, which is also one of the aspirations of this work.
To sum up, for $L_G = 25$ nm and $V_{DS} = 0.5$ V, by combining the advantages of using an ultra-thin 2.4 nm channel and removing the spacers **the InAs and InGaAs devices can achieve a performance of SS $\simeq$ 71 mV/dec and $I_{on} \simeq$ 1800 A/m**.
In addition, by taking advantage of the better V_{DD} scalability of InGaAs, one can design a 25 nm-long InGaAs device without spacer that reaches $I_{on} \simeq 4000$ A/m and SS $\simeq 71$ mV/dec for $V_{DS} = 0.7$ V (which would not have been possible with InAs).
We can also mention the impact of some additional parameters that have been omitted in the discussion:

- First, we have deliberately chosen to work with a state-of-the-art EOT in order to study cutting edge devices. By example, with the choice of a thicker EOT of 0.8nm, the performance of the device drops from 67.5 to 70 mV/dec in terms of SS and from 490 to 380 A/m in terms of I_{on}.

- Second, we have not discussed the effect of the strain on the quality of the transport. In a test simulation where the strain was removed, the performance seemed to be slightly degraded. However, more extensive work would be required to draw any conclusion on that matter.

- Third, we have not mentioned the possible impact of interface traps and other

defects such that surface roughness. The first issue has already been reviewed in [12], while the second will be addressed in next chapter, in the case of nanowire transistors.

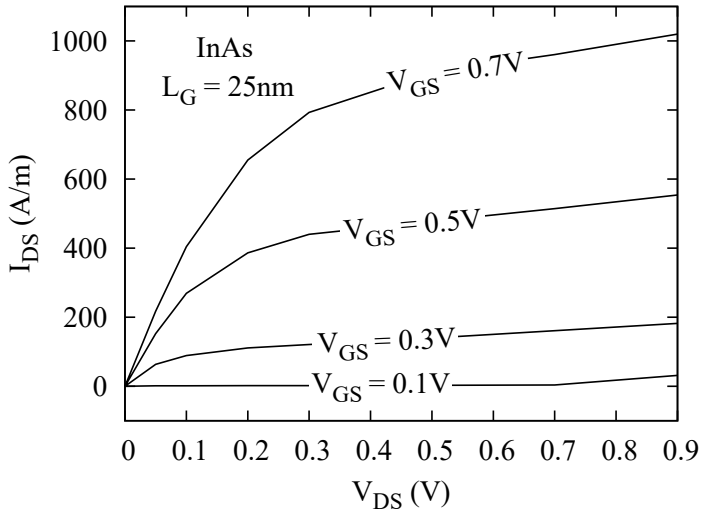


Figure 4.23: Output characteristics a V_{GS}=0.1, 0.3, 0.5 and 0.7 V for an InAs device with L_G=25 nm.

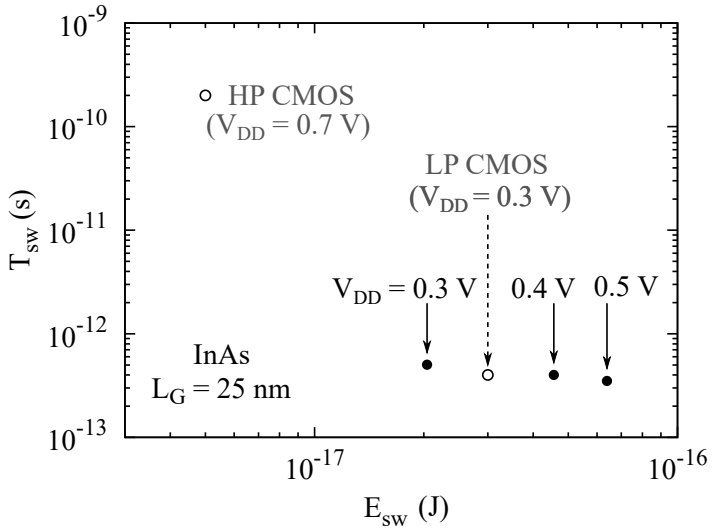


Figure 4.24: Intrinsic switching time T_{sw} versus switching energy E_{sw} of a InAs device (black dots)with L_G=25 nm, estimated for different supply voltages and assuming I_{off}=0.1 A/m, and compared with typical CMOS technologies (white dots), extracted from Ref. [13]. The best performances correspond to the bottom left corner.

On a more applicative point of view, Fig.4.23 shows the output characteristics of a device with L_G=25 nm presenting an ideal linear behavior at small V_{DS} and high saturation currents, demonstrating the excellent transport properties of the ultra-thin III-V MOSFET. One can also estimate the intrinsic switching time T_{sw} and switching

energy E_{sw} of a digital inverter, defined as

$$T_{\mathrm{sw}} = \frac{Q_{\mathrm{on}} - Q_{\mathrm{off}}}{I_{\mathrm{on}}}, \tag{4.9}$$

and

$$E_{\mathrm{sw}} = V_{\mathrm{DD}} \times \big(Q_{\mathrm{on}} - Q_{\mathrm{off}}\big), \tag{4.10}$$

where Q_{on} and Q_{off} are the channel charges in the *on* and *off* states, obtained by integrating the charge density in the device. As expected, in Fig.4.24 the switching delay of the 25 nm InAs MOSFET decreases with V_{DD}, while the switching energy increases (even if more factors, such as the parasitic capacitances, should be taken into consideration to realistically predict these metrics).
As stated before, one of the main impediments in the design of efficient III-V MOSFETs is to obtain an acceptable electrostatic control. Here, one of the main hurdles to overcome is the fact that the gate controls only one side of the channel. To address this problem, we will move to 3D simulations in the next chapter and study III-V nanowire FETs.
On a final note, we inform the reader that the work presented in this chapter has resulted in a journal publication [14].

Bibliography

[1] S. M. Sze. *Physics of Semiconductor Devices*. Wiley, 2007.

[2] International technology roadmap for semiconductors. http://www.itrs.net/2013ITRS/Summary2013.htm, 2013.

[3] J.A. del Alamo. Nanometre-scale electronics with III-V compound semiconductors. *Nature*, 479(7373):317–323, nov 2011.

[4] S. Datta M.S. Lundstrom A. Rahman, J. Guo. Theory of ballistic nanotransistors. *IEEE Transactions on Electron Devices*, 50(9):1853–1864, sep 2003.

[5] T. Rollo D. Esseni, M.G. Pala. Essential physics of the OFF-state current in nanoscale MOSFETs and tunnel FETs. *IEEE Transactions on Electron Devices*, 62(9):3084–3091, sep 2015.

[6] S. Lee, V. Chobpattana, C.-Y. Huang, B. J. Thibeault, W. Mitchell, S. Stemmer, A. C. Gossard, and M. J. W. Rodwell. Record I_{on} 25nm-gate-length ZrO_2/InAs/InAlAs MOSFETs. In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*. Institute of Electrical and Electronics Engineers (IEEE), jun 2014.

[7] L.R. Ram-Mohan I. Vurgaftman, J.R. Meyer. Band parameters for III–v compound semiconductors and their alloys. *Journal of Applied Physics*, 89(11):5815–5875, jun 2001.

[8] M. Lundstrom. *Fundamentals of carrier transport*. Cambridge University Press, 2 edition, 2000.

[9] D. Esseni M.G. Pala. Interface traps in InAs nanowire tunnel-FETs and MOSFETs. *IEEE Transactions on Electron Devices*, 60(9):2795–2801, sep 2013.

[10] M. Luisier, M. Lundstrom, D. A. Antoniadis, and J. Bokor. Ultimate device scaling: Intrinsic performance comparisons of carbon-based, InGaAs, and si field-effect transistors for 5 nm gate length. In *2011 International Electron Devices Meeting*. Institute of Electrical and Electronics Engineers (IEEE), dec 2011.

[11] J. Lin, D.A. Antoniadis, and J. A. del Alamo. Novel intrinsic and extrinsic engineering for high-performance high-density self-aligned InGaAs MOSFETs: Precise channel thickness control and sub-40-nm metal contacts. In *2014 IEEE International Electron Devices Meeting*. Institute of Electrical and Electronics Engineers (IEEE), dec 2014.

[12] D. Esseni M.G. Pala. Interface traps in InAs nanowire tunnel-FETs and MOSFETs - parts i and ii. *IEEE Transactions on Electron Devices*, 60(9):2795–2801, sep 2013.

[13] Dmitri E. Nikonov and Ian A. Young. Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. *Proceedings of the IEEE*, 101(12):2498–2533, dec 2013.

[14] C. Grillet and M.G. Pala. VDD scaling of ultra-thin InAs MOSFETs: A full-quantum study. In *EUROSOI-ULIS 2016*. Institute of Electrical and Electronics Engineers (IEEE), jan 2016.

CHAPTER 5

InAs Nanowire-FET

In which we study nanowire architectures for InAs nanotransistors and compare them to their silicon counterparts – with a specific focus on the influence of surface roughness.

In the previous chapter, we have shown that the electrostatic integrity was a vital aspect of transistor design. A possible solution to improve this feature without having to increase the gate length consists in using a nanowire (NW) architecture. The efficiency of the device comes from the fact that the semiconductor channel can be surrounded by the gate oxide – unlike a "classical" MOSFET for which only one side of the channel is exposed to the gate. The aim of this chapter is thus to study the behavior of NWFETs and determine if they can perform better that the planar architecture. Since the experimental demonstration of such devices is still difficult as of today, we lack a basis of comparison for our III-V devices. For that reason, we will also perform similar simulations for strained silicon devices.
As already mentioned, III-V materials present both advantages and drawbacks with respect to Si. On one hand, they have a smaller effective mass and energy bandgap resulting in higher injection velocity and bulk mobility, but on the other hand, they present a smaller DOS and are more sensitive to quantum effects such as phase-coherent tunneling and lateral confinement. Therefore, at short gate lengths, these materials are expected to present a higher ballisticity [1], but also larger SCEs leading to a reduced electrostatic control of the gate on the channel region [2]. The relative impact of such effects is strongly dependent not only on the gate length, but also on the applied bias voltage.
In the second half of this chapter, we will also study whether III-V based transistors are similarly affected by device variability [3] induced by surface roughness at the semiconductor-oxide interface.

5.1 Description of the device

Various designs of multi-gate transistors exist [4–6]. We choose to focus on square gate-all-around (GAA) NWFETs with a 5×5 nm^2 cross section. Even though actual NWs usually feature circular cross sections [7], the finite-difference discretization employed in this work (see Sec.2.6) is more suitable for rectangular shapes. Fig.5.1 shows a scheme of the device, and its geometrical parameters are listed in Tab.5.1. We considered a gate length L_G ranging from 5 to 20 nm in order to study the impact of SCEs, ballisticity and quantum tunneling on these devices. Since a better electrostatic integrity is expected, no spacer is used between the source/drain regions and the gate. This helps to further reduce the size of the nanodevice. The EOT was set to 0.6 nm, in accordance with the ITRS specifications for the 5 nm technological node and beyond [8]. The doping concentration of the source and drain extensions N_{SD} was chosen in order to enforce

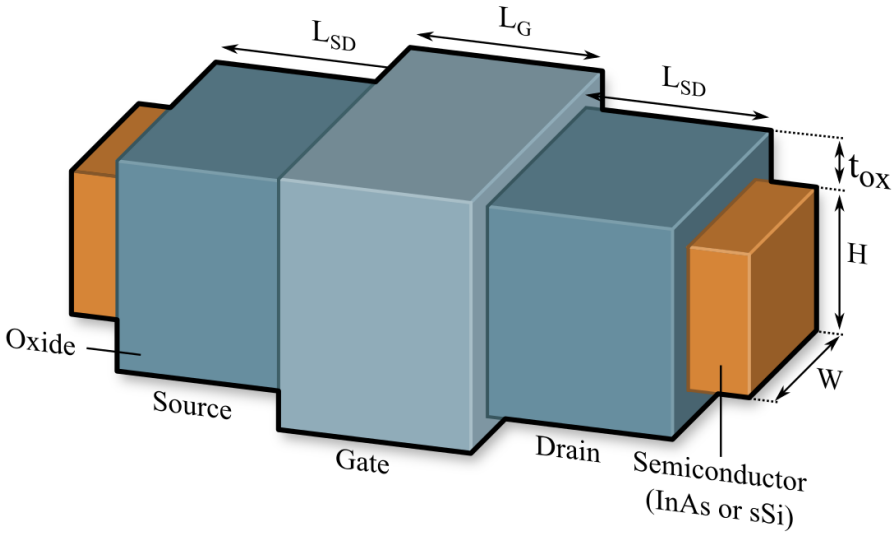


Figure 5.1: Scheme of the NWFET. The lengths of the source-drain regions and the gate (respectively L_{SD} and L_G) as well as the height of the channel and the oxide thickness (respectively H and t_{ox}) are detailed in Tab.5.1.

	L_G (nm)	$W = H$ (nm)	t_{ox} (nm)	L_{SD} (nm)	N_{SD} (cm^{-3})	Transport direction
InAs	5-20	5	1.2	20	$3 \cdot 10^{19}$	[100]
sSi	5-20	5	1.2	10	10^{20}	[110]

Table 5.1: Device parameters of InAs and Si square GAA NWFETs. The oxide dielectric constant is $\epsilon_{ox} = 7.8\ \epsilon_0$. $W(H)$ is the width (height) of the semiconductor, L_{SD} is the length of the source and drain extensions, and N_{SD} is their donor concentration. A compressive uniaxial stress of -1 GPa along the [110] direction is applied to the silicon devices, now denoted "sSi".

degenerate conditions for the semiconductor in the doped regions and mimic small resistance contacts. Since InAs presents a smaller DOS that Si, we set $N_{SD}^{Si} > N_{SD}^{InAs}$, in order to make both semiconductors degenerate. Moreover, the value of N_{SD} used for the InAs devices is limited by the technological difficulties in doping III-V materials [9] and by the necessity to suppress the tunneling leakage reported for InAs nanowires with higher doping concentrations [10]. The reader may also notice that the source/drain lengths L_{SD} are not identical in both cases. This comes from the fact that a shorter distance is required for silicon to reach charge neutrality (*i.e.* flat electric potential), due to its larger DOS and subsequent shorter screening length. Therefore, reducing L_{SD}^{Si} from 20 nm to 10 nm allows us to save computational time, without distorting the results (since the contacts are considered as a periodic prolongation of the source and drain regions).

It has been shown that the inclusion of strain could improve the performance of silicon devices. For that reason, we apply a compressive uniaxial stress of -1 GPa along the [110] to the silicon devices, now denoted as "sSi". The InAs NWFETs are left unstrained, as the effect of strain on III-V materials is not the purpose of this study. Such an InAs/sSi comparison is similar to the benchmarking simulations done in [11]

for MOSFET devices.

	Phonons			Modes		Temp.
	D_{ac}(eV)	D_{opt}(10^8 eV/cm)	$\hbar\omega_{\text{opt}}$(meV)	CB	VB	(K)
InAs	5.8	2	30	10	10	300
sSi	14.7	11, 2, 2	63.3, 63.3, 47.5	20	0	

Table 5.2: Simulation parameters for InAs and sSi NWFETs. Due to the large gap of silicon, the valence band of the sSi is not taken into account since it does not contribute to transport. The phonon parameters are taken from [12] for InAs and from [13] for sSi.

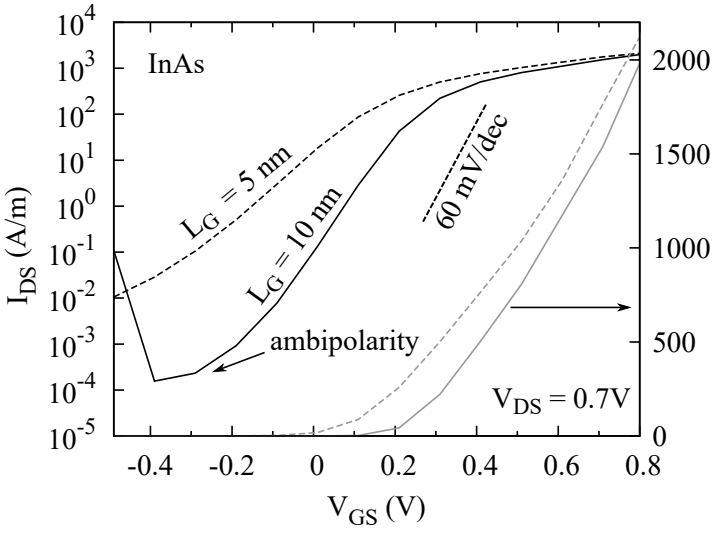
The sSi devices are simulated with the 2-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian detailed in Ref. [14], where only the conduction band is taken into account. Indeed, the effect of the valence band can be neglected in silicon, due to its large bandgap energy E_G. The Si phonon parameters are shown in Tab.5.2, as well as the values used for InAs (that are the same than in Chap.4). In both simulations, 20 modes are used and are equitably distributed between the CB and the VB in the case of InAs.

5.2 Gate-length scaling

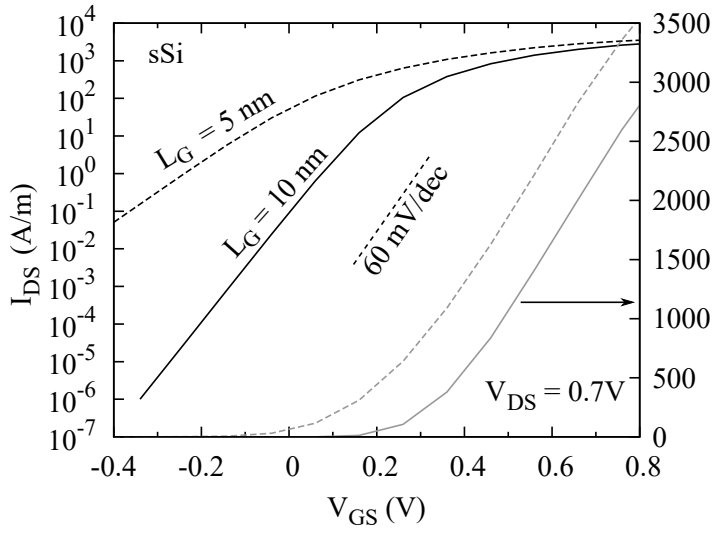
As expected, the performance of the device depends on the gate length. Fig. 5.2 shows the transfer characteristics of InAs and sSi NWFETs with perfect channel-oxide interfaces for L_{G} = 5 nm and 10 nm. The I_{DS} versus V_{GS} curve of the InAs device with L_{G} = 10 nm shows an ambipolar behavior at negative gate voltage values due to the activation of BTBT when the HV subband in the channel gets higher than the LC subband in the source region [2]. The sSi devices do not exhibit any ambipolarity due to their larger energy bandgap that suppresses BTBT. Similarly to the n-MOSFET from Chap.4, the device reaches the I_{off} value required for HP applications, but is still unable to fulfill the low power (LP) requirements (that necessitate $I_{\text{off}} = 10^{-5}$ A/m). Since these devices are interesting only for high performance applications, we adopted an I_{off} target of 0.1 A/m and focused on V_{DD} of 0.5 and 0.7 V close to the ITRS specifications for the ultimate technological nodes [8].
As previously observed in the case of the UTB MOSFET, both InAs and sSi devices show a SS degradation with decreasing L_{G} due to the loss of electrostatic integrity and the increase of direct STDT in the *off* state (as detailed hereafter). Fig.5.3 illustrates the scaling of the SS with the gate length. For large L_{G}, the InAs device presents a better SS than the sSi one, whereas it is the opposite for short L_{G}. These results can be explained by analyzing the electrostatic properties of the two devices at large $L_{\text{G}}$and their sensitivity to electron tunneling at short L_{G}.

