Modeling and Design of GaN High Electron Mobility Transistors and Hot Electron Transistors through Monte Carlo Particle-based Device Simulations

by

Riccardo Soligo

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2016 by the
Graduate Supervisory Committee:

Marco Saraniti, Chair
Srabanti Chowdhury
Stephen Goodnick
Trevor Thornton

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

In this work, the insight provided by our sophisticated Full Band Monte Carlo simulator is used to analyze the behavior of state-of-art devices like GaN High Electron Mobility Transistors and Hot Electron Transistors. Chapter 1 is dedicated to the description of the simulation tool used to obtain the results shown in this work. Moreover, a separate section is dedicated the set up of a procedure to validate to the tunneling algorithm recently implemented in the simulator. Chapter 2 introduces High Electron Mobility Transistors (HEMTs), state-of-art devices characterized by highly non linear transport phenomena that require the use of advanced simulation methods. The techniques for device modeling are described applied to a recent GaN-HEMT, and they are validated with experimental measurements. The main techniques characterization techniques are also described, including the original contribution provided by this work. Chapter 3 focuses on a popular technique to enhance HEMTs performance: the down-scaling of the device dimensions. In particular, this chapter is dedicated to lateral scaling and the calculation of a limiting cutoff frequency for a device of vanishing length. Finally, Chapter 4 and Chapter 5 describes the modeling of Hot Electron Transistors (HETs). The simulation approach is validated by matching the current characteristics with the experimental one before variations of the layouts are proposed to increase the current gain to values suitable for amplification. The frequency response of these layouts is calculated, and modeled by a small signal circuit. For this purpose, a method to directly calculate the capacitance is developed which provides a graphical picture of the capacitative phenomena that limit the frequency response in devices. In Chapter 5 the properties of the hot electrons are investigated for different injection energies, which are obtained by changing the layout of the emitter barrier. Moreover, the large signal characterization of the HET is shown for different layouts, where the collector barrier was scaled.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

THE MONTE CARLO METHODS FOR DEVICE SIMULATIONS

1.1 Monte Carlo

From the invention of the transistor in the early 1950s, semiconductor devices have revolutionized all fields of the modern industry. The demand for performance coming from the military and aerospace sectors, motivated the interest of developing faster, smaller, less noisy, more powerful devices. Parallel to the efforts in the manufacturing sector, the study of the transport properties of semiconductors played a fundamental role in the development of these devices. An important contribution to the understanding of semiconductor transport properties has been given by transport simulations. In particular, the flexibility offered by the computational environment where control is extended over all variables, allows isolating causality effects between the device layout and performance. On the other hand, the validity of computational analysis is only as good as the model that is implemented. Therefore, simulations programs have been subjected to continuous improvements implementing more complicated and computationally challenging models. In this regard, the increase of the computer speed and the availability of memory to store data represented an important factor for the development of simulators.

A complete description of the carrier transport is provided by the Boltzmann Transport Equation (BTE) Hamaguchi (2001):

$$\frac{\partial f}{\partial t} = -\vec{v} \cdot \nabla_{\vec{r}} f - \dot{\vec{k}} \cdot \nabla_{\vec{k}} f + \left(\frac{\partial f}{\partial t} \right)_{coll}, \tag{1.1}$$

where $\vec{v}$ and k represent the carrier velocity and momentum respectively. The variable of this equation is represented by $f(\vec{r}, \vec{k}, t)$ that in its purest meaning represents the probability of on electron to be at the coordinate $\vec{r}$ in real space and coordinate $\vec{k}$ in k-space at the time t. On the other hand, the meaning of $f(\vec{r}, \vec{k}, t)$ can be extended to a distribution function describing the properties in real and k-space of an ensemble of non-interacting particles Ferry (1991). The distribution function provides a complete description of the ensemble of particles, and all the important observable quantities, such a velocities or currents, can extracted through averages on the ensemble. Therefore, device simulators are built around methodologies to calculate a solution for the BTE. However, despite a large volume of research on this matter, analytical solutions are not available unless severe approximations are made. Such approximations are rarely compatible with the physical situation that is intended to be modeled, so that the results tend to be just the outcome of a mathematical exercise instead of a description of the microscopic processes Jacoboni and Reggiani (1983). A popular technique for device simulation is the drift and diffusion method Selberherr (1984) based on the first momentum of the BTE. This approach allows implementations without much computational burden, but fails to describe non-equilibrium phenomena in short channel devices. In order to capture these phenomena, the drift and diffusion method is coupled with the energy balance equation in hydrodynamic simulations Forghieri *et al.* (1988). In this approach, the particle distribution to be in equilibrium with the lattice, but a relaxation time is considered for the energy to relax. However, both these models are based on approximations on the BTE and rely heavily on empirical parameters, such as mobility or relaxation times, for their operation. A different approach is used by Monte Carlo simulators which do not rely on approximations to provide a deterministic solution of the BTE, but are based on stochastic procedure to provide an exact solution. The treatment of transport in

Monte Carlo is semiclassical with the particles moving classically following Newton's law until a scattering event occurs according to a quantum mechanical treatment. The implementation of the Monte Carlo algorithm for device simulations is shown in the flow chart in Figure 1.1.

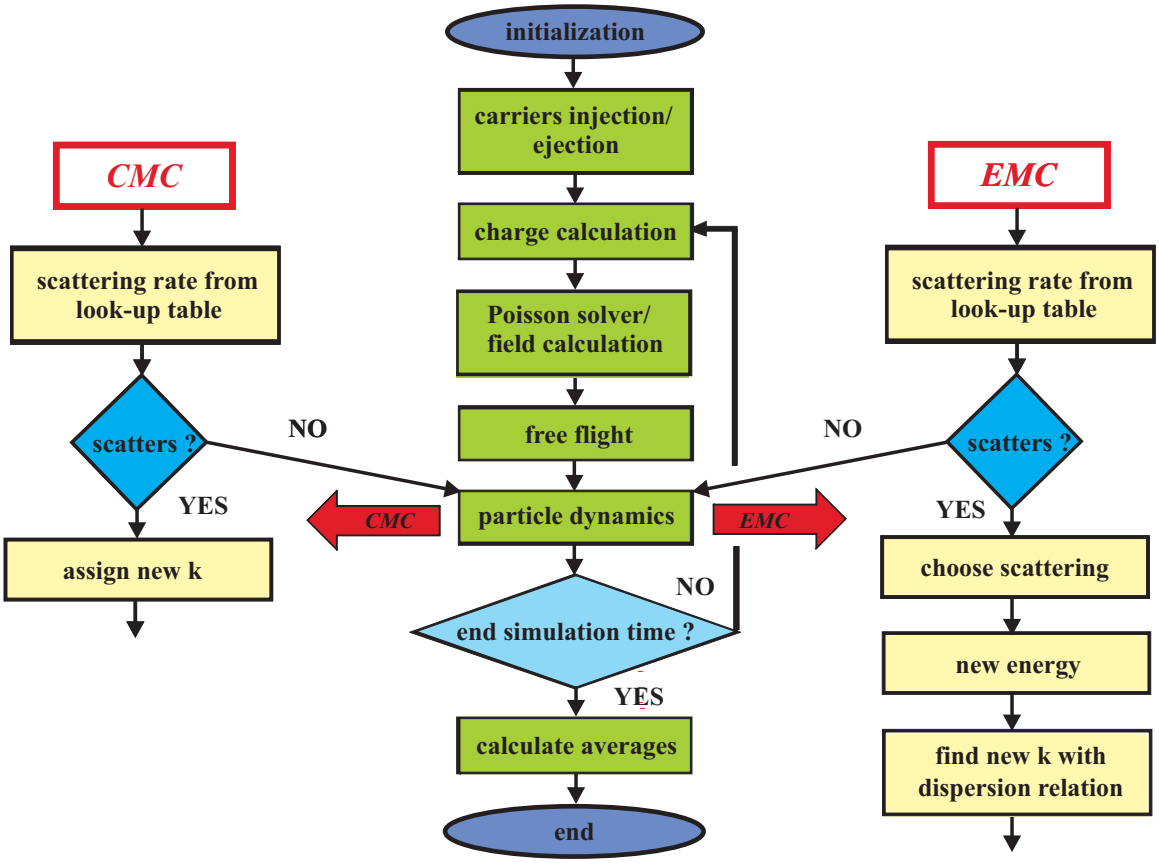

Figure 1.1: Flowchart describing the main steps of a particle based Monte Carlo for device simulation. The scattering part is expanded to highlight the difference between the EMC and CMC approaches.

The algorithm starts with the calculation of the electric field by the Poisson equation according to the real time distribution of charge inside the device. The electric field is then assumed constant for a short amount of time, called the free flight time step, in which the particles travel ballistically with a velocity computed from the their dispersion relationship. At the end of free flight a stochastic procedure determines whether a scattering event occurs, in which case the momentum of the particle is

changed. Usually 4 or 5 free flight time steps are repeated before running another cycle of the Poisson solver to update the field according to the new position of the particles.

The treatment of scattering represents the part of the algorithm that is more challenging from the computational point of view. In particular, the probability of a certain particle to scatter is calculated as summation over all possible final states, for all the possible scattering mechanism. This calculation, that is very expensive computationally, is repeated for every particle at every free flight time step during the whole time of the simulation. The total scattering probability for each state due to all scattering mechanisms is therefore often precomputed and stored in a look-up table available at runtime in Ensemble Monte Carlo (EMC). However, when a scattering occurs, the scattering probability for each mechanism has to be recomputed to calculate the final momentum of the particle among the states that conserve the energy. A faster (up to 25 times Saraniti and Goodnick (2000)) version of the EMC code called Cellular Monte Carlo is obtained expanding the scattering table, saving the transition probability for each couple of initial and final state. However, the improvement in terms of time is achieved with a higher memory consumption and also losing information on the specific mechanism responsible for the scattering. In the present CMC simulator, the scattering tables include the scattering probability due to each scattering mechanism as well as the total probability. This not only allows to recover the accuracy of the EMC method while maintaining the speed of the CMC approach, but it also enable the possibility to change the scattering probability due to a particular mechanism. In particular, this is the case of the screening effect that electrons have on scattering with ionized impurities Chattopadhyay and Queisser (1981). In fact, the screening can be implemented in the scattering algorithm as a rejection probability to apply to the tabulated rates depending on the local concentration of electrons

and to the mechanism responsible of the scattering. Moreover, since a rejection can only decrease the initial scattering probability, the tables have to be calculated for a doping concentration higher than what is present in the device so that the scattering from ionized impurity can be reduced according to the local concentration of doping.

Beside the above mentioned scattering mechanisms due to ionized impurities, and polar interactions, which include polar optical phonon and piezoelectric scattering, other mechanisms included in the tables are: acoustic and optical deformation potential, dislocation scattering and impact ionization. The description of each scattering is done within a full band framework, where the electronic band structure is defined in the whole First Brillouin Zone using the Empirical Pseudopotential Method Kittel (1976). The full band approach is particularly useful to describe the dynamics and kinetics of electrons at high energies where the parabolic bands approximation fails to provide the correct density of states to calculate accurate scattering rates, as well as the appropriate velocity to determine the particle trajectory. Quantum phenomena are also included in our code through the effective potential method Ferry (2000). In this approach the wave package nature of the particles is taken into account calculating a smoothed version of the potential. The difference between the classical and the effective potential calculation of conduction band confining the electrons in the channel of a GaN HEMT is shown in Figure 1.2.

In this picture, the channel is located in the lower part of the conduction band where electrons accumulate as if they were balls under the effect of gravity. Comparing the two bands, it can be seen that two quantum phenomena are described by the effective potential calculation: the charge setback and the quantization of energy. The first phenomenon is important because it shifts the position of the channel a few nanometers away from the barrier, and therefore away from the gate. It will be shown later in this work that the channel, to gate distance influences the control of the

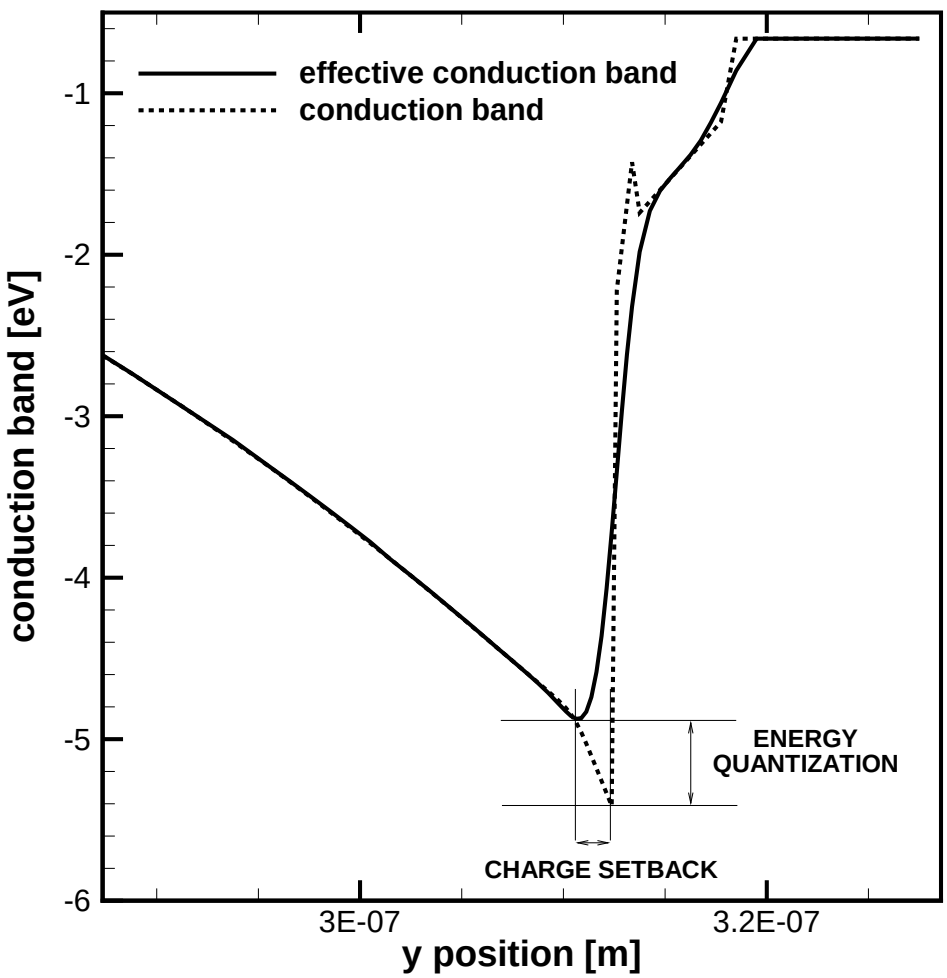

Figure 1.2: Quantum confinement effects at the AlN/GaN interferface illustrated through the comparison of the conduction band calculated with of without the effective potential.

gate on the channel so that neglecting the charge setback leads to an overestimation of quantities like gm and other figures of merit. Similarly the upward shift of the effective conduction band represents the reduction of the electron density due to the quantization of energy in confined systems, which reduces the density of states in the channel Knezevic *et al.* (2002).

1.2 Tunneling Algorithm

A tunneling algorithm is implemented in the Monte Carlo simulator to simulate direct tunneling across arbitrarily shaped potential barriers. This algorithm is integrated with the simulation positional kernel that activates the tunneling function when a particle encounter a potential barrier along its trajectory. A new position on the other side of the barrier is assigned to the particle when tunneling occurs otherwise the position is not updated and the particle is reflected. After tunneling, the initial state of the particle is updated in order to preserve the total (non kinetic) energy and at the same time preserve the components of the momentum parallel to the barrier interface. The critical aspect of the algorithm is the calculation of the tunneling probability. The implementation available in our simulator is based on the transfer matrix method Ando and Itoh (1987), where the potential barrier is decomposed as a series of rectangular barriers, where the potential and effective mass are assumed constant as shown in Figure 1.3.

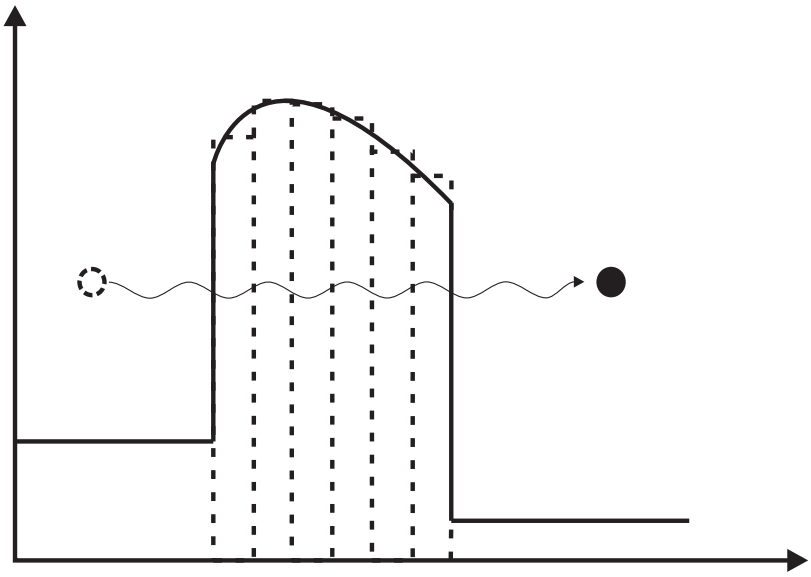

Figure 1.3: Schematic illustration of the decomposition of an arbitrary potential barrier in a series of rectangular barrier.

According to this method, a (2×2) transmission matrix can be associated to each rectangular barrier depending only on the barrier height and electron effective mass. The transmission coefficient through the whole barrier is then calculated by connecting momentum eigenfunctions multiplying all transmission matrices. Further details about the calculations and the implementation can be found in Marino (2010).

1.3 Tunneling Validation

Section 1.2 showed that besides the calculation of a probability, the implementation of the tunneling algorithm requires other operations such as the update of the momentum and position of the particle, the interpolation of the energy *etc.* Moreover, it is never trivial to integrate a rather complex function with the execution flow of the simulation, not to mention that numerical errors are also possible. Therefore, every time a model is implemented in the code it should be validated in terms of final output that may be affected, which in the case of tunneling is the current. For this purpose, a real device where tunneling is known to play a relevant role can be simulated, but the modeling of the device itself would add some uncertainties on its own. For example, in the simulation of the simplest structure based on tunneling, the Schottky diode, the physics of the metal-semiconductor interface is not yet completely understood, and more important, our code is not designed to simulate transport in metals. In this Section we show a possible validation of the tunneling algorithm, in all its components, using an alternative approach. In particular, instead of modeling a real device, we managed to design a layout comparable with the model of the transmission of the single electron through a finite rectangular potential barrier, for which an algebraic solution is available. Indeed, in Figure 1.4 it can be seen that the band profile of the simulated device is analogous to the picture of the tunneling across a rectangular barrier developed in any quantum mechanics book.

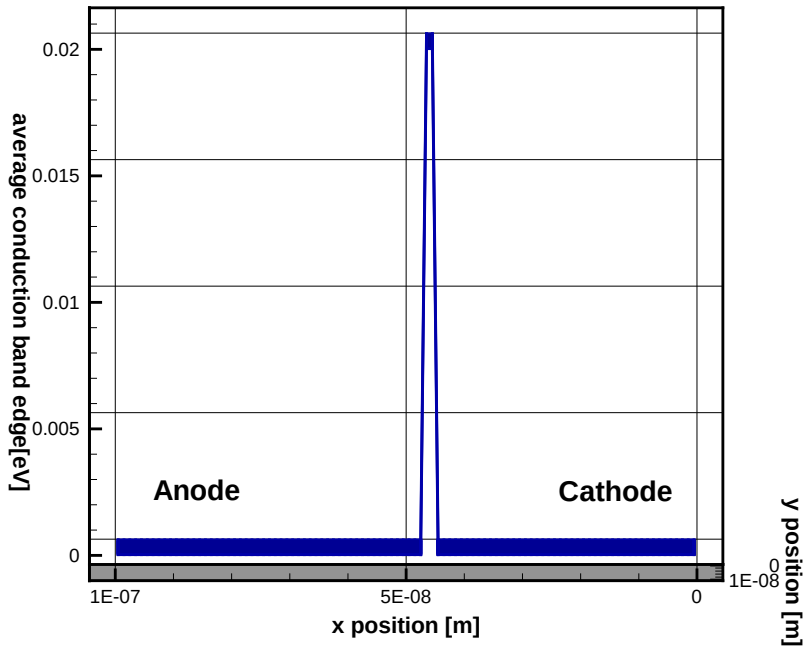

Figure 1.4: Band diagram of the simulated 0.02eV, 2nm potential barrier.

In order to obtain this band diagram, some of the functions that are commonly required for device simulations have to be modified or totally disabled. The latter is the case of the Poisson solver that adjusts the potential according the distribution of charge. In this test layout, the profile of the potential is set by 5 equipotential regions that constitute the device. The regions at the two ends are contacts that beside pinning the potential at the two ends, inject and eject carriers. The adjacent regions are common equipotential regions kept at the potential of the contacts. The central region represents the barrier and is defined by setting a band discontinuity of 0.02 eV. Eq 1.2 reports the transmission coefficient through a rectangular barrier Griffiths and Harris (1995) that we want to reproduce to validate our tunneling function:

$$T = \left(1 + \frac{V_0^2 \sinh^2(ak)}{4E(V_0 - E)}\right)^{-1} \tag{1.2}$$

where T is the transmission or tunneling probability, E is the energy of the electron, V_0 is the barrier height, a is the barrier thickness and k is the wave vector of the electron calculated as:

$$k = \sqrt{2m(V_0 - E)/\hbar^2} \tag{1.3}$$

where m is the effective mass of the electron, $\hbar$ is the reduced plank constant.

In particular the variables of interest in Eq 1.2 and Eq 1.3 are the energy and the momentum of the electrons. However, in order reproduce the transmission calculated from Eq 1.2 with the simulation output calculated averaging over a sample of particles, all the electrons have be to introduced in the device with the same energy and momentum. On the other hand, the standard injection algorithm initializes carriers in the contacts according to a truncated Maxwellian distribution in energy and isotropic in k-space to emulate the metal contact. This algorithm is changed in the current simulation so that all electrons are initialized at a 0.01 eV energy respect to the bottom of the conduction band, and with only the component of the k-vector in the direction of transport different from zero, to reproduce the 1D character of Eq 1.2. Moreover, the injection is only enabled in one contact (cathode) while the other (anode) can only eject. The time evolution of the current calculated in two regions defined in between the contacts and the barrier called anode and cathode region according to the closest contact are shown in Figure 1.5.

It can be seen in Figure 1.5 that the currents are not constant in time. In particular, at the beginning of the simulation both currents are zero because all the electrons are initialized in the contacts so that there is no charge in both regions. At the instant t_1 shown in Figure 1.5, the front of the electrons injected by the cathode reaches the cathode region. The current associated with the injected electrons can be written as:

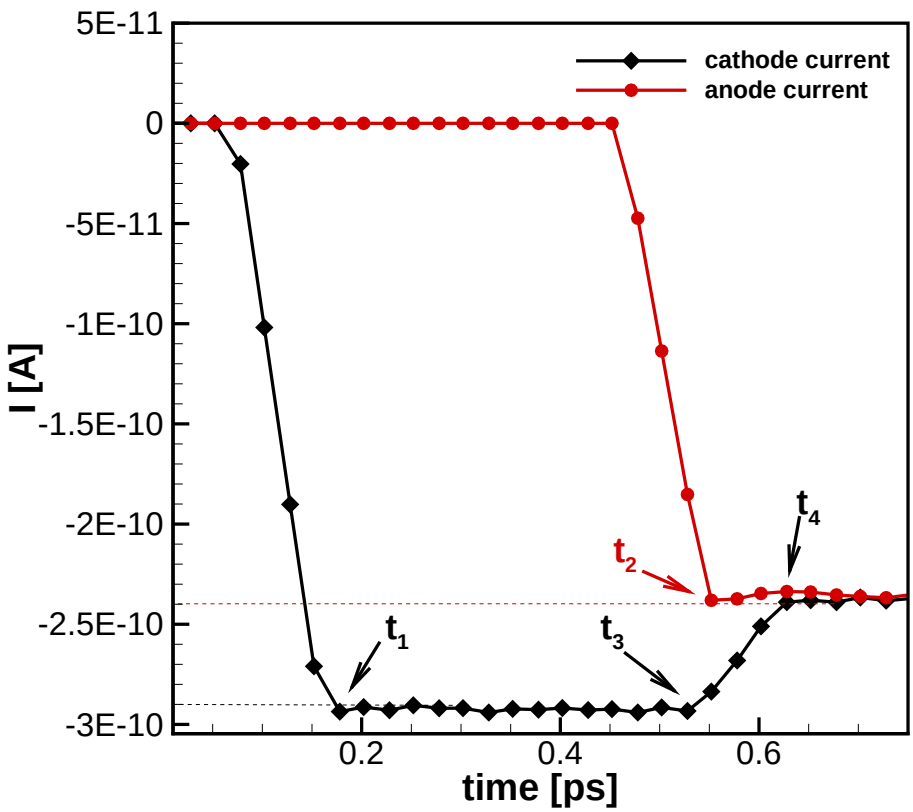

Figure 1.5: Time evolution of the current across a region between the injecting contact and the barrier (cathode, back diamonds) and a region between the barrier and the ejecting contact (anode, red circles).

$$I_{inc} = L \sum_{i}^{N_{ca}} e v_i = L N_{ca} e v \tag{1.4}$$

where I_{inc} is the current incident to the barrier, N_{ca} is the number of electrons in the cathode region, e is the electron charge and v_i is the injection velocity of the electrons. Assuming the electrons have the same velocity allows simplifying the sum with the simple product on the right of Eq 1.4. Eventually, this flux of electrons hits the barrier splitting in a transmitted component, that keeps traveling at the same velocity towards the anode, and a reflected current that heads back towards the cathode. When the transmitted front of electrons reaches the anode region the anode current increases until it reaches the final value corresponding the to transmitted cur-

rent at time t_2. The decrease of the cathode current at time t_3 is due to the electrons reflected by the barrier reaching back to the cathode region. The situation at time t_3 is captured in Figure 1.6 where the monoenergetic flux of electrons is represented by the black spheres. In particular, looking at the electron density contour, the front of the reflected electrons can be seen as the increase of electron density extending from the barrier until approximately half way to the cathode electrode.

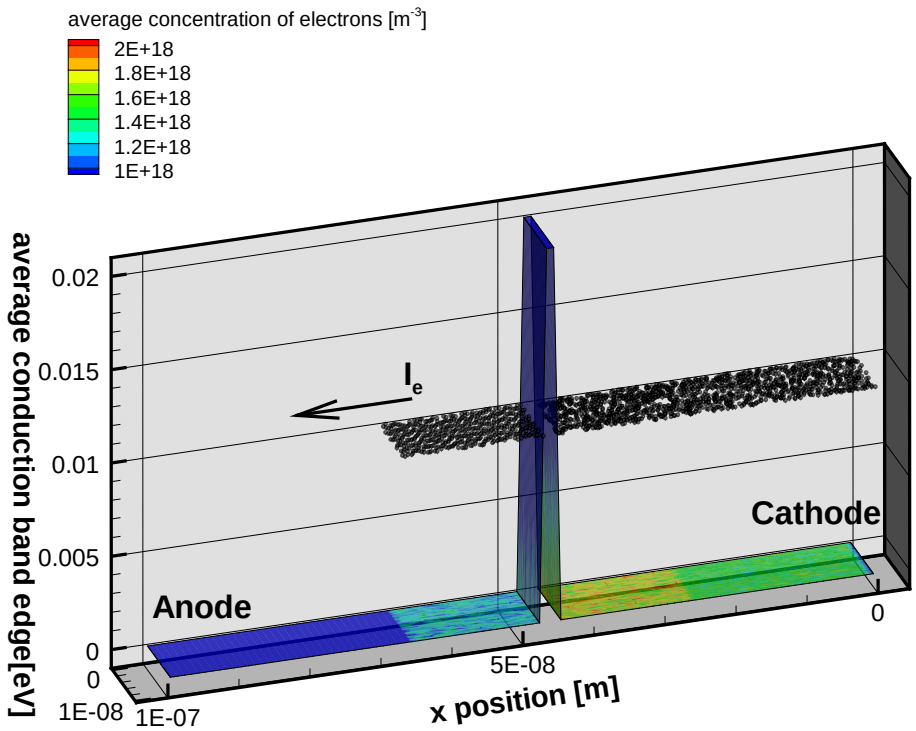

Figure 1.6: Conduction band profile colored according to the concentration of electrons. The spheres represent electrons in the simulation and the arrow points towards the injection direction.

Finally, at time t_4, the steady state situation is reached when, since the current is conserved, its calculation on the two regions gives the same value. In the analysis of the time evolution of the current, the incident current was defined as the cathode current between time t_1 and t_3, while the transmitted current corresponded to the

anode current after time t_2. Since the transmitted current is calculated with the same formula of Eq 1.4, just substituting N_{ca} with N_{an}, the ratio between the two currents corresponds to the ratio between the electrons that tunnel through the barrier over the electrons that hit the barrier. This number corresponds to the tunneling probability and can be compared with the analytical result given by Eq 1.2. To compare the analytic transmission to what is shown in Figure 1.5, the calculation is performed using the effective mass m for silicon, which is the material used in the simulation, a barrier thickness of 2 nm and E and V_0 equal respectively to 0.01 eV and 0.02 eV which results in a transmission value of 0.819 which very close to 0.827 given by the current ratio. Another test of the tunneling algorithm is shown in Figure 1.7 where the transmission coefficient is calculated for different barriers.

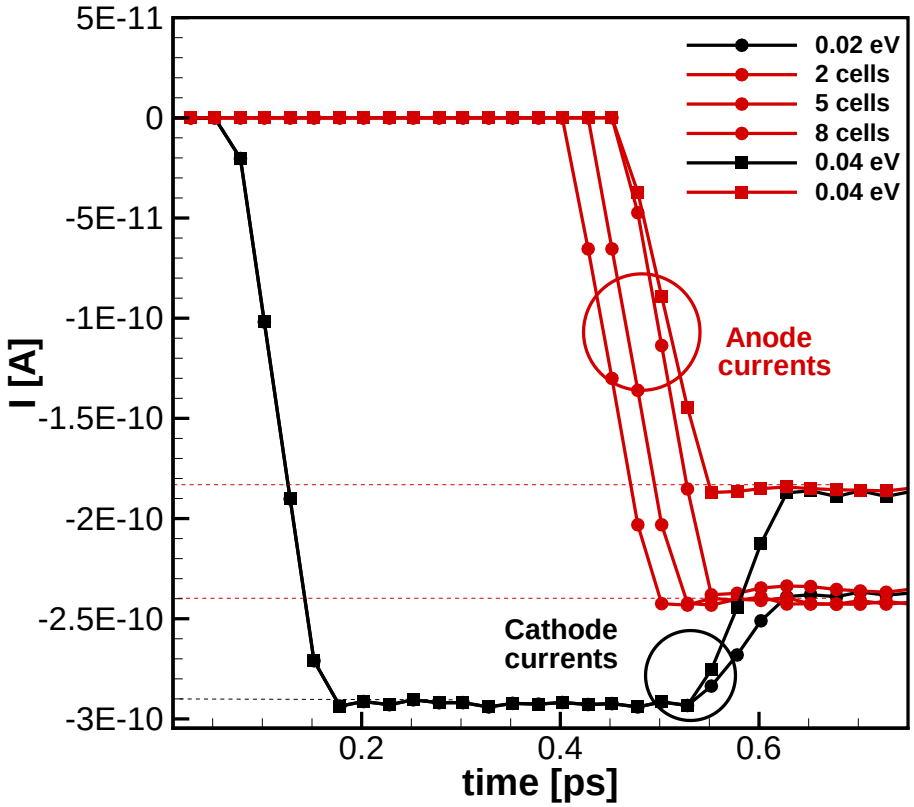

Figure 1.7: Time evolution of the current for: (circles) 0.02 eV, 2 nm barrier composed of a different number of cells indicated in legend; (squares) 0.04 eV, 2 nm barrier composed of 2 cells. The injection energy is 0.01 eV in all cases.

In particular, the red curves with circle symbols represent the current transmitted through barriers with identical width and height, but defined over a different number of grid cells. Since all these simulations represent the same physical situation, the transmitted current is correctly the same in all the cases apart from the noise. The only systematic difference that can be observed is the time when the anode current increases, which changes only because the cathode to barrier distance has been slightly modified changing the cell dimensions. The curve with square symbols in Figure 1.7 is instead related to a barrier height of 0.04 eV, for which the analytical transmission coefficient is 0.6156 while the ratio of the current gives 0.6207. The excellent agreement found in all tests validates the implementation of tunnel which will be essential for the simulation work show in chapter 4.

Chapter 2

MODELING OF GAN HEMTS

2.1 Introduction

One of the highest performance devices for high power and high frequency applications is the high electron mobility transistor (HEMT). The fabrication of these devices began in the early eighties after the successful implementation of the modulation doping in AlGaAs/GaAs heterostructures Dingle *et al.* (1978). In these structures, the electrons supplied by a thin doping layer in the barrier are separated from the parent donor atoms in a potential well introduced by the heterostructure. In this way, the conduction of electrons along the quantum well occurs without the detrimental effect of the scattering from the ionized impurities leading to high mobilities. With advancements in epitaxial growth techniques, new heterostructures have been available to design HEMTs optimized for specific applications. The use of narrow gap material like $InAs$ in ternary alloys with GaAs was found to increase the electron mobility above 15000 cm^2/Vs Bollaert *et al.* (2001). This heterostructure features an undoped $In_{0.53}Ga_{0.47}As$ channel with buffer and barrier layers of $In_{0.52}Al_{0.48}As$ grown on a InP substrate. The precise control of the In and Al molar fraction is essential to guarantee the lattice constant matching of all the layers. In particular, the growth of layers with different lattice constants can lead to generation of defects in the relaxation process. However, it is possible to grow defects free mismatched epilayers as long as their thickness is kept below the critical value when the crystal relaxes. This category of HEMTs called pseudomorphic HEMTs currently holds the record for f_T of 710 GHz among all transistors with a 60 nm gate InAlAs/InAs device Chang *et al.*

(2013). The applications of pseudomorphic HEMTs are mainly related to the design of high frequency low power sub-millimeter-wave amplifiers and Terahertz electronics Kim and del Alamo (2008b) Guerra *et al.* (2011a). On the other hand, despite the high values of current and transconductance, these devices are not suitable candidates for high power applications. In fact, in order to maximize the output power, the device has to be biased with high voltages, which is not sustainable by low bandgap materials. In this regard, HEMTs based on heterostructure of GaN and its alloys with other nitrides, especially AlN and InN, have been developed. In particular, the large band gap of GaN, 3.4 eV as opposed to 1.42 eV for GaAs and 0.36 eV for InAs, allows this material to withstand fields up to 5 MV/cm, orders of magnitude higher than GaAs (0.4 MV/cm) and InAs (0.04 MV/cm) Mishra and Singh (2007). Another interesting property of GaN and other nitrides with Wurtzite crystal structure is the presence of a spontaneous polarization within the crystal. When a heterostructure of nitrides is grown, the polarization discontinuities at the interfaces can induce an accumulation of charge confined by the band discontinuity between the layers. In particular, interfaces between GaN and AlN can result in a sheet of electrons with density 2×10^{13} cm^{-2}, approximatively 20 times higher than traditional GaAs HEMTs, without the introduction of any delta layer of doping Mishra *et al.* (2002). Moreover, the high quality of the AlN/GaN interface and the lack of alloy scattering due to an inserting AlN interlayer, allows for mobilities as high as 2000 cm^2/Vs which is only six times lower than in InGaAs HEMTs Mishra *et al.* (2008). However, due to the higher separation between the gamma and satellite valleys, the peak velocity in GaN is achieved at much higher fields than in GaAs and InAs, which means that the maximum output current of the device is delivered for higher drain to source bias. As a result, power densities in AlGaN/GaN HEMTs have demonstrated levels up to 69 W/mm, approaching a one-order of magnitude improvement over conventional

HEMTs Bengtsson *et al.* (2009). On the other hand, f_T up to 370 GHz Yue *et al.* (2012) was reported in GaN HEMTs with a layout dedicated to maximize the frequency response. These characteristics of GaN HEMTs makes them ideal candidate for for the next generation of high-power and high-frequency applications, such as radar and satellite communication front-ends with particular relevance for the design of millimeter-wave power amplifiers in the Ka and X-band Mishra *et al.* (2008). In this chapter, we discuss the DC and small signal AC operation of a state-of-art GaN HEMT, providing first a background of the modeling techniques and then showing the results obtained with the Monte Carlo simulator introduced in Chapter 1. In particular, Section 2.2 describes the theoretical basis to model the polarization in nitride heterostructures. Applications of the polarization theory is found in Section 2.3 to realize a complete model of a GaN HEMT for Monte Carlo simulation. The simulation approach is validated, showing the match of the DC characteristic in Section 2.4, while in Section 2.5, the operation of the HEMT will be discussed in terms of the dynamics of electrons in the channel. In Section 2.6 the two port model is introduced to study the small signal behavior of devices, providing general relations that will be applied to an equivalent circuit of the FET intrinsic area in Section 2.7. Section 2.8 describes the most recent model for the HEMT's cutoff frequency in terms of the dynamics of the electrons in proximity of the gate, which will be further developed in Section 2.9. Finally, the new cutoff frequency model is tested in Section 2.9.1, where properties of the electron dynamics are related to the dielectric constant of the passivation layer.

2.2 Theory of Polarization in Nitrides

One of the advantages of GaN HEMTs compared with GaAs HEMTs is the presence of a polarization induced two dimensional electron gas that can act as a channel.

The formation of this highly conductive sheet of electrons is due to the charge accumulation induced by the pyroelectric (spontaneous) and piezoelectric polarization properties of GaN and other nitride compounds. These polarization properties are a consequence of the reduced symmetry of the Wurtzite crystal structure typical of nitrides, as opposed to the Zincblende structure that is the most common form of crystallization for GaAs. Moreover, this form of polarization is always present with the same sign regardless to any external electric fields, in contrast to ferroelectric materials. The magnitude of the spontaneous polarization changes from nitride to nitride and increase as their crystal structure deviates from the ideal Wurtzite structural parameters Pearton *et al.* (1999). In particular, Figure 2.1 reports the spontaneous polarization of the most common nitrides, among which AlN has the highest polarization charge. The high polarization charge of AlN is in fact one of the main reasons why this compound and its alloys with GaN are found as a layer in the layout of several devices Nidhi *et al.* (2009) Chung *et al.* (2009).

The piezoelectric polarization occurs when piezoelectric crystals are strained by pseudomorphical growth over substrates with mismatched lattice and thermal expansion coefficient. In particular, during epitaxial growth the crystal structure of the material grown over the buffer is strained along the basal plane causing stress components along the same plane, whereas along the growth direction, the stress is 0 since is free to grow and no external force is applied. The relation between the strain ϵ and the stress γ along one of the basal plane direction is given by Ambacher *et al.* (2002):

$$\sigma 1 = \epsilon_1 \left(C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} \right), \tag{2.1}$$

where C_{ij} are elastic coefficients specific of the material derived from the elastic tensor in Hook's law. The strain ϵ_1 is calculate from the mismatch of the lattice constants:

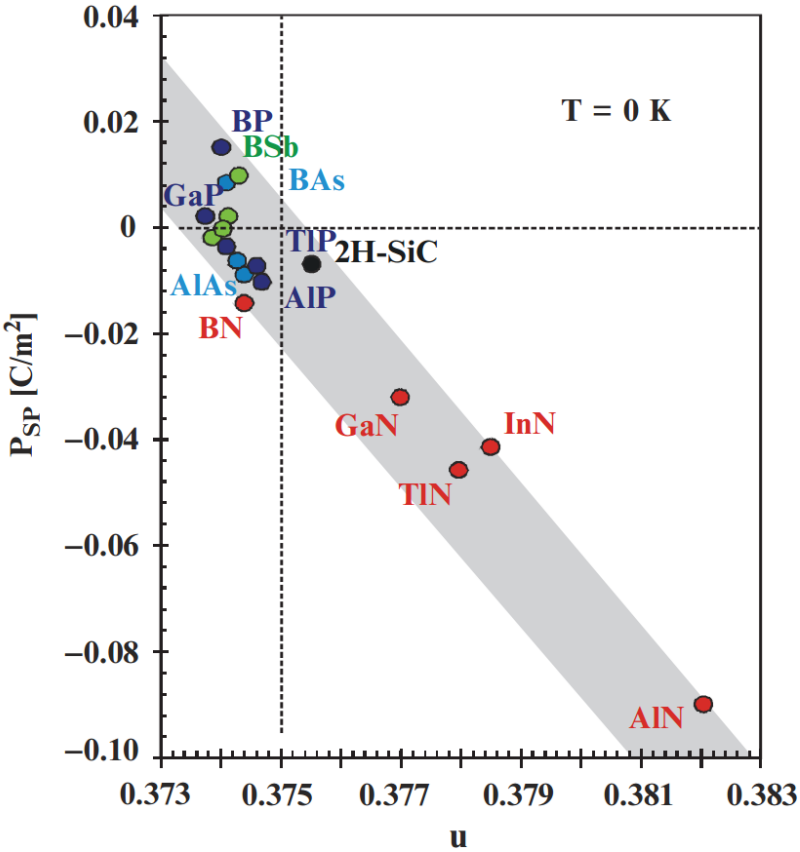

Figure 2.1: Wood and Jena (2007) Spontaneous polarization in wurtzite compound semiconductors.

$$\epsilon_1 = \frac{a - a_0}{a_0}, \tag{2.2}$$

where a is the strained lattice constant (corresponding the the buffer lattice constant) and a_0 is the relaxed one. The stress components are related to the piezoelectric polarization by coefficients d_{ij} called piezoelectric moduli. Considering the case of hexagonal crystals, the piezoelectric polarization induced by the biaxial strain due to pseudomorphic growth has only one component parallel to the growth direction which is calculated as:

$$P_3 = d_{31}\sigma_1 + d_{32}\sigma_2, \tag{2.3}$$

Moreover, due to the symmetry of the crystal $d_{32} = d_{31}$ and since the biaxial strain is the same in both the direction σ_1=σ_2, Eq. 2.3 can be rewritten as:

$$P_3 = 2d_{31}\sigma_1. \tag{2.4}$$

Finally substituting Eq. 2.1 and 2.2 in 2.4:

$$P_3 = 2d_{31}\epsilon_1 \left(C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} \right), \tag{2.5}$$

which relates the piezoelectric charge directly to the lattice constant of the buffer and the grown material. Since the following relation between elastic coefficient holds:

$$C_{11} + C_{12} - 2\frac{C_{13}^2}{C_{33}} > 0, \tag{2.6}$$

the direction of the piezoelectric and spontaneous polarization are parallel when $\epsilon_1 > 0$ (*i.e.* tensile strain) and antiparallel when $\epsilon_1 < 0$ (*i.e.* compressive strain). The total polarization is then calculated as:

$$P = P_{PE} + P_{SP}, \tag{2.7}$$

where P_{PE} is the piezoelectric polarization and P_{SP} is the spontaneous one. The charge density associated with a non uniform polarization is given by:

$$\sigma_P = - \bigtriangledown P. \tag{2.8}$$

Therefore, at the interfaces of pseudomorphic $A_xB_{1-x}N/GaN$ heterostructure, where the polarization changes abruptly, the polarization charge can be calculated as:

$$\sigma_{ABN/GaN} = (P_{GaN}^{SP} + P_{GaN}^{PZ}) - (P_{ABN}^{SP} + P_{ABN}^{PZ}). \tag{2.9}$$

The spontaneous and piezoelectric polarization of ternary nitrides, $A_xB_{1-x}N$, can be found by linear approximation of the binary compounds according the molar fraction (Vegard's law) Denton and Ashcroft (1991). However, non-linear behavior was observed and models based on a quadratic approximation Eq.2.9 are available in literature Deger *et al.* (1998). As an example the quadratic interpolation used to determine the piezoelectric polarization of $Al_xGa_{1-x}N$ grown on relaxed $/GaN$ substrates is:

$$P_{AlGaN/GaN}(x) = [-0.0525x + 0.0282x(1-x)]\, C/m^2. \tag{2.10}$$

So far the direction of the polarization was described using the crystallographic direction [0001] as a reference. However, the [0001] direction in Wurtzite compounds can be inverted according to the growth procedure. In fact, compounds with Wurtzite crystal structure like GaN, exhibit two different sequences of atomic layering in the two opposing directions parallel to the [0001] and $(000\bar{1})$ axes, so that they do not present inversion symmetry. Along these directions, atoms are arranged in bilayers of closely spaced hexagonal layers, one formed by the anion (N) and the other from by cation (Ga). According to which of the two atoms constitutes the the basal plane 0001, two different types of GaN can be distinguished: N-face and Ga-face. In Ga-face the bottom layer is made of N atoms and the [0001] direction, defined by a vector pointing from a Ga atom to the closest N atom, points toward the surfaces. Vice versa, in N-face GaN the bottom layer is occupied by Ga atoms and the [0001] direction points towards the substrate as shown in Figure 2.2.

2.3 GaN HEMT Layout

The approach adopted in this work to model devices for Monte Carlo simulation is shown in detail for a state-of-art GaN-HEMT. The layout of the device fabricated

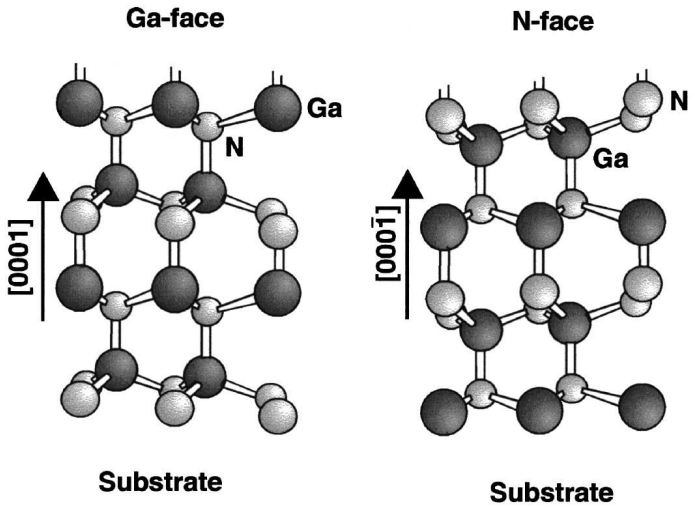

Figure 2.2: Ambacher *et al.* (2000) Schematic of the crystal structure of wurtzite Ga-face and N-face GaN indicating the different direction of the polarization in the two cases.

by Palacios *et al*. Lee *et al.* (2011) is shown in Figure 2.3.

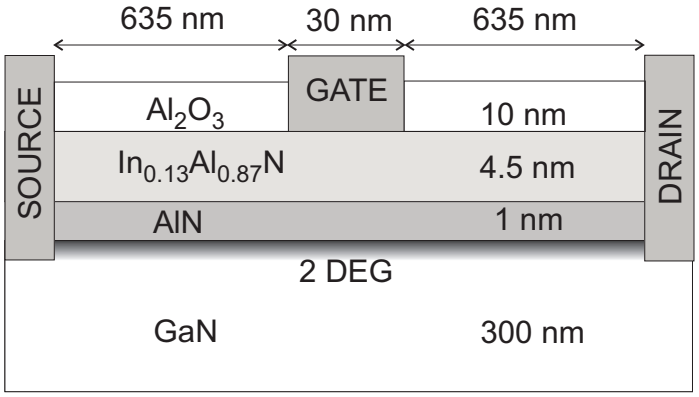

Figure 2.3: Layout of the simulated HEMT. The name of every layer is reported together with its thickness.

The HEMT heterostructure consists of a 4.5 nm lattice matched $In_{0.17}Al_{0.83}N$ layer separated by a 1 nm AlN interlayer from the Ga-face GaN bulk grown on a SiC substrate. On top of the $In_{0.17}Al_{0.83}N$ layer a 30 nm long gate metallization is applied in the middle of the device, equally spaced by 635 nm from the source and drain contact. Finally, the passivation of the $In_{0.17}Al_{0.83}N$ surface is obtained through a plasma treatment and the deposition of a 10 nm Al_2O_3 layer. The two

layers of $In_{0.17}Al_{0.83}N$ and AlN, which interface with GaN, constitute the barrier for this device. The order in which these two materials are deposited is very important for two reasons. First, as anticipated in Section 2.2, the AlN is the nitride compound with the highest spontaneous polarization which introduced to a high polarization discontinuity and consequently electron density at the Gan/AlN interface. Secondly, the Al layer reduces the effect alloy scattering in ternary alloys that has shown to degrade the electron mobility at low fields Wong *et al.* (2008). In fact, although the channel of electrons is formed in the GaN layer of the InAlN/GaN heterojunction, the wavefunction of electrons partially penetrates into the alloy layer feeling the effect of the perturbation of the periodic potential due to the substituting atom in the alloy crystal structure. The calculation of the polarization induced charge at each interface is show in Figure 2.4. Since the polarity of the crystal is Ga-face it can be noted that all the spontaneous polarization vectors point towards the bottom. Moreover, the piezoelectric polarization is present only in the AlN layer that is under tensile strain, while only the spontaneous polarization has to be considered for the relaxed GaN bulk and the lattice matched $In_{0.17}Al_{0.83}N$ layer.

In the device model for the simulation, sheets of charge are introduced to take into account the polarization discontinuity at each interface. Particular attention is due to the polarization charge at the free surface of the $In_{0.17}Al_{0.83}N$ layer where the polarization charge is partially neutralized by the passivation. This process is believed to be essential for the formation of the 2DEG, since the donor states at the free surfaces provide the electrons of the 2DEG Bernat *et al.* (2003). As a consequence the exact extent of the passivation, and consequently the value of charge to assign to the $In_{0.17}Al_{0.83}N$ free surface is determined by a fitting the charge density of the 2DEG electron gas. The physical properties that defines each material are initialized in a 144×95 inhomogeneous tensor-product grid that discretized the simulation domain

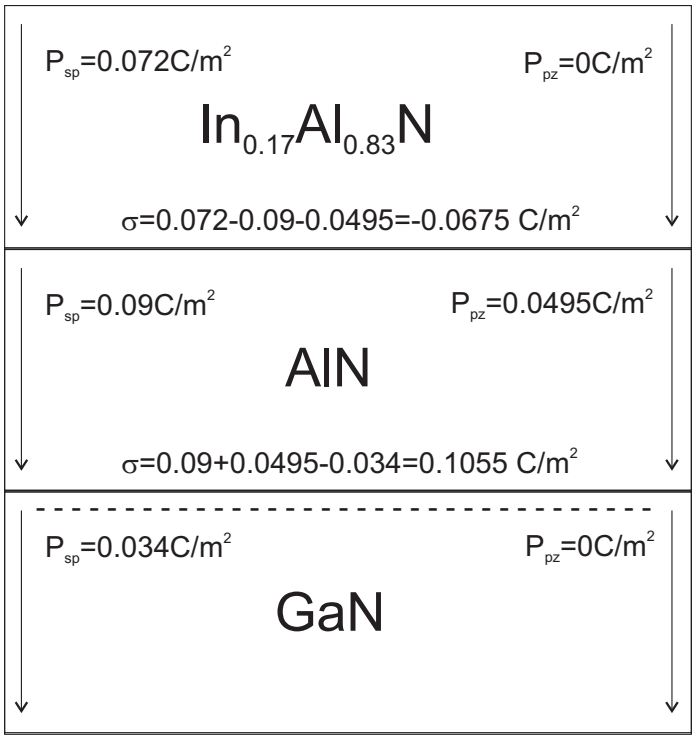

Figure 2.4: Schematic of the calculation of the spontaneous and piezoelectric polarization charge associated with the material interfaces of the HEMT in Figure 2.3. The 2DEG is formed between the AlN and GaN layer where the symbols "-" are drawn.

in real space shown in Figure 2.5.

The same grid is used by the multigrid Poisson solver with relaxation cycles both for the pre- and the post-smoothing Saraniti *et al.* (1996), to calculate the potential at each Poisson time step. The spacing of the grid and the transitions between the cell dimensions are fundamental for the Poisson solver to reach a converged solution Fischetti and Laux (1988). In particular, the dimension of adjacent cells was never varied of more than a factor of 2 and the length of the cell along the transport direction is kept smaller than the Debye length calculated at the local carrier concentration. Along the perpendicular direction, the dimension of the cells was calibrated to resolve the finer features of the device stack (*i.e.* 1 *nm* interlayer and 2DEG). Contacts (red boxes in Figure 2.5), are defined laterally to directly contact the 2DEG,

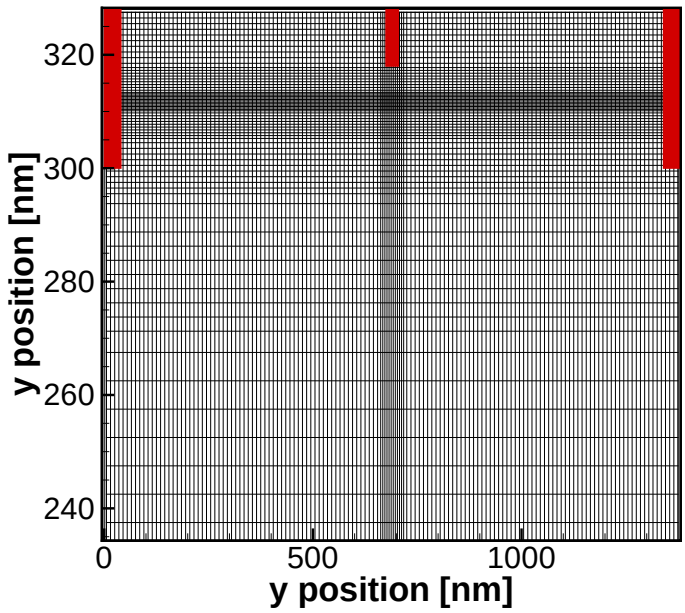

Figure 2.5: Real space mesh used to simulate the HEMT in Figure 2.3.

and their doping is used as a fitting parameter to reproduce the resistance of the source and gate ohmic contacts. The electron confinement induced by polarization charge and band discontinuity at the AlN/GaN interface lead to quantum phenomena (see Section 1.1) that are accounted by the calculation of the effective potential in the area of the channel. The band diagram and electron density in a vertical cross section of the device is shown in Figure 2.6.

2.4 DC Characteristics

The device model described in Section 2.3 is validated in Figure 2.7 and Figure 2.8, where the simulated $I_d - V_d$ and $I_d - V_g$ are compared with the experimental measurements.

In both cases, the fit results are excellent for lower currents, while for currents above 800 mA/mm simulations overestimate the current compare to measurement. This difference is thought to be be due to self heating in analogy to the observations reported in Guerra *et al.* (2011b) for the same range of current. In particular, for

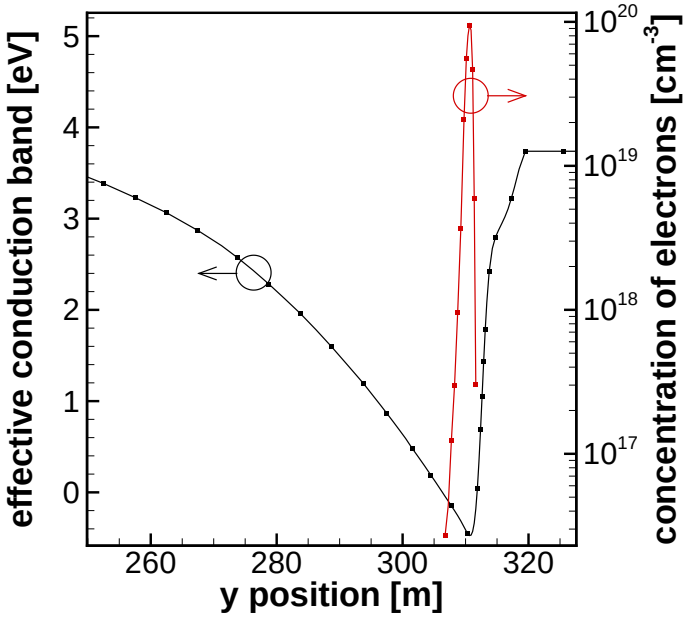

Figure 2.6: Band diagram of a vertical slice cut along the source access region. The effective conduction correspond to the black line with circles and the density of electron is shown in red with diamond symbols. The device bias is $Vgs = 2\ V$, $Vd = 5\ V$.

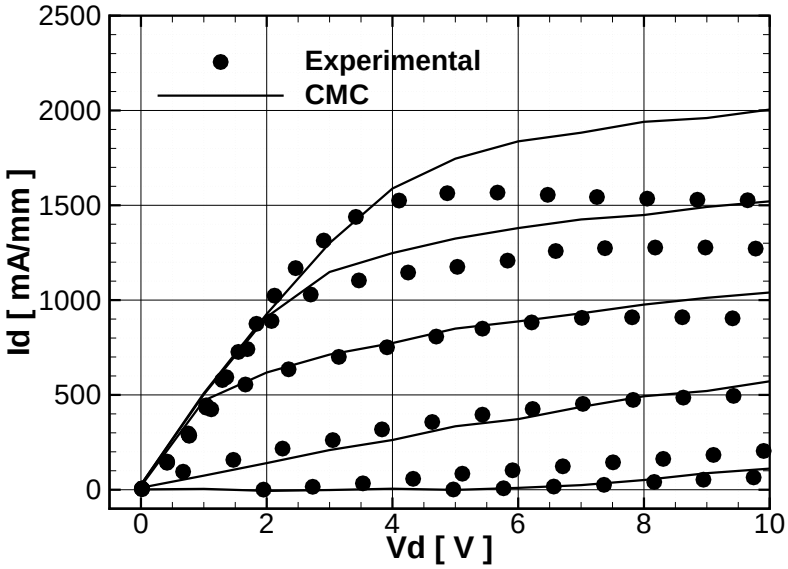

Figure 2.7: Comparison of the experimental (dots) and simulated (line), $I_d - V_d$ characteristic of the GaN HEMT in Figure 2.3. The lower curve is calculated at $Vgs = -4\ V$ while for the others Vgs was increased at steps of $1\ V$. The gate Schottky barrier is $0.8\ V$.

high values of current, more electrons lose part of their energy interacting with the lattice. This phenomenon can be modeled by the generating phonons, that can ac-

cumulate if the rate of generation is higher than the diffusion of phonons away from the active regions. As a result of this local accumulation, the scattering rate due to the absorption of phonons increases, reducing the simulated current towards the values that are measured experimentally. However, this effect is not accounted in the simulations work shown in Figure 2.7 and Figure 2.8. In fact, in order for self heating to be accurately modeled, the electrons dynamics should be coupled with the phonon dynamics, and this is not yet been implemented in our code.

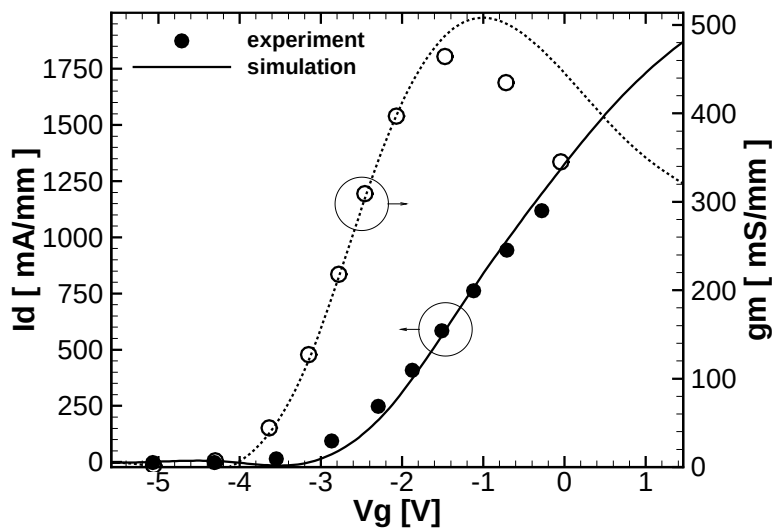

Figure 2.8: Comparison of the experimental (dots) and simulated (line), $I_d - V_g$ characteristic of the GaN HEMT in Figure 2.3. The gate Schottky barrier is 0.8 V and $V_d = 5\ V$.

2.5 Dynamics of Electrons in the Channel

The description of the microscopic dynamics of the carriers is essential to understand the macroscopic phenomena that influence the operation of devices. In particular, in HEMTs the device behavior is regulated by the transit of the electrons that are accelerated by the electric field in their path along the channel. However, since the device is not uniform along the transport direction, as a resistor would be,

the plot of the average properties of the electrons along the channel is often useful to understand the device DC and AC performance. Figure 2.9 shows the conduction band profile calculated along the channel of the GaN HEMT described in Section 2.3. Three areas can be distinguished along this plot: the central part, also called intrinsic area, with a small potential barrier followed by a large slope of the conduction band, and two lateral portions where the conduction band shows a constant slope in a resistor like behavior called access regions.

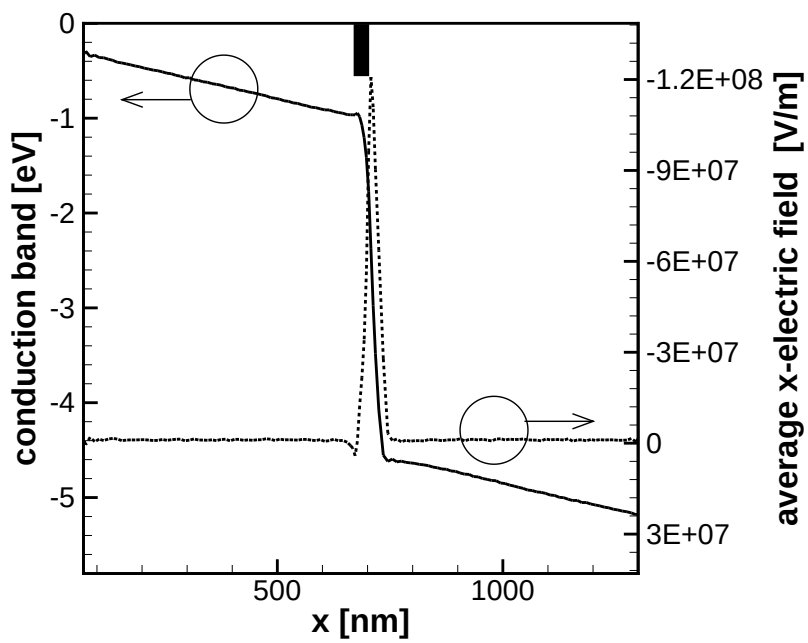

Figure 2.9: Profile of the conduction band and electric field component plotted along the transport direction in the channel. The gate is represented by the black rectangle on the top of the figure. Device biased at $Vgs = -2\ V\ Vd = 5$.

The big slope in the intrinsic part of the device is associated with a spike of the electric field component along the transport direction, as can be seen in Figure 2.9. As a result of this sudden increase in the electric field, electrons are accelerated up to velocities that are well above the saturation value of $3.5 \times 10^5\ m/s$ measured in GaN bulk. Figure 2.10 shows a zoom into the intrinsic region, were it can be seen that the

spike of the electric field occurs right in the proximity of the gate as does the velocity in the same plot. However, it can be seen that the peak of the velocity spike occurs before the peak of the electric field. This observation, that would be contradictory if electrons were moving in vacuum, is indeed explained by the distribution of the electrons in the momentum space.