5.2.1 Electrostatic integrity

One would expect a worst electrostatic integrity of InAs as a consequence of the higher permittivity of this material with respect to sSi. This can be quantified by using the concept of the natural transistor length λ, which estimates the length of the region electrostatically influenced by the drain contact [6].



(a)



(b)

Figure 5.2: Transfer characteristics at V_{DS}= 0.7 V of (a) the InAs NWFET and (b) the sSi NWFET with L_G= 5 and 10 nm. The current is normalized with respect to the lateral width W = 5 nm. Other device parameters are listed in Tab.5.1. The right y-axis shows the current on a linear scale.

For a rectangular GAA nanowire, it reads

$$\lambda = \sqrt{\frac{\epsilon_{sc}}{4\epsilon_{ox}}\left[1+\frac{\epsilon_{ox}}{4\epsilon_{sc}}\frac{W}{T_{ox}}\right]T_{ox}W} \tag{5.1}$$

where ϵ_{sc} is the dielectric constant of the semiconductor (11.7 ϵ_0 for sSi and 14.6 ϵ_0 for InAs). According to Eq.(5.1), the natural length equals 2.09 nm for InAs and 1.95 nm for sSi transistors.
Despite this prediction, it has been observed that the InAs MOSFETs could achieve a

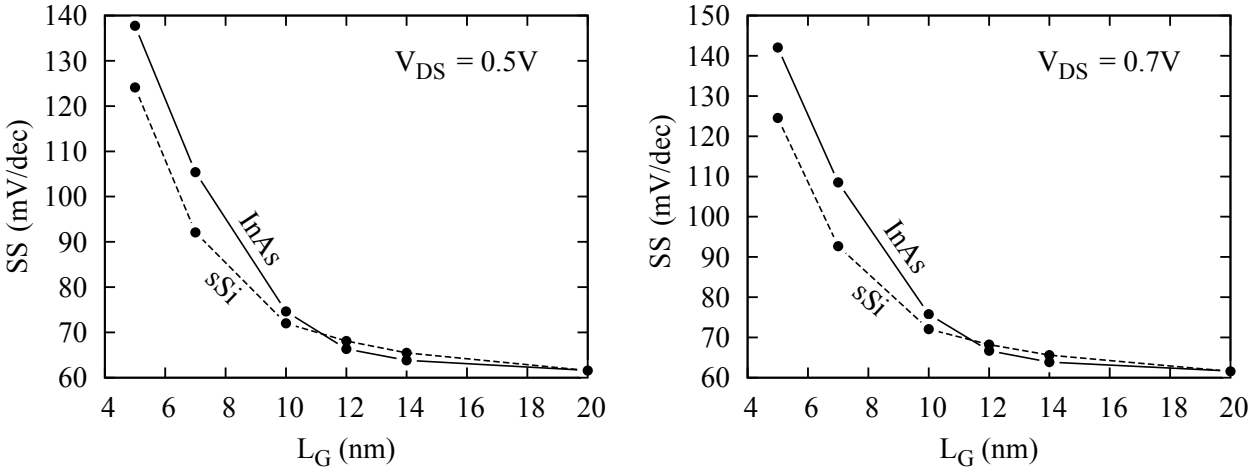


Figure 5.3: Subthreshold swing as a function of L_G for InAs and sSi NWFETs with (a) V_{DS}= 0.5 V and (b) V_{DS}= 0.7 V. SS is computed by averaging the I_{DS} between 0.01 and 1 A/m. Off-state tunneling is the main reason for the difference observed at short L_G.

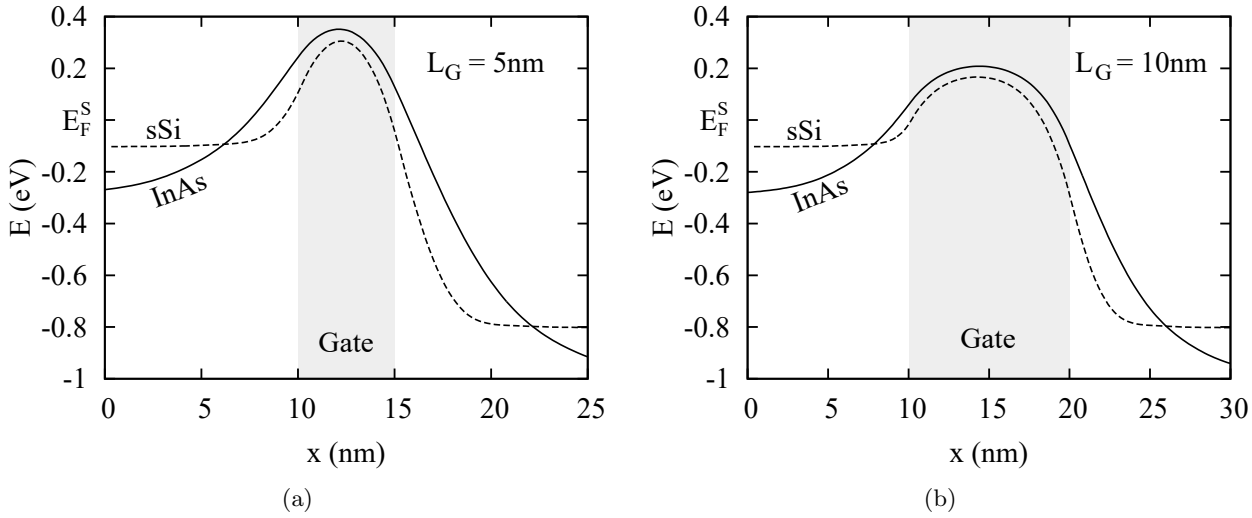
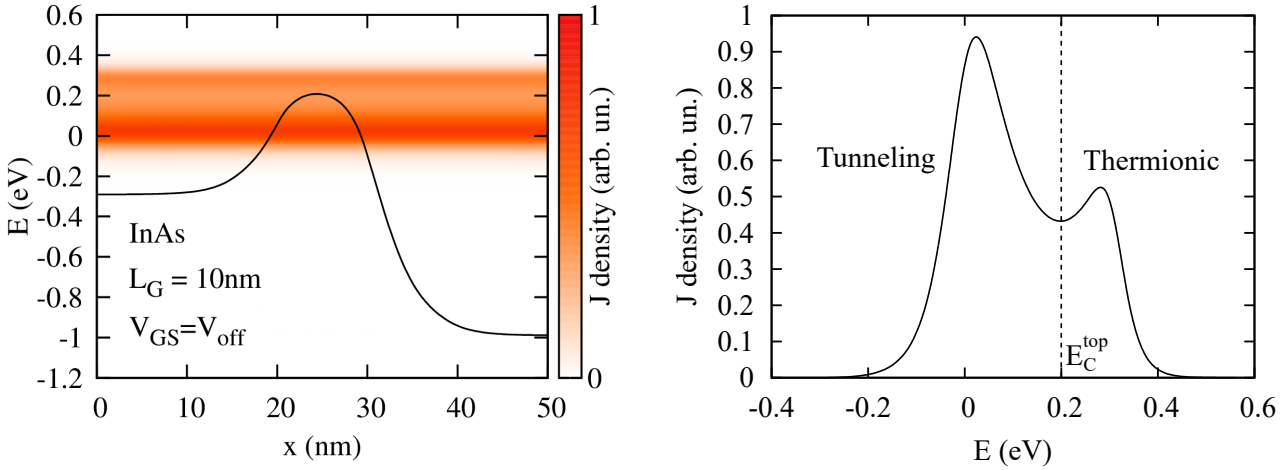


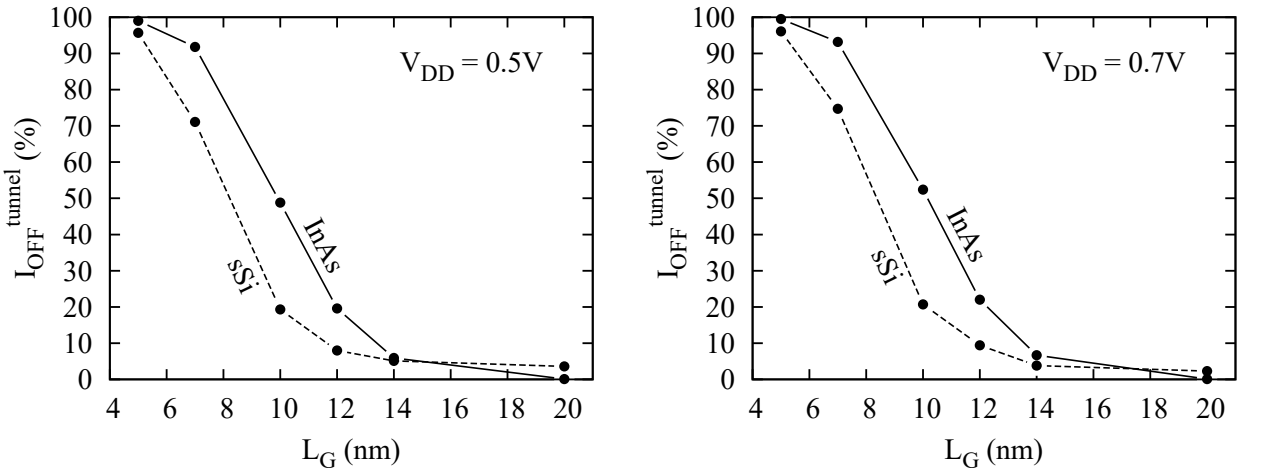
Figure 5.4: Spatial profile along the transport direction (x-axis) of the lowest conduction subband of InAs and sSi NWFETs in the off state (V_{GS} = 0 V and V_{DS} = V_{DD} = 0.7 V). The Fermi level at source is E_{Fs} = 0 eV. The gated region extends from x = 10 nm to x = 20 nm. The source and drain doped regions of the InAs NWFET further extend for 10 nm beyond the interval considered in the plot in order to attain the charge neutrality at the source and drain contacts.

better SS, due to another competing phenomenon: the smaller DOS exhibited by InAs implies a longer screening length [11] and hence requires a smaller doping concentration to obtain degenerate subbands (see Tab. 5.1). A comparison of the lowest conduction subband in the *off* state for the InAs and the sSi nanowires is shown in Fig. 5.4. In the InAs nanowire, the lower doping concentration in the source and drain extensions



(a) (b)

Figure 5.5: (a) Off-state spectral current through the LC subband, in the InAs NWFET. Direct STDT is observed, as a portion of the current tunnels through the potential barrier. (b) The off current exhibits two spikes: a thermionic component and a tunneling component, located respectively above and below the top of the LC subband, denoted E_C^{top}.



(a) (b)

Figure 5.6: Percentage of the tunneling component in the off current versus L_G for InAs and sSi NWFETs with (a) V_{DS}= 0.5 V and (b) V_{DS}= 0.7 V. InAs is more sensitive to STDT when the gate length is shorter than 14 nm

allows the gate to wield a stronger electrostatic control over the channel and hence to assure a better SS. As can be seen, this has also an influence on the shape of the potential experienced by electrons traveling from the source to the drain, which is wider in InAs devices and thus also helps in reducing the tunneling component of the current in the subthreshold regime. This effect explains how the InAs NWFETs can achieve a slightly better SS than the sSi devices for some values of L_G. However, the low effective mass and high mobility of InAs act against this advantage, by increasing

the electron tunneling, especially at short L_G.

5.2.2 Off-state tunneling

Fig.5.5 illustrates the *off*-state spectral current in the InAs device and confirms the presence of STDT. The *off* current can be split into two components: a thermionic contribution, that consists of electrons from the tail of the Fermi distribution jumping above the barrier; and a tunneling (STDT) contribution, which is due to the electrons tunneling through the potential barrier. For short L_G, the SS of both InAs and sSi NWFETs starts to be dominated by this tunneling effect. This is quantified in Fig. 5.6, which shows the tunneling component of the I_{DS}, computed by integrating the current spectrum from the minimum of the conduction band up to $E_C^{top} - 2k_BT$, where E_C^{top} is the top of the lowest conduction subband [2]. Figure 5.6 shows that, in the InAs NWFETs, the percentage of tunneling current is substantially higher than in the sSi ones already at $L_G = 12$ nm, thus indicating that the faster SS degradation for shorter gate lengths in InAs devices is related to the increased STDT.

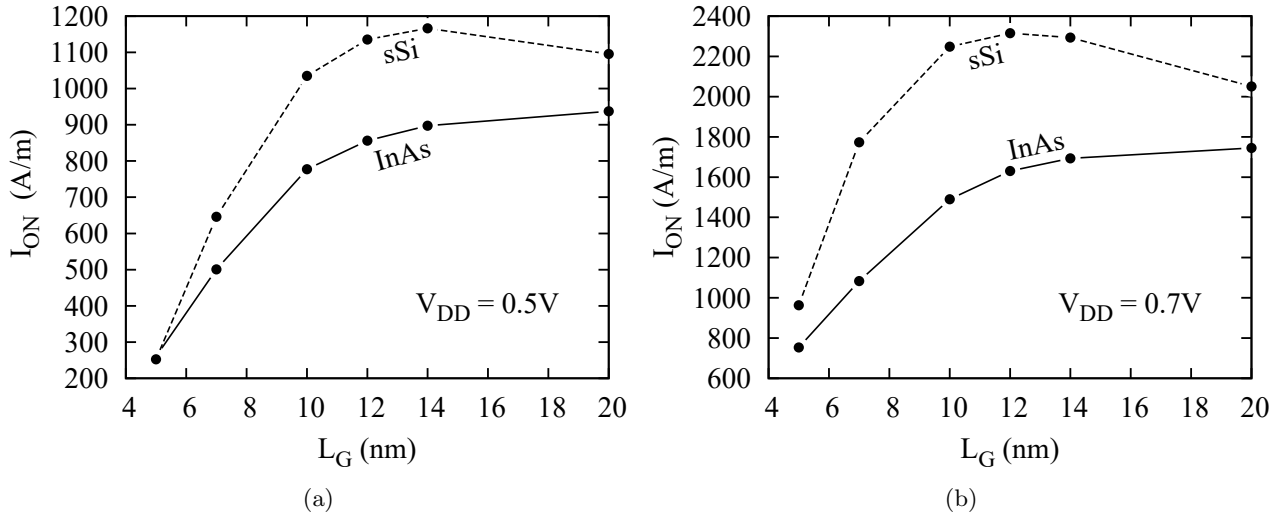


Figure 5.7: On current versus L_G for InAs and sSi NWFETs for (a) $V_{DD} = V_{DS} = 0.5$ V and (b) $V_{DD} = V_{DS} = 0.7$ V. $I_{off} = 0.1$ A/m. Decreasing the gate length causes I_{on} to drop. The sSi devices also show signs of non-balisticity $L_G \gtrsim 14$ nm, due to their lower electron mobility.

5.2.3 I_{on} degradation

Fig.5.7 presents the I_{on} as a function of L_G for InAs and sSi NWFETs at $V_{DD} = 0.5$ V and 0.7 V. This figure presumes a large flexibility of the gate work-function, which is implicitly modified to attain the value of $I_{off} = 0.1$ A/m for the different gate lengths. As can be seen, the maximal I_{on} values are obtained by sSi NWFETs at around $L_G \cong 12$ nm, while the I_{on} drops down when L_G becomes smaller than 10 nm. For $L_G > 12$ nm, the I_{on} of the sSi devices decreases due to the increasing impact of phonon scattering on the channel resistance, while the InAs devices keep experiencing an increase of I_{on} with L_G due to their higher ballisticity. Despite this different behavior, for large values of L_G the InAs devices still provide an *on* current smaller

than the sSi devices. The latter can sustain a larger I_{on} thanks to the higher density of states, implying a larger number of conducting modes at high V_{GS}. The electrical performances of the two devices get closer only for small values of L_G, when the gate overdrive of the *on* state is reduced as a consequence of the SS degradation analyzed in the rest of this section and the number of conducting channels becomes comparable. The dramatic drop of I_{on} visible in Fig. 5.7 for $L_G \lesssim 10$ nm is therefore essentially due to the SS deterioration induced by STDT.

5.2.4 Comparison with the MOSFET

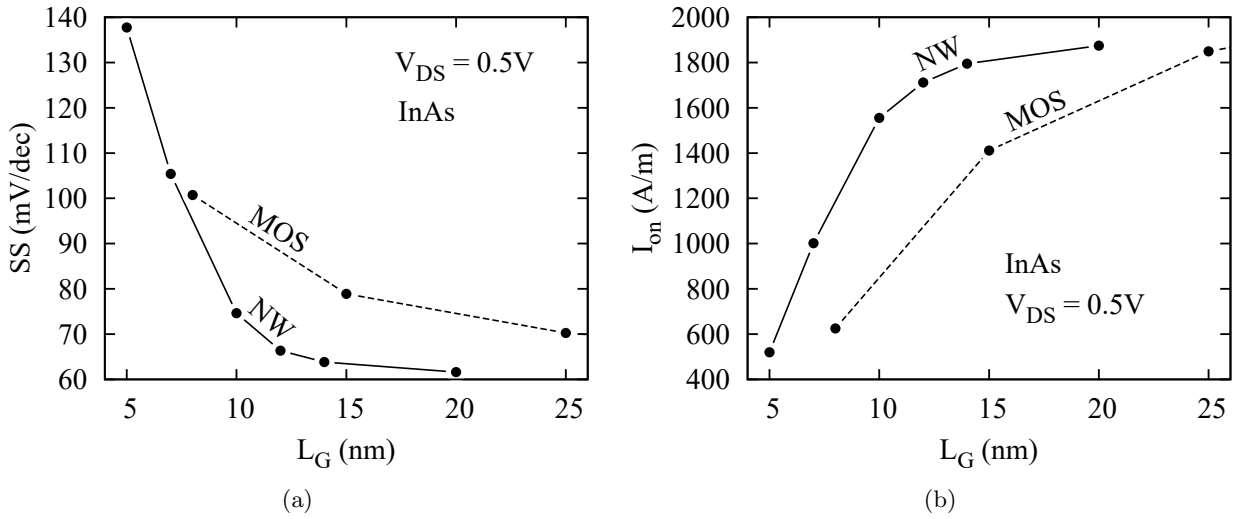


Figure 5.8: (a) SS and (b) I_{on} comparison in an InAs NWFET and a 2.4 nm thick InAs MOSFET without spacers (from Chap.4) at V_{DS}=0.5V. The current in the NW has been renormalised according to the configuration represented in Fig.5.9.