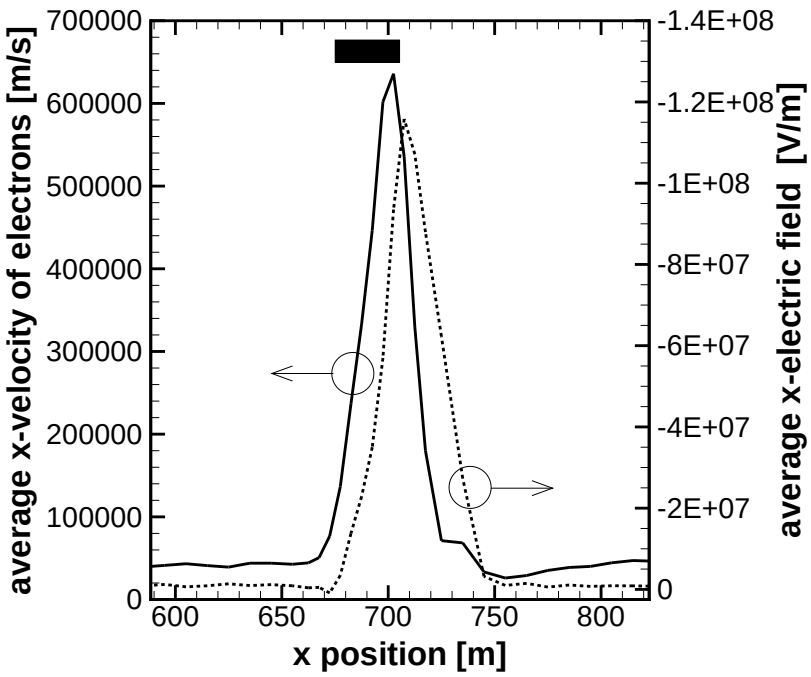

Figure 2.10: Profile of the x-velocity and x-electric field along the channel plotted around the gate represented by the black rectangle. x is the direction of transport. Device biased at $Vgs = -2\ V\ Vd = 5$.

In particular, in order to analyze the effects of the electric field on the average electron momentum, the position of the electrons in k-space is recorded in three areas of the device around the position of the peak electric field. However, since the momentum space is three dimensional, the electron distribution is shown only in two planes perpendicular to the k_z direction of the GaN Brillouin zone, corresponding to the position of the gamma and satellite valleys. By comparing the momentum distribution before, Figure 2.11a, and just after the peak of the electric field, Fig-

ure 2.11b, it can be noticed that part of the electrons in the gamma valley transfer to the satellite valleys. Moreover, since the electron effective mass is lower in satellite valleys than in the gamma valley, the transition of part of the electron population to the satellite valleys results in a decrease of the average electron velocity as observed in Figure 2.10 Choudhry and Arora (1986).

On the other hand, it can be seen in Figure 2.11c that the electron population of the satellite valleys does not vanish as fast as it forms. In fact, while the distribution in Figure 2.11c still show a moderate occupation of the satellite valleys, the distribution in Figure 2.11 that is calculated at the same distance from the peak electric but on the source side, does not show any electrons in the satellite vallyes. This asymmetry is due to the higher density of states of the satellite valleys compared to the gamma valley that favors scattering within the upper valleys. This behavior is reflected in Figure 2.12 by the profile of the energy along the channel, whose peak is delayed compared to the peak of the electric field. In fact, for the average energy of electrons to decrease, the rate of electron transitions towards the gamma valley has to exceed the rate at which the electric field pushes electrons into the satellite valleys, which in general can occur somewhere after the peak of the electric field.

The profile of the total scattering rate along the channel is shown in Figure 2.13, where it can be noticed that its peak is delayed compared to the electric field, similarly to the energy profile. In fact, as for the energy, the scattering increases with the population of the satellite valleys where the final states available for a scattering event are increased by the high density of states in the satellite valleys.

The more detailed analysis of the scattering in Figure 2.14 shows which of the mechanisms are the dominant. The importance of distinguishing between the scattering mechanisms is related to the thermal management of the device. In particular, it can be seen in Figure 2.14 that 20 nm aside from the drain end of the gate a large

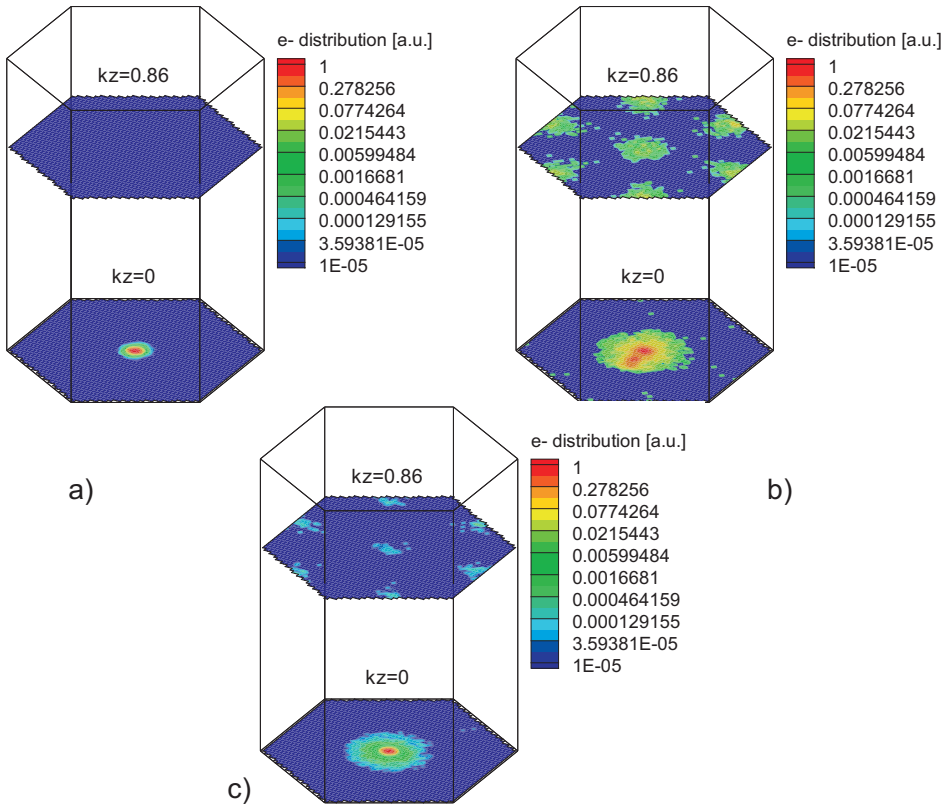

Figure 2.11: Average carrier distribution density in the momentum space in different positions long the channel. Subfigure a) is calculate from 50nm to 40nm before the electric field peak, Subfigure b) from the electric field peak to 10nm after it, and Subfigure c) 50nm to 40nm after the electric field peak. The 3D Brillouin zone (the extension is normalized in the kz-direction from -1 to +1) is sliced on a plane corresponding to the minimum of the gamma valley (kz = 0), and on another plane corresponding to the minimum of the satellite valleys (kz = 0.87). Device biased at $Vgs = -2\ V\ Vd = 5$.

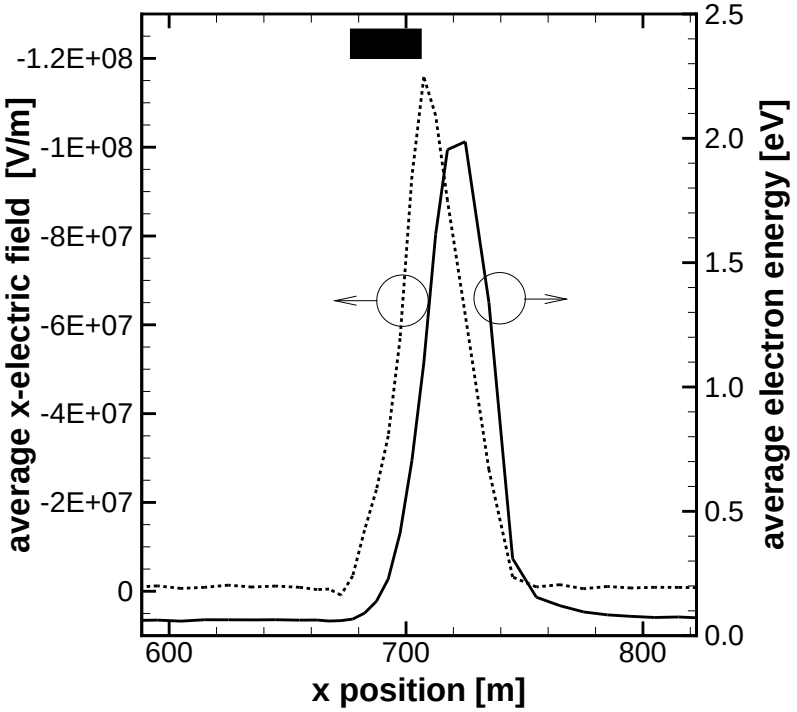

Figure 2.12: Profile of the average electron energy and x-electric field along the channel plotted around the gate represented by the black rectangle. x is the direction of transport. Device biased at $Vgs = -2\ V\ Vd = 5$.

emission of phonons takes place. This emission of phonons represents the energy that the energetic electrons due to the spiking electric field give up to the lattice in the form of heat. When this heat is not dissipated, but keeps accumulating in an area of the device, the local temperature increases and a hot spot is formed. The dissipation of the energy can be described by the diffusion of the phonons which varies according the mechanism that generated the phonon. In particular, the polar optical phonons that dominates in this device, are characterized by a low group velocity Pop *et al.* (2003) such that most of the heat is dissipated only after the optical phonons decay in acoustic modes Matulionis *et al.* (2004). Since the formation of hot spots in devices have been shown to negatively impact the transport of electron Rowlette *et al.* (2005), Ardaravicius *et al.* (2003), simulators of the phonon population and dynamics are under development, but are not available for this study.

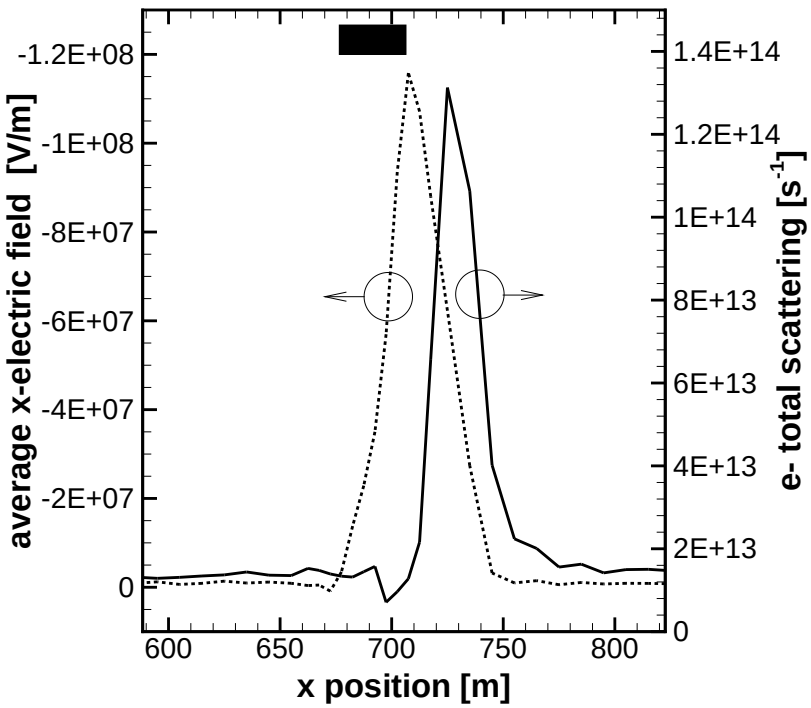

Figure 2.13: Profile of the total electron scattering and x-electric field along the channel plotted around the gate represented by the black rectangle. x is the direction of transport. Device biased at $Vgs = -2\ V\ Vd = 5$.

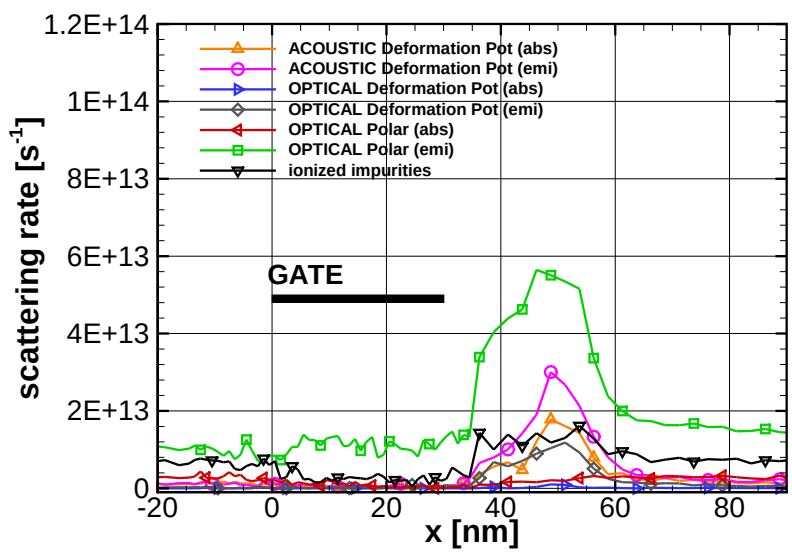

Figure 2.14: Profile of the scattering rates of the relevant scattering mechanisms along the channel. The gate represented by the black rectangle. Device biased at $Vgs = -2\ V\ Vd = 5$.

2.6 Two Port Network Small Signal Model

One representation that can be given to a device is the two port model, where each port is characterized by a current and a voltage between its terminals Pozar (2005). The relation between these four quantities, i_1,i_2,v_1,v_2, represents a black box description of the device and can be provided in different forms. One of the most common forms is the admittance matrix that relates the currents to the voltages of the two ports:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \tag{2.11}$$

where the elements of the matrix, called Y-parameters, are defined as:

$$y_{11} \equiv \frac{i_1}{v_1}\bigg|_{v_2=0}, \tag{2.12}$$

$$y_{12} \equiv \frac{i_1}{v_2}\bigg|_{v_1=0}, \tag{2.13}$$

$$y_{21} \equiv \frac{i_2}{v_1}\bigg|_{v_2=0}, \tag{2.14}$$

$$y_{22} \equiv \frac{i_2}{v_2}\bigg|_{v_1=0}. \tag{2.15}$$

When this representation is applied to a FET in common source configuration, port number 1 is referred to the gate and port number 2 to the drain. The Y-parameters thus become:

$$y_{11} \equiv \frac{i_g}{v_{gs}}\bigg|_{v_{ds}=0}, \tag{2.16}$$

$$y_{12} \equiv \frac{i_g}{v_{ds}}\bigg|_{v_{gs}=0}, \tag{2.17}$$

$$y_{21} \equiv \frac{i_d}{v_{gs}}\bigg|_{v_{ds}=0}, \tag{2.18}$$

$$y_{22} \equiv \frac{i_d}{v_{ds}}\bigg|_{v_{gs}=0}. \tag{2.19}$$

The value of each Y-parameter is in general a complex number that is a function of the frequency of the applied AC signal, as well as the bias point of the device. Since in this representation, the device is considered to behave linearly, the amplitude of the AC signal has to be small enough to avoid non linear response; hence the name "small signal parameters" given to Y and equivalent sets of parameters. A quantity of interest that can be calculated from the Y-parameter is the short circuit current gain:

$$h_{21} = \frac{Y_{21}}{Y_{11}} = \frac{i_d}{i_g}. \tag{2.20}$$

In particular, from the short circuit current gain the cutoff frequency (f_T) can be extracted as the value of frequency where the magnitude of the short circuit current gain is one. The cutoff frequency represents a common figure of merit used to evaluate and compare the microwave performance of devices. In order to calculate f_T, a plot of the short circuit current gain versus frequency is necessary, which means that the Y-parameters have to be calculated over multiple frequencies. However, it can be seen from the definition of the Y-parameters in 2.16 that two simulations are required to calculate a full set of Y-parameters: one applying the AC signal to gate while keeping the drain potential constant, which is equivalent to a short circuit in AC, and the other applying the AC signal to the drain while keeping the gate potential constant. Moreover, since the Y-parameters are defined as a function of the AC amplitude, the Fourier transform is applied to the voltages and currents in the simulation to calculate the short circuit current gain as in 2.16. This analysis can be repeated for all the different frequencies of interest using a single sinusoid as excitation signal but, as a condition to maintain the linear response of the device, it is possible to use a

polychromatic excitation signal and then separate the contributions at different frequencies with the Fourier decomposition method Laux (1985). However, since the amplitude of the input AC signal has to remain small, the amplitude of each single sinusoid composing the signal has to decrease as the number of sinusoids increases. This can represents a problem when the amplitude of each sinusoid is comparable to the amplitude of the noise. Therefore, a trade-off is present in the choice of the excitation signal between the better signal to noise ratio achievable simulating signal with reduced spectral content, and the extent of spectrum coverage given by signal of infinite spectral content like a step perturbation.

In Figure 2.15 the short circuit current gain derived from the Y-parameters is calculated for the GaN HEMT described in Section 2.3. In particular, the Y-parameters are calculated with two different excitation signals, a step perturbation a multisinusoid signal with the fundamental harmonic at 12.5 GHz and the first 3 harmonics. It can be seen that the short circuit current gain calculated with the two excitations overlap almost completely, which confirms that the amplitude of the excitation signal have been appropriately chosen according to the criteria discussed above. However, the main difference using one or the other excitation regards the post simulation processing required to obtain the results in Figure 2.15. In the case of the multisinusoid signal the current gain at each frequency is derived directly from the Fourier decomposition as described above. On the other hand, when the Fourier decomposition method is applied to the step perturbation, the algorithm shows numerical instabilities. For the resolution of these instabilities, a windowing technique is applied to the signals to decompose, consisting in eliminating part of the samples. The uncertainty introduced by this process is usually overcome knowing the type of frequency response to expect, which can be suggested by experimental measurement, when available, or by calculations based on equivalent circuits (see Section 2.7). In the case of FETs,

for example, the short circuit gain has been consistently measured to decrease with a -20dB/dec slope as can be seen also in Figure 2.15. Considering Eq. 2.20, this behavior can be explained in terms of an increase of the current through to the gate capacitor whose associated impedance decreases when the frequency is increased.

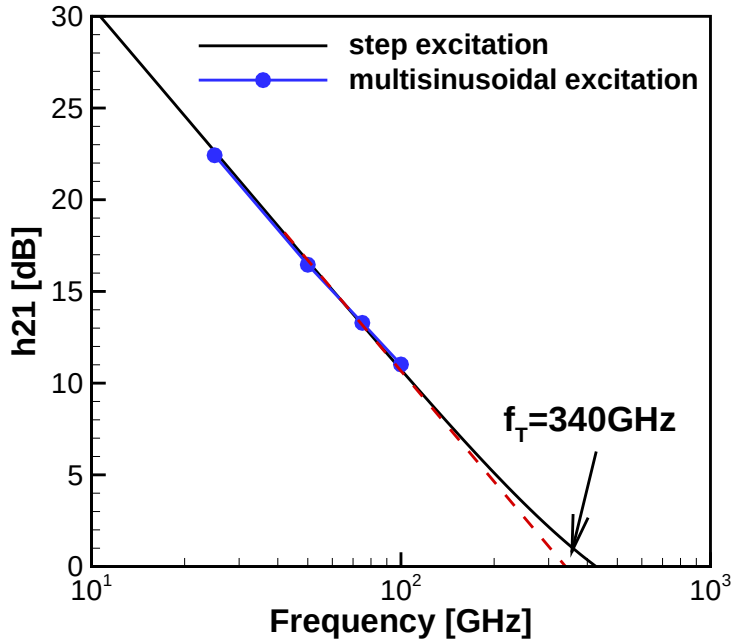

Figure 2.15: Short circuit current gain calculated with AC simulations using a step like perturbation (black solid line) and multisinusoidal excitation (blue circles). The red dashed line shows the -20dB/dec slope line used for the extraction of f_T.

2.7 HEMT Small-Signal Equivalent Circuit

Another microwave model used in the design of analog and integrated circuit is the small signal equivalent circuit. This model substitutes the black box representation of the device by an equivalent circuit of physically meaningful elements like resistors, capacitors and generators. The most common equivalent circuit Berroth and Bosch (1990) conceptually valid for both FETs and BJTs, is shown in Figure 2.16.

The idea behind the equivalent circuit is to represent the nanoscale processes oc-

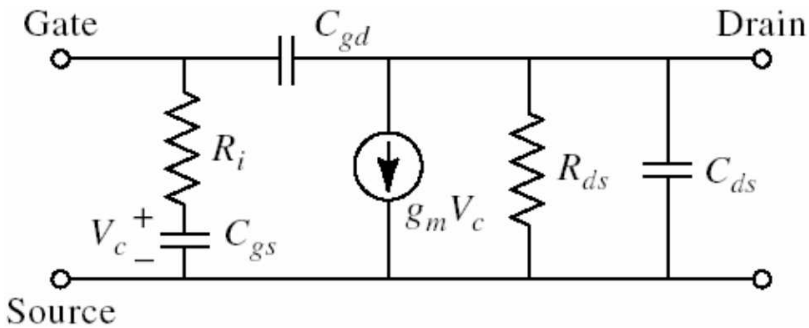

Figure 2.16: Small ignal circuit of the intrinsic FET.

curring inside the device in terms of circuit elements. For example, in Figure 2.16, a model of the intrinsic part of the device is shown, where the effect of the contact resistances and other extrinsic parameters is not considered. In this description, the gate capacitance is associated with the gate oxide, while the gate resistance represents the losses along the current path across the barrier. The drain to gate capacitance represents the electrostatic influence of the drain, and the source to drain resistance takes into account the lack of current saturation for Vd. As anticipated, the same equivalent circuit is valid for BJTs as well. However, since the processes inside the device are different, the circuital elements assume different meanings. From an operative point of view, the values of the circuital elements can be determined by experiments or simulations once the equivalent circuit is chosen. For example, the circuital elements in Figure 2.16 can be related to the Y-parameters by:

$$Y_{11} = \frac{j\omega C_{gs}}{1 + j\omega R_{gs} C_{gs}} + j\omega C_{gd}, \tag{2.21}$$

$$Y_{21} = \frac{g_m}{1 + j\omega R_{gs} C_{gs}} - j\omega C_{gd}, \tag{2.22}$$

$$Y_{12} = -j\omega C_{gd}, \tag{2.23}$$

$$Y_{22} = \frac{1}{R_{ds}} + j\omega(C_{ds} + C_{gd}). \tag{2.24}$$

Using the definition of f_T given as a function of the Y-parameters in 2.20, the

short circuit current gain of the equivalent circuit in Figure 2.16 can be written as:

$$h_{21} = \frac{Y_{\frac{g_m}{1+j\omega R_{gs}C_{gs}} - j\omega C_{gd}}}{\frac{j\omega C_{gs}}{1+j\omega R_{gs}C_{gs}} + j\omega C_{gd}}. \tag{2.25}$$

and f_T can be approximated to Sze and Ng (2007):

$$f_T \approx \frac{g_m}{2\pi(C_{gs} + C_{gd})} \approx \frac{g_m}{2\pi C_G}, \tag{2.26}$$

where C_G is the total FET gate capacitance that can be rewritten as a function of the oxide capacitance and device dimension as:

$$C_G = ZLC_{ox}, \tag{2.27}$$

.

Using the typical of relation for FET gm=$C_{ox}v/L$ and substituting 2.27 into 2.26 f_T can be written as:

$$f_T = \frac{g_m}{2\pi ZLC_{ox}} = \frac{v}{2\pi L}, \tag{2.28}$$

where v is the average velocity of electrons along the gate length L.

2.8 Physical Meaning of f_T

Section 2.7 was concluded obtaining Eq 2.28 that relates the cutoff frequency to the average transits time of electrons under the gate. This relation predicts f_T to increase linearly with the inverse of the gate length. However, significant departures from this prediction have been seen for deep submicrometer devices Kim and del Alamo (2008a). It is now accepted that such departures are due to the fringing fields of the gate electrode whose charge control extends laterally compared to its vertical

projection on the channel. In order to take this into account Eq. 2.28 was modified in a more accurate formulation as Akis *et al.* (2008):

$$\tau_T = \frac{1}{2\pi f_T} = \int_{L_{eff}} \frac{1}{v_{ave}(x)} dx, \tag{2.29}$$

where L_{eff} is the effective gate length, and $v_{ave}(x)$ is the velocity calculated along the channel (here assumed as the x-direction) obtained as a weighted average along the y-direction with respect to the carrier density. Beside the introduction of the integral, due to the fact that the velocity is not constant under the gate (see Figure 2.10), the element of interest in the formulation in Eq 2.29 is the domain of integration which occurs along an effective gate length instead of the metallurgic gate length. The determination of L_{eff} is thus the crucial element for the reliability of Eq 2.29. This topic has been source of controversy with a first approach that considered the full depletion of carrier underneath the gate as effective gate length Wu *et al.* (2006) and a following study by Akis *et al*. that showed this definition to underestimate the cutoff frequency Akis *et al.* (2008) as can be seen in Figure 2.17.

According to Akis, the effective gate length is calculated from the acceleration point of the electrons on the source side of the gate to the peak of the velocity. In particular, the choice of the acceleration point is justified because it shows where electrons start to respond to the gate modulation, while the velocity peak is found to define L_{eff} such that the f_T calculated with Eq 2.29 matches the value calculated from the short circuit current gain. An explanation that can justify the end point of the effective gate length is that once an electron reaches that the position of the velocity peak it is already subjected to the strong electric field of the drain so by that point the gate has already lost the property to modulate the current. Moreover, since the peak velocity is associated with considerable valley transfer, it is possible that the higher effective mass of electrons decreases to capability to be modulated by the

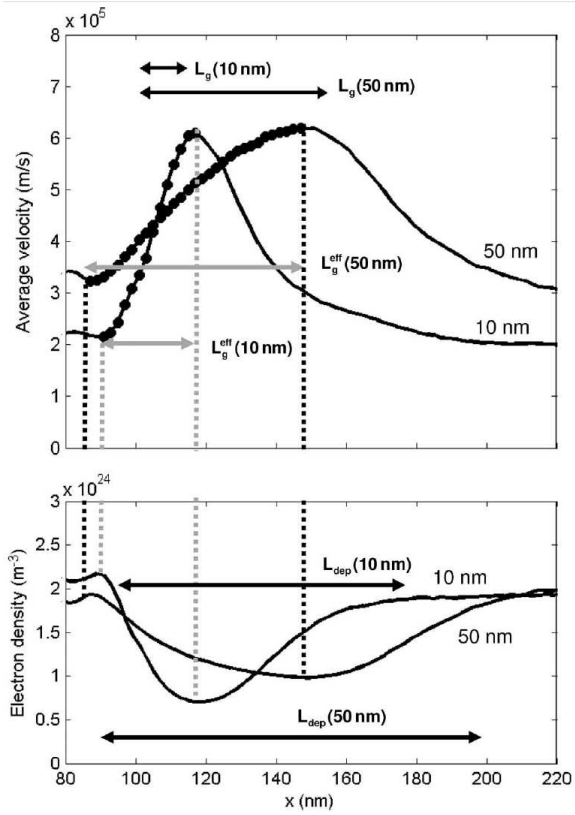

Figure 2.17: Upper panel: Average electron velocity along the channel of a 10nm and 50nm InP HEMT. The gray and black dashed lines indicate the beginning and end points of the effective gate length calculated according the criteria suggested by Akis in his work. Lower panel: The corresponding average electron density with estimated depletion lengths indicated. Akis *et al.* (2008)

gate.

2.9 Empirical Determination of the Effective Gate Length

While the ending point of the effective gate length is defined with precision in Section 2.8 the noisy profile of the velocity in the access region can makes it difficult to see where the acceleration point is located. For this purpose, an alternative method based on the modulation of the electric field orthogonal to the channel is proposed in this section. The idea of the following analysis derives from the very fundamental description of the operation of the intrinsic part of the HEMT: a gate controlled

potential barrier that modulates the flux of electrons flowing from the source to the drain. Within this frame, the DC component of a signal applied to the gate sets the height of the potential barrier, while the AC component periodically raises and lowers the barrier around the DC value. The coupling between the gate contact and the barrier in the channel occurs through the electric field, and in particular the y component, transverse to the channel, since the gate is placed above the structure. When an AC signal is applied to a device biased in common source configuration, the gate potential oscillates between a maximum and a minimum and so does the electric field. Therefore, the change of the y-electric field along the channel can be seen as the direct effect of the perturbation of the gate potential. Figure 2.18 shows the profile along the channel of the difference between the y-electric field calculated at $-2\ V$, which would corresponds to the DC level of the signal, and the y-electric field calculated at the upper and bottom peak of the AC signal, respectively $-1.9\ V$ and $-2.1\ V$.

As it may be expected, the two curves peak in proximity of the gate, where its influence is larger, and then they converge to zero towards the contacts, with a significant noise overlapped. The arrow in Figure 2.18 shows where the beginning of the gate length has to be for the f_T calculated from Eq 2.29 to match what is calculated from the Y-parameters. It is important to notice that this values is the reference to validate any methods developed to find the beginning the effective gate length. Figure 2.19 shows a close up around the gate of the field modulation shown in Figure 2.18.

The reference effective gate length is indicated in the plot and it can be seen that it corresponds to the intersection of the parabolic interpolation of the field modulation. This empirical observation as been validated by several simulations of different layouts (some reported in Section 2.9.1), and represents a univocal way the calculate L_{eff}.

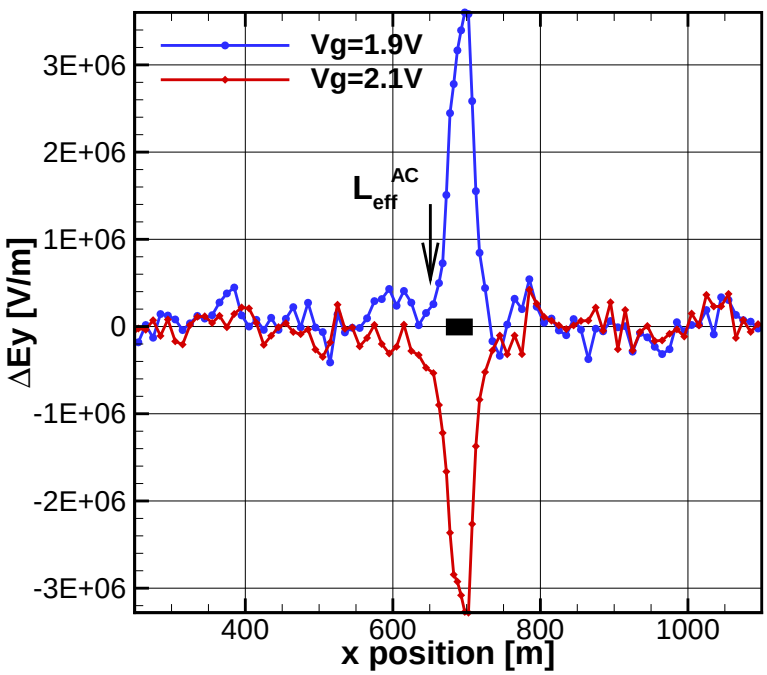

Figure 2.18: Profile of the modulation (see text) of the y-electric field along the channel. The reference beginning of the effective gate length is shown with the black arrow while the black rectangle represents the metallurgic gate. Device biased at $Vgs = -2\ V\ Vds = 3.5\ V$.

As a comparison, the velocity profile is plotted as well and it can be seen that using the electron acceleration point would lead to significantly underestimate the effective gate length.