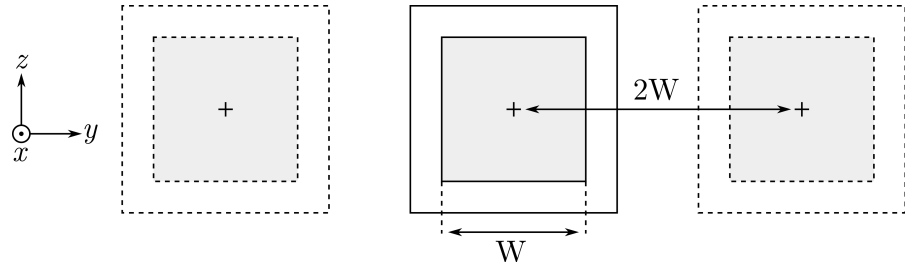


Figure 5.9: Possible spacial arrangement of parallel nanowires, that would occupy the same wafer area than the planar MOSFET. A spacing of 2W (10 nm) corresponds to an optimistic case of density of integration.

As already stated, the NW devices are improved versions of the MOSFETs, as they feature a more effective gate, wrapping around the channel. In Fig.5.8-a, we compare the SS of the present device with the InAs MOSFET from the previous chapter. For this comparison to be fair, we selected the UTB InAs MOSFET without spacers (since the NWFET does not feature any spacer either). At long gate lengths, the NWFET exhibits a better SS than the MOSFET due to its enhanced electrostatic integrity.

However, it appears that the GAA architecture is not able to ensure a significantly better SS than the planar architecture when L_G is reduced – that is, when the *off*-tunneling ratio gets close to 100%. In Fig.5.8-b, we compare the evolution of the *on* current in both architectures. In order to make this examination possible, the I_{on} in the NWFET had to be renormalised. The MOSFET has been simulated as a 2D system, periodic in the y direction. An equivalent design would be to build an array of nanowires, placed side by side in the (x,y) plane. Depending on the spacing between these NWs, the renormalised *on* current will differ. Even though the experimental fabrication of well-arranged NWFETs is still hard to achieve, we assume that a lateral spacing close to $2 \times W$ could ultimately be reached, as illustrated in Fig.5.9. Under this assumption, the I_{on} obtained would be greater than that of the planar MOSFET on the studied L_G range.

5.3 Surface roughness

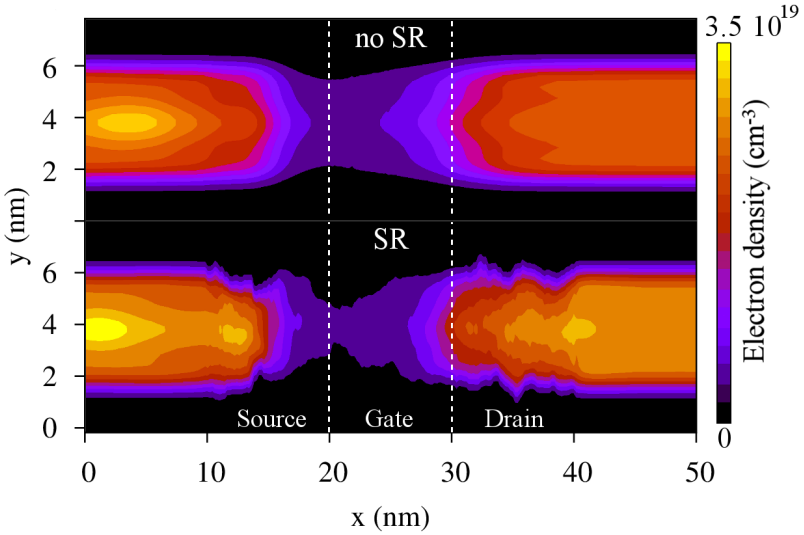


Figure 5.10: Comparison of the electron density at the on state in an InAs NWFET, with and without surface roughness (SR).

The growth of nanowires requires methods such as metal-organic chemical vapor deposition [15], that can not yet achieve the precision of traditionnal etching techniques. In realistic conditions, the devices are thus likely to exhibit surface roughness (SR) [16]. This variability is especially relevant at the channel-oxide interface [17], since it modifies the shape and the cross-section of the channel. In this work, the impact of surface roughness was evaluated by computing the transfer characteristics of InAs and sSi NWFETs for 50 spatial realizations of the rough channel-oxide interfaces. Fig.5.11 corresponds to the rough InAs and sSi devices with $L_G = 5$ nm and 10 nm. Rough interfaces induce a positive threshold-voltage shift (ΔV_{th}) in these devices due to the increase of the top of the lowest conduction subband, which finally determines the threshold voltage [18]. In Fig.5.11, a larger dispersion of V_{th} in InAs devices is found since they are more sensitive to quantum confinement effects and develop higher amplitudes of subband fluctuations. This can be intuitively explained by considering that

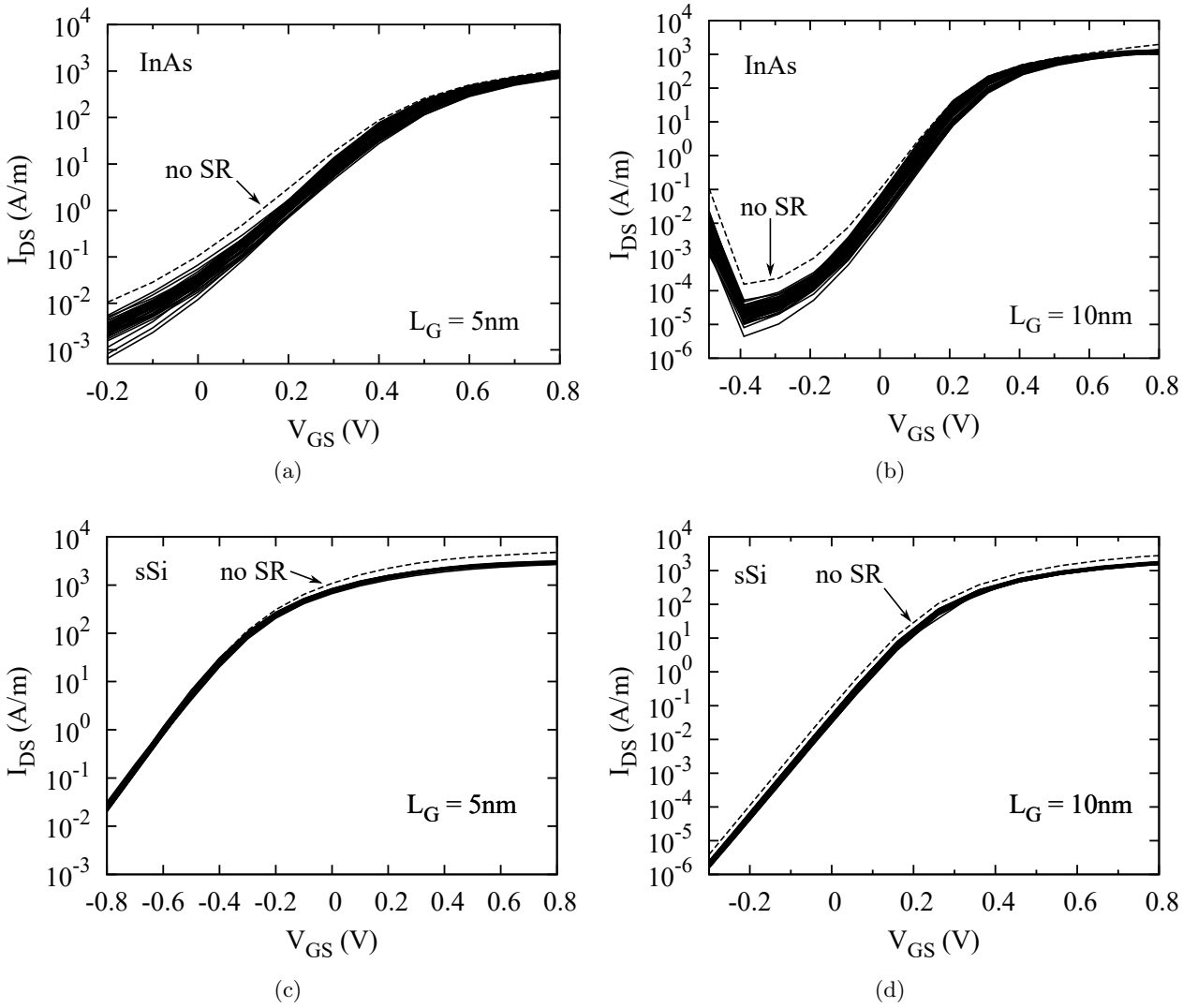


Figure 5.11: Transfer characteristics at V_{DS} = 0.7 V of (top) InAs and (bottom) sSi NWFETs with L_G = 5 nm (right) and 10 nm (left), for different realizations of rough interfaces. The RMS of SR is Δ_R = 0.4 nm. The curves of the devices with no surface roughness are also shown with a dashed line for comparison. The dispersion is more important in the InAs devices.

subband fluctuations are, at the first order in the lateral size variation, inversely proportional to the effective mass [19].

Fig.5.12 illustrates the SS and I_{off} variability of InAs and sSi NWFETs for different realizations of the rough interfaces. For the devices including surface roughness, the I_{off} has been determined as the current at the V_{GS} value for which the device with perfect interfaces features $I_{DS} = 0.1$ A/m. As it can be seen, rough interfaces induce a significantly larger variability of SS in InAs NWFETs. Moreover, Fig.5.11 shows that the SS of InAs devices is typically improved by roughness. According to the discussion of Fig.5.13, this can be explained by considering that the SR tends to induce an increase of the potential barrier in the *off* state and hence, a decrease of the STDT. To confirm this statement, a correlation between the *off*-state tunneling process and the SS can be observed in Fig.5.14. The figure reveals that the L_G=5 nm and 10 nm

InAs devices exhibit less STDT when exposed to surface roughness and that the SS is subsequently improved. We note that STDT is a predominant phenomenon in the 5 nm InAs NWFETs, since the *off* state tunneling ratio in these devices is close to 100%. This also explains the I_{off}/SS correlation observed in Fig.5.12-a for InAs.

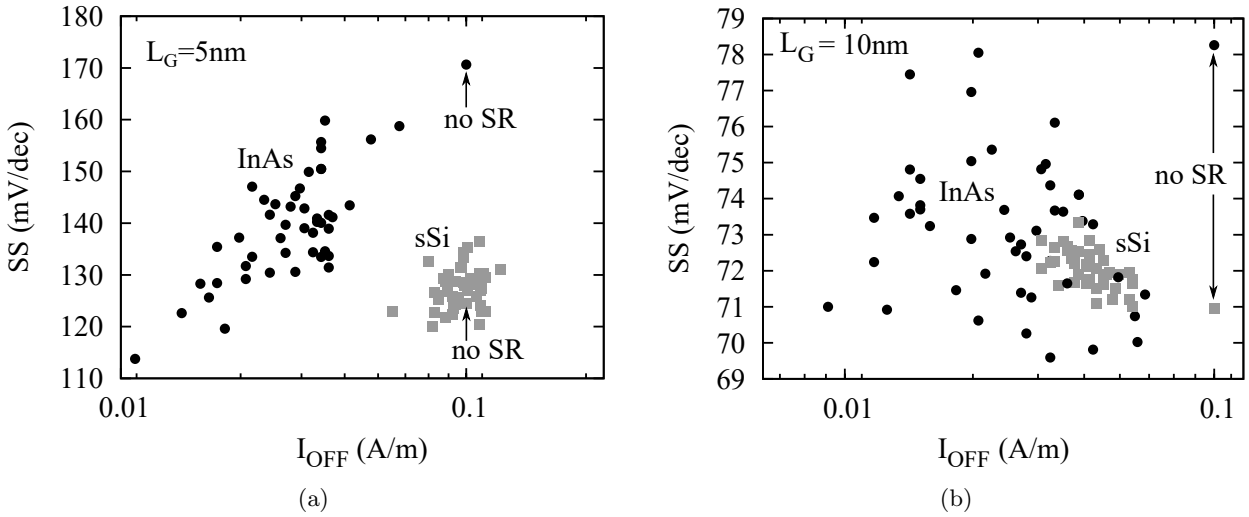
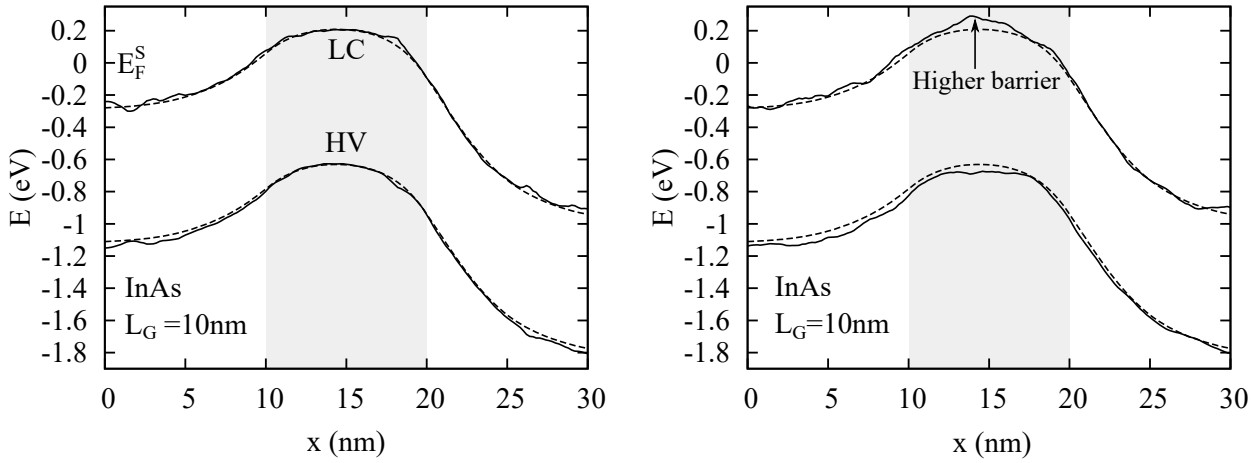


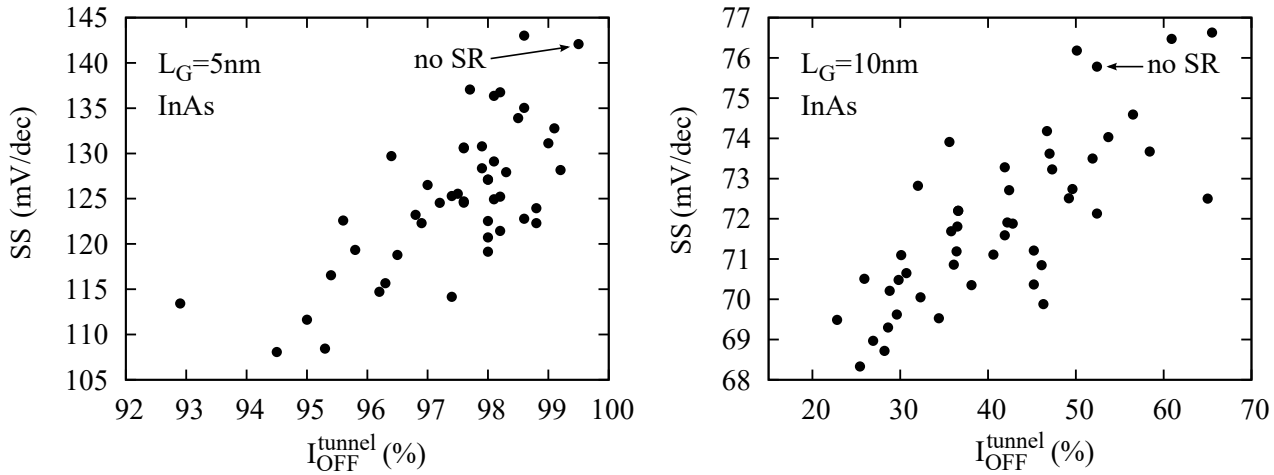
Figure 5.12: Subthreshold swing versus off current at V_{DS} = 0.7 V of (circles) InAs and (squares) sSi NWFETs with L_{G} = 5 nm (a) and 10 nm (b), for different realizations of rough interfaces. The RMS of SR is Δ_R = 0.4 nm. The values obtained for the devices with perfect interfaces are also shown for comparison. The InAs NWFETs present larger I_{off} and SS dispersions than the sSi devices. However, the the SS of the sSi NWFETs tend to be degraded by the addition of SR, whereas the opposite is observed for InAs.

Finally, Fig.5.15 shows the I_{on} versus the I_{off} of InAs and sSi NWFETs extracted from the data in Fig. 5.11. As already observed in Fig.5.12, the I_{off} variability is much larger in InAs nanowires due to their larger subband fluctuations, whereas the I_{on} variability is similar. Another interesting finding is that, in sSi nanowires, the roughness is responsible for a stronger I_{on} reduction with respect to the device without roughness. At short gate length (L_{G}=5 nm), the rough InAs devices can even outperform their rough sSi counterpart in terms of I_{on}. This is a consequence, on one hand, of the SS improvement induced by SR in InAs NWFETs and, on the other hand, of the different lateral distribution of the electron charge at high gate overdrives: back-scattering due to roughness is stronger in sSi nanowires because their *on*-state transport involves higher-order transverse modes. This means that carriers in sSi NWs are closer to the interfaces, whereas they are more localized in the center of the cross-section in InAs NWs, as shown in Fig.5.16. Hence, as illustrated in Figs.5.12 and 5.15, surface roughness has both negative and positive effects on InAs NWFETs: it induces a significant device variability but also an improvement of the SS, which results in I_{on} values closer to those of the sSi counterpart for L_{G}=10 nm and even (on average) better at L_{G}=5 nm. The carrier distribution is further investigated in Fig.5.17, which presents a sample of the rough nanowire realizations. On one hand, at the *off* state, the InAs NWFETs show more fluctuations that the sSi ones. On the other hand, at the *on* state, the InAs and sSi carrier distributions are similarly affected by spatial fluctuations. This observation confirms that most of the differences between these materials



(a) Small subband perturbations, SS weakly improved (b) Large subband perturbations, SS strongly improved

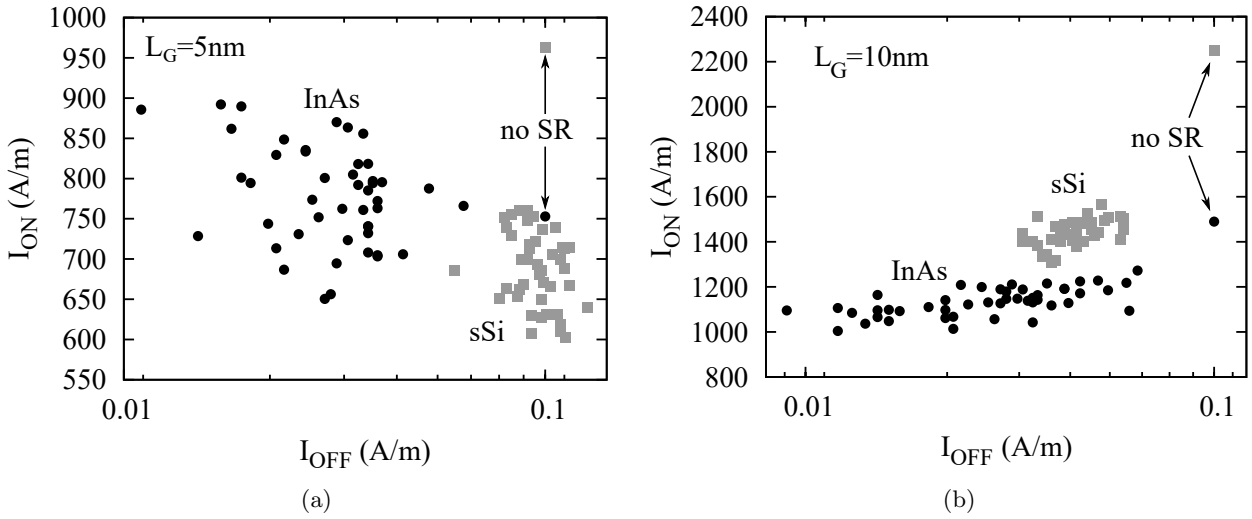
Figure 5.13: Spatial profiles along the transport direction (x-axis) of the subbands of the InAs NWFET with (dashed lines) perfect and (solid lines) rough interfaces in the off state (V_{GS}≐0 V and $V_{DS} = V_{DD} = 0.7$ V). In example (a), the LC subband is only slightly modified by the roughness and the corresponding SS is close to that of the clean device. Example (b) corresponds to a rough device with a greatly improved SS, where the SR induces an increase of the height of the LC barrier.



(a) (b)

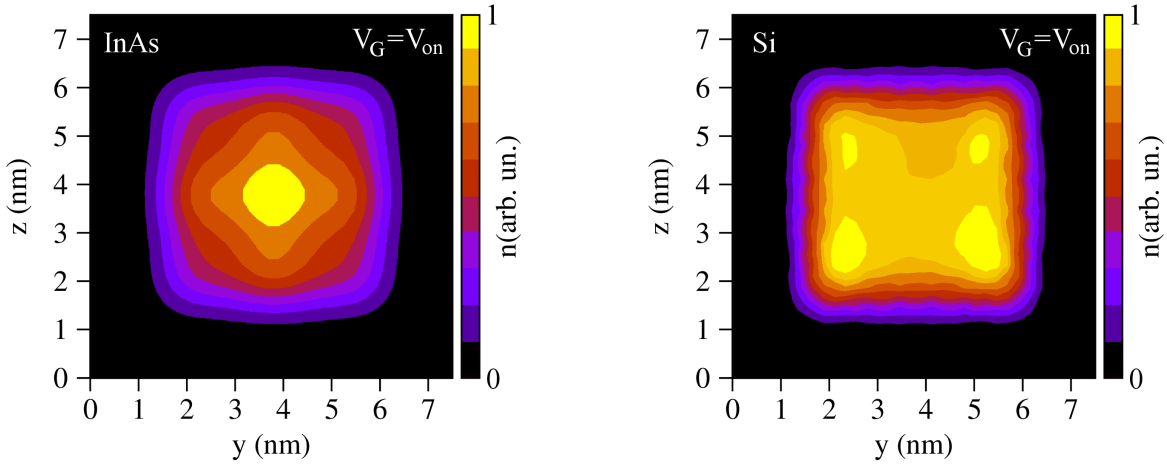
Figure 5.14: Subthreshold swing versus off-tunneling current ratio in InAs NWFETs with surface roughness, for L_G=5 nm (a) and 10 nm (b), at V_{DS}=0.7V. The SR tends to reduce off-state tunneling (via the mechanism described in Fig.5.13). Such a reduction is correlated with an enhancement of the SS.

arise from their *off*-state behavior.



(a) (b)

Figure 5.15: On current versus off current at V_{DS} = 0.7 V of (circles) InAs and (squares) sSi NWFETs with L_G = 5 nm (ab) and 10 nm (b) for different realizations of rough interfaces. The on current in rough InAs devices is less degraded than in sSi. It can even be improved at L_G=5 nm, where the InAs NWFETs can actually outperform the sSi devices.



(a) (b)

Figure 5.16: (a) Cross-section of the charge density the channel of an InAs NWFET, at the on state and without SR. (b) Similar plot for the sSi device

5.4 Conclusion

We have performed full-quantum simulations aimed to benchmark the use of InAs against strained Si as channel material in sub-20 nm gate-length GAA NWFETs. The performance of these devices was evaluated in the context of high power applications, with I_{off}= 0.1 A/m and V_{DS}= 0.5 V and 0.7 V. Even though the sSi devices reached better overall I_{on} and SS values, our results indicate that a slightly better electrostatic integrity could be obtained in InAs NWFETs with $L_G > 10$ nm. An increased STDT

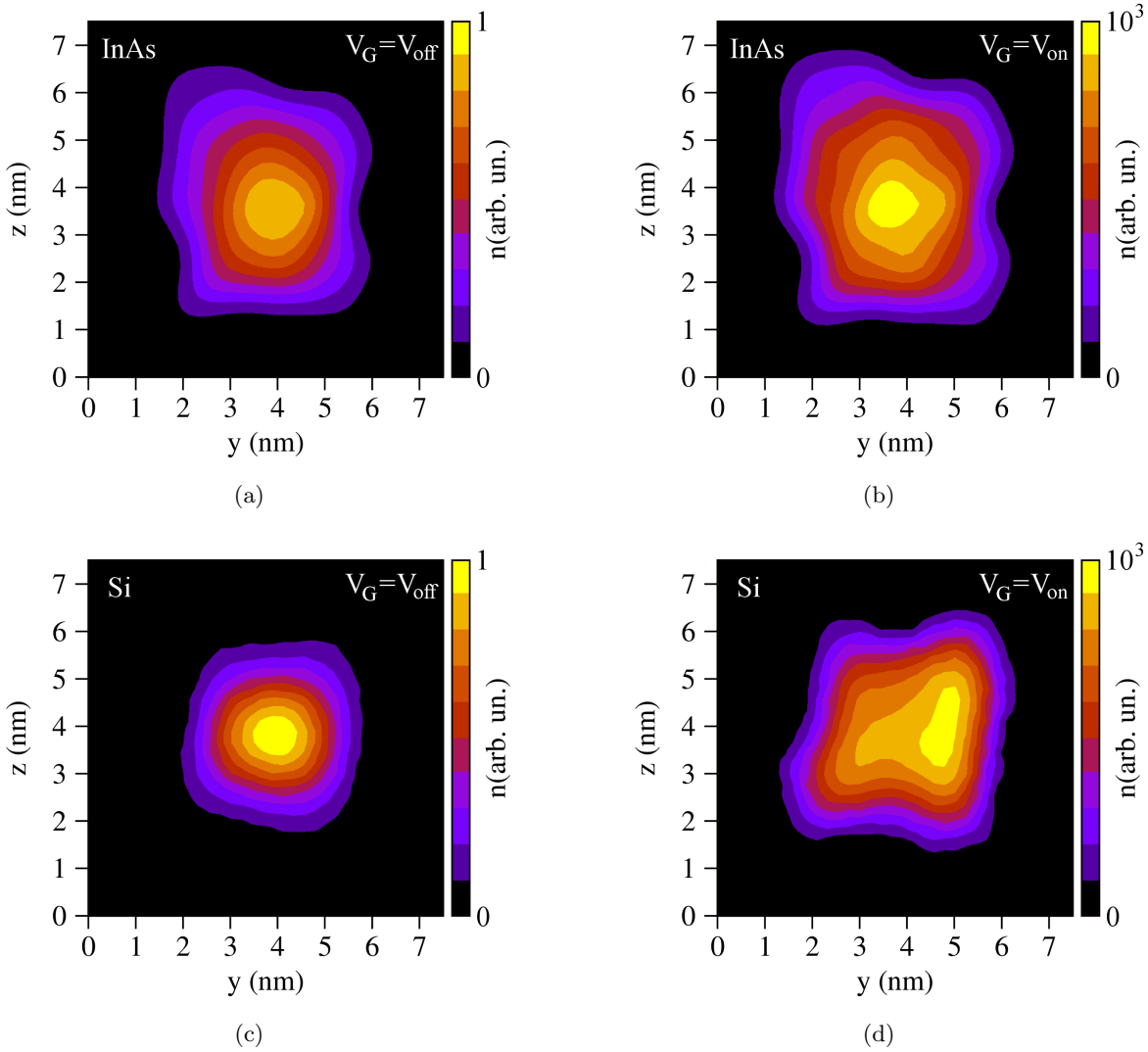


Figure 5.17: Cross-section of the charge density in a rough InAs NWFET (top) and in a rough sSi NWFET (bottom), at the off state (left) and at the on state (right). We have selected specific examples that are representative of the overall behavior.

is the main reason that explains why the InAs devices hardly outperform the sSi ones, despite benefiting from a lower source doping and a longer natural gate length. The sSi devices also exhibit signs of non-ballisticity at long gate lengths, whereas the InAs NWFETs do not present such a degradation, due to their higher electron mobility.
Compared to the UTB InAs MOSFET from the previous chapter, the GAA architecture presented better SS and current, with a subthreshold swing close to the 60mV/dec limit for $L_G \gtrsim 12$ nm and a I_{on} greater than that of the planar device, if we assume an optimistic density of integration.
The second part of this chapter was focused on the effect of surface roughness at the semiconductor-oxide interface. We have simulated about ∼200 realizations of rough sSi and InAs devices with $\Delta_R = 0.4$ nm, with $V_{DS} = 0.7$ V and $L_G = 5$ nm and 10 nm. We have found that the I_{off} and SS variability of InAs was greater that in sSi. An interesting result is that the swing of the InAs NWFET actually improves when a SR

is applied. The effect is due to a change in the shape of the subbands, that tends to reduce the occurrence of STDT. However, it is not sufficient for the average rough InAs device to outperform a typical rough sSi device. The reduction of the value of the *on* current induced by the surface roughness is typically lower in InAs devices than in sSi ones. At $L_G = 5$ nm, when the electrical performance of both devices is degraded by STDT and poor electrostatics, the rough InAs devices even exhibits (on average) a better I_{on} than the sSi one.
At the time of writing, these results have been submitted to the Transactions on Electron Devices and are currently being reviewed to be published soon [20].

5.4.1 Perspectives

In this chapter, we have covered a wide range of gate lengths, but the effect of the lateral and vertical dimensions has not been treated. This study could indeed be extended to thinner or wider devices. The implementation of a finite elements discretization would also allow us to investigate more complex shapes, such as circular nanowires. In order to make the simulations more realistic, we could also evaluate the impact of other non-idealities, such as the presence of traps. Similarly to the investigations conducted in Chap.4, it would also be possible to simulate NWFETs based on other III-V materials, such as InGaAs (or GaSb, for p-type devices).

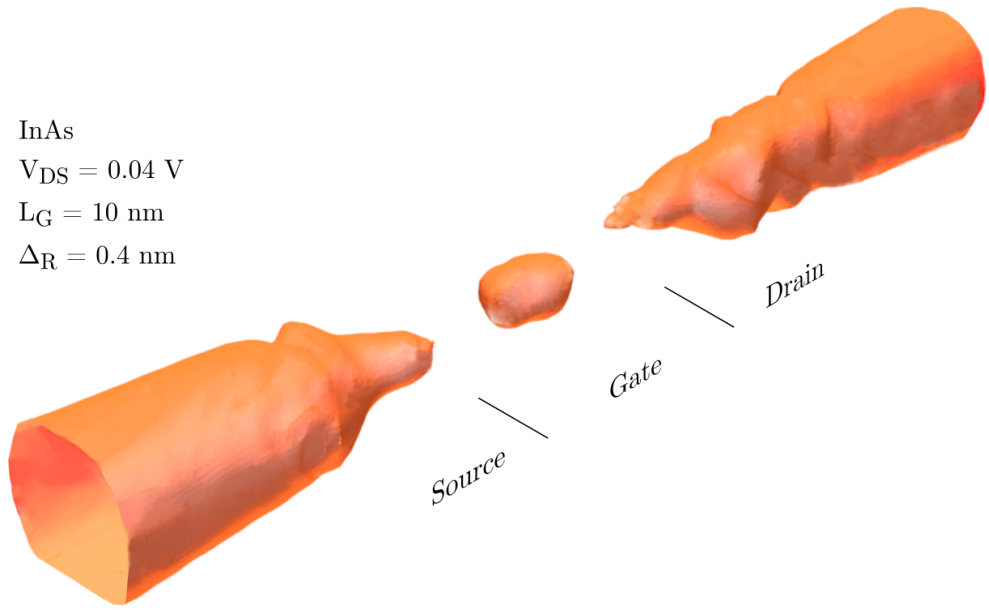


Figure 5.18: Equicharge envelope in a rough InAs NWFET with a 3×3 nm² cross section (the x, y and z axis are not represented proportionally). This specific realization of SR gives rise to a 3 nm-long dot in the channel. The total charge integrated in the channel is close to the elementary charge e.

Studies have also shown that surface roughness could actually be useful to generate quantum dots in thin-channel Si NWFETs [21]. As shown in Fig.5.18, we have been able to obtain similar results with rough InAs NWFETs, where a quantum dot was formed in the middle of the channel, as a result of the roughness. Such "artificial atoms" could prove themselves useful in possible future applications, like single electron transistors or quantum bits.

Bibliography

[1] S. Datta M.S. Lundstrom A. Rahman, J. Guo. Theory of ballistic nanotransistors. *IEEE Transactions on Electron Devices*, 50(9):1853–1864, sep 2003.

[2] T. Rollo D. Esseni, M.G. Pala. Essential physics of the OFF-state current in nanoscale MOSFETs and tunnel FETs. *IEEE Transactions on Electron Devices*, 62(9):3084–3091, sep 2015.

[3] J.J. Gu, X. Wang, H. Wu, R.G. Gordon, and P.D. Ye. Variability improvement by interface passivation and EOT scaling of InGaAs nanowire MOSFETs. *IEEE Electron Device Letters*, 34(5):608–610, may 2013.

[4] M. Radosavljevic, G. Dewey, and J. M. Fastenau et al. Non-planar, multi-gate InGaAs quantum well field effect transistors with high-k gate dielectric and ultra-scaled gate-to-drain/gate-to-source separation for low power logic applications. In *2010 International Electron Devices Meeting*. Institute of Electrical and Electronics Engineers (IEEE), dec 2010.

[5] Y.Q. Wu, R.S. Wang, T. Shen, J.J. Gu, and P. D. Ye. First experimental demonstration of 100 nm inversion-mode InGaAs FinFET through damage-free sidewall etching. In *2009 IEEE International Electron Devices Meeting (IEDM)*. Institute of Electrical and Electronics Engineers (IEEE), dec 2009.

[6] J.P. Colinge. Multiple-gate SOI MOSFETs. *Solid-State Electronics*, 48(6):897–905, jun 2004.

[7] W. Lu, P. Xie, and C. M. Lieber. Nanowire transistor performance limits and applications. *IEEE Transactions on Electron Devices*, 55(11):2859–2876, nov 2008.

[8] International technology roadmap for semiconductors. http://www.itrs.net/2013ITRS/Summary2013.htm, 2013.

[9] R.J. Harrison and P.A. Houston. LPE growth and characterization of n-type InAs. *Journal of Crystal Growth*, 78(2):257–262, nov 1986.

[10] U. E. Avci R. Kim and I. A. Young. Source/drain doping effects and performance analysis of ballistic III-v n-MOSFETs. *IEEE Journal of the Electron Devices Society*, 3(1):37–43, jan 2015.

[11] D. Lizzit, D. Esseni, P. Palestri, P. Osgnach, and L. Selmi. Performance benchmarking and effective channel length for nanoscale inas, ingaas, and ssi n-mosfets. *IEEE Transactions on Electron Devices*, 61(6):2027–2034, jun 2014.

[12] M. Lundstrom. *Fundamentals of carrier transport*. Cambridge University Press, 2 edition, 2000.

[13] M. Lenzi, P. Palestri, E. Gnani, S. Reggiani, A. Gnudi, D. Esseni, L. Selmi, and G. Baccarani. Investigation of the transport properties of silicon nanowires using deterministic and monte carlo approaches to the solution of the boltzmann transport equation. *IEEE Transactions on Electron Devices*, 55(8):2086–2096, aug 2008.

[14] J. C. Hensel, H. Hasegawa, and M. Nakayama. Cyclotron resonance in uniaxially stressed silicon. II. nature of the covalent bond. *Physical Review*, 138(1A):A225–A238, apr 1965.

[15] Y.-C. Chou, K. Hillerich, J. Tersoff, M. C. Reuter, K. A. Dick, and F. M. Ross. Atomic-scale variability and control of III-v nanowire growth kinetics. *Science*, 343(6168):281–284, jan 2014.

[16] A. Martinez, N. Seoane, Andrew R. Brown, John R. 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*, 57(7):1626–1635, jul 2010.