2.9.1 Effect of Dielectric on the Effective Gate Length

Despite the introduction of the parabolic interpolation, which did not find a rigorous theoretical explanation, the procedure described in Section2.9 showed to be a reliable empirical method to predict the extension of the effective gate length on the source side. The new definition of effective gate length is applied in this section to different device layouts to understand the impact that the dielectric constant of the passivation have on the cutoff frequency. In all the cases studied in this Section the

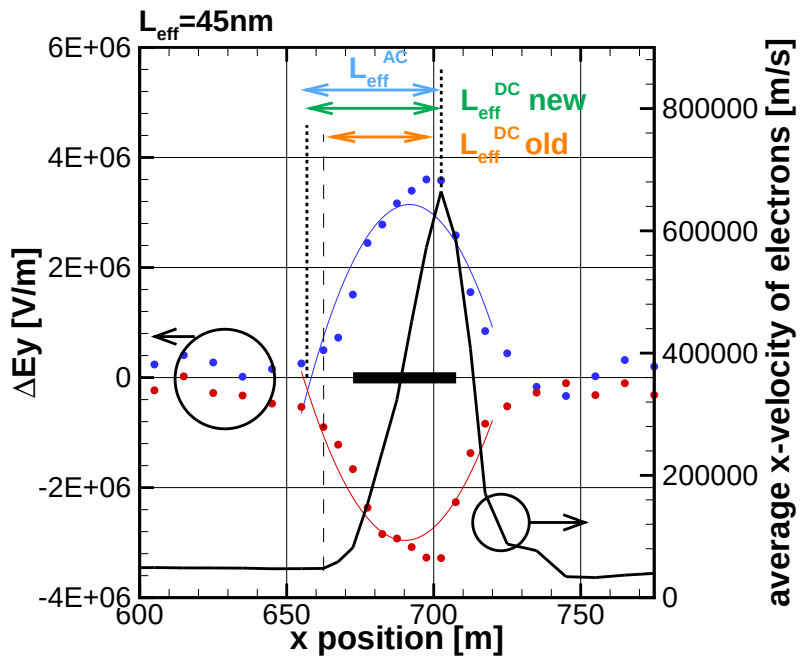

Figure 2.19: (Symbols) Focus around the gate area of the profile of the modulation (see text) of the y-electric field along the channel. The red and blue lines represents a quadratic interpolation of the modulation under the gate. The black line shows the velocity profile in the direction of transport (x direction). L_{eff}^{AC}, $L_{eff}^{DC}new$, $L_{eff}^{DC}old$ represent respectively: the reference effective gate length extracted by AC simulations, the effective gate length calculated from the modulation of the y-electric field, and effective gate length calculated from the electrons acceleration point. Device biased at $Vgs = -2\ V$ $Vds = 3.5\ V$.

effective gate length calculated from the modulation of the electric field shows an excellent match with the reference effective gate length as in defined in Section2.9, therefore only one L_{eff} will be indicated in the plots. The layouts analyzed in this section consists of six different modifications of the HEMT in Figure 2.3 where the dielectric constant of the 10 nm Al_2O_3 layer was triplicated, or reduced to a third, first only on the source side, then only on the drain side and finally on both sides. The related f_T calculated with AC simulations are shown in Table 2.1:

As it can be seen, both the reduction of the dielectric constant on the source side and on the drain side increases f_T, and the effect is maximized when the dielectric

	ϵr source	ϵr drain	ϵr AR
3	382 GHz	389 GHz	407 GHz
27	345 GHz	321 GHz	313 GHz

Table 2.1: f_T of the layouts where the dielectric contact was changed from 9 to 3 or 27 only on the passivation on the source side ϵr source, drain side ϵr drain, or in both sides ϵr AR.

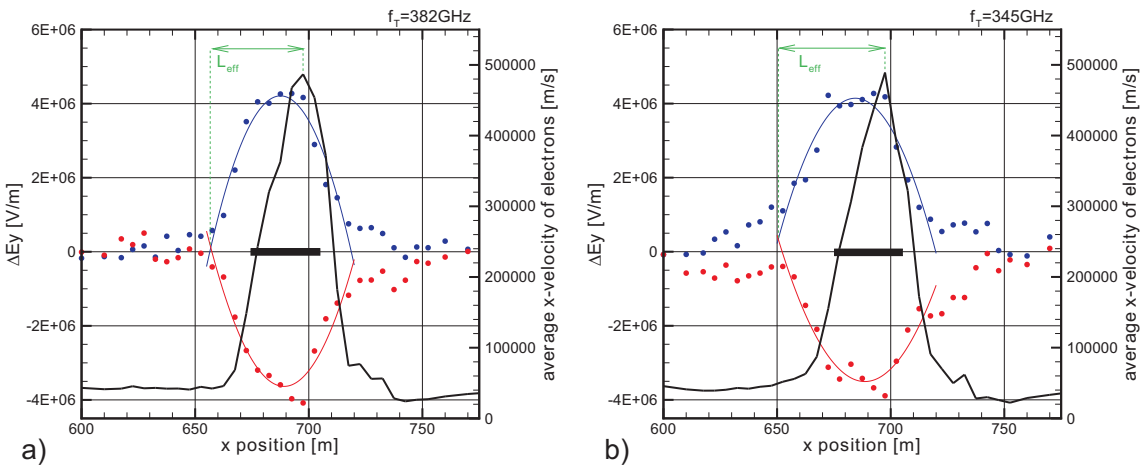

Figure 2.20: Influence of the dielectric constant of the passivation layer on the source side of the gate on the effective gate length calculated as in Section 2.9. The dielectric constant of the passivation layer on the source side is 3 in subfigure a) and 27 in figure b) while on the drain side it is 9 for both. The blue and red lines represent the parabolic interpolation of the modulation of the y-electric field (see Section 2.9). The velocity profile (black solid) is also shown. Device biased at $Vgs = -2\ V$ $Vds = 3.5\ V$.

constant is reduced on both sides. Vice versa, the increase of the dielectric constant causes f_T to decrease. The effect of the dielectric constant on the source side on L_{eff} is investigated in Figure 2.20.

Comparing Figure 2.20a with Figure 2.20b it can be seen that an increase in the dielectric constant expands the extension of the effective gate length towards source which explain the decrease f_T seen in Tab 2.1. The reason for the extension of L_{eff} is due to the reduction of confinement of the electric field propagating in materials with

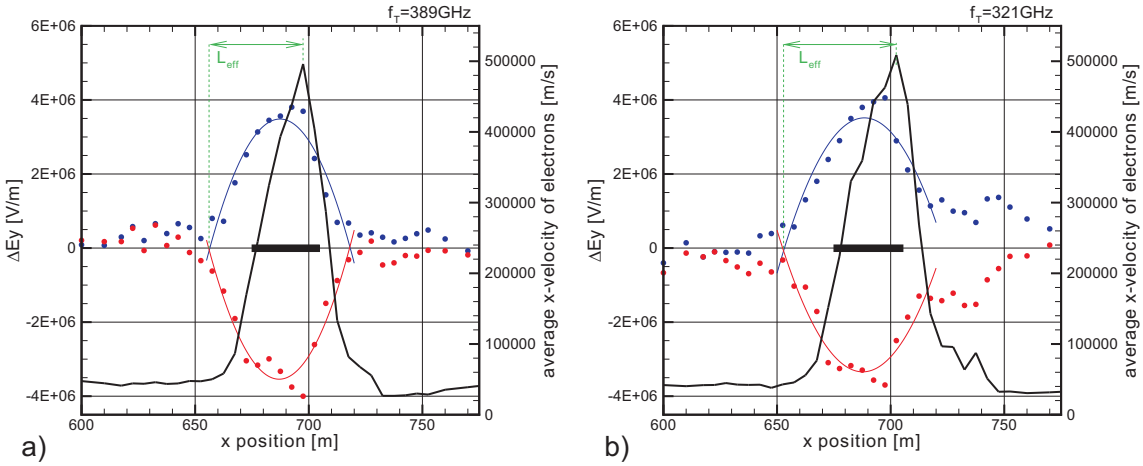

Figure 2.21: Influence of the dielectric constant of the passivation layer on the drain side of the gate on the effective gate length calculated as in Section 2.9. The dielectric constant of the passivation layer on the drain side is 3 in subfigure a) and 27 in figure b) while on the source side it is 9 for both. The blue and red lines represent the parabolic interpolation of the modulation of the y-electric field (see Section 2.9). The velocity profile (black solid) is also shown. Device biased at $Vgs = -2\ V$ $Vds = 3.5\ V$

higher dielectric constant, and therefore the increase of the extension of the fringing field. Figure 2.20 reports the comparison between layouts where the passivation dielectric constant has been changed only on the drain side.

In this case, the beginning of the effective gate length does not seem to change, meanwhile the peak of the velocity can be seen to shift towards the drain end of the gate when the dielectric constant is increased. The increase of the f_T calculated with the reduction of the passivation dielectric constant on drain side can therefore be related again to a decrease of the effective gate length. The last two layouts, where the dielectric constant was changed in all the passivation layer is shown in Figure 2.20.

As it can be seen, the two subfigures in Figure 2.22 differ in both the beginning and the end of the effective gate length combining effects seen in Figure 2.20 and Figure 2.21 for the single side alteration. This result is coherent with the f_T reported in Table 2.1 showing that maximum effects are obtained when the passivation dielectric constant is reduced or increased on both sides, and the in order to maximize f_T

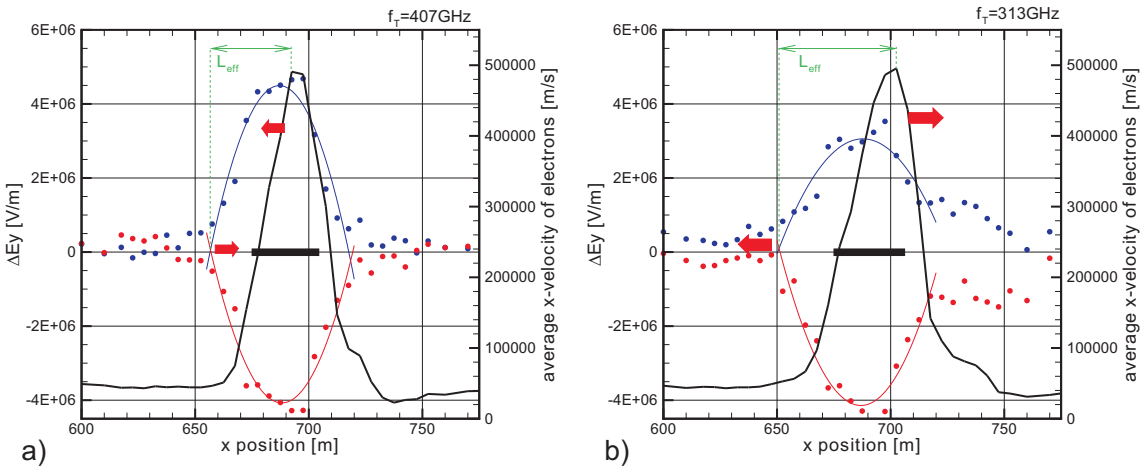

Figure 2.22: Influence of the dielectric constant of the passivation layer on both sides of the gate on the effective gate length calculated as in Section 2.9. The dielectric constant of the passivation layer is 3 in subfigure a) and 27 in figure b).. The blue and red lines represent the parabolic interpolation of the modulation of the y-electric field (see Section 2.9). The velocity profile (black solid) is also shown. Device biased at $Vgs = -2\ V\ Vds = 3.5\ V$

the passivation with the lower dielectric constant should be adopted to reduce the effective gate length as much as possible.

2.10 Conclusions

In this chapter the fundamental aspects of GaN HEMTs modeling for Monte Carlo Simulations are described, proposing a simulation approach that is validated with the analysis of the simulation results. The first part of the chapter is dedicated to the device modeling for DC performance analysis. The crucial aspect of polarization in GaN heterostructures is initially introduced within a theoretical frame to be then applied to the specific HEMT considered in this chapter to test the simulation approach. The experimental layout of this HEMT is explained in detail together with the numerical issues deriving from a particles that have to be considered in the simulation setup file. The simulation approach if then validated through the fit of the experimental IdVd and IdVg curve, and the observation of phenomena related the carrier dynamics along

the channel. The second part of the chapter is dedicated to the modeling of HEMT frequency response. A small signal description was introduced based on the two port y-parameters. The concept of cutoff frequency was introduced in respect to the y-parameters first and then to an alternative representation of the device microwave behavior based on an equivalent circuit. A physical interpretation of f_T was reported based on the argument of the gate controlled area called effective gate length. This topic was analyzed in detail and a new method to define the effective gate was presented. Based on this definition the effect of the passivation dielectric constant was studied and it was found that higher cutoff frequencies can be obtained reducing the extension of the fringing field associated with a low dielectric passivation.

Chapter 3

EFFECT OF ACCESS REGION SCALING IN GAN HEMTS

3.1 Introduction

For the last decades the scaling of the gate length has shown to be provide the largest increase in the cutoff frequency of HEMTs. However, as the gate reaches deep sub micrometer dimensions, shorts channel effects have shown to deteriorate the performance of the device is not addressed properly Kim and del Alamo (2008a). The suppression of the short channel effects was calculated to be possible maintaining the ratio between the gate to channel distance and the gate length below certain thresholds depending of the materials of the stack Guerra *et al.* (2010). This implies that when the gate is shortened the thickness of the dielectric layer has to be decreased as well, which has the undesired effect to increase the gate leakage. Moreover, the reduced gate to channel distance is found to induce depletion of electrons in the channel 2DEG Tilak *et al.* (2001). Therefore, alternative approaches to increase the cutoff frequency in HEMT have became of interest; examples are the scaling of source to gate and gate to drain distance, called access regions. Experiments by Shinohara *et al.* measured a 25% increase in f_T by shortening the gate to drain access region from 1 μm to 170 nm attributing this improvements to the reduction of the drain delay time Shinohara *et al.* (2010). Palacios *et al.* experimented the effect of shortening the source to gate distance with the intent of reducing the dynamic resistance to improve the lineary of f_T for high drain currents Palacios *et al.* (2005). Matsui *et al.* compared the the g_m, f_T and f_{MAX} of a self aligned T-gate GaN HEMs differing among them only by the suorce to drain spacing Yamashita *et al.* (2011). They

found that g_m increases monotonically decreasing the access regions length while f_T and f_{MAX} peaked for the intermediate value of the source to drain spacing. A Monte Carlo simulation study was performed by Di Carlo *et al.* with the aim of understanding the influence of the access regions length on the performance of AlGaN/GaN HEMTs Russo and Di Carlo (2007). While they found the source access region to considerebly increase the current and g_m, the drain access region did not show the any impact. The influence on the AC performance was not studied by them with equal detail. Only a qualitative observation was presented claiming that despite increasing g_m the scaling of the source access region was not belived to increase f_T because of the simultaneous increase of the gate to source capacitance. In this chapter the effect of symmetrically and asymmetrically reducing the access regions length is studied in terms of DC and AC performances. Afterwards, the effect of lateral scaling will be exploited in all its potential to derive a limiting cutoff frequency for GaN-HEMT. The limiting value we calculate, completes the estimate suggested by Akis *et al.* Akis *et al.* (2008) including the effect of the access region. Moreover, the limiting frequency provided here is a more accurate value than previous estimetes based on overshoot velocities Foutz *et al.* (1999) or effective gate length corresponding to the length of the whole depletion under the gate Wu *et al.* (2006).

3.2 DC Scaling

The impact of scaling the access region has been initially evaluated in relation to the DC characteristic. In particular, Figure 3.1 shows the evolution of the $Id - Vg$ as the source access region is decreased. In this set of simulations, the geometry of the device was unchanged (*i.e.* same gate length and gate to channel distance) except for the source to gate and gate to drain distance that was symmetrically reduced from the original 635 nm to 50 nm at steps of 100 nm. As it can be seen Figure 3.1

the current increases monotonically shortening as the access regions. Moreover, the threshold voltage remains constant with the scaling and so g_m increases with the current. On a quantitative note, both the current and g_m increase of almost a factor of 3 when the access regions are reduced from 635 nm to 50 nm.

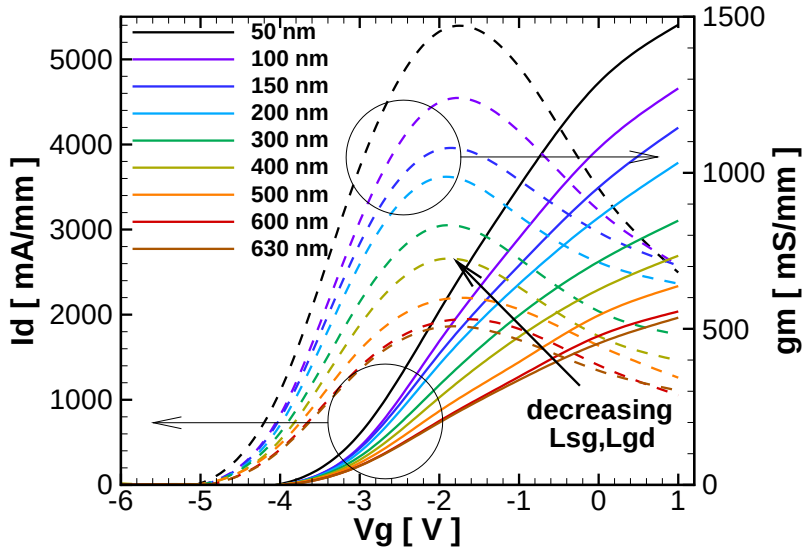

Figure 3.1: Comparison of $Id - Vg$ and g_m for different source-gate, gate-drain access region lengths corresponding to different colors in the legend and decreasing according to the arrow. The devices are biased in saturation at $V_{DS} = 5\ V$.

In order to get a better inside on the effect of the access regions on the DC performance, we simulated the $I_d - V_g$ for asymmetrically scaled devices. In particular, Figure 3.2 and Figure 3.3 shows the effect of shortening respectively only the source to gate and the gate to drain distance. While in Figure 3.2 the scaling of the source access region shows the same qualitative effect of the symmetrical scaling, Figure 3.3 does not show any effects resulting from shortening the drain access region. However, a quantitative analysis reveals that the increase of current and g_m resulting from scaling the source access region alone is only 70% the incresed obtained by a symmetrical reduction of both the access regions. Therefore, despite not showing any effect in Figure 3.3 the scaling of the drain access region becomes relevant when it is done in

conjunction with the scaling of the source access region.

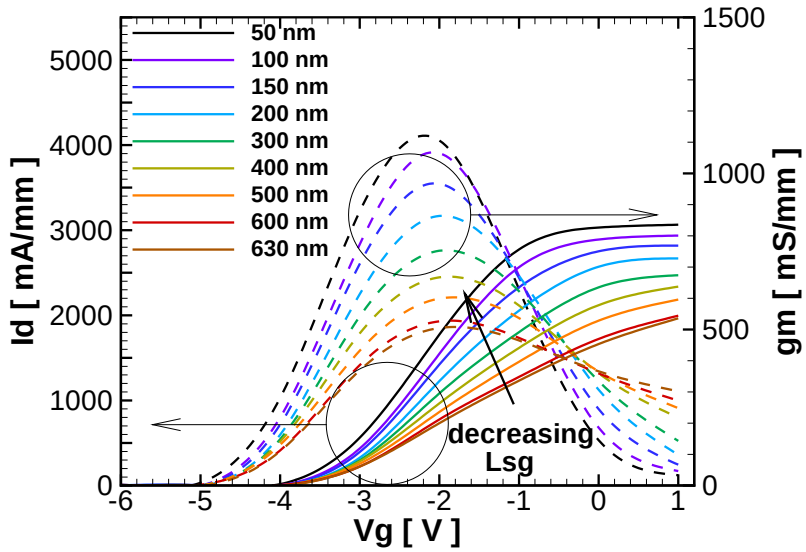

Figure 3.2: Comparison of $Id - Vg$ and g_m constant gate to drain distance and different source-gate access regions lengths corresponding to different colors in the legend and decreasing according to the arrow. The devices are biased in saturation at $V_{DS} = 5\ V$.

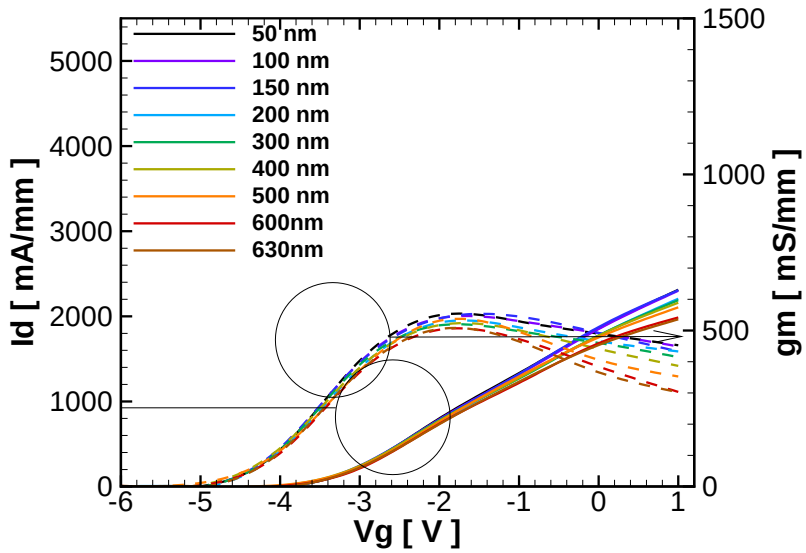

Figure 3.3: Comparison of $Id - Vg$ and g_m constant gate to source distance and different gate-drain access regions lengths corresponding to different colors in the legend and decreasing according to the arrow. The devices are biased in saturation at $V_{DS} = 5\ V$.

In order to understand these results, the access region of the HEMT are modeled

as series resistors between the contacts, where the potential that is shown in the x-axes of the $I_d - V_g$ is applied, and the intrinsic part of the device where a potential barrier modulates the flux of electrons. According to this model, the threshold voltage correspond to a barrier small enough that electrons with higher energy can overcome it. The height of the barrier depends in first approximation, *i.e.* neglecting second order effects like DIBL, on the source to gate voltage in the intrinsic area of the device. In general, the intrinsic source to gate voltage is smaller than the what is directly applied to the contact because, a fraction of the voltage is dropped on the source access resistance which depends on the source to gate distance. However, when the device is off, there is no current flowing through the device and no voltage is dropped across the access regions. In this condition, the intrinsic voltage is equal to the voltage applied at the contacts and the barrier that modulates the electrons is not influenced by the access region, which confirms what was observed in Figure 3.1 and Figure 3.2. On the other hand, when current flows through the device the resistive behavior of the access regions is expected to be relevant and as in fact is seen by the increase g_m. In particular, the increase of g_m decreasing the source resistance is well know and described by the equation:

$$g_m = \frac{g_{m0}}{1 + g_{m0} R_{si}} \tag{3.1}$$

where $g_m 0$ is the transconductance of the intrinsic device, and R_{si} is the resistance of the source access region. For a deeper understanding Figure 3.4 shows the effect that downscaling the access region has on the $I_d - V_d$. In fact, the $I_d - V_g$ can be seen as an $I_d - V_d$ calculated at a constant V_{DS} to focus the attention on the variation of the spacing between the $I_d - V_d$ curves as a function the potential applied to the gate, V_{GS}. The black curve with diamond symbols in Figure 3.4 represents an $I_d - V_d$ curve corresponding to a certain source to gate potential applied at the contacts and an

intrinsic source to gate potential that is influenced by the length of the source access region. In particular, when the source access region is reduced the intrinsic source to gate voltage corresponding the same applied source to gate potential increases as it can be seen to do the current shown in blue with circle symbols. However, $I_d - V_d$ curves with higher current have higher saturation voltages, and in fact we can see in Figure 3.4 that the device does operate in saturation after the source access region is shortened. Therefore, the $I_d - V_g$ calculated at $Vd = 3.5\ V$ (green line in Figure 3.4) is calculated with the device operating outside the saturation regime. However, it can seen that shortening the gate to drain access region pushes back the $I_d - V_d$ curve(red curve with square symbol), moving the saturation point before $Vd = 3.5\ V$. This effect can be explained in terms of the resistive behavior associate whit the access regions. In particular, when the drains access region is decreased, the voltage drop across this region is decreased and this results in a higher intrinsic drain to gate that can lead to the saturation of the device. Therefore, the scaling of the drain access region is effective in increasing the current and g_m whenever the device is operating outside the saturation regime. This explains why Di Carlo in Russo and Di Carlo (2007) did not find the drain access region to have any influence, meanwhile in this work, where a more aggressive scaling was performed, we could observe its effect.

3.3 AC analysis

In order to study the effect of the access regions on the frequency response the cutoff frequency of the geometries characterized in Section 3.2 was calculated from the short circuit current gain as described in Section 2.6. In particular, Figure 3.5 shows the current gain when the length of the access regions is symmetrically decreased from 635 nm to 50 nm as shown in the legend. It can be seen that while the response of each device displays the same -20dB/dec trend, f_T which is defined as the 0 dB

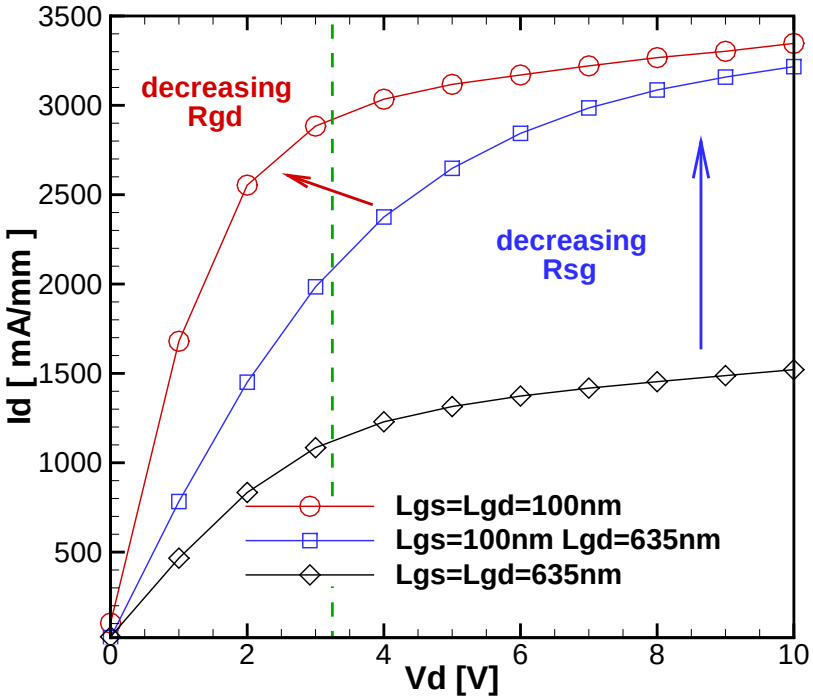

Figure 3.4: $I_d - V_d$ curve calculated at $Vg = 0$ for symmetrical 635 nm access regions(black diamonds), $Lsg = 100$ nm and $Lsg = 635$ nm (blue squares), and $Lsg = 100$ nm and $Lsg = 100$ nm (red circles).

crossing, monotonically increases from 340 GHz to 860 GHz as the access regions are shortened.

The effect of asymmetrical scaling where only one access region is scaled while the other is kept constant is shown in Figure 3.6 compared to symmetrical scaling. The cutoff frequency for each configuration is shown as a function of the source to drain distance of the device so that the effectiveness of symmetrical and asymmetrical scaling can be directly compared. In particular, it can be seen that the simultaneous scaling of both the access regions remains the more effective strategy to increase the cutoff frequency. The only exception is represented by the last point of the curve related to the scaling of the drain access region in Figure 3.6, where the gate to drain distance is reduced to 50 nm. However, in general simulations for very short distance between the contacts and the intrinsic part of the device must be considered with

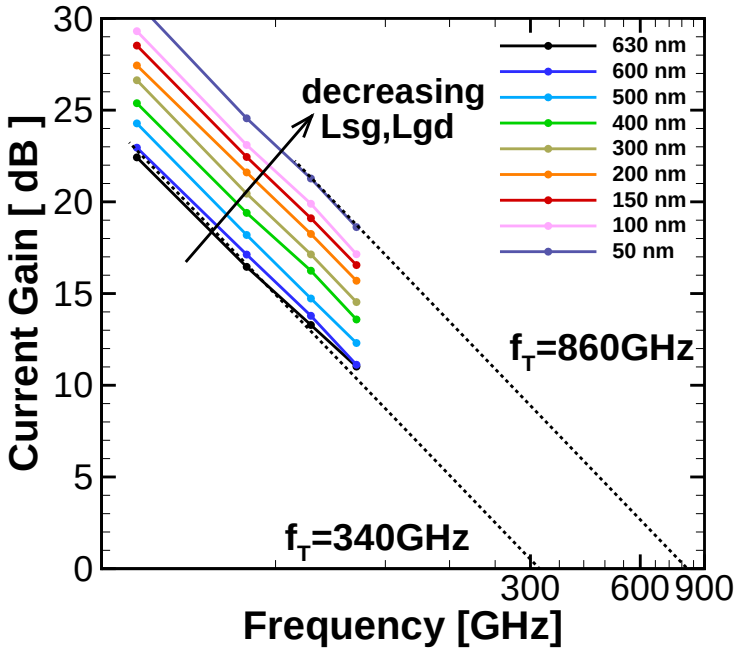

Figure 3.5: Short circuit current gain calculated for decreasing symmetrical access region according to the arrow. The devices are biased for maximum f_T at $V_{GS} = -2\ V$ and $V_{DS} = 3.5\ V$.

attention because of the artifacts can be introduced by the model used to emulate the ohmic contacts.

3.4 Velocity profiles in devices with scaled access regions

In Section 3.3 we showed that scaling the access region allows a noticeable increase in the cutoff frequency. In particular, f_T was increased of almost a factor of 3 when the access region where shortened from 635 nm to 50 nm, similarly to what was observed on g_m in Section 3.2. A microscopic explanation is provided in terms of the dynamics of the carriers in the channel using the physical interpretation of the cutoff frequency. In particular, it was shown in Section 2.8 that the cutoff frequency of a device can be calculated from the average velocity of electrons in the area under the electrostatic control of the gate which was called effective gate length, L_{eff}. The

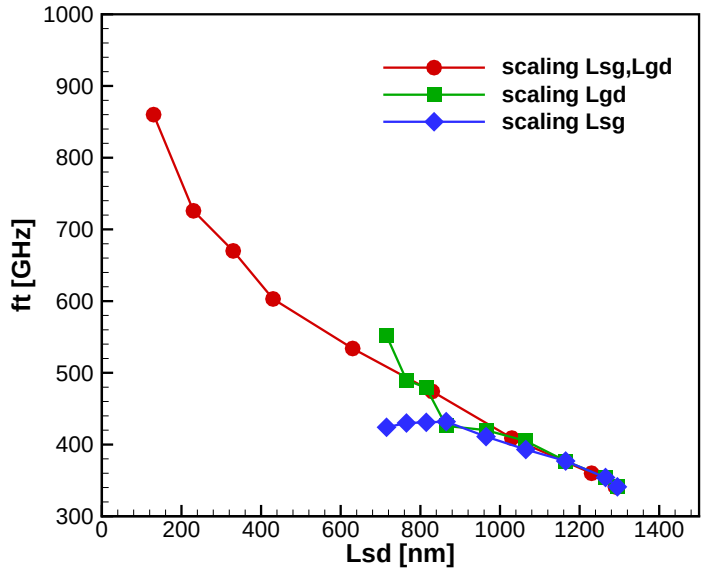

Figure 3.6: f_T plotted as a function of the total source to drain distance for symmetrically scaled access regions (red circles), source access region scaled only(blue diamonds), and source access region scaled only(green squares). f_T have been extracted from the short circuit current gain. The devices are biased for maximum f_T at $V_{GS} = -2\ V$ and $V_{DS} = 3.5\ V$.

electron velocity profile is shown in Figure 3.7 for different lengths of the access region, which have been symmetrically scaled. In particular, Figure 3.7 focuses on the area near the gate, represented by the black rectangle, which includes the effective gate length, shown by the green rectangle.

The main difference between the velocity profiles can be seen near the source side of the effective gate length, where the velocity monotonically increases from $4 \times 10^4\ m/s$ to $1 \times 10^5\ m/s$. On the other hand, it can be seen that the position and the height of the velocity overshoot is not influenced by the scaling. Therefore, the access regions determine the average velocity of the electrons when they enter the effective gate length. The reason for this phenomena can be seen in Figure 3.8 where the electric field profile in the direction of transport is plotted for the same simulations used to generate Figure 3.7.

As it can be seen, the electric field increases shortening the access regions as it was

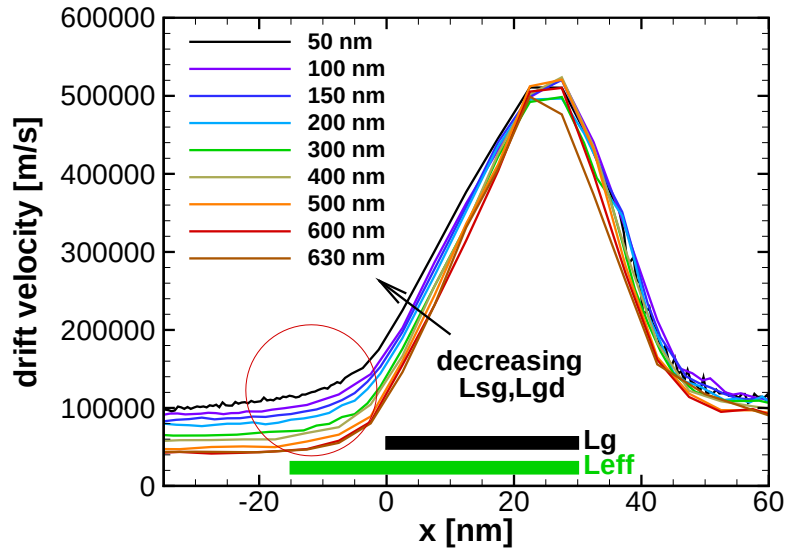

Figure 3.7: Average electron velocity profile for different access region lengths decreasing according to the arrow in the plot. The gate source-end is set as origin of the axis. Metallurgic gate from 0 nm to 30 nm.

observed for the velocity profile. The increase of the electric can be directly related to the shortening of the distance where the same source to gate voltage drops.

It may still sound dubious that a considerable improvement of the cutoff frequency as it reported in Figure 3.6 my be due to such a localized variation of the velocity profile as shown in Figure 3.7. In order to clarify this observation we can recall the definition of cutoff in relation with the transit time:

$$\tau_T = \frac{1}{\omega_T} = \frac{1}{2\pi f_T} = \int_{L_{eff}} \frac{1}{v_{ave}(x)} dx, \tag{3.2}$$

where L_{eff} is the effective gate length, and $v_{ave}(x)$ is the velocity calculated along the channel (x-direction) which is obtained averaging the velocity weighted with respect to the carrier density along the depth of the device. In this way the value of the velocity takes into account the contribution to the conduction from electrons outside the channel, weighted for their concentration. Looking at Eq. 2.8 the cutoff frequency

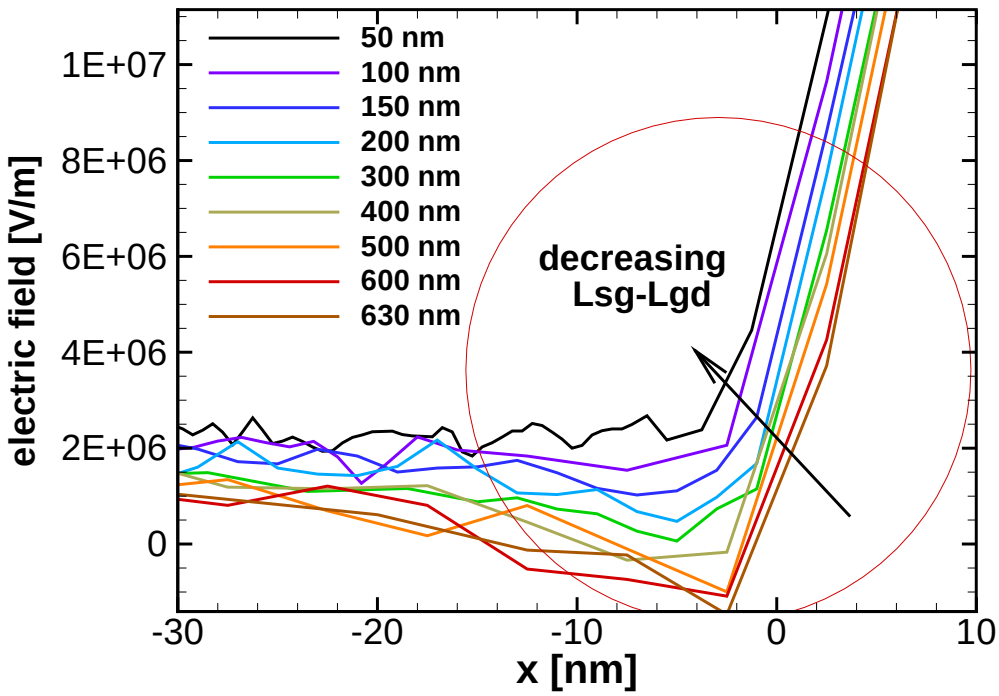

Figure 3.8: Profile of the electric field component along the transport direction for access region lengths decreasing according to the arrow in the plot. The gate source-end is set as origin of the axis. The red circle highlights the source end of L_{eff} which corresponds to the same red circle in Figure 3.7.

is inversely proportional to the total transit time of electrons under the effective gate length. However, during this path the velocity of the electrons is not constant so that the transit time does not increase uniformly while crossing the effective gate length. In Figure 3.9 the total transit times of 3 layouts with different access region length, have been decomposed in partial transit times calculated in 5 nm sections of the effective gate length. In particular, it can be seen that the partial transit times towards the source side of the effective gate length are the ones mostly reduced in devices with scaled access regions. Furthermore, these partial transit times represent the major component of the total transit time, therefore the their reduction allow to greatly decrease the total transit time explaining the significant increase of f_T seen in Figure 3.7.

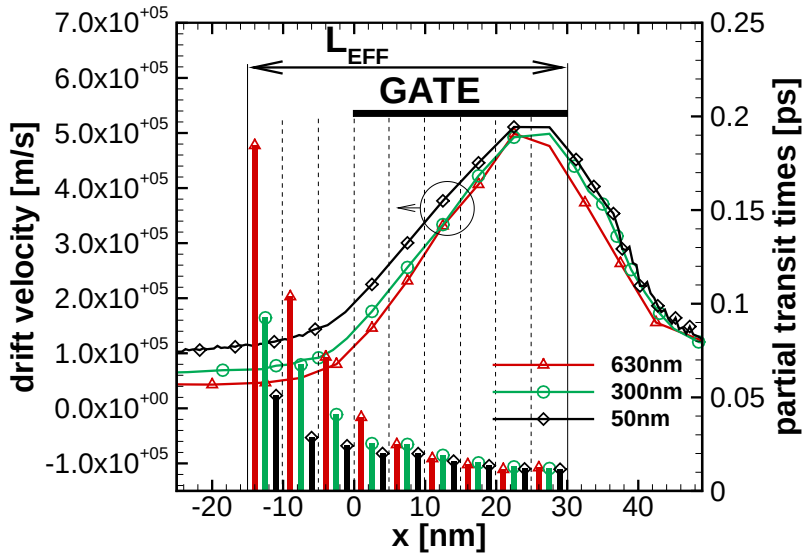

Figure 3.9: Average electron velocity profile for three different L_{sg} L_{gd}. The histograms represents the partial contribution to the total transit time of 5 nm sections of L_{eff}, for different symmetrical access region lengths. The gate source-end is set as origin of the axis. Metallurgic gate from 0 nm to 30 nm.

3.5 Calculation of the Limiting Cutoff Frequency Due the Lateral Scaling

The analysis in Section 3.3 showed that the joint scaling of both the access regions is the most effective to increase the frequency response. In fact, in Figure 3.6 it can be seen that the cutoff frequency reached, decreasing both the access regions to 50 nm, is almost double than what can be achieved scaling a single access region. Therefore, symmetrical scaling is performed for the purpose of studying the limits of the HEMT frequency response. However, as anticipated in the introduction, the vast majority of the research aimed to improve the frequency response is related to the reduction of the metallurgical gate length. It is interesting to notice that the scaling of the access region and the scaling of the metallurgic gate length are not mutually exclusive but they complement each other. In fact, Section3.4 showed that the scaling of the access regions increases the average velocity of electrons, while it is natural to think that shortening the gate decreases the effective gate length. For this reason, in this section

the two scaling strategies are combined to calculate the limiting cutoff frequency for the a GaN HEMT with the restriction of preserving the vertical structure shown in Figure 2.3. In Figure 3.10, the f_T of devices with three different access region lengths, each corresponding to a different color and symbol, is shown as a function of the metallurgic and effective gate length. In particular, the cutoff frequency of each configuration is plotted versus the inverse of the metallurgic gate length in solid symbols and versus the inverse of the effective gate length in the empty symbols.

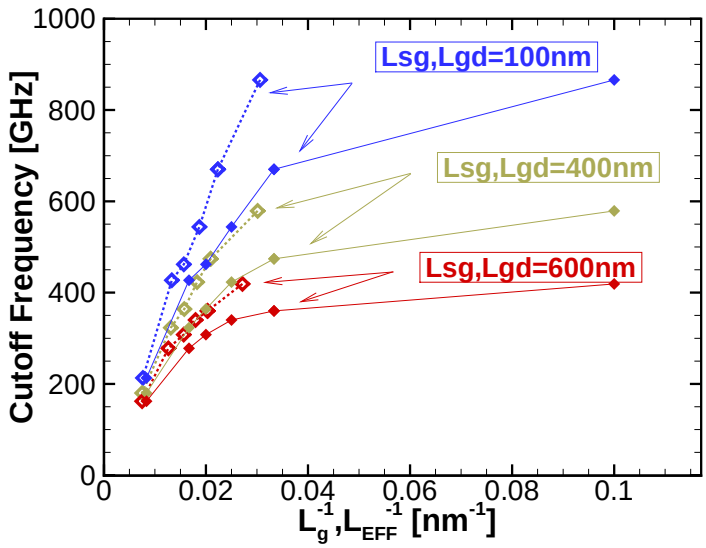

Figure 3.10: f_T versus the inverse of metallurgical (solid line solid symbols) and effective (dashed line empty symbols) gate length obtained at f_T peak bias $V_{DS} = 3.5\ V$, $V_{GS} = -2\ V$

The effective gate length reported is determined matching the f_T calculating through the integration of the velocity profile with the f_T obtained from the AC analysis. As it can be seen in Figure 3.10 f_T increases sub linearly decreasing the metallurgic gate length, especially for shorter gate (*i.e.* on the right side of the plot). This trend, which had been extensively observed experimentally over the last decades, is attributed to the increasing relevance of the area controlled by the gate laterally compared to its vertical projection on the channel, which was explained in Section 2.8

to be due to the fringing fields. In fact, when f_T is plotted versus the inverse of the effective gate length the plot becomes linear. Moreover, looking at Figure 3.10 the slope of this linear relation increases when the access regions are shortened. This slope, which has the dimensions of a velocity, can be seen as the average velocity of the electrons transiting through the effective gate length. Therefore, the increase of the slope is consistent with the increase of the velocity upon reduction of the access regions observed in Figure 3.7.

3.5.1 Calculation of the Limiting Velocity

The average velocity extracted from the slope of the f_T - L_{eff}^{-1} plot is reported in Figure 3.11 versus the length of the access regions. It can be seen that the velocity increases linearly with the scaling so that the intersection between the y-axes and the linear interpolation of the data corresponds to the velocity for vanishing access region. This value of the velocity is an important parameter to describe the potential of a technology. In particular, this limiting value for the average transit velocity is independent to the geometry of the specific device and therefore it is a function only of the material of the channel. This conclusion is drawn from 2 different set of calculations. In this work, it was shown that this limiting velocity does not depend on the horizontal geometry since the value is extrapolated for vanishing access regions and arbitrary gate length. Moreover, Guerra showed in Guerra *et al.* (2010) that the the slope of the f_T - L_{eff}^{-1} plot is not influenced by the device vertical structure (*i.e.* the gate to channel distance). Therefore, it can be stated that the velocity extrapolated in Figure 3.11 is only dependent on the material while the velocity found in Guerra *et al.* (2010) was still depending on the particular access region length chosen for the calculation.

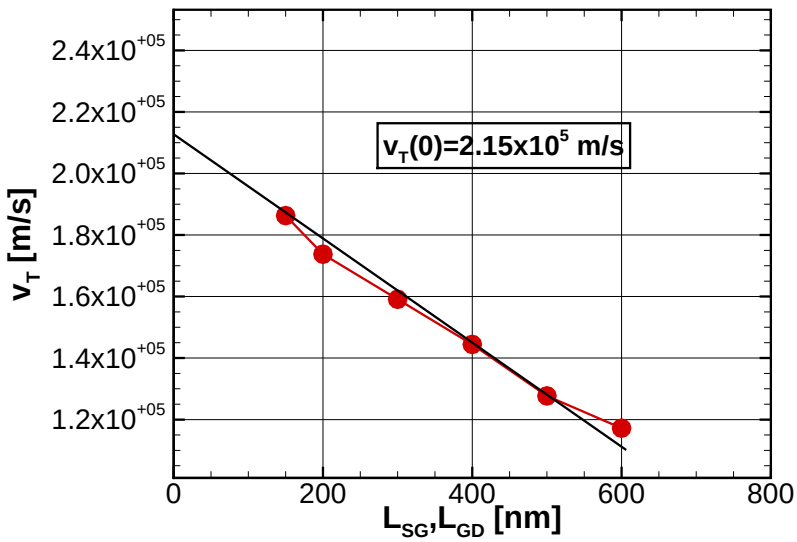

Figure 3.11: Average transit velocity plotted as a function of the access regions length. The limiting transit velocity is extrapolated for vanishing access regions. Bias: $V_{DS} = 3.5\ V$, $V_{GS} = -2\ V$

3.5.2 Limiting Effective Gate Length and Maximum Cutoff Frequency

According to the physical interpretation of the cutoff frequency discussed in 2.8 the velocity and the effective gate length are the two terms determining the value of f_T. Therefore, in order to calculate an upper limit for the cutoff frequency, the shortest effective gate length has to be estimated as well as the highest achievable velocity. Since the limiting velocity was calculated in Section 3.5.1, in this section a calculation of the limiting effective gate length is proposed. Ideally, one expects this value to be a property of the material in analogy with what was found for the limiting velocity. However, the effective gate length is a concept based on geometrical considerations and it does not make sense to attribute its value to a material without a geometrical contest. In particular, the calculation of the limiting effective gate length will be done preserving the channel to gate distance, so that the limiting cutoff frequency that will be calculated is not absolute but it is valid for GaN HEMTs with a 5.5 nm channel to gate distance. The idea to find the limiting effective gate length is suggested again

looking at Figure 3.10. In fact, while it can be seen that the scaling access regions does not influence the effective gate length, the scaling of the metallurgic gate does. In Figure 3.12 the effective of the gate length calculated for devices with different access regions is plotted as a function of the metallurgic gate length.

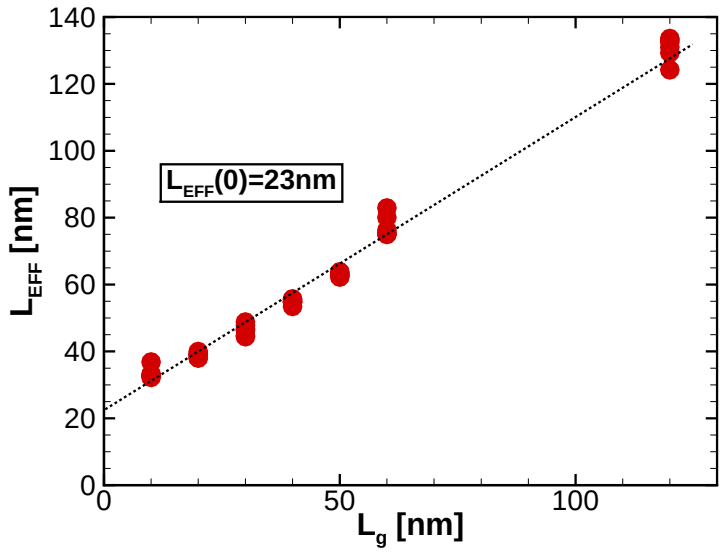

Figure 3.12: Effective gate length plot as a function of the metallurgic gate length. The limiting effective gate length is extrapolated for vanishing access regions. Bias: $V_{DS} = 3.5\ V$, $V_{GS} = -2\ V$

As it was mentioned before the length of the access regions does not change the effective gate length as all the points for a certain metallurgical gate length are very close to each other. However, it can be seen that the effective gate length decreases linearly with the metallurgic gate length. Therefore, similar with what was done for the velocity, the limiting value of effective gate length can be extrapolated from the linear interpolation at vanishing metallurgical gate length. With the value of the limiting velocity of $2.15 \times 10^5\ m/s$ calculated in Figure 3.11 and the limiting effective gate length of 23 nm, the transit time is calculated to be 0.11 ps corresponding to a limiting cutoff frequency of 1.49 THz. It is interesting to point out what would be the outcome if the constrain regarding the preservation of the gate to channel

distance was dropped. According to the analysis in Guerra *et al.* (2010) the effect of the fringing field decreases reducing the gate to channel distance until the effective gate length converges to metallurgical gate length. Therefore, in this case, when the limiting value of the effective gate length is extrapolated for vanishing metallurgic length it results to be 0. The cutoff frequency associated with a 0 effective gate length is obviously infinite, which despite being a reliable value for an upper limit is of little use.

3.6 Conclusions

In this chapter the effect of downscaling the access regions was studied in regards to the DC and AC performance of a GaN-HEMT. The drain current and g_m have been found to increase when the source to gate distance was decreased regardless the fact that the gate to drain distance was decreased as well. On the other hand, shortening the gate to drain distance did not influence the DC characteristic unless it was done in conjunction with the reduction of the source access region. Indeed, the major improvement in g_m was observed in devices where both the access regions have been shortened. No changes were observed in the turn-on voltage shortening the access region length. The influence of the access regions length on the frequency performance was evaluated in relation to the cutoff frequency. The conjunct scaling of both the access regions led to an improvement of a factor of 3 of the cutoff frequency calculated for the unscaled layout. Both the improvement of the DC and AC performance are explained in terms of the higher injection of electron due the increased electric field developing over shorter access region. In the last part of the chapter, the scaling of the access regions was coupled with the reduction of the gate length to calculate a limiting cutoff frequency for a GaN HEMT with a channel to gate distance of 5.5 nm. The limiting transit velocity of 2.15×10^5 m/s was calculating by extrapolating for

vanishing access region length the velocity obtained from the slope of the relation between the cutoff frequency of the inverse of the effective gate length. Similarly, the limiting effective gate length of 23 nm was extrapolated from the plot of the effective length as a function of the metallurgic gate length for vanishing value of the latter. The corresponding value of limiting cutoff frequency is 1.49 THz.

Chapter 4

GAN HOT ELECTRON TRANSISTORS

4.1 Introduction

Hot Electron Transistors (HETs) are devices consisting of three regions: emitter, base and collector, separated by two barriers as can be seen in Figure 4.1. When the emitter-base region is forward biased, electrons are injected into the base and gain kinetic energy. Part of this energy is lost through inelastic scattering processes while electrons cross the base. When the loss of energy is small, electrons can overcome the base-collector barrier and reach the collector, otherwise they thermalize to the bottom of the base and are collected by the base contact.

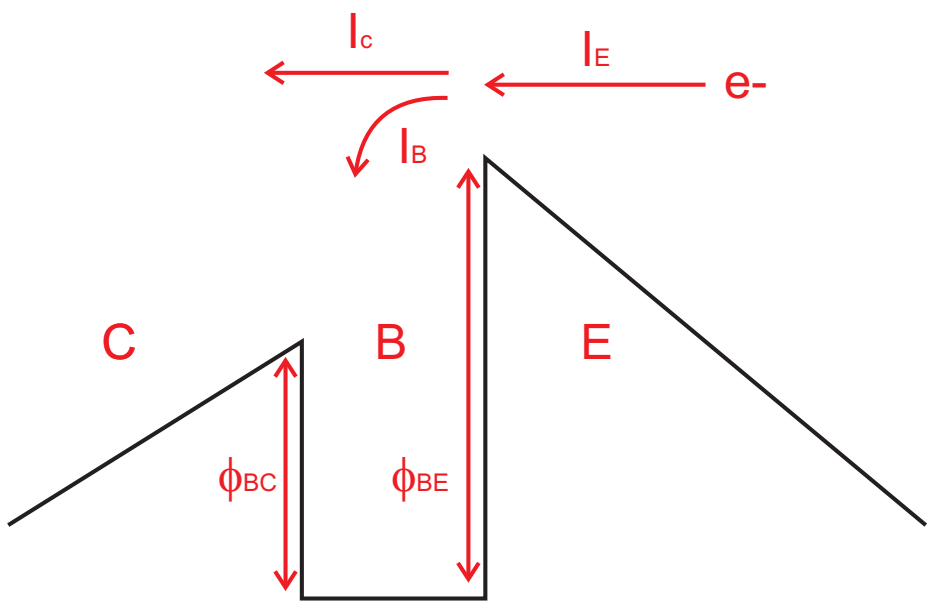

Figure 4.1: Schematic of a HET band diagram, showing the important parameters and quantities.

Precursors of HETs were introduced in the sixties and called metal-oxide-metal (MOMs) Mead (1961). The structure a MOMs is made of two metal-oxide-metal

stacks as shown in the band diagram of Figure 4.2.

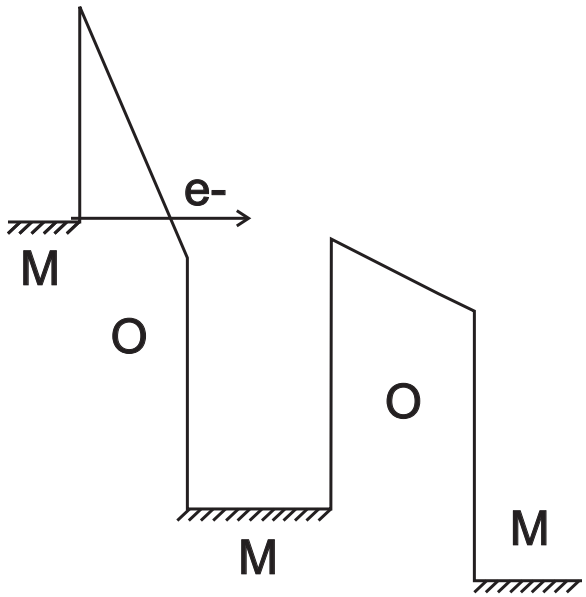

Figure 4.2: Schematic of the band diagram of a MOM device.

The first metal-oxide-metal system worked as emitter by injecting electrons, while the second was designed to prevent the cold electrons of the central metal layer to flow towards the collector. The applications of MOMs were mainly related to the study of the transport properties of materials, in particular ballistic transport. However, because of the short mean free path of electrons in metals, the measured current gain defined as collector current over base current was very low. In the eighties, new implementations were proposed where the metal oxide structure was substituted by a semiconductor material and the barrier was obtained modulating the conduction band with p-doping Shannon and Gill (1981). The structure of these devices, called camel diodes, is shown in Figure 4.3.

Indeed, it was the introduction of new material systems like GaAs/AlGaAs that gave the substantial push to the development of several HET layouts Muto *et al.* (1985). In fact, the possibility to finely tune the alloy molar fraction during the fabrication process allowed the modulation of the conduction band for different ap-

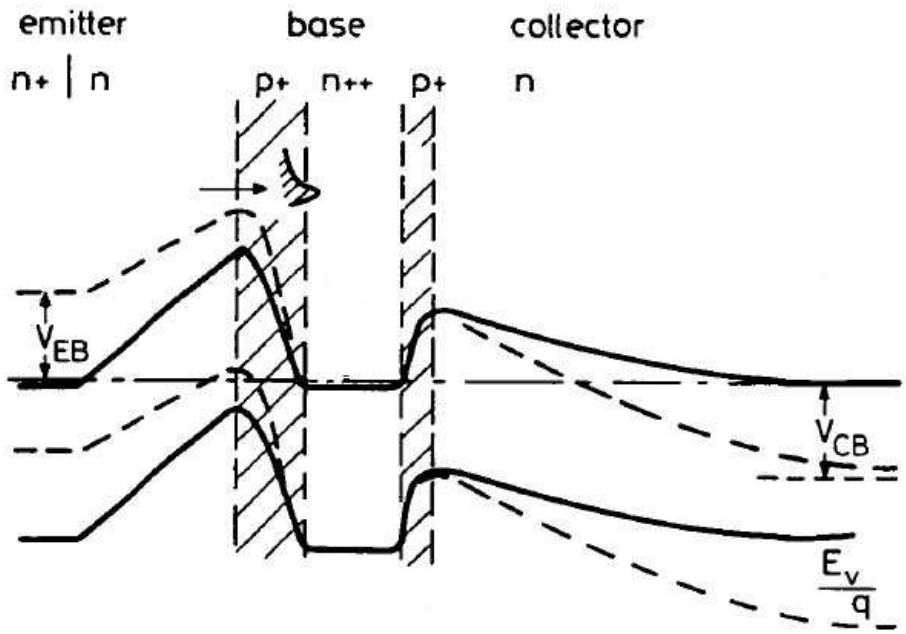

Figure 4.3: Band diagram of the conduction and valence bands of a monolithic Si camel diode Shannon and Gill (1981).

plication. For example, new emitter structures were developed to increase the energy selectivity in the injection of electrons, leading to a variation of HETs known as resonant HETs or RHETs Yokoyama *et al.* (1985). Soon, the potential of HET for high frequency amplifiers became of interest. The critical aspect to obtain amplification is a current gain larger than one, which simply means that more electrons reach the collector than thermalize intothe base. In order to have that, the thickness of the base layer has to be in the order of few decades of nanometers, which initially represented a major manufacturing challenge for the realization of HET amplifiers. However, several groups managed to obtain considerable current gains, initially at low temperature but eventually also at room temperature Levi and Chiu (1987).Despite the progress achieved in terms of current gain, a direct measurement of the cutoff frequency of a HET has not been reported until 2012 van Nguyen *et al.* (2012). In the meantime, different estimations of the cutoff frequency were proposed based on the total delay

time approach Heiblum and Fischetti (1990) Sze and Gummel (1966). In this model, the cutoff frequency is calculated from the summation of the characteristic times that describe different processes affecting the frequency response of the devices. The transit times are related to the velocity of the electrons transiting through different areas of the device. These times are very small in HETs due to the high group velocity of the hot electrons in the thin base and in the collector regions, making HETs good candidates for high frequency applications. However, when an AC signal is applied, both the intrinsic and parasitic capacitances dynamically charge and discharge at each cycle. Since the transit times are very small, the delay times introduced by the capacitances are dominant in the total delay time in HETs. However, to the best of our knowledge a unique formulation of these times has not been clearly defined in literature. In this chapter, we model a novel GaN-HET device and we validate our simulation approach with the experimental current-voltage curves. Subsequently, in Section 4.4 we investigate the effect of the thickness of the different layers of the base stack on the current gain. In Sections 4.5 we choose few of the layouts that exhibit current gain suitable for amplification purposes and we calculate the cutoff frequency. A model for the cutoff frequency is then proposed in Section 4.6. Finally Section 4.6.2 and Section 4.7 show the analysis of the HET capacitance with a graphic approach and its application for the interpretation of the C-V curve.

4.2 Modeling GaN-HETs

The layout of the basic device described in this work is illustrated in Figure 4.4, and is made of an $AlGaN/GaN/InGaN/GaN$ heterostructure grown on a GaN bulk substrate by metal-organic chemical vapor deposition.

This device was designed and fabricated at UCSB by the Mishra group and follows the structure of the devices published in Gupta *et al.* (2014a) Gupta *et al.*

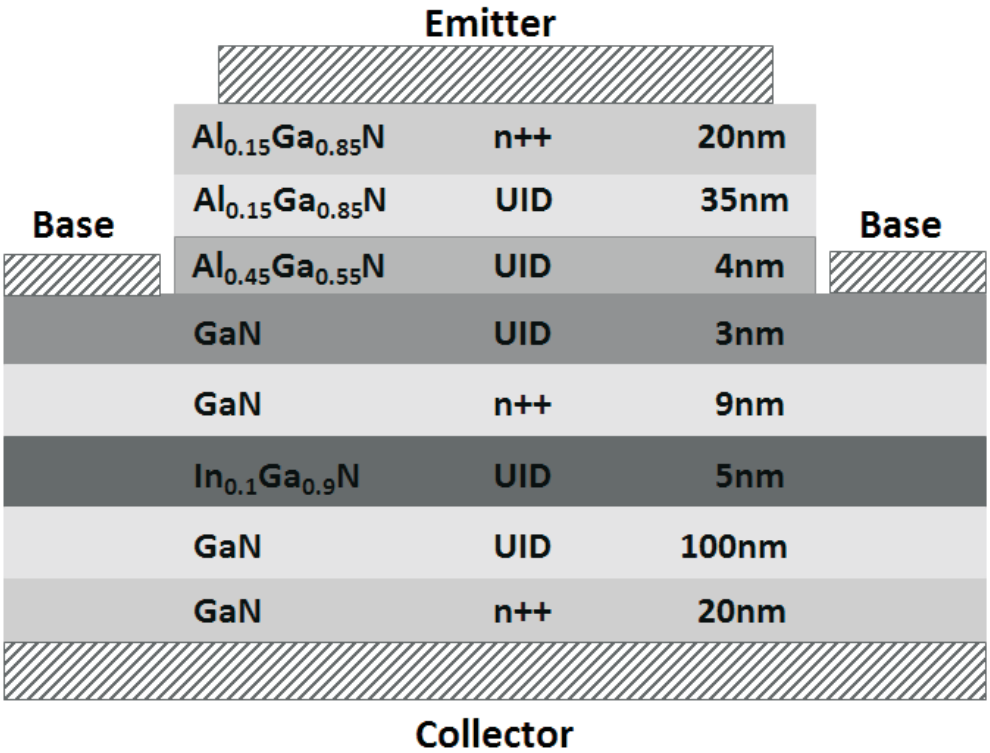

Figure 4.4: Layout of the simulated GaN HET. The material of every layer is shown with its doping and thickness.

(2015b) Gupta *et al.* (2015a). The emitter region consists of two layers of $AlGaN$ with different molar fraction. The $Al_{0.15}Ga_{0.85}N$ layer covers most of the emitter thickness, while the last 4 nm of the $Al_{0.45}Ga_{0.55}N$ generates a thin barrier for the electrons. On the bottom side of the Al-rich layer, the heterojunction with the GaN base is characterized by a large polarization discontinuity Ambacher *et al.* (2000). To account for this property in our model of the device, a sheet of positive charge is introduced at this interface. The effect of this polarization charge can be seen in the band diagram in Figure 4.5, where the notch a the bottom of the emitter barrier confines electrons in a highly conductive 2 dimensional electron gas. The 2DEG is formed in the undoped 3 nm layer of GaN, while in the following 9 nm towards the collector the base is highly doped to decrease the base resistance. Finally, a 5 nm $In_{0.1}Ga_{0.9}N$ separates the base from the GaN collector. The polarization disconti-

nuity between InGaN and GaN creates a polarization dipole across the InGaN layer and induces a barrier that confines the cold electrons in the base. The polarization dipole is modeled with two sheets of charge of equal magnitude but different sign: positive towards the base and negative towards the collector. An alternative design for the base-collector barrier is explored in Gupta *et al.* (2014b) where the InGaN layer is substituted by an AlGaN layer, similar to the emitter barrier. However, due to the opposite polarity of the polarization charge introduced by the AlGaN layer as opposed to the InGaN layer, the alternative design proved to deplete the base, increasing the base resistance, while the InGaN barrier induces an accumulation of electrons as can be seen in Figure 4.5.

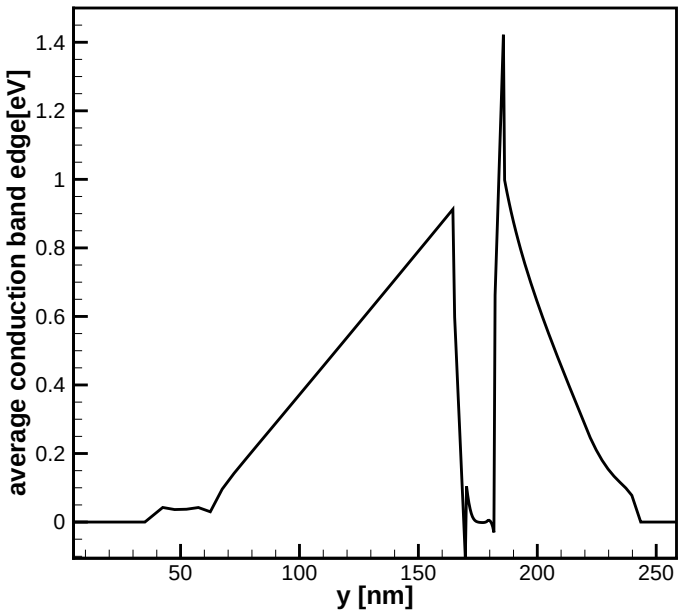

Figure 4.5: Equilibrium band diagram of the HET in Figure 4.4 extracted at the center of the device.