[17] F. Conzatti, M.G. Pala, and D. Esseni. Surface-roughness-induced variability in nanowire InAs tunnel FETs. *IEEE Electron Device Letters*, 33(6):806–808, jun 2012.

[18] C. Buran, M.G. Pala, M. Bescond, M. Dubois, and M. Mouis. Three-dimensional real-space simulation of surface roughness in silicon nanowire FETs. *IEEE Transactions on Electron Devices*, 56(10):2186–2192, oct 2009.

[19] A. Cresti, M.G. Pala, S. Poli, M.Mouis, and G. Ghibaudo. A comparative study of surface-roughness-induced variability in silicon nanowire and double-gate FETs. *IEEE Transactions on Electron Devices*, 58(8):2274–2281, aug 2011.

[20] C. Grillet, D. Logoteta, A. Cresti, and M.G. Pala. Assessment of the electrical performance of short channel InAs and strained Si nanowire FETs. *IEEE Transactions on Electron Devices*, 2017 (submitted).

[21] S. Barraud A.Corna X. Jehl M. Sanquer J. Li A. Abisset I. Duchemin Y.M. Niquet R.Lavieville, F. Triozon. Quantum dot made in metal oxide silicon-nanowire field effect transistor working at room temperature. *Nano Letters*, 15(5):2958–2964, may 2015.

CHAPTER 6

Vertical Tunnel-FET

In which we take advantage of the quantum nature of the transport to design a tunneling transistor, made of stacked III-V semiconductors

Chapters 4 and 5 dealt with the optimization the current field-effect transistor architecture, by studying the impact of the device dimensions, the channel material and the geometry of the gate. We have shown that some quantum effects, such as BTBT or STDT, are detrimental for the device operation. In this present chapter, we focus on a novel kind of architecture that instead does take advantage of the quantum tunneling effect.
A tunnel-FET (TFET) presents the same overall structure as a standard MOSFET: it has a source, a channel and a drain, and the current is controlled by the gate. The major difference comes from the nature of the channel barrier. The switching no longer consists in a modulation of the conduction band, but rather in a squeezing of the band-to-band tunneling window between the source VB and the drain CB. This is very convenient in order to achieve a very low SS, which is no more limited by thermionic emission as in standard FETs. The working mechanism is detailed in Fig.6.1, where a typical configuration with p-doped source, n-doped drain and intrinsic channel is considered. Thanks to the p-type donors, the Fermi level in the source is located in (or close to) the VB, while the CB is not populated. On the contrary, in the n-doped drain the Fermi level is located in (or close to) the conduction band. The gate acts on the intrinsic channel by increasing or reducing the width of the BTBT barrier with the source. When the gate voltage is low, the large width of the barrier prevent electrons from tunneling, thus suppressing the *off* current. When the gate voltage is high enough, the electrons start to tunnel towards the lower energy n-doped drain. Remarkably, the current is thus entirely due to BTBT, since no thermionic emission is possible in such a setup, where the gap above the source valence band cut the density of states in the energy region of the Fermi distribution tail. For that reason, the TFETs are expected to feature subthermionic slopes, which means that they can break the 60 mV/dec SS limit.
Due to its indirect gap, silicon is not the ideal candidate for the realization of such a device. However, III-V semiconductors are especially suitable for the design of TFETs, since they offer a wide variety of direct gap compounds with different band structures, which allows one to obtain the desired band profile by combining them properly.

6.1 Description of the device

Various III-V TFET architectures have been considered in previous studies [2–10]. The most intuitive way to design a tunneling transistor is to mimic the shape of a MOSFET [2–4] and replace the source, channel and drain materials in order to obtain a band profile similar to that depicted in Fig.6.1. As in Chap.5, it is also possible to resort

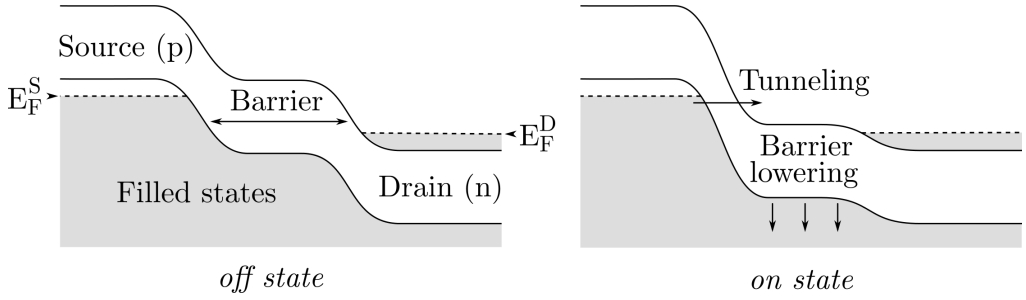


Figure 6.1: General working principle of a TFET. At the off state, a wide barrier prevents the electrons from tunneling from the source VB to the drain CB. When the barrier is modulated by the gate, it becomes thin enough to allow BTBT from the VB to the CB. Note that this picture, which depicts filled states below the Fermi energy E_F, is only true at $T = 0$ K. At room temperature, the tails of the Fermi functions are essential in the device operation.

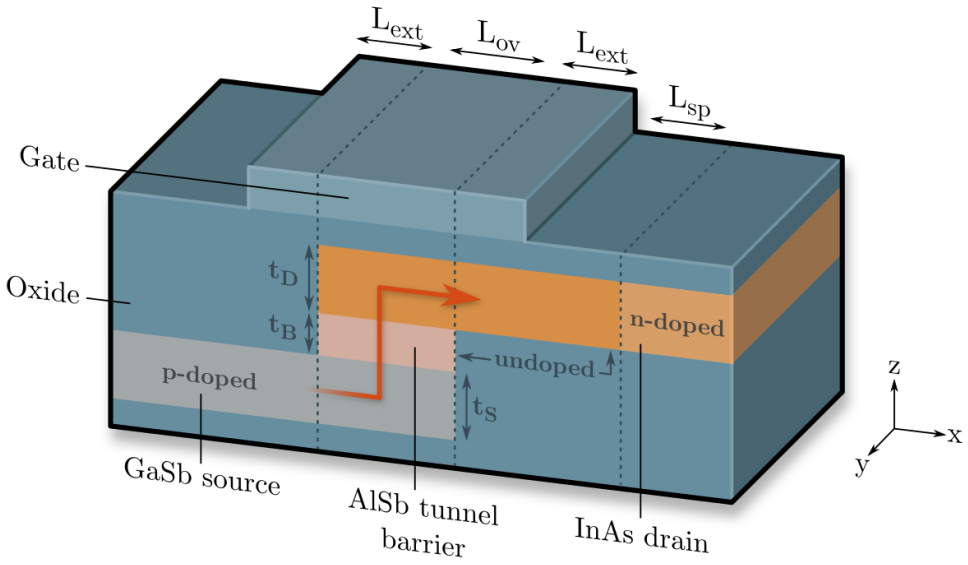


Figure 6.2: Sketch of the vertical GaSb/AlSb/InAs Tunnel-FET. The "default" device uses $t_S = 4.8$ nm, $t_B = 1.2$ nm and $t_D = 4.8$ nm. L_{ov}, L_{ext} and L_{sp} are all 20 nm long. The source layer is p-doped with an acceptor concentration $N_S = 5 \times 10^{19} cm^{-3}$, while the drain layer is n-doped on a 20 nm length, with a donor concentration $N_D = 2 \times 10^{19} cm^{-3}$. Most of the parameters will be modified in search of the best design options, as shown in Tab.6.1

L_{ov} (nm)	L_{ext} (nm)	L_{sp} (nm)	t_B (nm)	t_S (nm)	t_D (nm)	EOT (nm)	N_D (10^{19}cm^{-3})	Transport direction
10-30	0-30	0-20	0-10.4	3-6.1	3-4.8	0.6	0.5-2	[100]

Table 6.1: Geometry of the heterojunction TFET. All the values written as a range will we subjected to a specific investigation (where at least the minimum and the maximum of each given range will be simulated). The thicknesses t_B, t_S and t_D correspond respectively to multiples of the AlSb, GaSb and InAs lattice parameters.

to multigate or gate-all-around architectures [5–8] to improve the electrostatic control. However, these longitudinal TFET devices are sensitive to trap-assisted tunneling,

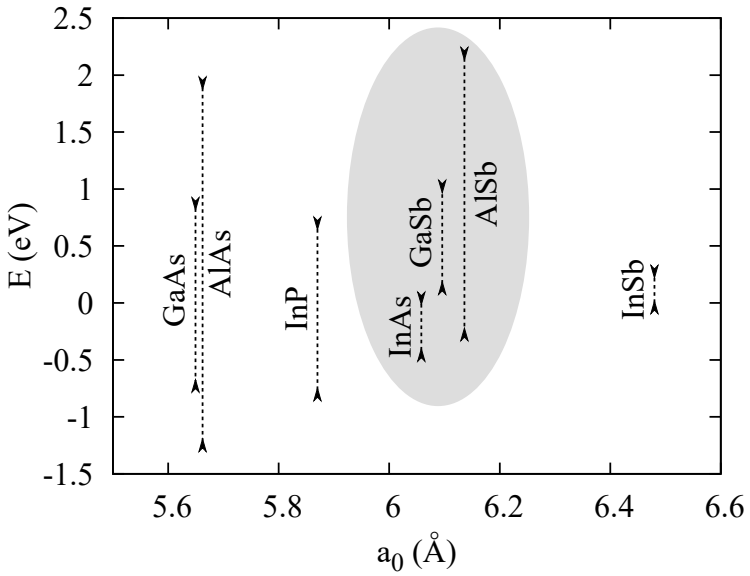


Figure 6.3: Energy of the CB (down arrows) and the VB (up arrows) at the Γ *point for different III-V compounds, arranged according to their lattice parameter. The vertical lines correspond the energy gaps. The offset between the conduction band of different materials given by their electron affinity [1], where we have taken the CB of InAs as a reference. The semiconductors highlighted in the figure present several advantages for the design of a vertical TFET: InAs and GaSb are broken-gap, AlSb has a large gap, and these three materials present a similar lattice parameter.*

which can degrade the *off* state [11]. The alternative solution investigated in this work is to simulate a 2D vertical heterojunction TFET [9, 10], that consists of stacked semiconductor layers. In the present case, three III-V compounds are considered: GaSb, AlSb and InAs.

As illustrated in Fig.6.2, the source is the bottom p-doped layer of GaSb, while the drain is the top layer of InAs with a n-doped access region. A thin interlayer of AlSb is sandwiched between top and bottom layers. This material is ideal as a tunnel barrier because of its large gap. BTBT occurs through the overlap (OL) region controlled by the gate, which, at high gate voltages, induces the crossing of the conduction band in the top layer and of the valence band in the bottom layer. At low gate voltages, on the contrary, the conduction band minimum of the top layer is higher than the valence band maximum of the bottom layer inside the overlap region and the tunneling current is suppressed. Fig.6.3 shows that GaSb, AlSb and InAs have similar lattice constants [1] and can be pseudomorphically grown in order to achieve defect-free interfaces. Indeed, the layers are deposited in the (x,y) plane, which is perpendicular to the growth direction and should lead to high quality interfaces. For that reason, we also neglect the influence of strain in this study. Another advantage of such a configuration, also illustrated in Fig.6.3, is that the GaSb/InAs heterojunction is broken-gap (the GaSb VB is energetically higher the the InAs CB) and can therefore provide large *on*-current values due to the short tunneling path occurring between source and channel regions [12]. Finally, the fact that the gate is perpendicular to the direction of the current in the channel (see the red arrow in Fig.6.2) should offer an improved control of the tunneling barrier, since the gate potential acts nearly uniformly

	InAs	AlSb	GaSb	
E_G	0.417	2.386	0.812	(eV)
E_P	18.0	13.0	22.0	(eV)
γ_C	2.25	2.10	2.91	
γ_1^L	20.00	5.18	13.4	
γ_2^L	8.50	1.97	4.7	
γ_3^L	9.20	1.97	6.0	
m^*	0.026	0.14	0.039	(m_0)
Δ_{SO}	0.39	0.68	0.76	(eV)
ϵ	15.5	12.04	15.7	(ϵ_0)

Table 6.2: $\boldsymbol{k}\cdot\boldsymbol{p}$ parameters used in the simulation for the three semiconductors layers, extracted from [1].

on the entire length of the AlSb layer. As V_{GS} gets larger, the tunneling probability increases in the whole overlap length at once. We are thus dealing with a so called *line tunneling* (as opposed to the *point tunneling*, found in longitudinal TFETs). The aim of this chapter is to assess the effect of the various geometrical parameters of this III-V vertical TFET (see Tab.6.1), among which we find the AlSb barrier thickness t_{B} the overlap length L_{ov}, the gate-extension length L_{ext}, or the spacer length L_{sp}. We will also investigate the role of the drain doping N_D and the impact of phonons on the transport through the barrier.

6.2 Simulations

The layers are grown along the z direction, while the transport takes place along the x direction (see Fig.6.1). Similarly to the device of Chap.4, the system is periodic in the y direction and we resort to 2D simulations. Even though 3D simulations can be useful to account for possible defects, surface roughness or lateral confinement, we will not consider these aspects here. Although the electron flow is vertical in the OL region, the problem is solved by slicing the TFET in the x direction. This approach is still valid, since the i-th slice is described by a submatrix $\hat{\mathcal{H}}_{i,i}$ of $\hat{\mathcal{H}}_{\mathrm{tot}}$ that accounts for heterostructures in the vertical direction (see the discretization method, in Sec.2.6).

Phonons			Modes		Temp.
D_{ac}(eV)	$D_{\mathrm{opt}}(10^8$ eV/cm)	$\hbar\omega_{op}$(meV)	CB	VB	(K)
5.8	2	30	4	16	300

Table 6.3: Simulation parameters used for the TFET device. The phonon deformation potentials are taken from [13]

The band profile evolution along the vertical direction is shown in Fig.6.4. In the source, the Fermi level E_F^S is energetically lower than the VB. It ensures that the electrons do not populate the CB, where they could easily pass through the interlayer barrier. As the device is switched on, the band profile in the drain is strongly shifted towards lower energies. The InAs CB in now below E_F^S and a tunneling current can

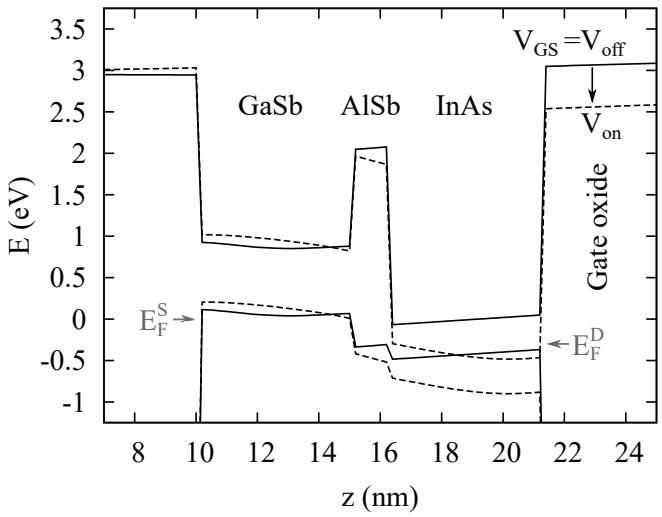


Figure 6.4: Off-state (plain line) and on-state (dashed line) band profiles in the middle of the OL region (vertical cut along the z direction), for the device with t_B= 1.2 nm.

flow through the AlSb layer.
Unlike the previous MOSFET and NWFET devices, such a nanotransistor is especially suitable for low power (LP) applications [14]. Indeed, the presence of a tunneling barrier does not allow for the high current densities required in the HP specifications. However, as stated before, TFETs can benefit from an improved SS and very small *off* current. For these reasons, the transistor will be operated with a bias voltage V_{DS} = 0.3 V and a *off* current $I_{off} = 10^{-5}$ A/m, in accordance with the ITRS requirements.

6.3 Drain spacer

As explained in Chap.4, a spacer region can be useful to reduce the coupling between the contacts and the gate, thus enhancing the electrostatic integrity. In our case, such a spacer is only necessary in the drain, since there is no need for the gate to affect the GaSb layer. We first simulate a TFET with t_B = 1.2 nm, L_{ext} = 20 nm and L_{ov} = 20 nm (the rest of the default parameters are listed in the caption of Fig.6.2). Its transfer characteristics with and without spacer are plotted in Fig.6.5. The figure shows that the addition of a 20 nm spacer in the drain region allows the device to reach a smaller *off*-state current, which subsequently improves its SS. Thanks to this spacer, the slope is steeper than the 60 mV/dec mark up to $V_{GS} \simeq 0.22$ V. The *on* current only reaches 18 A/m, but this still represents an I_{on}/I_{off} ratio greater than 10^6. This confirms that the TFET device is well suited for LP applications, where it can outperform the MOSFET architecture in terms of SS. The TFET with a L_{sp} = 20 nm and t_B = 1.2 nm will be our reference device for the rest of this chapter's investigations.

6.4 Gate geometry

There are two ways to extend the gate of the device. First, one can increase the overlap length L_{ov}, that is defined as the region where the gate covers all three semiconductor

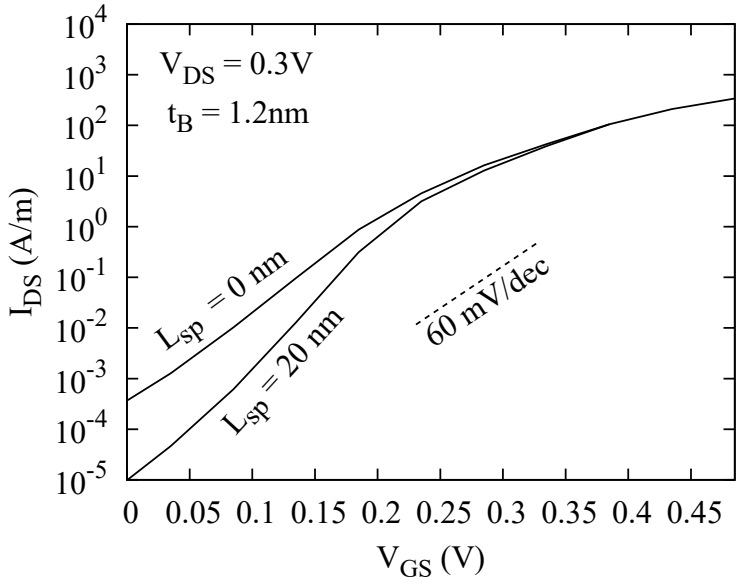


Figure 6.5: Transfer characteristic of the vertical TFET, with and without spacer. The spacer improves the swing and allows the device to perform beyond the 60 mV/dec mark. Its SS (extracted between $V_{GS} = V_{off}$ and $V_{GS} = V_{th} = 1$ A/m) is equal to 42 mV/dec.

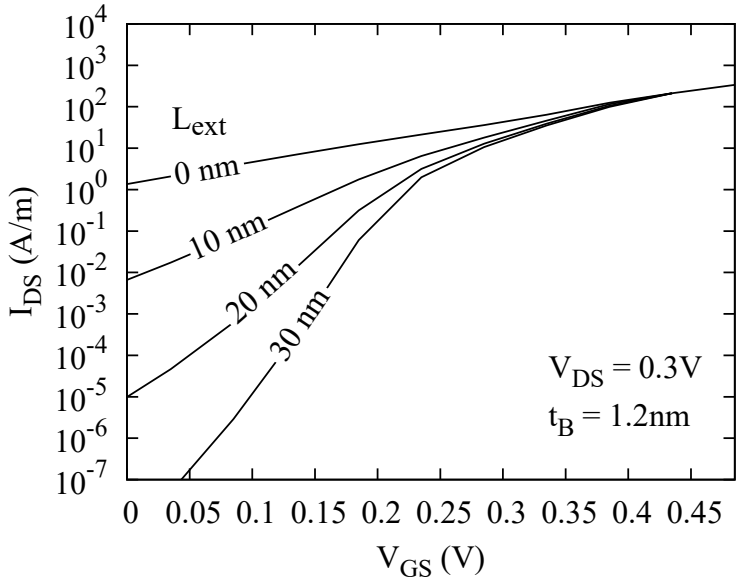


Figure 6.6: Transfer characteristic of the device for different values of extension lengths. The SS is improved as L_{ext} is increased, due to a reduction of the off-state tunneling.

layers at once. Second, one can vary the gate extension length L_{ext}, which means that the gate electrode is stretched on both sides of the overlap region. Note that changing the extension length does not modify the size of the OL region.
Fig.6.6 shows the I(V) characteristic of the TFET for L_{ext} ranging from 0 to 30 nm. It appears that longer extension lengths greatly improve the *off* state and the SS. Indeed increasing L_{ext} also extends the length of the BTBT path, as can be seen in Fig.6.7.

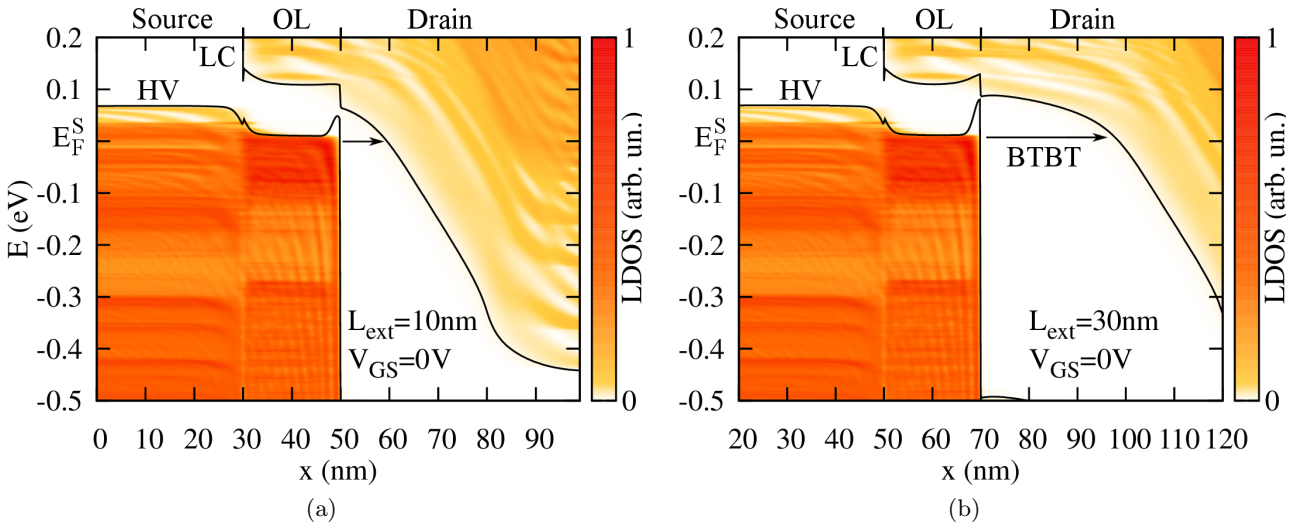


Figure 6.7: (a) LDOS integrated over the z direction in the device with L_{ext} = 10 nm. The white area is the gap, while the available states are represented by the color map. (b) Same plot with L_{ext} = 30 nm. The BTBT distance seen by the carriers has increased and the off tunneling current is consequently reduced.

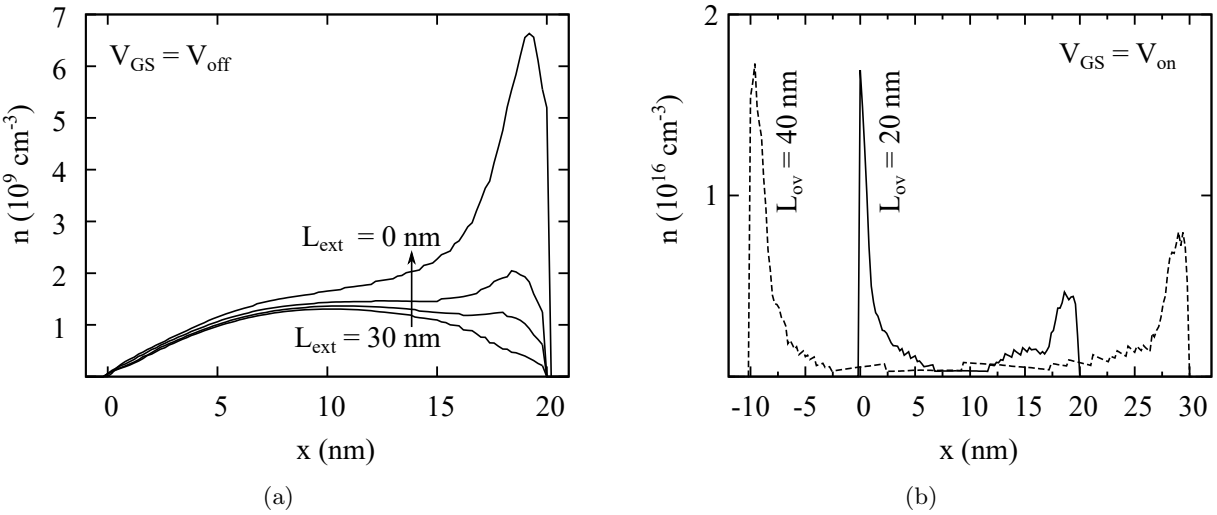


Figure 6.8: (a) Off-state electron density along the x direction, inside the AlSb interlayer, for different values of L_{ext}. Reducing the extension lengths increases the carrier density along the right edge of the AlSb interlayer. The coordinate x = 0 nm denotes the beginning of the overlap region, that ends at x = 20 nm. (b) On-state electron density inside the AlSb layer for different overlap lengths. At the on state, the charge is accumulated along the edges of the overlap region. This behavior is almost independent of the value of L_{ov}.

The tunneling takes place between the HV and the LC subbands, at energies close to the source Fermi level. Indeed, BTBT occurs where the tunneling path is the shortest and decreases exponentially when this paths gets longer. As a result, the TFETs with $L_{ext} \geq 20$ nm reach easily the desired I_{off} value of 10^{-5} A/m and present a better SS. As shown in Fig.6.8-a, the additional *off*-tunneling contribution for small values

of L_{ext} appears along the right edge of the OL region. This SS improvement also leads to a shift of V_{off} which, in turn, increases V_{on} and I_{on}. For example, when L_{ext} goes from 20 to 30 nm, the SS goes from 42 to 25 mV/dec and I_{on} also gets dramatically boosted from 18 to 120 A/m. The L_{ext}=30 nm device thus presents an I_{on}/I_{off} ratio of more than 10^7.
Contrary to L_{ext}, which is extremely beneficial for both the SS and the I_{on}, the overlap length L_{ov} very weakly affects the behavior TFET (if not at all). We could have expected that a larger L_{ov} would enhance the value of I_{on}, but this is not actually the case. Between L_{ov}=10 nm and L_{ov}=40 nm, the I_{on} and the SS stay nearly identical, at 18 A/m and 42 mV/dec, respectively. This is due to the fact that most of the electrons are packed along the edges of the AlSb interlayer at the *on* state, as shown in Fig.6.8-b. Unlike what has been postulated in this chapter's introduction, the TFET actually appears to exhibit point tunneling (instead of line tunneling). Since both the *off* and *on* current are mostly governed by edge effects, L_{ov} does not impact the overall transfer characteristic of the TFET. For that reason, L_{ov} will be limited to 20 nm in the rest of this work to avoid any further increase in the size of the device.

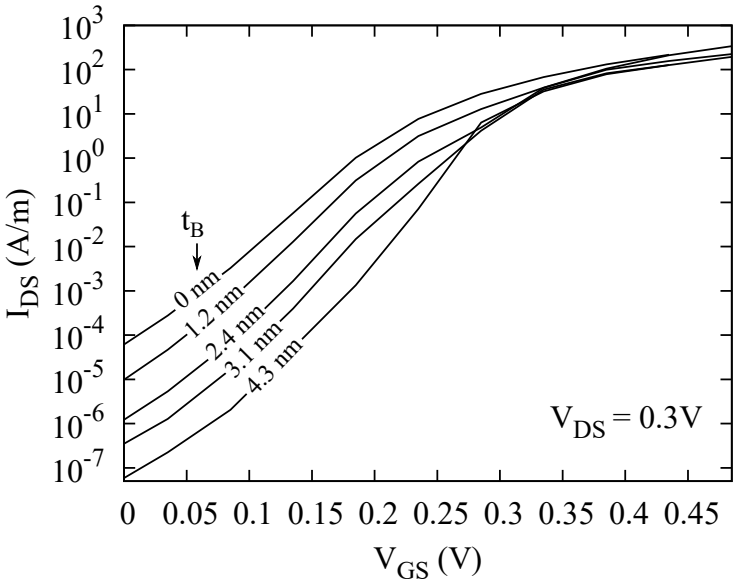
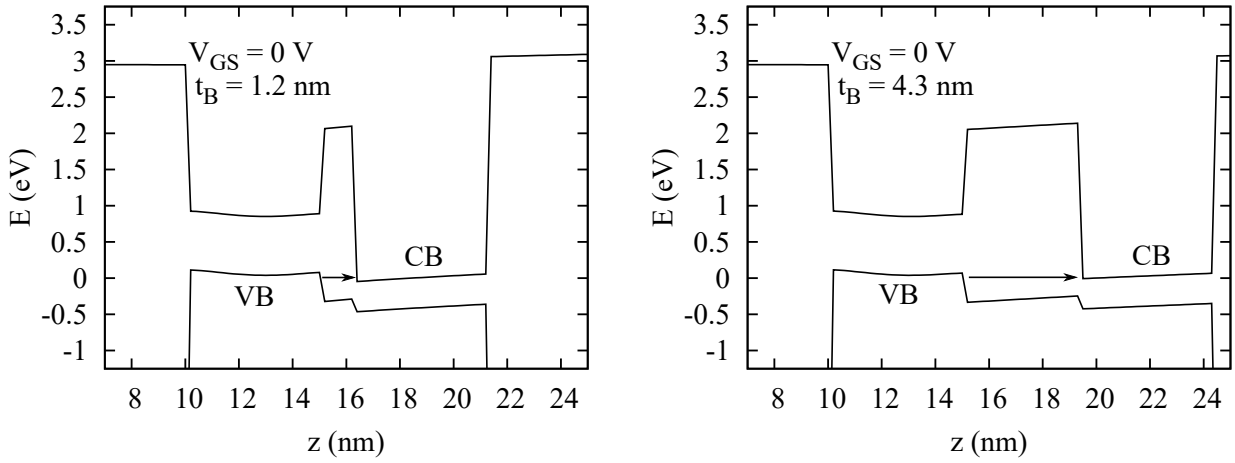


Figure 6.9: Transfer characteristic of the TFET device for different barrier thicknesses. Increasing t_B improves the SS, without strongly affecting the on state. The curves for t_B = 4.9 nm and 6.1 nm, are not shown here, but nearly overlap with that of t_B = 4.3 nm. Above t_B = 7.4 nm, the SS starts to visibly degrade again.

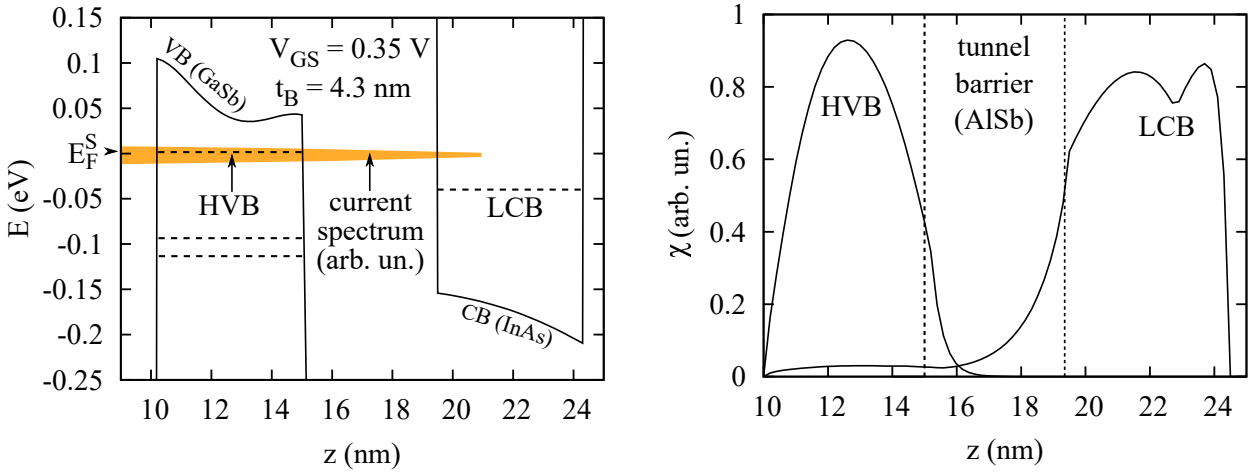
6.5 Tunnel barrier scaling

We now study the effect of the AlSb barrier thickness t_B, by simulating TFET devices with t_B ranging from 0 to 10.4 nm. The chosen values for the thickness of this layer correspond to multiples of the AlSb lattice constant, that is close to 0.61 nm [1]. Fig.6.9 demonstrates that the devices with larger t_B values exhibit a lower *off*-state current and a better SS. We plot the *off*-state band profile of a thin (t_B= 1.2 nm) and a thick (t_B= 4.9 nm) AlSb layer in Fig.6.10. As it could be expected, increasing the thickness



(a) (b)

Figure 6.10: Off-state band profiles along the z direction in the OL region, for t_B = 1.2 nm (a) and t_B = 4.3 nm (b). Increasing the barrier thickness also increases the length of the BTBT path (depicted as arrows) and subsequently reduces the off current.



(a) (b)

Figure 6.11: (a) On-state band profile performed on the edge of the OL region and zoomed-in at the extremities of the bands. The dashed lines represent the energies of the modes. The HVB mode in the source is located above the LCB mode in the drain. (b) Vertical profile of the transverse modes corresponding to the HVB and LCB modes. A strong mixing can be observed in the AlSb layer.

of this layer directly affects the width of the tunneling barrier. Since the BTBT current is exponentially suppressed with when the barrier is enlarged, the devices with a larger t_B present en smaller *off* current

More unexpectedly, increasing t_B appears to be beneficial in terms the *on* current, at least for $t_B \leq 4.9$ nm. To gain more insight on this behavior, we examine the individual transverse modes in the CB and VB. As detailed in Tab.6.3, the simulation

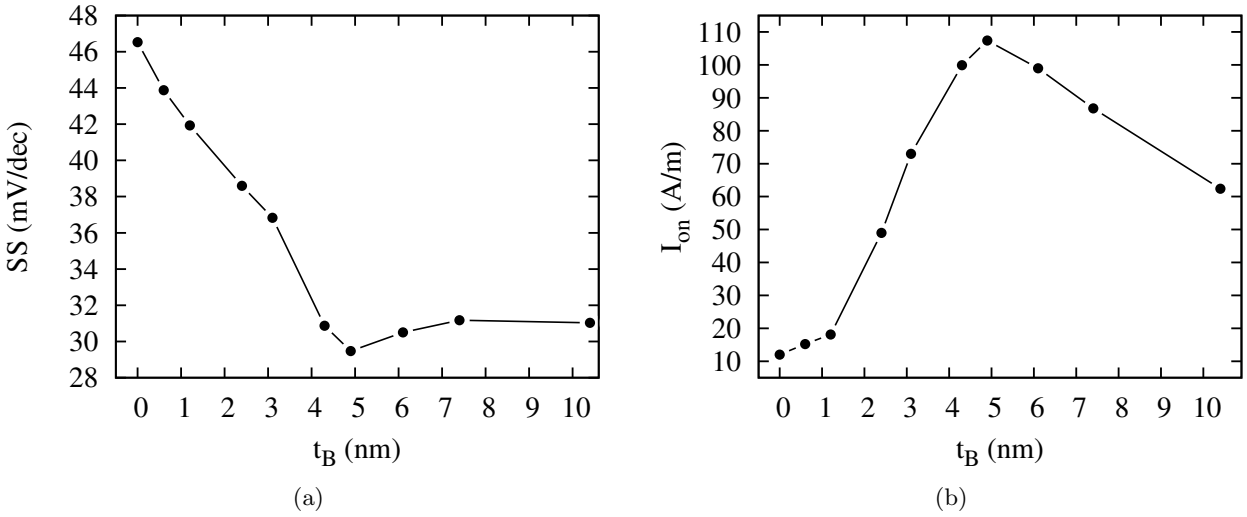


Figure 6.12: SS and I_{on} evolution as a function of t_B. Thicker barriers reduce the off-state current, while the on-state current remains high due to resonant tunneling. This leads to a improved SS and a subsequent increased I_{on} (due to the V_{th} shift). The on current starts to be degraded for $t_B \gtrsim 4.9$ nm

uses 16 valence modes and 4 conduction modes. The most significant modes for the transport are the (energetically) highest VB (HVB) mode and the lowest CB (LCB) mode. Fig.6.11 shows that, in the OL region, the HVB mode is located above the LCB mode, which is a consequence of the crossing of the CB in InAs and VB in GaSb. By analyzing the vertical profile of the transverse modes in Fig.6.11-b, it can be seen that that these modes are strongly mixed among them and they easily penetrate into the AlSb barrier. As a consequence, the transmission probability through the AlSb layer remains relevant when the thickness of the barrier is increased and does not prevent the achievement of high on currents. At the same time, in the device with thicker barrier, the gate voltage leaves almost unaltered the VB in the GaSb, which implies a more efficient crossing of the HVB and LCB modes and an improved SS with respect to thinner barrier devices. This means that it exists a tradeoff between the decrease of the transmission probability through the barrier and the SS enhancement given by increasing t_B. The best tradeoff appears to be a barrier thickness close to 5 nm, which is large enough to allow an efficient crossing/uncrossing the of VB and CB, but it still leads to a a strong transverse mode mixing able to provide a high tunneling probability through the AlSb barrier. For that reason, the performance enhancement stops when $t_B \gtrsim 4.9$ nm, as shown un Fig.6.12. Around this value, the TFET exhibits very good performances, with I_{on} >100 A/m and SS = 30 mV/dec.