4.2.1 Transport Modeling

Figure 4.6 shows a snapshot of the electrons in the device biased in amplification mode, with the emitter-base junction forward biased and the base-collector junction reverse biased. Electrons in the emitter are pushed towards the thin $Al_{45}Ga_{55}N$ layer without gaining considerable energy but rather accumulating at the bottom of the barrier, suggesting that the dominant mechanism to cross the barrier is tunneling rather than thermionic emission. In fact, it can be noticed that most of the hot electrons near the base side of the barrier have an energy comparable with the bottom of the barrier as opposed to the top of it. Moreover, Figure 4.7 shows that the energy distribution of the hot electrons in the base is characterized by a peak, which is reasonable to associate to the energy that all electrons have when they are injected into the base. In particular, the energy corresponding to the peak in Figure 4.7 is 1.22 eV, which is comparable with the difference between the energy at the bottom of the injection barrier and the one in the base, whereas the top of the barrier is at 1.7 eV. It is important to highlight that the distribution shown in Figure 4.7 is not calculated over all the electrons in the base, but only considering electrons that satisfy two conditions: being injected by the emitter and having an energy above 0.3 eV. These conditions allow to avoiding counting the cold electrons at the bottom of the base that, being 4 orders of magnitude as many as the hot electrons, would shift the distribution towards the thermal energy.

A further check to confirm tunneling as the dominant injection mechanism was performed by running simulations without enabling tunneling through the emitter barrier and no current was measured. The electrons in the base shown in Figure 4.6 are characterized by a wide variety of colors corresponding to different kinetic energies. However, it is possible to distinguish two populations, the cold electrons depicted in

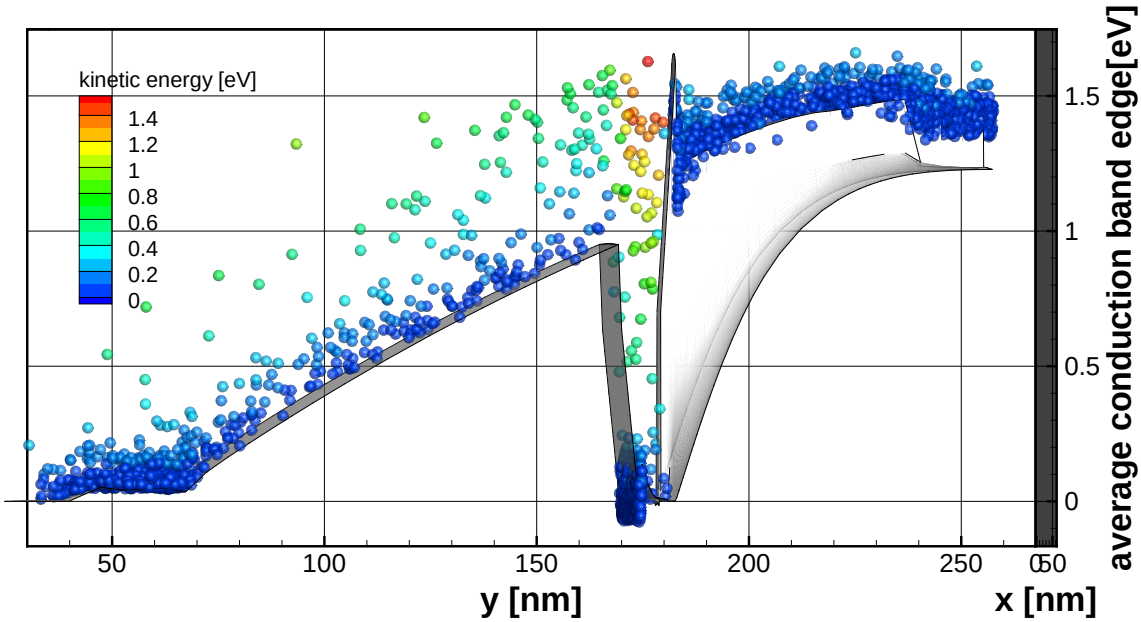

Figure 4.6: Band diagram of the HET in Figure 4.4 calculated at $Ie = 0.25\ mA/\mu^2$ and Vc =2 V. A snapshot of the superelectrons in the device is plotted and colored according to their kinetic energy.

blue at the bottom of the conduction band and the ones with higher energy and different colors. The red and yellow particles correspond to electrons that scattered inelastically only a few times, therefore still preserving enough energy to overcome the collector barrier. The particles depicted in green and light blue color correspond to electrons with energy below the collector barrier that are trapped into the base and keep scattering until they thermalize. The mechanism responsible for the energy loss for the electrons in the base is inelastic scattering, however, elastic scatting is also relevant in determining the percentage of electrons that reach the collector. In particular, electrons are injected into the base collimated by the electric field of the emitter, with the main component of their momentum along the longitudinal axes of the device. Along this emitter-to-collector direction, the length of the path that electrons have to travel to cross the base is simply the distance between the emitter and the collector barrier. However, only ballistic electrons preserve the direction they

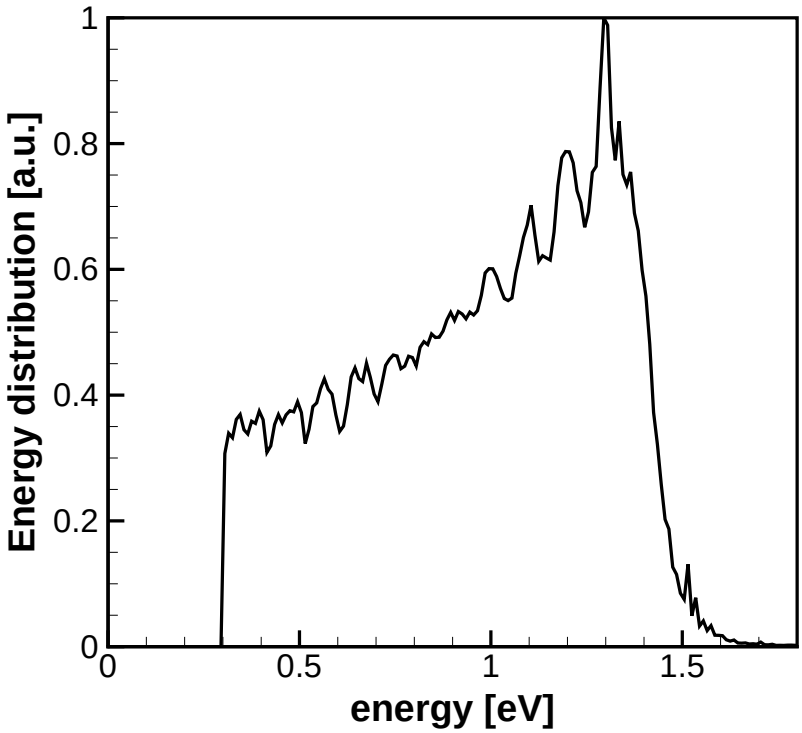

Figure 4.7: Energy distribution calculated for the electrons in the base of the GaN HET that have an energy larger than 0.3 eV. The device is bias at $Ie = 0.25\ mA/\mu^2$ and Vc =2 V.

had at the injection because scattering events tend to randomizes the momentum and thus the trajectory of electrons. The main scattering mechanism of doped GaN are shown in Figure 4.8 with their rates as a function of the energy.

It can be seen that ionized impurities are the dominant scattering at low energies, but around the injection energy, the rate of emission of polar optical phonons is twice the impurity scattering rate, which is twice that of deformation potential scattering. Among these scattering mechanisms, the emission of polar phonons is mainly responsible for the loss of energy of electrons, while the interaction with impurities is an elastic process, where the energy of the electron does not change. As for the momentum randomization, deformation potential scattering is much more effective than the other two mechanisms. This is because the strength of both the interaction

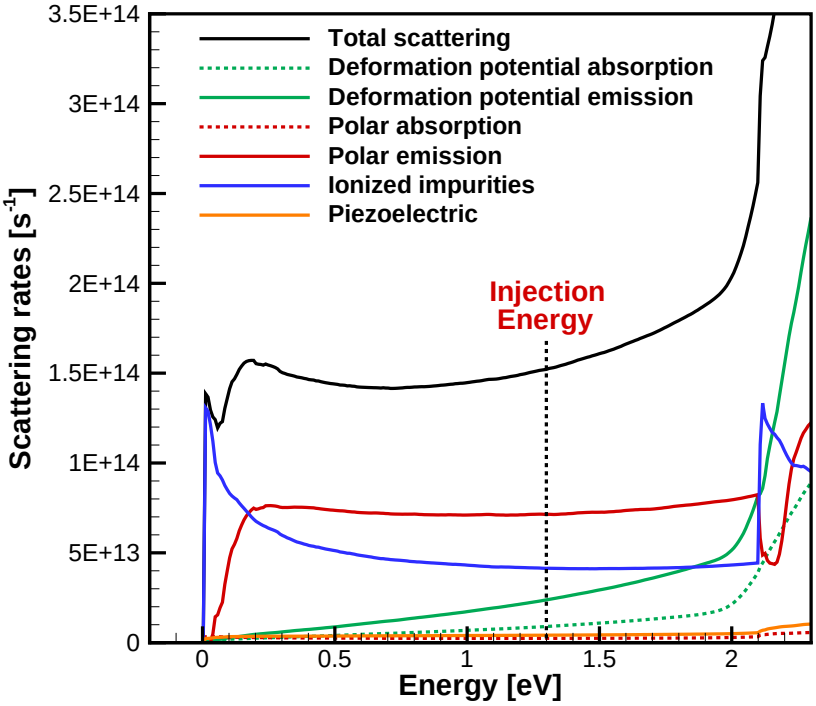

Figure 4.8: Rates of the dominant scattering mechanisms in GaN for $1e19\ cm^{-3}$ doping as a function of the electron kinetic energy. The average injection energy is shown in the plot for reference.

with ionized impurities and with polar phonons is inversely proportional to module of the vector representing the difference of the momentum of the electron before and after the scattering event Ferry (1991). On the other hand, the rates of deformation potential scattering show no dependence to the modulus of the phonon wave vector, and therefore any scattering angle has the same probability. Another observation that can be made in relation to Figure 4.8 is the injection energy. In particular, it can be seen that the scattering rates suddenly increase around an energy of 2 eV. This energy value corresponds to the separation between the bottom of the Γ valley and the bottom of the closest satellite valleys in GaN. Therefore, for high energies the satellite valley provide additional states that can preserve the energy in the scattering

process and the scattering rates increase as a consequence. The inter-valley separation is indeed an important parameter in the design of HET, posing a limit on the injection energy which ideally is wanted as high as possible to maximize the energy that electrons can loose without being trapped in the base. Figure 4.9 shows the distribution in momentum space of hot electrons in the base of the HET. In particular, the momentum distribution is shown in 3 slices of the first GaN hcp Brillouin zone (BZ1). The two slices at the top and at the bottom are extracted at three quarters of the distance between the Γ and the A point corresponding to the position of the lower satellite valleys, while the middle slice is extracted in center of BZ1. As it can be seen, almost all the electrons are in the Γ valley, since we saw in Figure 4.9 that the injection occurs around 1.22 eV, well below the energy of the satellite valleys. In more detail, it can be seen that electrons do not exactly occupy the center of the Gamma valley, but the peak of the distribution is located along the transport direction since the device is not at equilibrium and current is flowing.

One of the complications to accurately model the electron transport across HETs is the heterogeneity of the materials where the transport occurs. We saw in Chapter 1 that in our Monte Carlo simulator the transport properties of the material are contained in a scattering table that is loaded at the beginning at the simulation. This approach is ideal for HEMTs where electrons mainly flow in the channel corresponding to a single layer of the device stack. On the other hand, looking at Figure 4.4, we can see 4 different layer of different materials that are interested by the transport for the GaN-HET. However, as we saw in the section 4.1 the base is the crucial area of the device, therefore GaN tables are used in the simulation. As for the parameters of the GaN material, the doping is set to a value above the maximum doping in the device to allow the rejection algorithm to work properly as explained in Section 1.1. The electron screening effect explained in Section 1.1 is also very

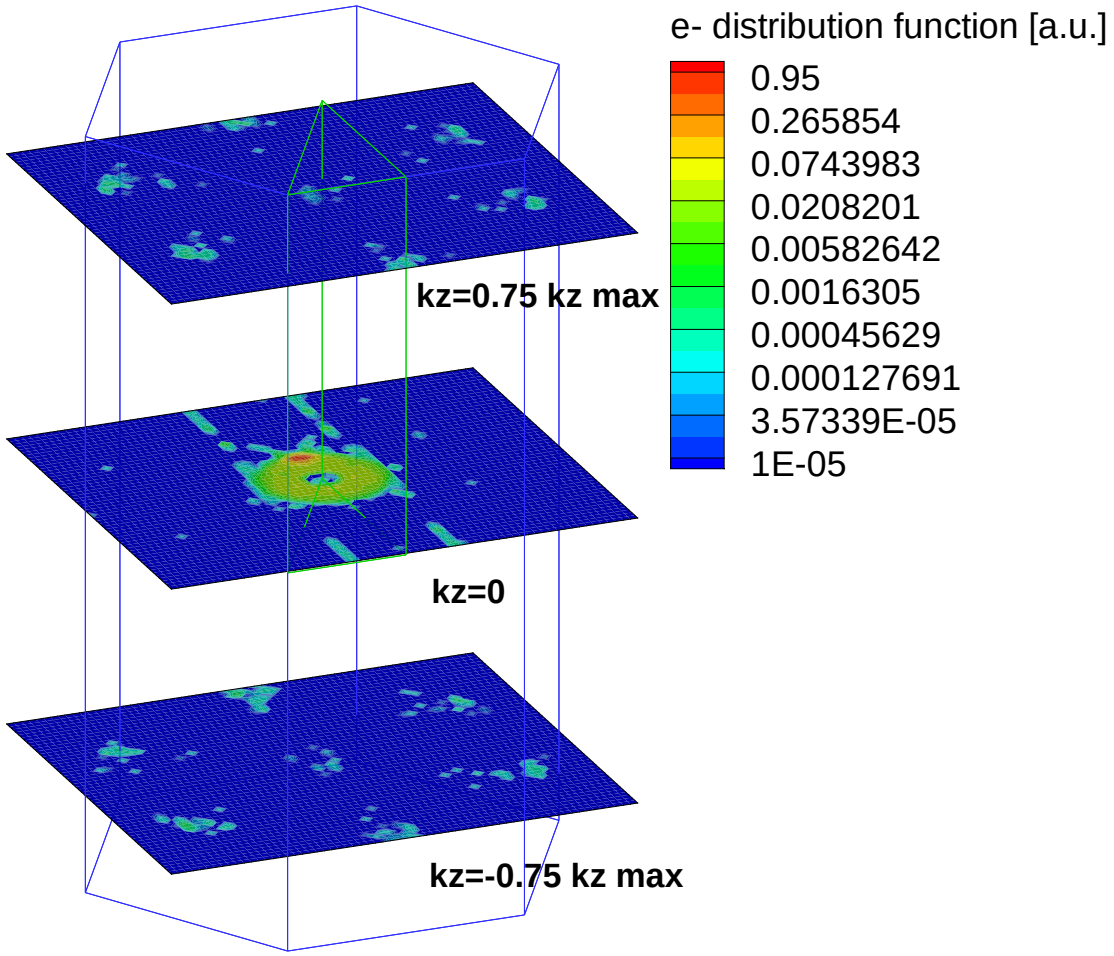

Figure 4.9: Momentum distribution calculated for the electrons in the base of the GaN HET that have an energy larger than 0.3 eV. $Ie = 0.25\ mA/\mu^2$ and $Vc = 2\ V$.

important to simulate the device, in particular to take into account the shielding of the cold electrons on the scattering affecting hot electrons, which decreases the original scattering rate. Moreover, since the screening is more effective on scattering with phonons with short wavelength, the randomization due to each scattering event is increased Ferry (1991). The last observation about scattering is related to the interaction of a single electron with the population of the cold electrons in the base. According to the quantum mechanical treatment, the collective vibration of an electron gas can be modeled with particles called plasmons. The energy of plasmons

is a function of the vibration frequency of the electron gas, which increases with the square root of the free electron density Diff and Brennan (1991). When the concentration of electrons is high enough, as it is the case in the HET base, the energy of plasmons is such that they can interact with electrons. The scattering rate due to plasmons can be calculated with a quantum mechanical treatment similar to what is used to calculate scattering rates from phonons Diff and Brennan (1991). However, a semiclassical approach exists in alternative to the quantum mechanical treatment. In particular, plasmons are a representation of the long distance coulombic interactions between electrons, that can be resolved by the Poisson Solver self consistently coupled with the Monte Carlo Fischetti and Laux (1988). The main requirements on the accuracy are: 1) updating the Poisson solution with a frequency higher than the plasma frequency, 2) a mesh spacing finer than the Debye length calculated from the local doping, and 3) having a number of particles sufficient the have a good statistics. The two approaches are compared in an article that exposes strengths and weakness of both Mansour *et al.* (1992). For our simulations, we chose the classic approach which is of more immediate implementation. Finally, another effect that plays a relevant role in the transmission of electrons over the collector barrier is the quantum reflection Gupta *et al.* (2015b). In our code this effect is included in the tunneling probability that can assume values below one even for electrons with energy is above the barrier.

4.3 DC Charateristics

In order to validate our simulation approach we computed the $Ic - Vc$ characteristics of the HET shown in Figure 4.4, and compared it with the experimental measurements in Figure 4.10.

The agreement between simulated and measured results is excellent. In particular,

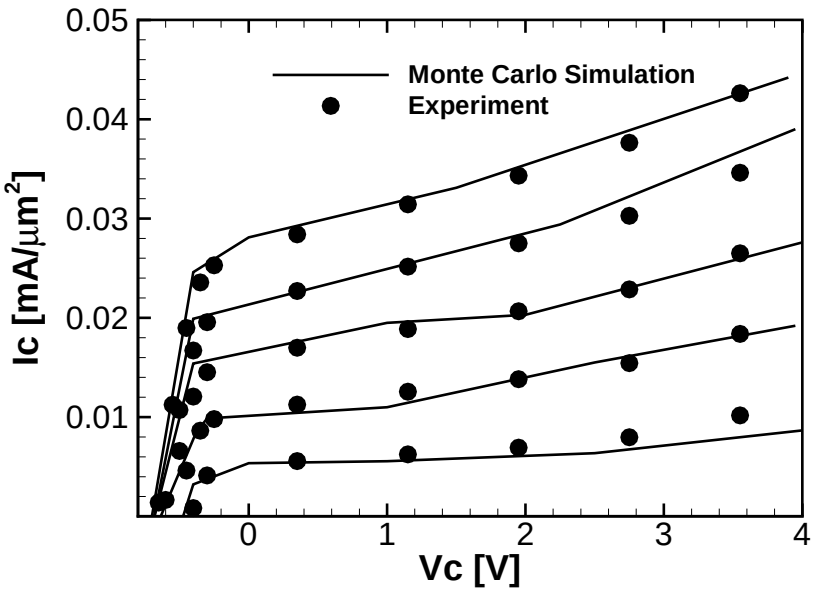

Figure 4.10: Comparison between the calculated (black) and the experimental (full circles) $I_c - V_c$.

we can see that the collector current at $Vcb = 0V$ is around one fifth of the emitter current that biases the device. This means that the current gain β is 0.25, well below the threshold for amplification which is 1. It can be seen that the $Ic - Vc$ characteristics displays a high output impedance which justifies the choice of $Vcb = 0V$ for the calculation of β.

4.4 Impact of the Base Stack on the Transmission Coefficient

In Section 4.2 we showed the simulation approach and the fit of the current characteristics that confirms the low current gain of this HET. However, in order to extend the analysis to the frequency response of HETs the layout of the device has to be modified to provide amplification at room temperature. Once again, the structure of the base is more likely to have an impact on the current gain. In particular, a close look to the band diagram shown in Figure 4.11 shows that the total 4 nm AlGaN

layer and the 5 nm InGaN layer contribute to the emitter to collector barrier distance.

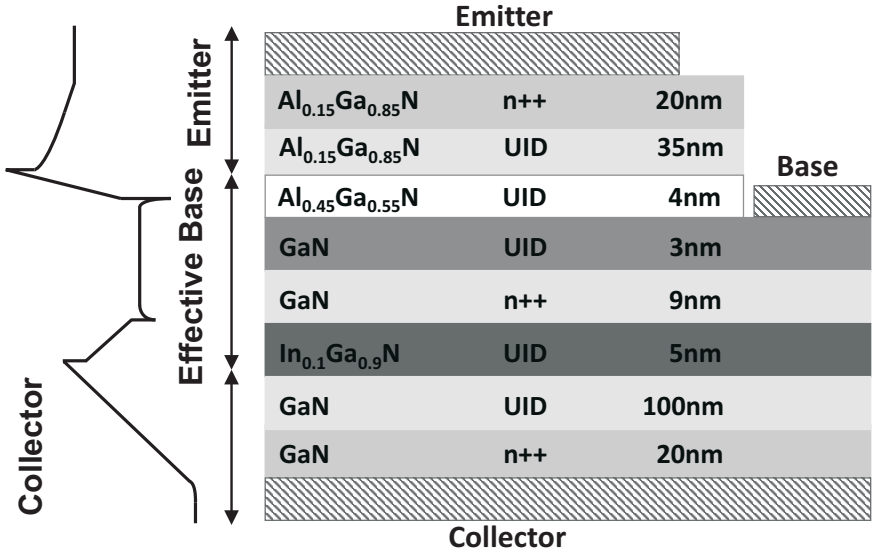

Figure 4.11: Schematic band diagram on the side of the GaN HET layout. The picture is out of scale to highlight all the layers included in the base stack.

In analogy with the concepts introduced in Chapter 2, we can say that the effective base length is 21 nm while the metallurgical base is 12 nm. Shortening any of the 3 three layers of what can be called the base stack, would therefore contribute decreasing the effective base length, but not necessarily to increase the current gain. In particular, the polarization dipole across the AlGaN layer is related to the height of the emitter-base barrier which defines the injection energy of the electrons. Shortening the AlGaN layer would decrease the energy of the injected electron and, as a consequence, fewer scatterings would be required to bring the electron energy below the level of the collector barrier. The second layer that can be scaled is the GaN base. In this case there is are no apparent trade-off; the reduction of the GaN thickness only reduces the length of the path of the electrons without altering the energy gap between the barriers. The last option for decreasing the effective base length is the InGaN layer. Similar to the scaling of the AlGaN layer, shortening the InGaN layer would decrease the barrier introduced by the polarization dipole. The

difference is that a smaller collector barrier increases the energy that electron can lose without being trapped in the base. Therefore, the reduction of the InGaN layer offers a double advantage instead of posing a trade off. In Figure 4.12 the two most promising approaches, GaN base scaling and InGaN scaling, are compared in terms of the transmission coefficient that is computed as the collector current over the emitter current.

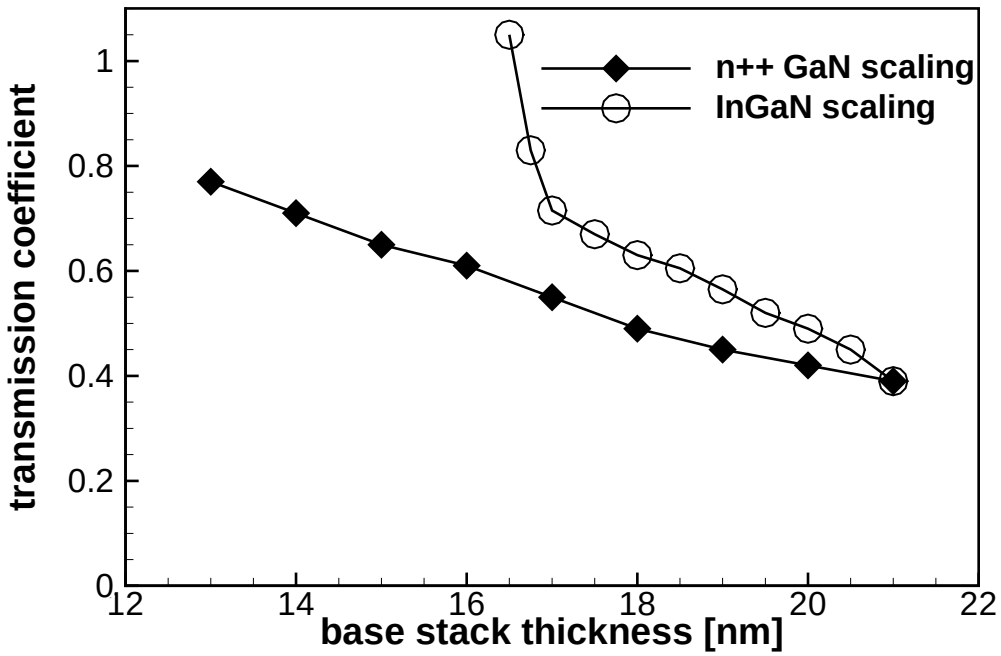

Figure 4.12: Transmission coefficient as a function of the base stack thickness. Solid diamonds show the effect of reducing the 9 nm doped GaN layer in Figure 4.11 to 2 nm, keeping the InGaN layer to 5 nm. Empty circles show the effect of reducing the 5 nm InGaN layer in Figure 4.11 to 1 nm, keeping the doped GaN layer to 9nm. The emitter current is 0.25 mA/μ^2 and $Vcb = 2\ V$.

The horizontal axes of the plot represents the total effective base length. To compare the two scaling approaches, only one layer has been scaled while the thickness of the other two layers of the base stack is kept contact. As expected, the scaling of the InGaN layer is the most effective in increasing the transmission coefficient. However, while the GaN curve shows a linear trend for all the thickness considered, the InGaN curve shows a sudden increase when the InGaN layer is reduced to 1 nm

and 0.5 nm corresponding to the 16 nm and 15.5 nm points in figure Figure 4.12. This effect is due to the spill over of the cold electrons over the collector barrier that has become too small and can be crossed by the most energetic "cold" electrons in the base.

4.5 Calculation of the Cutoff Frequency

The main result of the scaling studied in Section 4.4 is that a transmission coefficient above 0.5 can be achieved within realistic dimensions of the device. In particular, when the InGaN layer is reduced to 2.5 nm which is half of its original length, and the GaN layer is kept constant, the transmission coefficient reaches 0.6. This geometry is chosen as a starting point to study the frequency response of GaN HETs. The cutoff frequency of the new layout is calculated with a multisinusoid AC simulation as explained in Section 2.6. The device is biased to guarantee the class A operation of the amplifier, with the collector voltage far away from the knee voltage. As for the emitter current, we choose to bias the device at 0.25 mA/μ^2, close to the bias condition of the f_T measured in van Nguyen *et al.* (2012) for a different HET. The plot of the short circuit current is shown in Figure 4.13, where the cutoff frequency is seen to be 90 GHz.

Moreover, the frequency response shows a low pass filter behavior with a constant gain at low frequency and a -20dB/dec slope after the pole. This is due to the finite DC gain of HET as opposed to HEMT whose frequency response shows only a -20dB/dec slope. In fact, we can see in Figure 4.13 that the current gain converges at low frequency to 3.5 dB which corresponds the DC current gain of 1.5 calculated from the transmission coefficient shown in Figure 4.12.

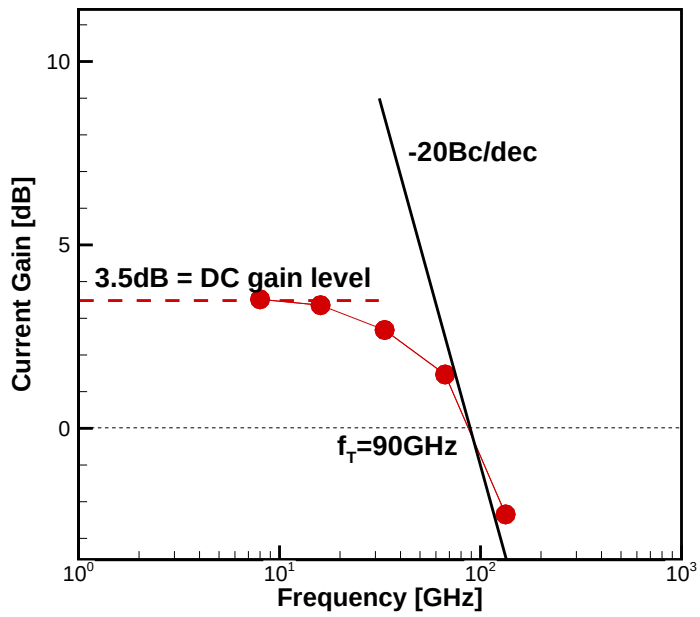

Figure 4.13: Short circuit current gain of the 2.5 nm InGaN variation of the GaN HET described in Section 4.5 calculated with multisinusoidal AC simulation(red dots). The emitter current is 0.25 mA/μ^2 and $Vcb = 2\ V$.

4.6 Cutoff Frequency Model

The common model of the cutoff frequency for any transistor is:

$$f_T = g_m/2\pi C_{in} \tag{4.1}$$

where g_m is the transconductance and C_{in} is the input capacitance. The calculation of the g_m for the HET geometry considered in Section 4.5 is shown in Figure 4.14 where g_m is seen to peak approximatively at 3 mS/μ^2 while at the bias considered in Section 4.5 its value is 1.8 mS/μ^2.

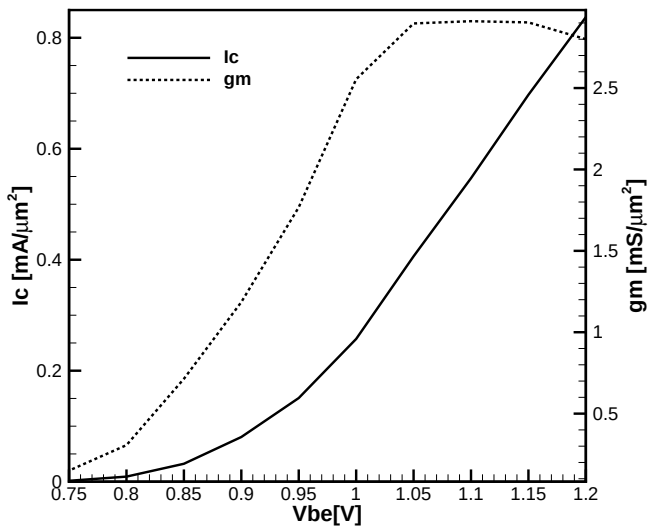

Figure 4.14: $I_c - V_be$ characteristic of the 2.5 nm InGaN HET described in Section 4.5 calculated $Vcb = 2\ V$. The g_m is shown with a dashed line while the collector current corresponds to the solid line.

4.6.1 Capacitance Calculation

The calculation of capacitances from particle-based simulations is not as straightforward as for the transconductance. The traditional approach consists in inverting Eq. 4.1 to calculate the input capacitance knowing f_T and g_m. Evidently, this approach is not applicable within the context of this work, since our aim is to provide a formulation for f_T. Therefore, we need an independent method to calculate the HET input capacitance, that corresponds to the base capacitance when the signal is applied to the base. An approach that is found in literature allows to calculate the gate-drain capacitance (Cgd) from the Y parameters after modeling the device with an equivalent circuit Berroth and Bosch (1990). The same approach could be used in principle to calculate base-collector capacitance (Cbc) in a HET but a formula to calculate base-emitter capacitance (Cbe) would still be needed in order to calculate total base capacitance (Cb) as the summation of Cbc and Cbe. In this section we

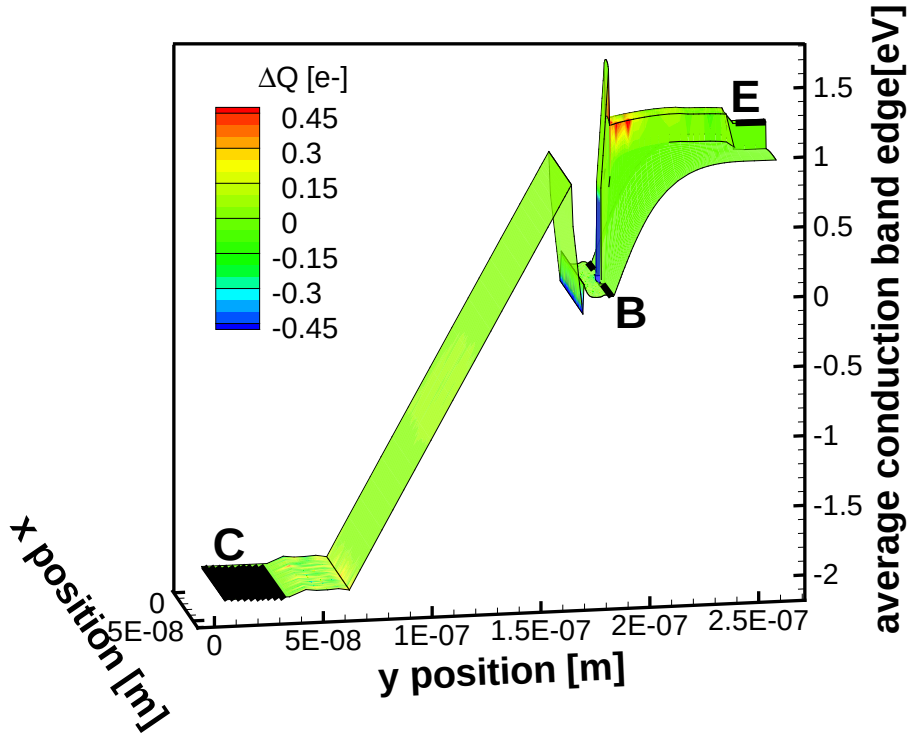

Figure 4.15: Band diagram of the HET the 2.5 nm InGaN HET described in Section 4.5 biased at $I_e = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$. The contour shows the accumulation (red) or depletion (blue) as a consequence of a perturbation of 0.025V applied to the base voltage.

show a method to calculate the capacitance of GaN HETs directly from its definition:

$$C = \frac{dQ}{dV}, \tag{4.2}$$

where dQ is the charge modulated by a perturbation dV of the potential. A quasistatic approach is used to determine the charge modulation corresponding to the potential perturbation. In particular, this approach consists in calculating the variation of the charge distribution resulting from a perturbation of the contacts bias. The first step of our approach consists in calculating the charge distribution at the bias point of interest and at a perturbed bias point. In particular, for the calculation

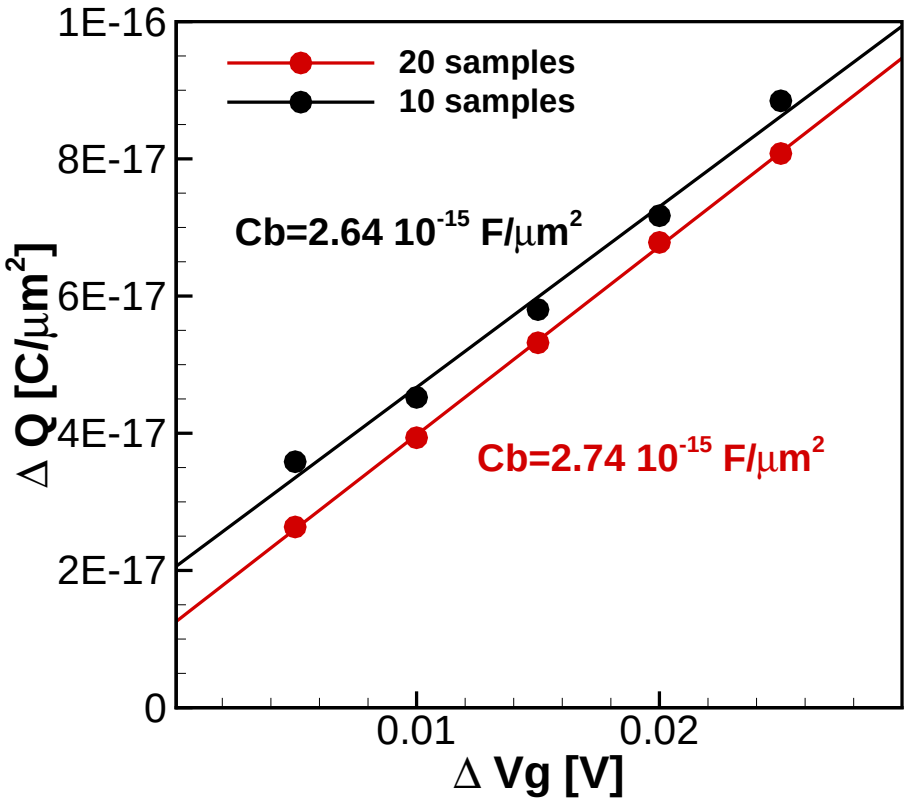

Figure 4.16: Charge density accumulation or depletion as a function of the amplitude of the perturbation applied to the base voltage. The capacitance corresponds to the slope of the linear fit. The black is calculated averaging each charge distribution over 10 simulations,while the red is obtaining averaging over 20 simulations. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5 biased at $Ie = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$.

of the base capacitance the bias point is changed by perturbing the potential applied to the base, while the other two electrodes maintain the same potential. When the potential variation is small, this perturbation can be compared to applying a small AC signal to the base while the other two electrodes are kept at a constant DC bias which, for AC analysis purposes, is equivalent to ground. When the difference between the unperturbed and the perturbed charge distributions is calculated, the result can be negative or positive in different areas of the device. Negative values mean that the effect of the perturbation is a depletion of electrons, while positive values represent an accumulation. Figure 4.15 shows the map of the charge modulation,

where the accumulation areas are colored in red while the depletion area are colored in blue. In particular, it can be seen that layers of accumulation alternate to layers of depletion separated by a barrier. This pattern of the charge reminds the analysis of the capacitance developed for pn junctions. In these structures, when the potential across the junction increases, an additional layer of charge is formed at both sides of the depletion region due to ionization of more dopant atoms. This additional charge constitutes the term dQ when Eq. 4.2 is used to calculated the capacitance for the pn junction while dV is the increase of potential across the barrier. However, the geometry of the HET is more complex than a pn junction, and the potential does not drop across a single junction, so the charge modulation has to be computed within the whole simulation domain. One way to perform this calculation is to consider the charge modulation in all the cells that discretizes the simulation domain, and sum all the positive or negative contributions. In fact, due to the charge neutrality, the summation of all the negative contributions is equal to the summation of all the positive contributions. A more practical way to calculate the total charge modulation is sum the absolute value the charge modulated in each cells and obtain the dQ to use in Eq. 4.2 dividing by 2 the result of the summation.

In Figure 4.16 the total charge variation is shown as a function of the amplitude of the bias perturbation. The relationship between the charge variation and the potential perturbation is linear, as expected from eq. 4.2, but an offset is present. Ideally, there should not be any offset since the charge variation is a consequence of the potential perturbation: when the potential is not perturbed the charge should not change. However, an additional term is introduced in the calculation of the variation of charge by the noise due to the random fluctuations of the charge during the simulation. This term is not related to the amplitude of the potential, and it can be seen as the noise floor associated with the calculation. In order to reduce

the effect of the random fluctuations of the charge, the charge distribution at each bias point is obtained averaging over multiple distributions calculated at the same bias condition. With this procedure we improve the statistics converting an average over time to an average over different samples, under the hypothesis of ergodicity. Obviously, in order to have independent calculations of the charge distribution, the seed of the pseudo-random number generator has to be changed for each simulation. The red and the black series of data in Figure 4.16 show the effect that the number of simulations used in the average has on the results. In particular, it can be seen that as the charge distribution is calculated between an increasing number of simulations the offset due to the noise decreases, while the slope of the curve, which represents the capacitance, remains the same.

4.6.2 Capacitance Visualization Through Charge Mapping

In the calculation of the capacitance described in Section 4.6.1 the total charge distribution is used to calculate the variation of charge. However, it is interesting to examine in detail where the accumulation or depletions occurs in order to associate the capacitive behavior to the features of the layout. Figure 4.17 shows the charge modulation when the potential perturbation is applied to the base, which is the case considered in Section 4.6.1. It can be seen that the charge variation is not uniform through the device, but it is more pronounced in proximity of the barriers. In particular, scanning Figure 4.17 from the emitter to the collector, a big depletion (blue) can be seen before the emitter-base junction where the injection barrier is located. Facing this depletion area, a very localized but intense accumulation (red) can be seen in correspondence to the 2DEG of the base. The pattern of this charge variation that occurs across the emitter-base barrier resembles two plates of a parallel plates capacitor and it can be associated with the capacitance of the emitter-base

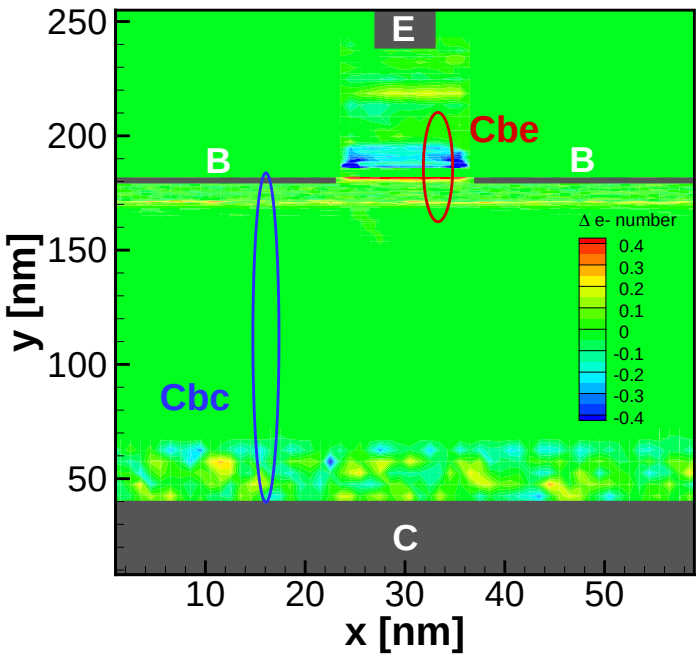

Figure 4.17: Contour plot of the accumulation / depletion of electrons following a 0.025V perturbation applied to the base which is related to the base capacitance. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5 biased at $Ie = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$.

junction. The lower part of the device is also characterize by an accumulation facing a depletion, this time across the base collector barrier. This charge modulation, which appears to be smaller than the one across the emitter barrier, can be associated to the collector-base junction. According to this result, when the base potential is perturbed both the base-emitter and the base-collector potential changes and the charge is modulated across both the barriers. However, when the potential perturbation is applied to one of the other two electrodes, say the collector, only the base-collector bias is changed, while the base-emitter potential remains unaltered, and therefore no charge should be seen across the emitter barrier. In fact, in Figure 4.18 that shows the effect of the perturbation applied to the collector side, the modulation of charge across the emitter junction has completely vanished. Similarly, Figure 4.19 shows the map of

the charge modulation when the potential perturbation is applied to the emitter, and it can be seen that the depletion-accumulation pattern across the collector barrier is not present. At last, Figure 4.20 shows the map of the charge modulation when no potential perturbation is applied. As expected, the accumulation-depletion pattern on the sides of the barriers are not visible. However, random charge variations can still be seen especially close to the collector, and represents the noise term discussed in Section 4.6.1.

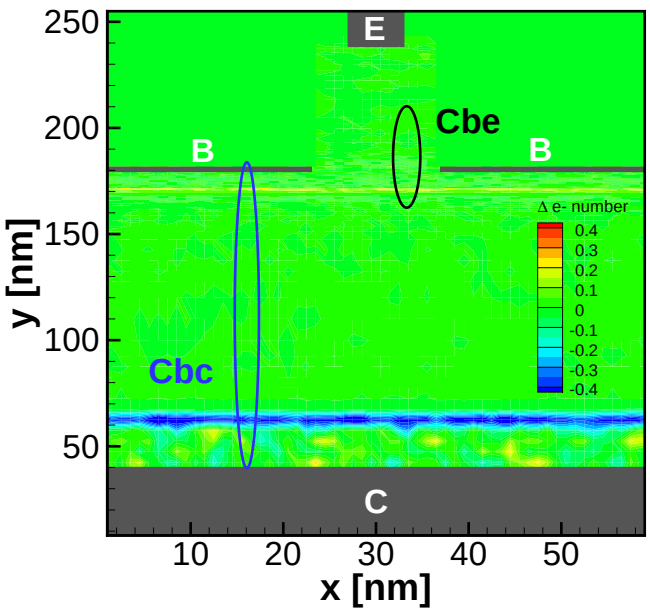

Figure 4.18: Contour plot of the accumulation / depletion of electrons following a 0.025V perturbation applied to the collector which is related to the collector-base capacitance. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5 biased at $Ie = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$.

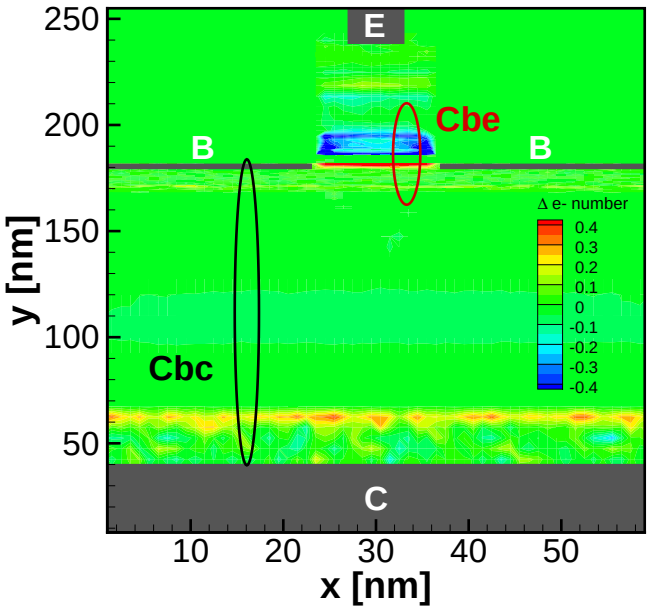

Figure 4.19: Contour plot of the accumulation / depletion of electrons following a 0.025V perturbation applied to the emitter which is related to the emitter-base capacitance. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5 biased at $Ie = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$.

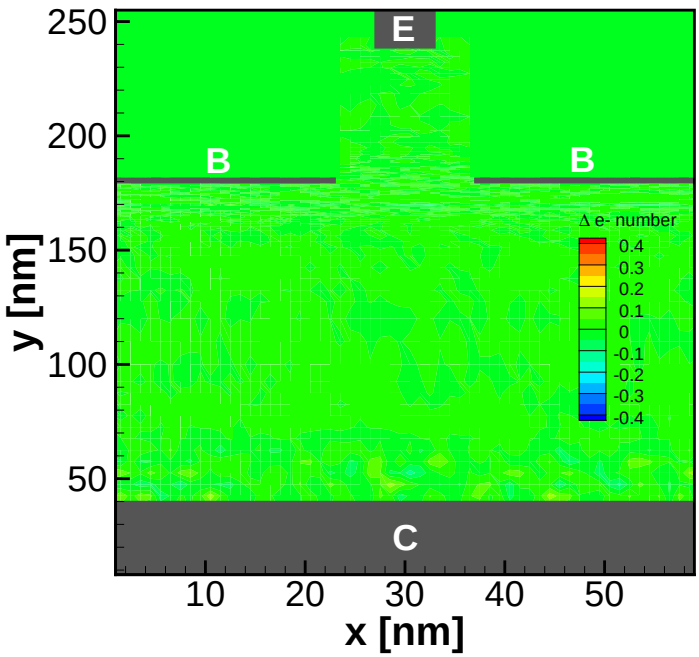

Figure 4.20: Contour plot of the variation of charge between simulation with same bias point but different seed of the random number generator. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5 biased at $Ie = 0.25\ mA/\mu^2$ and $Vcb = 2\ V$.

4.7 HET C-V Curve

An important metric used to characterize a device is the C-V relation. In particular, the measurement of the CV characteristic is very common in MOSFET, where the setup consists in short-circuiting the source and collector and applying the stimulus to the gate Chang *et al.* (1999). In a similar manner, the C-V characteristic of the GaN-HET described in Section 4.13 is calculated by grounding collector and emitter while perturbing the base and is shown in Figure 4.21. The capacitance is calculated with the method explained in Section 4.6.1, for different values of the base-emitter potential. As it can be seen from Figure 4.21, the capacitance is constant for small values of Vbe, and then it increases with Vbe until it saturates. An explanation of the C-V characteristic can be provided from the analysis of the charge modulation

introduced in Section 4.6.2. In particular, Figure 4.22 shows four maps of the charge modulation calculated for increasing values of Vbe according the the letter next to each map. From the comparison of the different plots it can be seen that as the emitter junction is more forward biased increasing Vbe, the depletion charge modulated across the emitter barrier moves towards the accumulated charge. Therefore, according to the parallel plate interpretation provided in Section 4.6.2, when the distance between the modulated charges decreases the capacitance increases, as it happens when the plates of the capacitor get closer. Eventually, for high Vbe the accumulation and depletion are separated only by the Al rich barrier and the capacitance value saturates as can be seen in Figure 4.21.

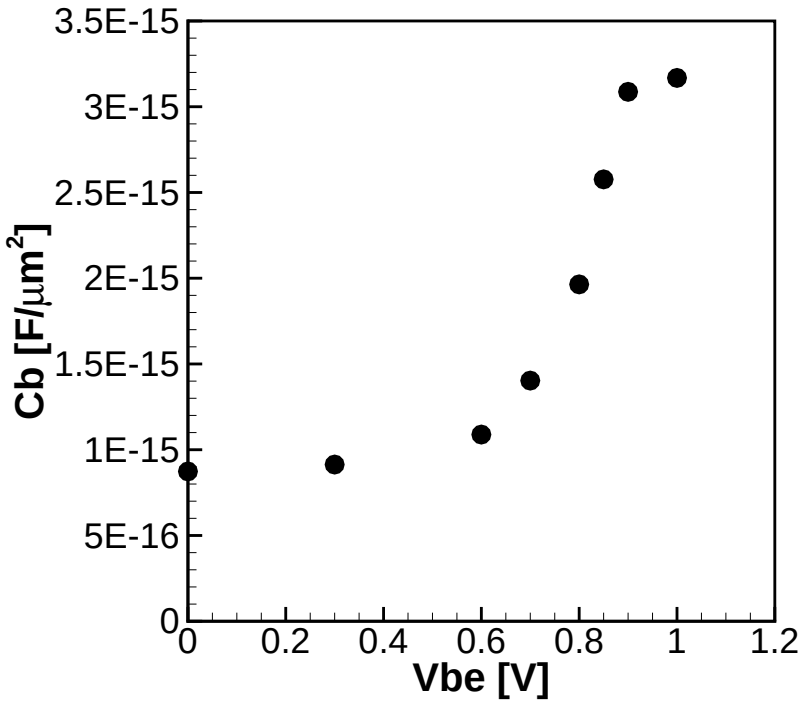

Figure 4.21: Capacitance-Voltage curve of a GaN HET calculated applying a 0.025 V perturbation to the base. Collector and emitter are kept grounded. The plot is calculated for the 2.5 nm InGaN HET described in Section 4.5.

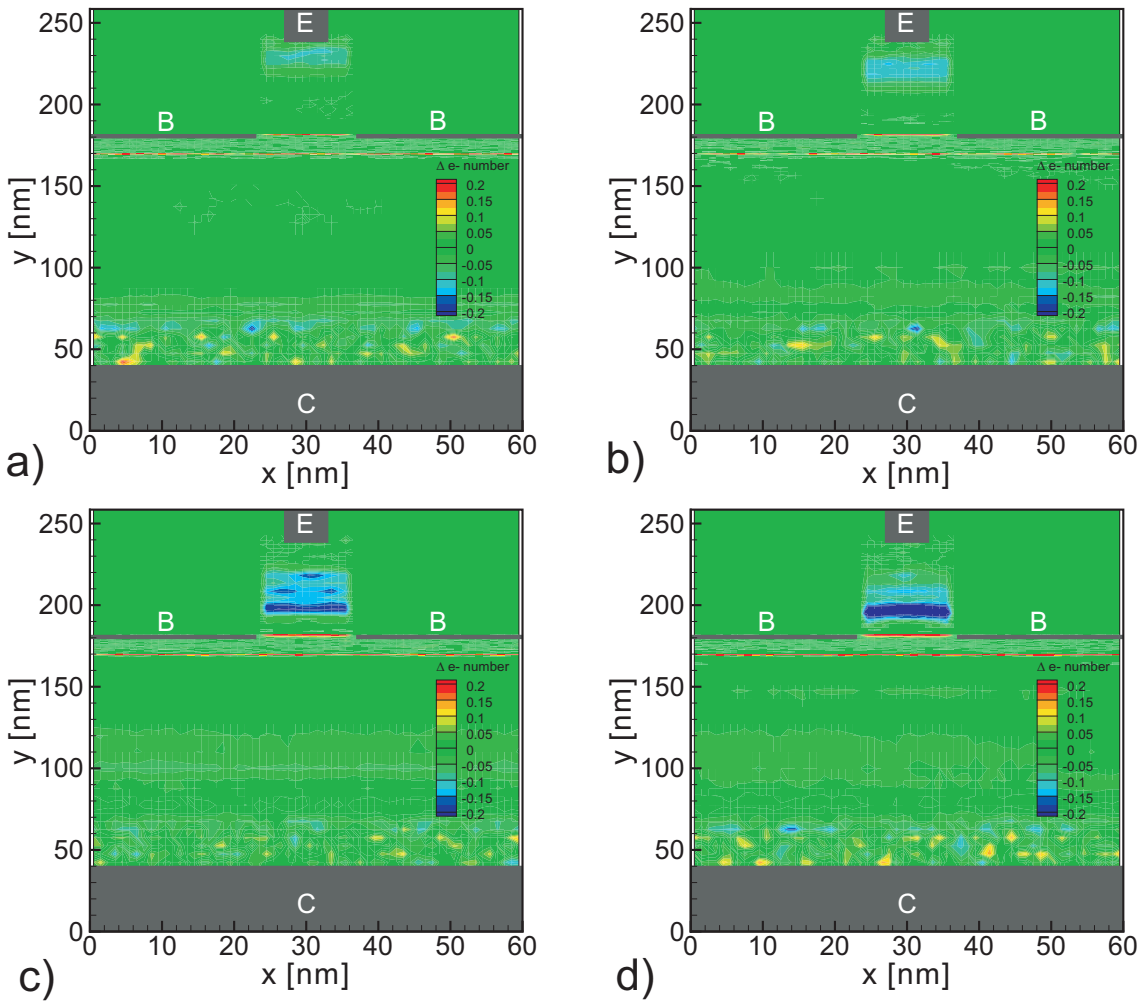

Figure 4.22: Contour plot of the accumulation / depletion of electrons following a 0.025V perturbation applied to the base. All the plots are calculated for the 2.5 nm InGaN HET described in Section 4.5 keeping the collector and emitter grounded. The base potential is 0.2 V in Fig. a), 0.6 V in Fig. b), 0.85 V in Fig. c), and 0.9 V in Fig. d).

4.8 Effect of the GaN Base Layer on f_T

Section 4.4 showed the calculation of the cutoff frequency when the size of the InGaN layer of the original GaN HET was reduced from 5 nm to 2.5 nm. This choice was motivated by the necessity of having a device with transmission coefficient larger than 0.5 but, as shown in Figure 4.12 more geometries were available. On the

other hand, we want to study devices with realistic layouts, therefore geometries with unrealistically thin layers are not considered in this work. Shortening the the InGaN layer was found to be the most effective modification to increase the transmission coefficient. In this section, the f_T of other geometries is calculated, to explore the potential of GaN HETs and to test the formulation of f_T developed in Section 4.6. Figure 4.23 shows f_T calculated for layouts where the InGaN layer is kept constant at 2.5 nm, while the size of the GaN layer is reduced. In the same plot, the transmission coefficient is shown for comparison. As it can be seen both f_T and the transmission coefficient increase in shorter bases, which suggests that the two quantities may be strictly related.

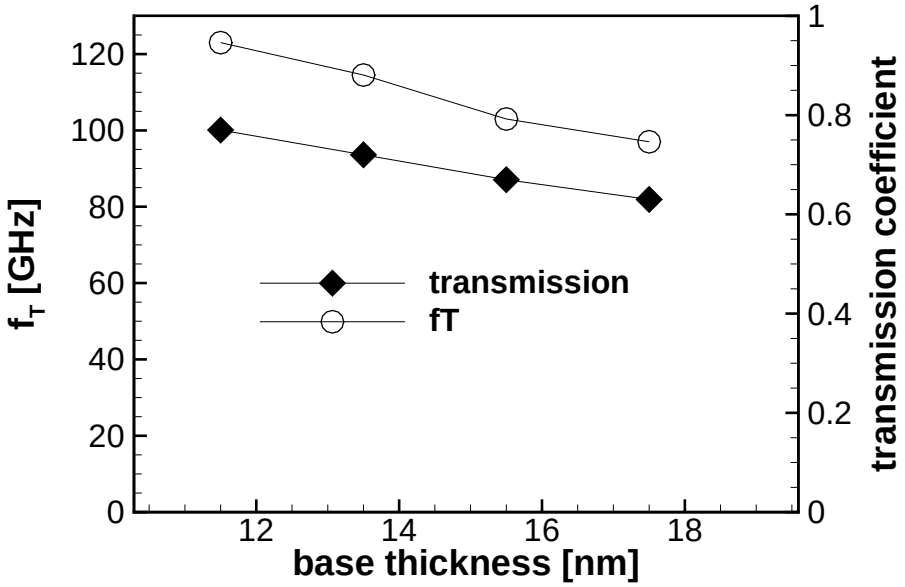

Figure 4.23: f_T (circles) and transfer coefficient (diamonds) as a function of the base thickness where the InGaN layer is kept constant at 2.5 nm and the doped GaN layer varied from 8 nm to 2 nm at steps of 2 nm. The emitter current is 0.25 mA/μ^2 and $Vcb = 2\ V$.

In order to verify this observation, we calculated the base capacitance and transconductance of the geometries considered in Figure 4.23. While the capacitance did not

show any particular trend, we can see in Figure 4.24 that g_m increases by shortening the base. This behavior of g_m can be explained by its definition:

$$g_m = \frac{dIc}{dV_{be}} \tag{4.3}$$

where Ic is the collector current and V_{be} the base-emitter voltage. Moreover, from the definition of transmission coefficient α:

$$\alpha = \frac{Ic}{Ie} \tag{4.4}$$

we can rewrite eq. 4.3 as:

$$g_m = \alpha \frac{dIe}{dVbe} \tag{4.5}$$

which shows the increase of g_m with the transmission coefficient. Finally, Figure 4.25 shows the comparison between the f_T calculated by AC simulations and the f_T calculated from the intrinsic delay time g_m/Cb. It can be seen that the two quantities are close for any of the base thickness as considered, which confirms the intrinsic delay time as dominant term in the determination of the cut off frequency.

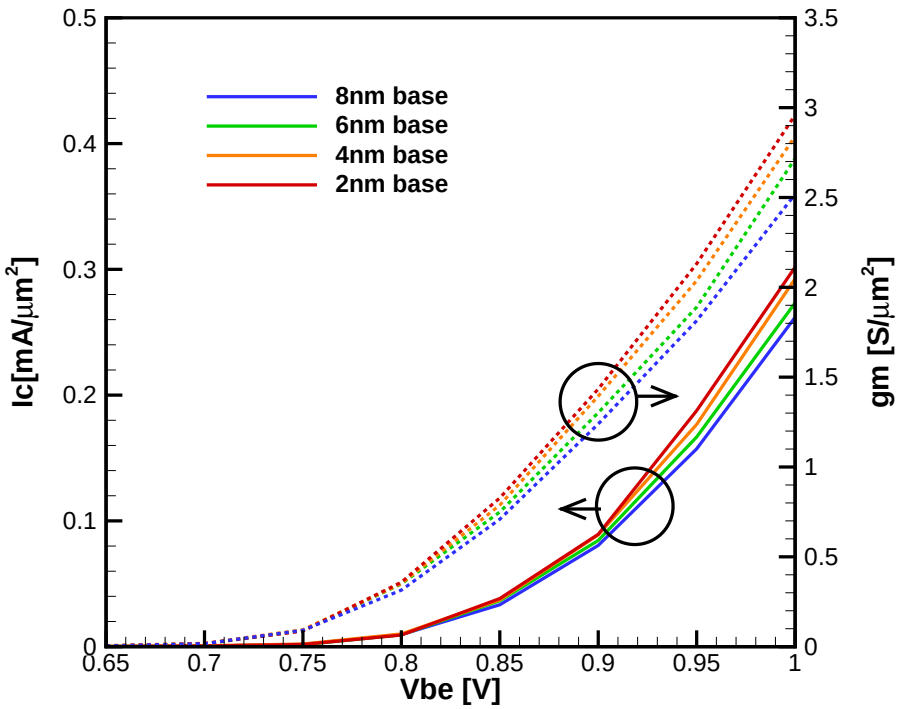

Figure 4.24: $I_c - V_b$ (solid lines) and g_m (dashed lines) calculated at $Vcb = 2\ V$ for the layouts considered in Figure 4.23.

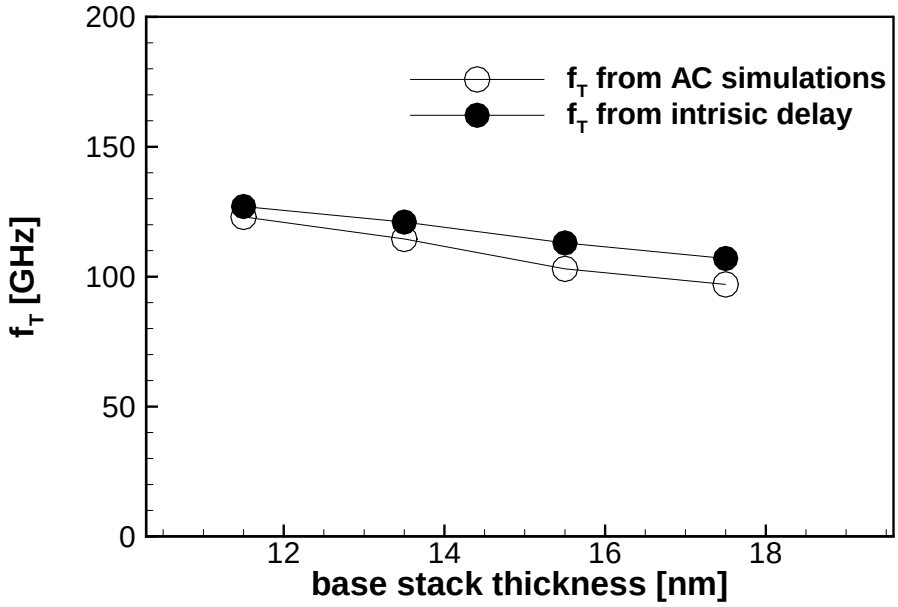

Figure 4.25: f_T calculate by AC simulation (empty circles) and f_T calculated from the intrinsic delay time g_m/Cb (full circles) as a function of the base thickness where the InGaN layer is kept constant at 2.5 nm and the doped GaN layer varied from 8 nm to 2 nm at steps of 2 nm. The emitter current is 0.25 mA/μ^2 and $Vcb = 2\ V$.

4.9 Conclusions

In this chapter, a GaN HET was modeled and simulated with Full Band Monte Carlo simulator. The simulation approach was validated by fitting the experimental $IcVc$ curves and the transmission coefficient is calculated to be 0.2, in agreement with the experimental results. The influence of the thickness of the base stack on the transmission coefficient was analyzed, allowing the design of a variation of the original GaN HET that allowed amplification. For this geometry, the frequency response was found to be a low pass filter with the flatband gain corresponding to the calculated DC current gain β. A quasistatic approach was developed for calculation of the capacitance in GaN HETs and was used to determine the base capacitance. Since the calculation of the base capacitance is independent of the cutoff frequency, we managed to prove that the intrinsic delay g_m/Cb is the dominant component of the total delay time. In particular, the intrinsic delay time was found to be approximatively 90% of the total delay time as extrapolated from the cutoff frequency calculated twith the AC simulations. Moreover, the C-V characteristic of the GaN HET was calculated and intrerpreted in terms of the accumulation and depletion patterns related to the capacitive behavior. In summary, in this chapter a series of numerical models and methods have been developed and applied to charaterize a GaN HET. These tools allowed to understand the operation of these devices and in future work they will be essential to systematiccaly explore the potential of GaN HET for high frequency operation.