6.6 Drain doping

A closer look at the subbands led us to consider a third approach to improve the slope of the device. The addition of a spacer or extension lengths create a stretch of the LC subband in the drain, that subsequently reduces the *off* tunneling. Another solution to increase the length of the BTBT path is to bend the LC subband towards higher energies. This can be achieved with a reduction of the drain doping N_D, as

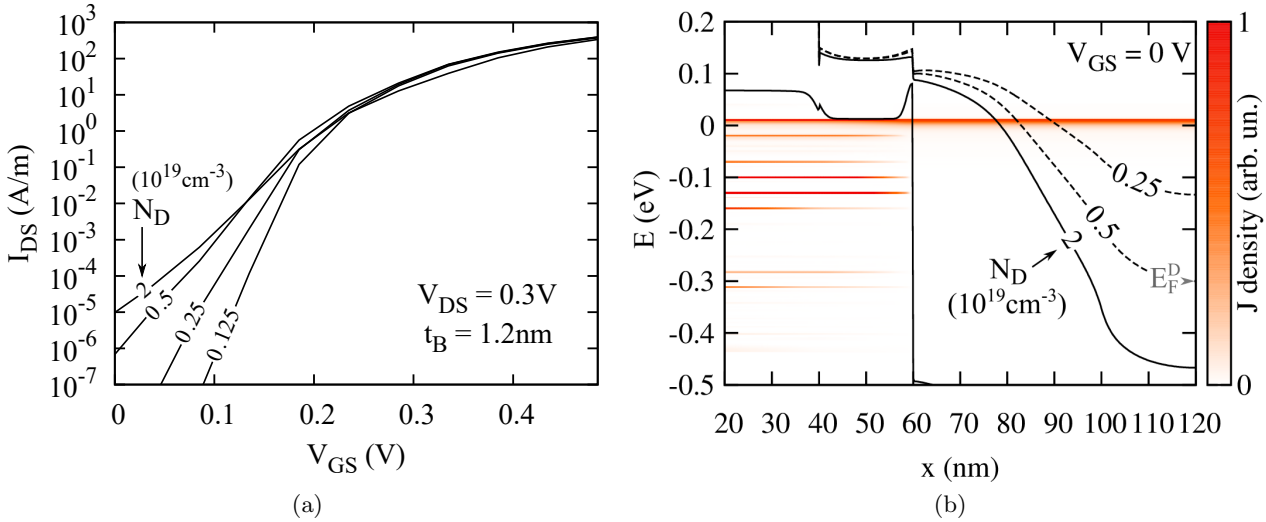


Figure 6.13: (a) Transfert characteristic of TFETs with N_D = 2 to 0.125×10^{19} cm^3. (b) Corresponding subband profiles along the x direction. The spectral current if shown for the device with N_D = 2×10^{19} cm^3. Decreasing the source doping brings the source LC towards higher energies, which increases the width of the BTBT barrier seen by the electrons. The off current is reduced and the performance of the TFET is improved.

illustrated in Fig.6.13. Low values of N_D lift the LC subband in the drain contact region, which also changes the height of the subband in the BTBT region. In practice, this optimization will be limited by the position of the drain Fermi level E_F^D. Indeed, the LC subband has to be above E_F^D at the contact, in order to avoid forming a Schottky barrier with the metallic contact. For that reason, the optimal choice for N_D is 0.5×10^{19}cm^{-3} (see Fig.6.13). This doping improves the SS from 42 to 32 mV/dec and the I_{on} from 18 to 71 A/m.

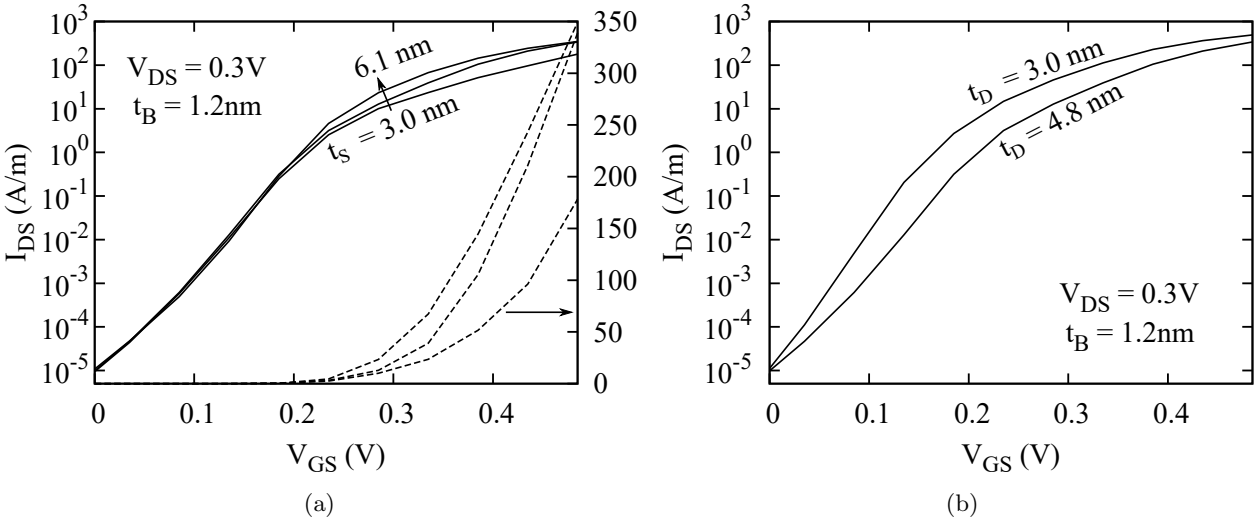
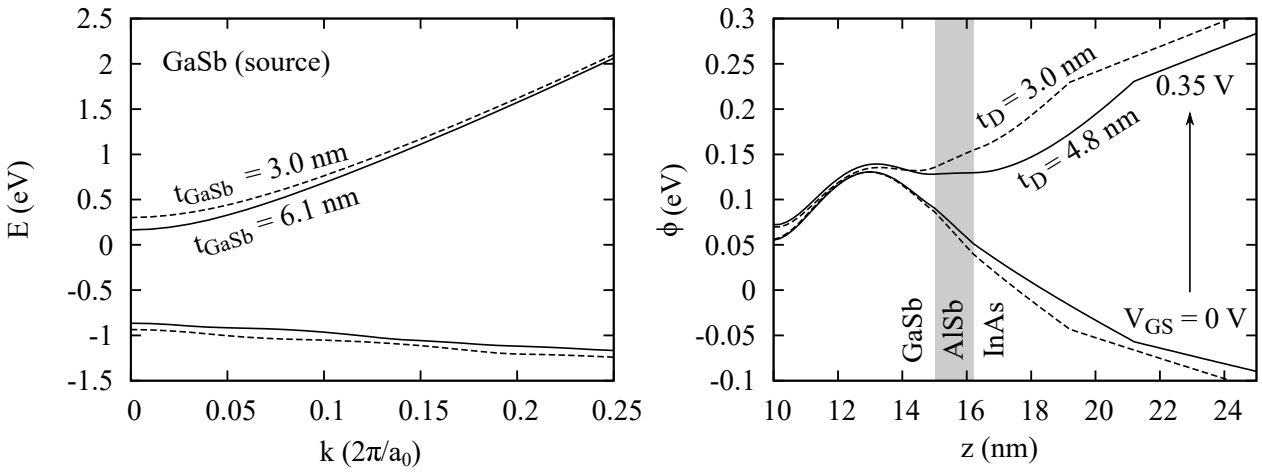


Figure 6.14: Transfer characteristics of a TFET with different source layer thicknesses (a) and drain layer thicknesses (b). The curve for t_D = 3 nm has been shifted from 0.1 V to the left



(a) (b)

Figure 6.15: (a) Effect of the confinement on the CB and the highest VB of an GaSb layer, plotted along the [100] direction. (b) Evolution of the electrical potential with V_{GS}, along the vertical direction, for different drain thicknesses t_D.

6.7 Source and drain thicknesses

We now consider the effect of the GaSb and InAs layers thicknesses, denoted respectively t_S (source) and t_D (drain). Fig.6.14-a shows that a thicker source layer can be beneficial for the *on* state of the vertical TFET device. Indeed, by increasing t_S, the quantum confinement in the GaSb layer is reduced, which, in turn, increases the density of states in the layer. This allows the device to conduct more current at the *on* state. Another contribution of t_S on I_{on} can be appreciated in Fig.6.15-a, which corroborates the fact that the bandgap E_G is smaller in the thick-source device. This amplifies the broken-gap nature of the bands and consequently enlarges the tunneling window. This feature is again beneficial for I_{on}, while having little, or no effect at all, on the *off* current. Compared to the normal device (t_S = 4.8 nm) whose I_{on} is 18 A/m, the TFET with t_S = 6.1 nm exhibits I_{on}= 31 A/m and no change in SS.
Conversely, it can be observed in Fig.6.14-b that a decrease of the InAs thickness t_D has a positive impact on both the swing and the *on* current. This effect can be directly connected to the improved electrostatic control present in the device with a thin drain, as illustrated in Fig.6.15-b. The potential profile, extracted along the vertical z direction, is substantially more affected by a given ΔV_{GS} in the device with small t_D. Indeed, the smaller thickness of the drain layer allows the gate to wield an enhanced electrostatic control on this part of the device. As a consequence, the SS goes from 42 to 33 mV/dec and the I_{on} jumps from 18 to 58 A/m.

6.8 Ideal configuration

In the previous sections, we have found efficient ways to enhance the SS and the I_{on} of the vertical TFET. We now propose to combine some of these solutions in order to design a more optimized device. For this device, we have selected L_{ov} =

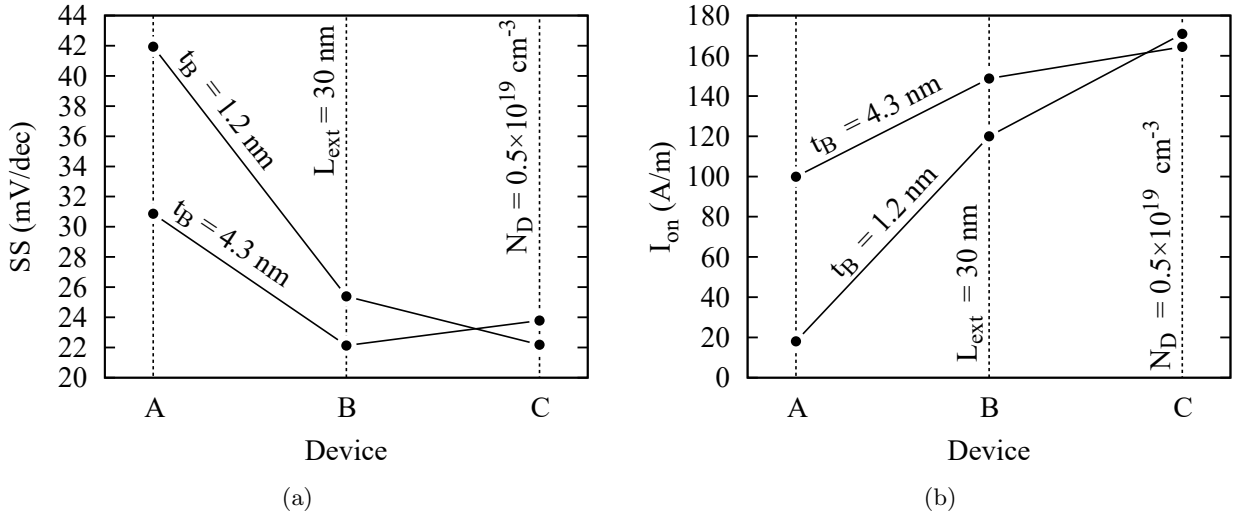


Figure 6.16: SS and I_{on} evolution as successive optimizations are added to the t_B = 1.2 and 4.3 nm devices. Case A corresponds to L_{ext} = 20 nm and $N_D = 2\times10^{19} cm^{-3}$. Case B corresponds to case A with an increased L_{ext} and case C corresponds to case B with a reduced N_D.

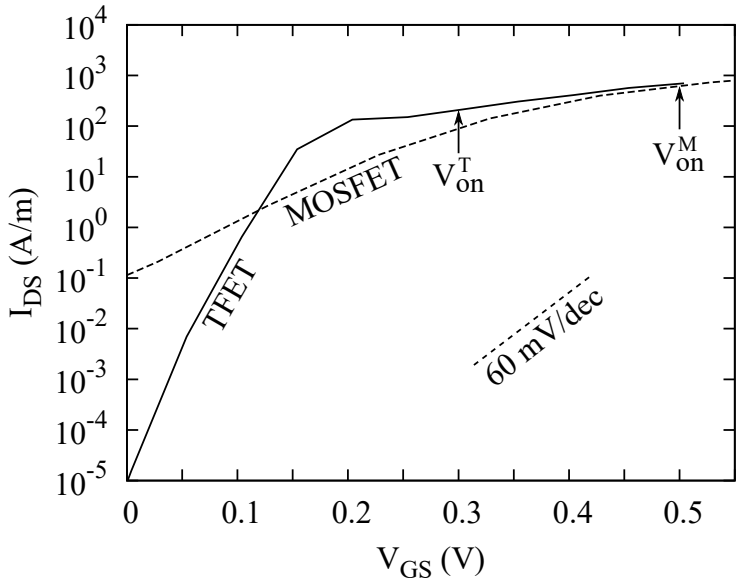


Figure 6.17: "Optimal" TFET device compared with the InGaAs MOSFET (V_{DS} = 0.5 V) of Chap.4. The curves have been shifted to obtain V_{off} = 0 V in both devices. Even though the current appears to be higher in the TFET, the MOSFET actually has a better I_{on} since its higher supply voltage allows it to reach a larger V_{on} (denoted V_{on}^{M} here).

20 nm , L_{sp} =20 nm and L_{ext} = 30 nm. Even though further increasing L_{sp} and L_{ext} could lead to even better results, we still try to design a device compatible with large scale integration, whose dimension does not exceed 100 nm. As explained in Sec.6.6, the ultimate drain doping N_D is limited to 0.5×10^{19} cm^{-3}. Fig.6.16 illustrates the cumulative effect of L_{ext} and N_D on the device performance, for two different barrier thicknesses. Even though the TFET with t_B = 4.3 nm initially presents better SS

and I_{on} than the device with $t_B = 1.2$ nm (as explained in Sec.6.5), increasing L_{ext} from 20 to 30 nm is more beneficial for the thinnest device. When N_D is reduced from 2×10^{19} to 0.5×10^{19} cm^{-3}, the performances of the thick and thin devices meet. We even observe a degradation of the $t_B = 4.3$ nm device in this third case, suggesting that we may be reaching a limitation case, that can be difficult to exceed. Furthermore, the weakly doped device is rather idealistic and could present contact resistances in practice. When looking for the best option, a pragmatic choice would be the thick device with a normal doping and an increased extension length (case B at 4.3 nm, in the figure) as it presents very good SS and I_{on}, without the risk of showing high contact access resistances.
The I-V curve of this optimal device is compared with that of the MOSFET from Chap.2 in Fig.6.17. The swing difference between the two kind of devices appears very clearly on this graph. We also notice that the *on*-state currents is in fact higher in the TFET than in the MOSFET. However, since the TFET is operated with $V_{DD} \tilde{=} 0.3$V, its I_{on} is actually still lower than in the MOSFET.

6.9 Phonons

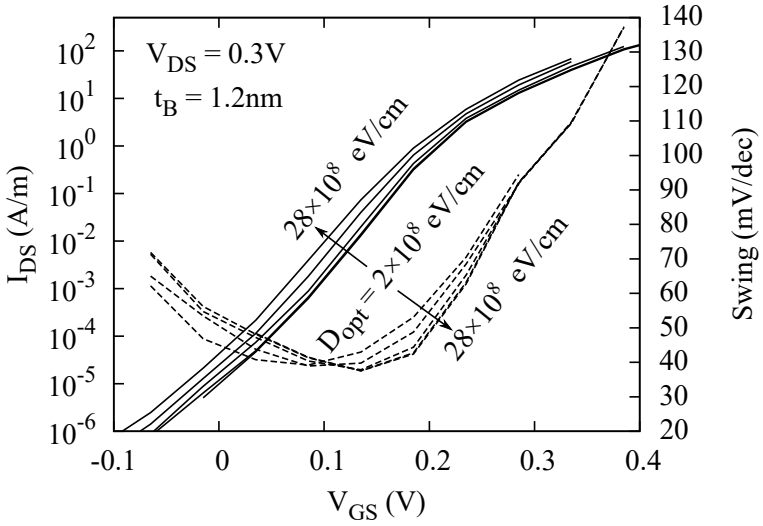


Figure 6.18: Evolution of the TFET's transfer characteristic (plain lines) and the local swing (dashed lines) with the optical phonons deformation potential.

In the *off* state and in the absence of inelastic phonon scattering, the current is determined by the electrons tunneling from the bottom layer to the top layer at energies close to the Fermi level at the source. In the presence of optical phonon scattering, however, electron tunneling from the top to the bottom layer can occur also at energies higher than the top of the valence band, due to optical phonon absorption, and such additional tunneling paths can increase the current. This effect is illustrated in Fig.6.19, where increased optical phonons enlarge the tunneling window, by allowing the electrons to tunnel at energies at least $3\times\hbar\omega_{opt}$ above the source Fermi level E_F^S. The relevance of this phenomenon depends on the intensity of the electron–phonon interaction, that is the optical phonon deformation potential and the temperature,

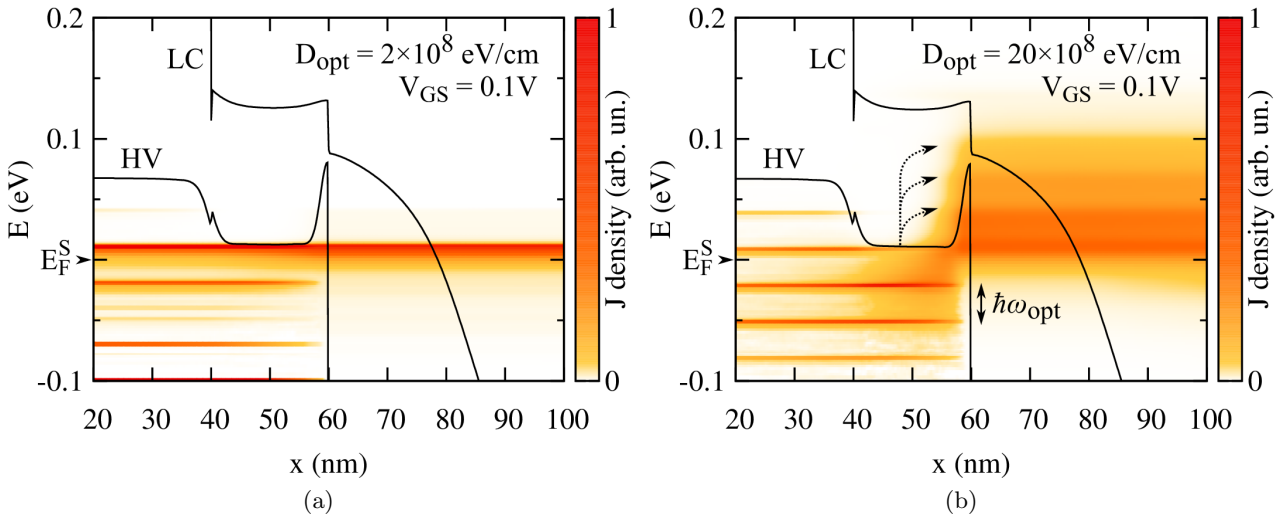


Figure 6.19: Spectral current at V_{GS} = 0.1 V in the initial device (a) and when D_{opt} is multiplied by 10 (b). Increasing the deformation potential enhanced optical phonon absorption in the OL region (symbolized as curved arrows), which facilitates the tunneling from the HV to the LC subband. Energy jumps of height $\hbar\omega_{opt}$ are clearly visible and correspond to phonon absorption/emission events.

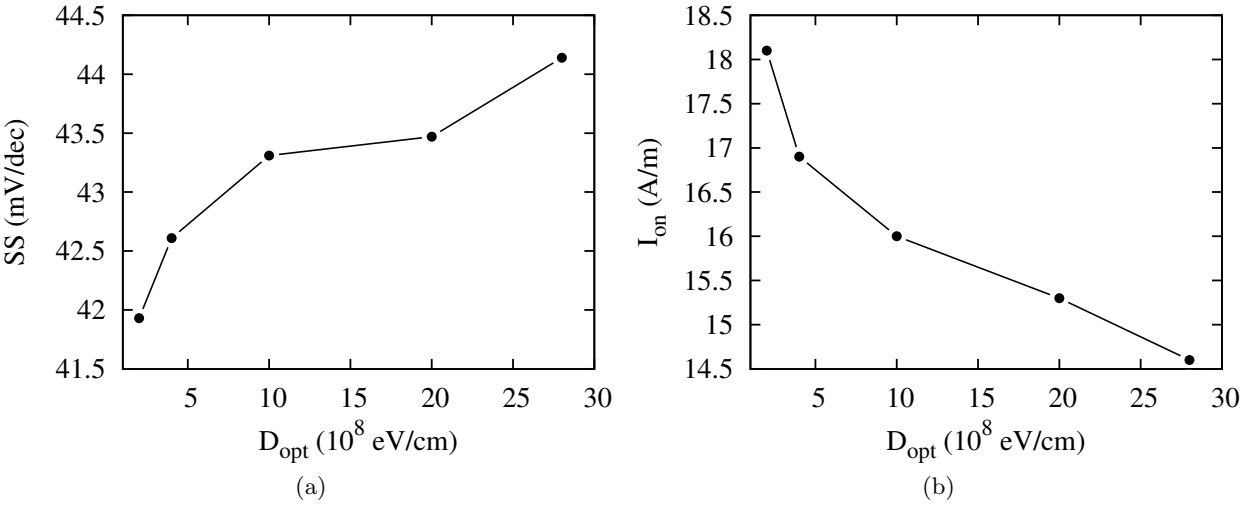


Figure 6.20: SS and I_{on} evolution as a function of the optical phonon deformation potential. By allowing more BTBT at the off state, the phonons degrade the SS. This detrimental effect is however rather weak in the studied device.

that we suppose to be the same for the optical phonon bath and the lattice. Fig.6.18 presents the effect of an increased optical phonons deformation potential on the transfer characteristic. The slope is degraded by optical phonons at low V_{GS}, but is then improved at higher gate voltages. The net effect of optical phonons on the SS and I_{on} is thus very weak, as shown in Fig.6.20. These results are consistent with a previous study performed on longitudinal III-V TFETs, where the value of D_{opt} did not strongly impact the performance of the device [15].

6.10 Conclusion

In this chapter, we have investigated the properties of vertical heterojunction tunnel-FETs. The GaSb/AlSb/InAs semiconductor combination offers both a broken-gap configuration and a large energy gap in the barrier, and is thus well suited for the design of such a device. This chapter also demonstrated the necessity of resorting to a quantum formalism to model innovative nanodevices, since a TFET device is inherently quantum.
We have shown that, with the right set of parameters, the vertical TFET could fulfil the ITRS requirements for LP applications, with V_{DS}=0.3 V. In order to reduce the *off*-state current and reach $I_{\mathrm{off}} = 10^{-5}$ A/m, one can resort to several tune-ups in the geometry of the device. The addition of spacer or longer gate extension regions both participate in the reduction of the *off* tunneling and in the improvement of the swing. This is also beneficial for the value of I_{on}. However, the overlap length L_{ov} had quasi non-existent effect on both the SS and the I_{on}.
Another way to improve the SS was to increase the barrier thickness t_{B}. Here, the most surprising result is that this modification only starts to degrade the *on* current when t_{B} gets greater than 5 nm. We have concluded that this was due to a stronger crossing of the HVB and LCB modes, resulting in a mixing of the confined states, in the devices with a thicker gate. This effect can compensate the decrease of the transmission probability due to the enlargement of the barrier, until the thickness of the latter reaches 5 nm.
Another slope enhancement was obtained by reducing the drain doping N_D. Indeed, this change in donor concentration lifts the LC subband in the drain, which greatly increases the *off* tunneling distance. However, the *on* state is not impacted by this change in doping, since the tunneling window becomes large enough at $V_{\mathrm{DS}} = V_{\mathrm{on}}$ to not be a limitation for the current flow. The practicality of this approach can however be debated, since the choice of N_D is limited by the Fermi energy at the drain contact.
The thicknesses of the source and the drain layers also impacted the behavior of the device, due to electrostatic effects in the case of the drain and confinement effects in the case of the source. The reduction of the drain thickness led to promising performance improvements, but has not been investigated more deeply due to convergence issues in the simulations for $t_{\mathrm{D}} \lesssim 2$ nm.
Finally, the effect of electron-phonon interactions has been investigated. We concluded that the phonons merely led to a shift of V_{th}, but nearly no performance degradation.
The work presented in this section has been partially included in a journal publication [11] and an additional, more complete, article is in preparation.

6.10.1 Perspectives

Our study of TFETs is far to be exhaustive. In particular, several disorder sources and optimization parameters should be investigated. For example, though interface traps can be strongly limited thanks to the vertical growth of the device, they could nevertheless affect the SS by promoting subthreshold tunneling. Their effect should be thus investigated in order to evaluate their possible impact on the performances.
Another aspect that could be studied is the role of strain, which could be exploited to optimize the devices by modulating the band structure in the different regions of the TFETs.

Bibliography

[1] L.R. Ram-Mohan I. Vurgaftman, J.R. Meyer. Band parameters for III–v compound semiconductors and their alloys. *Journal of Applied Physics*, 89(11):5815–5875, jun 2001.

[2] D Verreck, A.S. Verhulst, K.H Kao, W.G. Vandenberghe, K. De Meyer, and G.0 Groeseneken. Quantum mechanical performance predictions of p-n-i-n versus pocketed line tunnel field-effect transistors. *IEEE Transactions on Electron Devices*, 60(7):2128–2134, jul 2013.

[3] K. Jeon, W.Y. Loh, P. Patel, C.Y. Kang, J. Oh, A. Bowonder, and C.Park et al. Si tunnel transistors with a novel silicided source and 46mv/dec swing. In *2010 Symposium on VLSI Technology*. Institute of Electrical and Electronics Engineers (IEEE), jun 2010.

[4] M. Yokoyama O. Ichikawa T. Osada M. Hata M. Takenaka S. Takagi M. Noguchi, S.Kim. High ion/ioff and low subthreshold slope planar-type ingaas tunnel field effect transistors with zn-diffused source junctions. *Journal of Applied Physics*, 118(4):045712, jul 2015.

[5] A. nil W. Dey, B. M. Borg, B. Ganjipour, M. Ek, K.A. Dick, and E. Lind et al. High current density InAsSb/GaSb tunnel field effect transistors. In *70th Device Research Conference*. Institute of Electrical and Electronics Engineers (IEEE), jun 2012.

[6] F. Conzatti, M.G. Pala, D. Esseni, E. Bano, and L. Selmi. Strain-induced performance improvements in InAs nanowire tunnel FETs. *IEEE Transactions on Electron Devices*, 59(8):2085–2092, aug 2012.

[7] D. Leonelli, A. Vandooren, R. Rooyackers, A.S. Verhulst, S. De Gendt, M.M. Heyns, and G. Groeseneken. Performance enhancement in multi gate tunneling field effect transistors by scaling the fin-width. *Japanese Journal of Applied Physics*, 49(4):04DC10, apr 2010.

[8] M. Luisier and G. Klimeck. Simulation of nanowire tunneling transistors: From the Wentzel-Kramers-Brillouin approximation to full-band phonon-assisted tunneling. *Journal of Applied Physics*, 107(8):084507, apr 2010.

[9] G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu, Q. Zhang, H. Zhu, J.-M Kuo, and T. Kosel et al. Novel gate-recessed vertical inas/gasb tfets with record high ion of 180 ua/um at vds = 0.5 v. In *2012 International Electron Devices Meeting*. Institute of Electrical and Electronics Engineers (IEEE), dec 2012.

[10] Y. Zeng, C.I. Kuo, C. Hsu, M. Najmzadeh, A. Sachid, R. Kapadia, and C. Yeung et. al. Quantum well InAs/AlSb/GaSb vertical tunnel FET with HSQ mechanical support. *IEEE Transactions on Nanotechnology*, 14(3):580–584, may 2015.

[11] M.G. Pala, C. Grillet, J. Cao, D. Logoteta, A. Cresti, and D. Esseni. Impact of inelastic phonon scattering in the off state of tunnel-field-effect transistors. *Journal of Computational Electronics*, 15(4):1240–1247, sep 2016.

[12] S. Brocard, M.G. Pala, and D. Esseni. Design options for hetero-junction tunnel FETs with high on current and steep sub-threshold voltage slope. In *2013 IEEE International Electron Devices Meeting*. Institute of Electrical and Electronics Engineers (IEEE), dec 2013.

[13] M. Lundstrom. *Fundamentals of carrier transport*. Cambridge University Press, 2 edition, 2000.

[14] International technology roadmap for semiconductors. http://www.itrs.net/2013ITRS/Summary2013.htm, 2013.

[15] F. Conzatti, M.G. Pala, and D. Esseni. Surface-roughness-induced variability in nanowire InAs tunnel FETs. *IEEE Electron Device Letters*, 33(6):806–808, jun 2012.

CHAPTER 7

General conclusion

7.1 Summary

In this PhD, we have implemented numerical methods to simulate III-V semiconductor transistors. Thanks to an eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian model and the non-equilibrium Green's functions (NEGF) formalism, we were able to account for the quantum effects that arise in these nanoscale devices. Indeed, in order to accurately predict their behavior, one must consider the impact of quantum confinement, tunneling, interferences, or electron-phonon interactions, to cite some of them. In our simulations, these phenomena manifest themselves in the form of short-channel effects (SCEs), energy band shifts, band-to-band and source-to-drain tunneling processes (BTBT and STDT) and also influence scattering events. Additionally, the model has been improved by the implementation of strain effects and surface roughness. This allowed us to formulate predictions about the characteristics of different logic devices in a realistic context. The aim of this work was to look for possible III-V based candidates that could outperform current silicon technology. To carry out this investigation, we considered different device architectures.

First, we simulated an ultra-thin body (UTB) n-type MOSFET with a III-V channel. We have shown that SCEs had a substantial impact on the subthreshold swing (SS) and the *on* current. In particular, when the gate length was reduced, we observed the presence of drain induced barrier lowering (DIBL) and *off*-state STDT. In addition, hole-induced barrier lowering (HIBL) occurred for high supply voltages, due to the small gap of In(Ga)As. We shall stress that this phenomenon is not present in silicon, which represents an important difference between these materials. Our results suggest that the III-V UTB device should perform better than a III-V bulk architecture. Another interesting result is that a removal of the spacers does not strongly affect the transistor, which is an encouraging finding for device size reduction. Since this kind of device can already be fabricated experimentally, these simulations simultaneously allowed us to verify the validity of the code, by comparing our results with experimental data. Overall, the III-V MOSFET presents promising performances when compared to silicon, but still needs to be improved in order to outperform it.

Second, we made a further step towards ultra-scaled devices, by simulating gate-all-around nanowire transistors (NWFETs). These devices can be seen as an improved version of the UTB MOSFET, where the gate has been reshaped to be wrapped around the channel. Due to that fact that the experimental realization of this architecture is still technically demanding, we also had to model its strained silicon (sSi) counterpart to be used as a benchmark. Since we are aware of the fabrication difficulties, we also investigated the influence of surface variability at the channel-oxide interface. On the computational point of view, these calculations were especially resource-demanding, as the absence of periodicity in the system necessitated 3D simulations. Even though the larger density of states and immunity to STDT in the sSi devices always resulted in

a higher I_{on}, we found that a better electrostatic integrity could be obtained in InAs NWFETs when the gate length is longer than 10 nm. InAs devices exhibited a larger I_{off}, I_{on} and SS variability when surface roughness was applied. Despite this strong dispersion, the overall performance degradation in InAs was weaker than in Si. We also made the surprising observation that roughness could even enhance the SS (at $L_G = 10$ nm) or the I_{on} (at $L_G = 5$ nm) of the InAs NWFETs. Despite these advantageous features, we had to conclude that the InAs NWFETs could hardly outperform their silicon counterpart.
Third, we moved to a totally different kind of architecture, by studying a vertical heterojunction tunnel-FET (TFET) device. Since this device has no exact silicon counterpart, this final work constituted a state-of-the-art study, that could only be compared with other types of pre-existing technologies. This last example demonstrates the interest of the simulation approach, as it enables us to investigate the properties of nanoscale structures that are intricate to study both analytically and experimentally. It also emphasizes the importance of resorting to full-quantum simulations, as a TFET is intrinsically based on quantum tunneling. III-V compounds are ideal channel materials for this type of device, since their variety of band alignments and gap tunability allows the engineering of heterostructures with band profiles that can only be achieved with high values of doping in silicon. We have shown that the presence of a spacer, the addition of gate extensions, the increase of the barrier and drain thicknesses, or the reduction of the drain doping could all be beneficial for the SS and the I_{on} of the device. By combining some of these optimizations, the vertical III-V TFET exhibited very promising performances for low power applications, with a SS close to 20 mV/dec and an I_{on}/I_{off} ratio greater than 10^7. Finally, the effect of acoustic and optical phonons has been investigated, as they strongly impact the *off*-state tunneling current, by allowing inelastic electron transitions from the source valence band to the drain conduction band. We should also stress that – as of today – the (possibly optimistic) simulation results for TFETs do not match the experimental observations. This can be due to the presence of trap assisted tunneling, material imperfections or structure irregularities that are not considered in our model.
In this work, the numerical approach gave us a chance to perform deep and precise physical analysis of the investigated devices and to identify the most important parameters that determine their performance. One of the difficulties of computational physics is distinguishing between the results that do describe the real behavior of the simulated system and those that are just a consequence of the model approximations or the computational limits, and are thus unphysical or misleading. For that reason, it is essential to adopt a global view, both on a physical and on a computational standpoints. Among the technical difficulties that had to be dealt with during these PhD years, we can cite the convergence problems due to numerical issues, the abundant adjustments that were required in the code before it could lead to satisfactory results and the regular need to discriminate between real and spurious solutions – that can arise from both code imperfections and flaws in the model.

7.2 Perspectives

We can foresee several perspectives for the work conducted during this thesis. With only minor code adjustments, the behavior of all the devices could, for example, be investigated under various crystal orientations, strain conditions, or with defects at dif-

ferent interfaces. We could also simulate the p-type counterpart of the UTB MOSFET and the NWFET. Besides, the diameter and the shape of the nanowire device could be the subject of an entire new study. Regarding the TFET, it is certain that many geometrical configurations still need to be investigated. Among the others perspective of this work, we can mention the necessity to include in our simulations a more accurate description of the system (atomistic effects through, for example, the empirical pseudopotential method), more disorder models (trap assisted tunneling mechanisms, non-local phonons) and other important phenomena (such as self-heating).
Despite the promising paths explored in this manuscript, it is still unclear whether it will be possible to keep improving transistors beyond the 5 nm node in the coming years or decades. As extensively discussed in this work, size reduction is undoubtedly a strong limitation to the performances of such devices. However, the strategies employed in this PhD work (*i.e.* acting on the materials or on the device architecture) are not the only valid approaches to improve nanoelectronic devices. Beyond the scope of this work, other solutions could come from the investigation of different variables to carry the information, such as the spin of the electron. Other types of carriers, like the photon or various quasi-particles, may also prove useful in the future. Finally, different paradigms of computation can also be explored, and an increasing effort is currently being made towards the realization of quantum computers or neuromorphic chips, to cite just a few examples. In any case, computational physics is likely remain essential in the design of novel nanodevices, as the investigation of more and more advanced phenomena may be necessary to open the doors to new ideas.