Chapter 5

ENGINEERING THE EMITTER BARRIER IN HOT ELECTRON TRANSISTOR

5.1 Introduction

In Chapter 4, the layout of HETs was modified in the GaN base and collector barrier layer length with the aim of decreasing the length of the effective base and reducing the collector barrier height. However, as it was shown in Section 4.3, at the emitter barrier layer is also included in the base stack, and its effect is studied in the present chapter. For this purpose, a more recent GaN HEMT was modeled Gupta *et al.* (2015a). In Section 5.2 the validation of the model is shown by comparing the simulated common emitter $I_c - V_{ce}$ characteristic with the experimental one. In Section 5.3 the properties of the high energy electrons crossing the base are characterized in terms of their momentum, energy and velocity distribution. In Section 5.4 the original device is modified to lower the emitter barrier. The distributions in momentum space, energy and velocity are computed to the ones calculated for the original layout. A systematic study involving thickness and composition of the emitter barrier is shown in Section 5.5. For this study, the impact of the emitter barrier is shown on the DC characteristic and the cutoff frequency. Finally, Section 5.6 shows the large signal characterization of HETs through the calculation of output power, power added efficiency and transducer gain. The large-signal analysis is repeated for different collector barriers that affect the maximum power that can be delivered by the device.

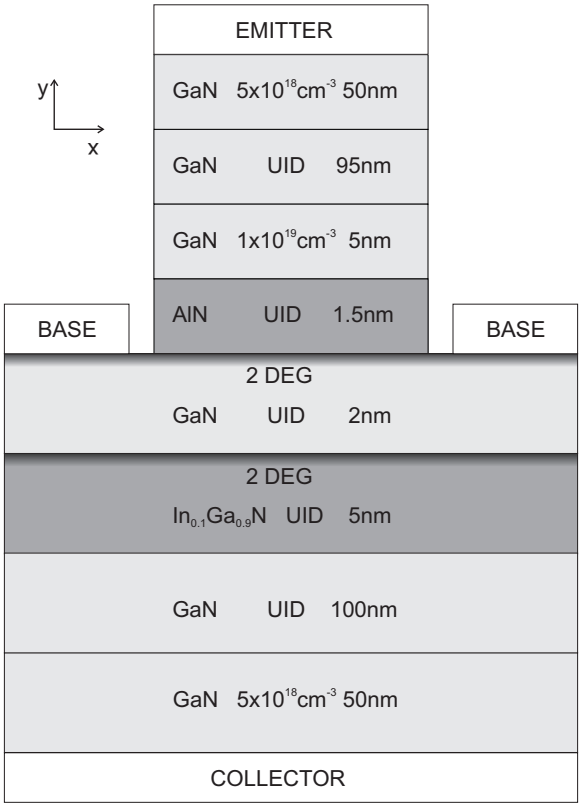

Figure 5.1: Layout of the simulated HET showing the thickness and doping of each layer Gupta *et al.* (2015a). Dimensions are not drawn in scale to highlight details.

5.2 Common Emitter DC Characteristic

An improved version of the GaN HET studied in Section 4.3 Gupta *et al.* (2014a) has been recently fabricated by the same authors Gupta *et al.* (2015a). The new layout, shown in Figure 5.1, presents the same 5 nm $In_{0.1}Ga_{0.9}N$ collector barrier, a shorter GaN base of 2 nm compared to 12 nm, and a 1.5 nm pure AlN layer for the emitter barrier. The polarization-induced 2DEG at the interface between the AlN barrier layer and the GaN base is very conductive so that the reduction of the base thickness down to 2 nm does not increase the base resistance Gupta *et al.* (2015a). The computational model of this HET follows the procedure reported in Section 4.2, where the geometry, polarization charge, and conduction band discontinuity have been updated to account for the different material stack.

The validation of our model is shown in Figure 5.2, where the the experimental common emitter $I_c - V_{ce}$ is compared with the simulated one. It can be noticed that the saturation part of the characteristic is well reproduced. Each curve in the $I_c - V_{ce}$ is calculated for the same base current that is reported experimentally. Therefore, Figure 5.2 also shows that the relation between the base current and the collector current is respected, meaning that the device current gain has been reproduced as well. As we saw in Chapter 4, this quantity is very representative of the operation of HETs, as it is the direct result of the dynamics of the hot electrons in the base. In particular, the value of the current gain of the simulation agrees to the experimentally measured value of 1, which corresponds to a transfer ratio of 0.5, meaning that only half of the electrons injected by the emitter manage to reach the collector.

On the other hand, it can be seen that the simulated HET shows a smaller Vce turn-on offset compared to the experimental results. This voltage offset is required to inversely bias the collector-base junction since at small Vce and positive Vbe both the junctions are forward biased and electrons flow from the collector contact towards the base giving a negative collector current. In particular, the collector-emitter voltage needed to reverse bias the base collector junction increases as the emitter-base forward voltage is increased. A quantitative estimation is shown in the band diagram of Figure 5.3 as:

$$V_{CEoffset} = V_{BE} - V_{BCbarrier} \tag{5.1}$$

Of the two terms, Vbe is what differs between the simulation model and the experiment. In fact, for the same biasing base current the base-emitter voltage reported in Gupta *et al.* (2015a) is 0.5V higher than what is required in the simulation. This difference is due to the emitter contact, which is treated as a low resistive ohmic contact in our simulation, while the real contact presents a higher resistance due to

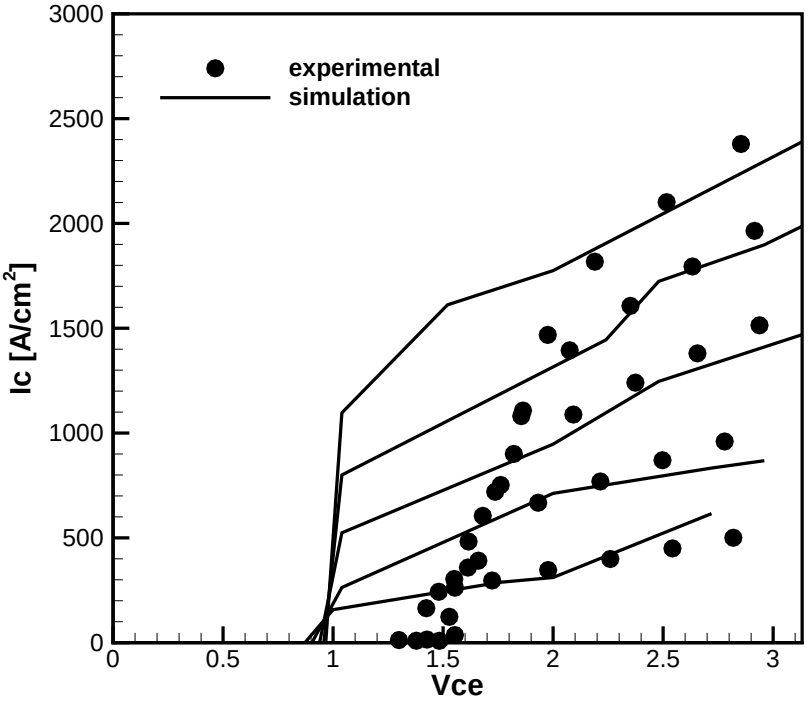

Figure 5.2: Comparison of the experimental Gupta *et al.* (2015a) and simulated $I_c - V_{ce}$ Common-Emitter characteristic. The base current is increased from 333 A/cm^2 to 1665 A/cm^2 at steps of 333 A/cm^2.

a Schottky barrier. This behavior was already observed in the emitter diode of the HET described in Chapter 4.

In particular, the diode current characteristics was calculated by Sentaurus energy balance simulations where the metal semiconductor interface is modeled by a Schottky or an Ohmic contact. By comparing these two characteristics with the experimental one, as shown in Figure 5.4, it can be seen that a good agreement is achieved for the Schottky contact model.

However, to the best of our knowledge, full simulations of Schottky diodes in Monte Carlo have never been reported, instead analytical approaches have been used to model these contacts Martín *et al.* (1996). In particular, one of the difficulties for

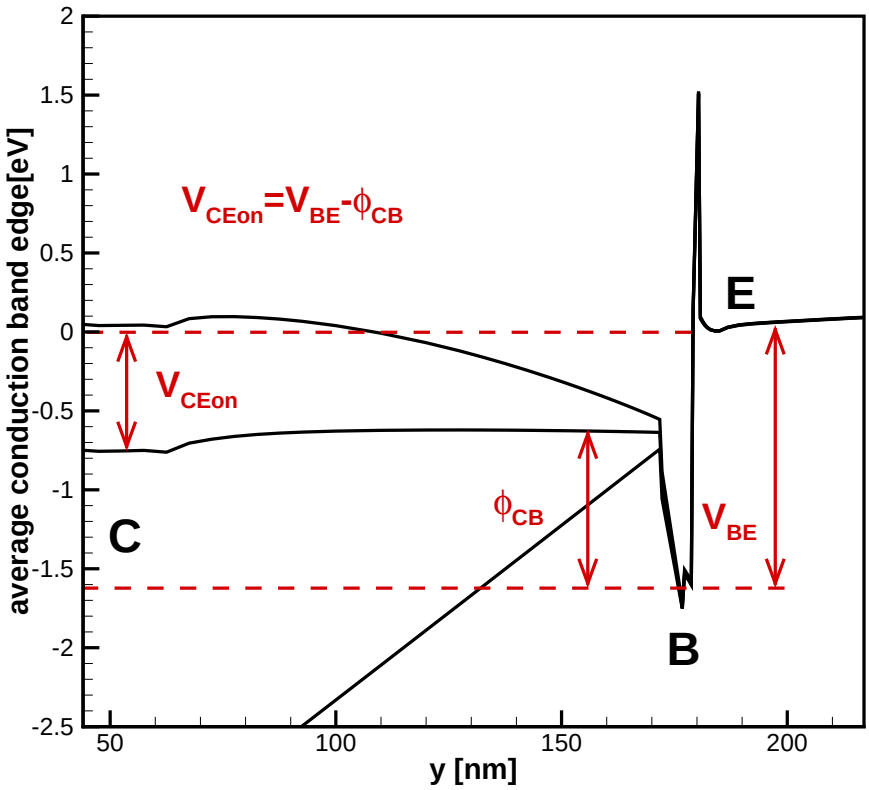

Figure 5.3: Conduction band edge diagram for the HET biased at Vce=0V, 0.5 V and 2V. Vbe=1.4V in all three cases. The collector flat band voltage is shown in the plot, and it corresponds to the bias condition when to the collector current inverts its sign .

the Full Monte Carlo modeling of the Schottky junction, is represented by the large difference in the number of electrons in the metal in respect to the semiconductor. This difference, that can be of several order of magnitudes, represents a problem since in Monte Carlo the operation of the device are described by the properties of the ensemble of electrons simulated. For example, the operation of HETs is determined by the transport of hot electrons through the base. Therefore, simulating the electrons in the metal and a statically relevant number of hot electron would require a very high number of particles in the simulation such that the calculation time would be be prohibitive. It should also be observed that the common base characteristic shown in Section 4.3 does not show discrepancies in the onset voltage as the current is plotted

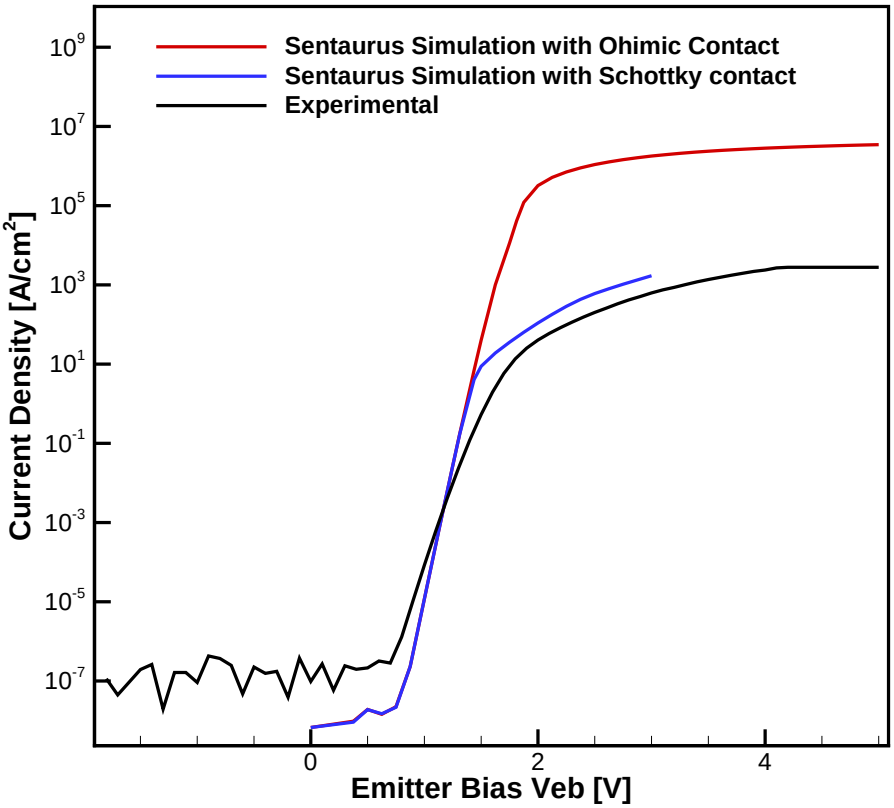

Figure 5.4: Comparison of the experimental current characteristic of the emitter diode shown in Gupta *et al.* (2014a) and two models implemented in Sentaurus, one with ohmic contacts (red) and the other with Schottky contact (blue).

against the base to collector voltage and not the emitter to collector voltage. In particular, the common base characteristic shows how the emitter current is modulated by the base collector voltage, without any direct correlation with the emitter to base voltage.

5.3 Hot Electrons Properties

Despite the substantial reduction of the base length, the current gain reported for the new HET layout is still far from the predictions shown in Section 4.4 for comparable base length. In particular, the current gain for a GaN base of 4 nm was predicted to be around 3 as opposed to the value of 1 in this layout. In order to

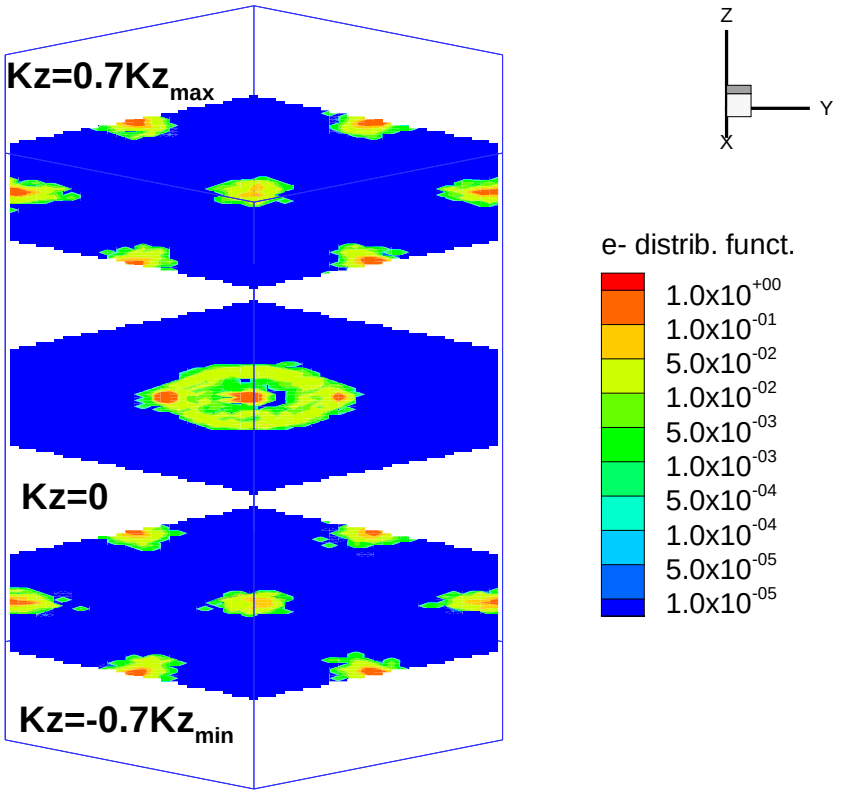

Figure 5.5: Distribution in momentum space of the hot electrons with energy higher than the collector barrier located in the GaN base. The three planes are slices of the 3D BZ1 for different values of Kz shown in the plot. The device is biased at Ib=1665 A/cm^2, Vbc=0.

identify the mechanism that is limiting the transfer of electrons, Figure 5.5 shows the distribution in momentum space of the hot electrons that are transiting across the GaN base region. It is evident that the satellite valleys are considerably populated. This condition needs to be avoided, because when electrons are injected in the satellite valleys only little part of the energy gained crossing the AlN/GaN heterojunction is converted in the kinetic energy that electrons need to overcome the collector barrier.

A graphical representation of this concept is shown in Figure 5.6. The two band diagrams represent the bottom edge of the Gamma valley and of the lower satellite valley (A valley) in GaN. As it can be seen, electrons that are injected at high energy, but below the satellite valley, are well above the edge of the collector barrier. However,

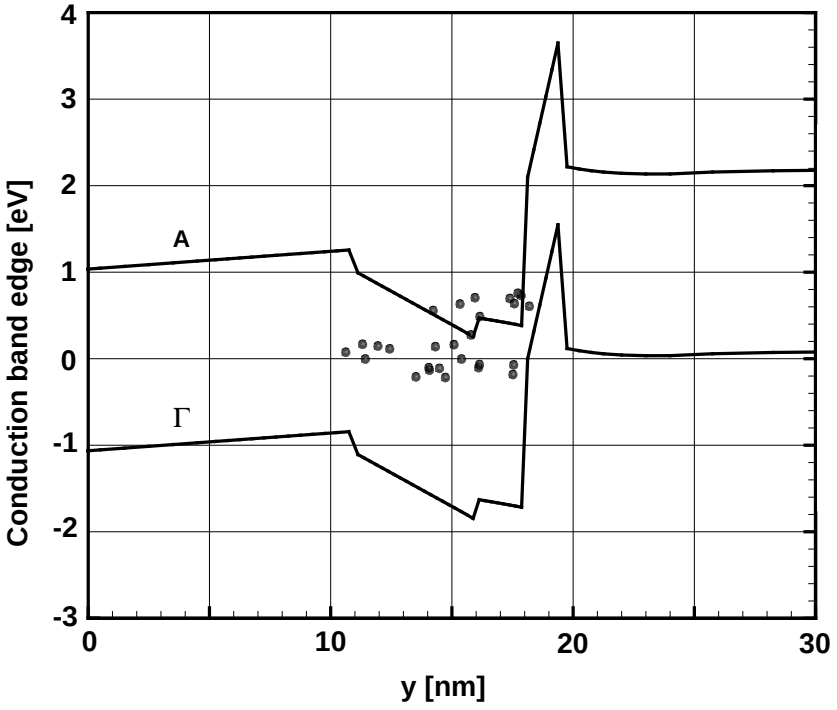

Figure 5.6: Snapshot of the hot electrons located in the GaN base. The two band diagrams are related to the bottom of the gamma and A valley in GaN. The device is biased at Ib=1665 A/cm^2, Vbc=0.

when electrons are in the satellite valley they have to overcome the barrier related to the satellite valleys that is much higher for the Gamma valley, and therefore they cannot overcome it.

Figure 5.8 shows the distribution of electrons according to their absolute energy referred to the Fermi level of the emitter contact, in different areas of the base. The same energy scale is used in the inset of Figure 5.8 to represent the HET band diagram, in such a way to provide a direct comparison of the energy distribution with the band diagram. The three distributions are normalized to the maximum value of the three, therefore in general only one distribution is normalized to one. Moreover, the dashed line shows the energy of the bottom of the satellite valleys in GaN which corresponds

to 2 eV from the bottom of the gamma valley. The red distribution, calculated in the pure GaN section of the base, shows a pronounced peak, which describes the injection energy of the electrons, at energy higher than the bottom of the satellite valley, in agreement with the satellite valleys population seen in Figure 5.5. On the other hand, it can be seen that the green distribution calculated in the first section of the $InGaN$ barrier shows a more relaxed distribution due to the high scattering probably for electrons in the satellite valleys. Finally, the blue distribution peaks at a value lower than the bottom satellite valleys, as the electric field due to the collector barrier has reflected most of the electrons of the satellite valleys. In fact, as can be seen in Figure 5.7, the satellite valleys population in the InGaN layer has almost disappeared.

The relaxation effect of the satellite valleys scattering can also be seen in the distribution of the velocity component along the direction of transport plotted for the hot electrons in Figure 5.9. The central bell-shape structure that characterizes the green and the red distribution represents the electrons in the satellite valleys where the intense scattering has randomized their velocity. The negative peak in the red distribution corresponds to the injection velocity and it can be seen in the green distribution how this peak is reduced by effect of the scattering. The shape of the blue distribution is completely different as the bell-shape feature attributed to the satellite valley electrons is not present anymore. Again, this is due the fact that only the hot electrons with high kinetic energy in the Gammma valley can reach this portion of the barrier. Note that the negative sign of the velocity is coherent with the coordinate system shown in Figure 5.1.

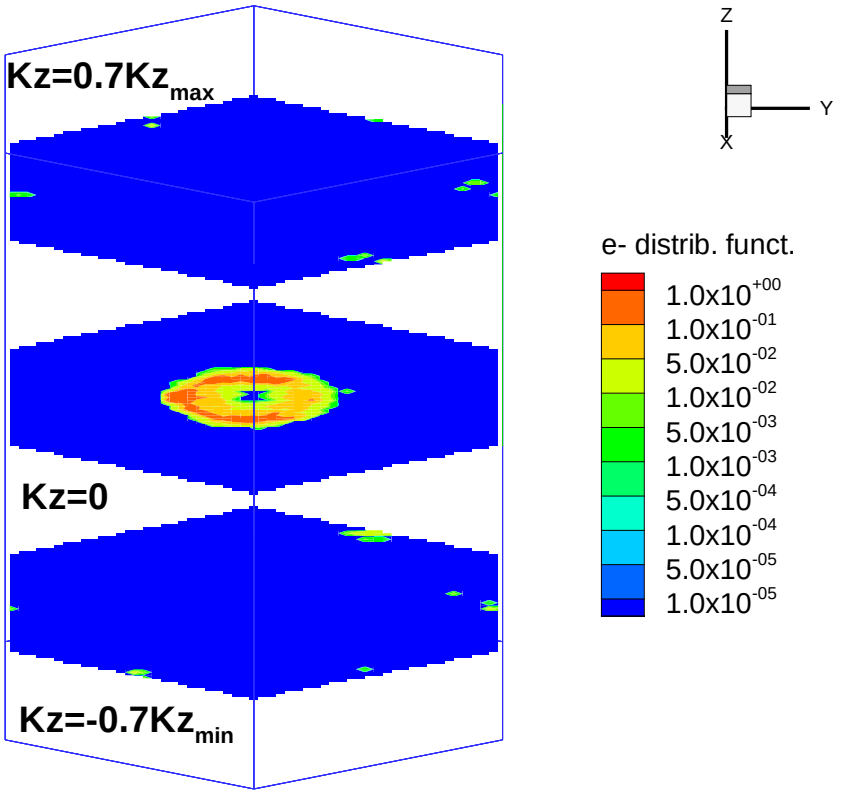

Figure 5.7: Distribution in momentum space of the hot electrons located in the collector side of the InGaN barrier (blue rectangle in Figure 5.8). The three planes are slices of the 3D BZ1 for different values of Kz shown in the plot. The device is biased at Ib=1665 A/cm^2, Vbc=0.

5.4 Hot Electrons Properties in Improved Layout

We saw that the injection energy plays a crucial role in determining the properties of the hot electrons and the operation of HETs. The main elements that control the injection energy is the design of the emitter-base junction and, to a second order, the emitter base bias, as it influences the distribution of electrons of the emitter barrier. In particular, the height of the emitter barrier can be decreased either by incorporating Ga in the AlN emitter barrier forming a $AlGaN$ barrier layer like in Gupta *et al.* (2014a), or by reducing the thickness of the AlN layer. In particular, Figure 5.10 shows the effect of a reduction of the AlN layer from 1.5 nm to 1 nm, in

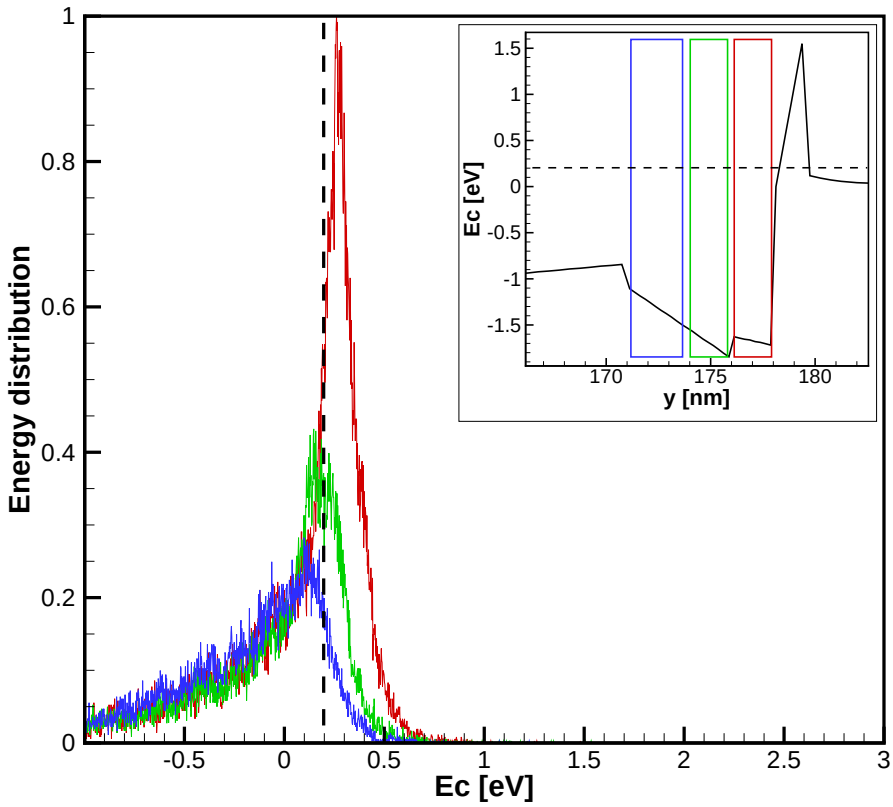

Figure 5.8: Energy distribution of the hot electrons in 3 areas of the HET base shown in the band diagram in the inset. The area and related distribution corresponds to the same color. The energy in the x-axes on the main figure corresponds to the energy shown in the y-axes of the inset. The dashed line indicates the bottom of the lowest satellite valley in the GaN base. Ib=1665 A/cm^2, Vbc=0.

the momentum space distribution of the hot electrons in the base. Indeed, the satellite valleys are depopulated as all the electrons have small momentum corresponding to the area near the center of the Brillouin zone shown in the middle of the 3 slices in Figure 5.10. More in detail, it can be noticed that a small area close to the middle of this central slice is empty. This is an artifact due to the energy threshold used to discriminate the hot conduction electrons from the cold electrons in the base, which are not considered in this plot.

The energy distribution of the 1 nm AlN device is shown in Figure 5.11. As opposed to what was seen in Figure 5.9, the energy of the electrons does not exceed

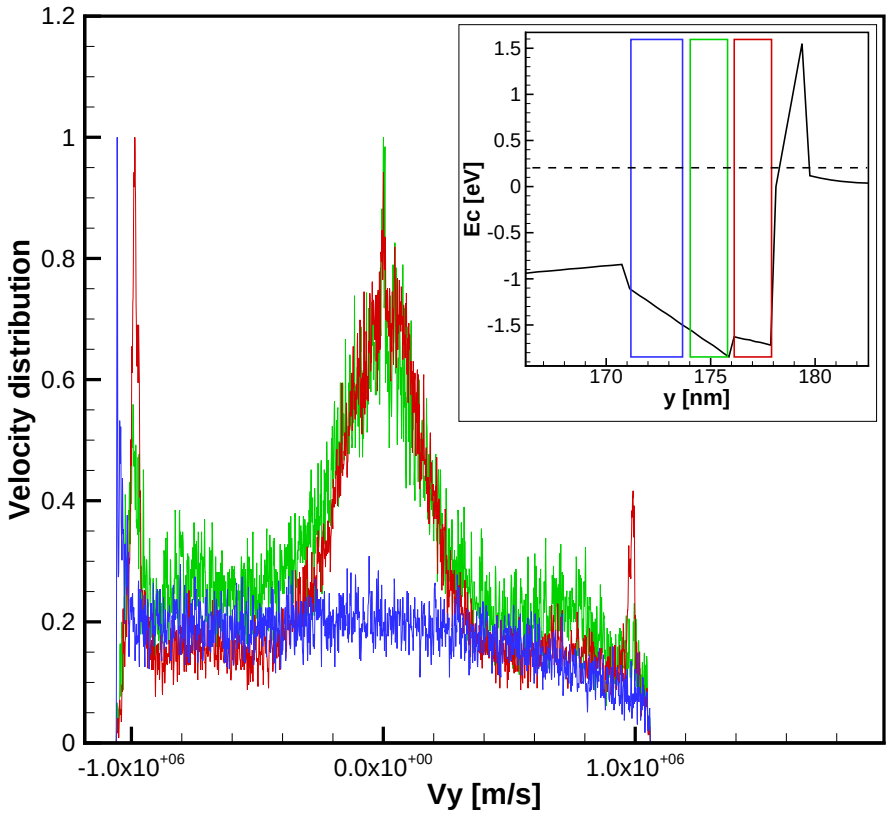

Figure 5.9: Distribution of the velocity component along the transport direction, calculated for the hot electrons in 3 areas of the HET base shown in the band diagram in the inset. The area and related distribution corresponds to the same color. Ib=1665 A/cm^2, Vbc=0.

the threshold voltage for satellite valley transfer. Moreover, the three distributions are more similar among them than in Figure 5.9 as a result of the lower number of scattering events. However, the energy relaxation effect of scattering can still be seen as the peak of the distributions shift towards smaller energies for the regions closer to the collector. The similarity among the three regions is even more evident in the velocity distribution shown in Figure 5.12. The peak corresponds to the velocity of most of the hot electrons, while smaller or negative velocities characterize the population of electrons that scattered multiple times. It may surprise that all the electrons in the GaN region and the beginning of the InGaN seem to have all the same velocity as the peaks in Figure 5.12 are very prominent. However, it can be

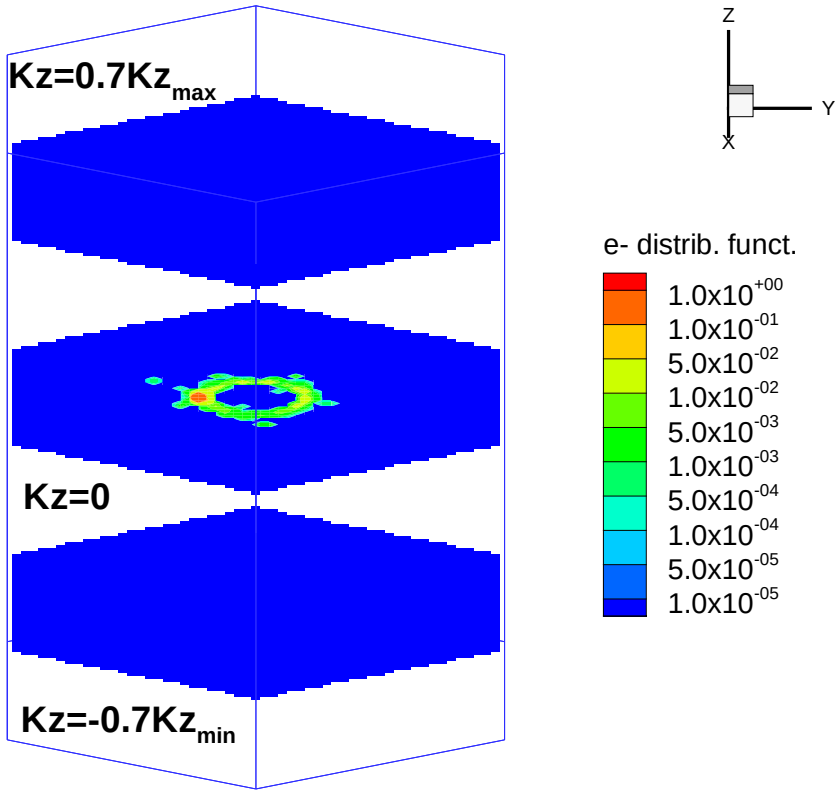

Figure 5.10: Distribution in momentum space of the hot electron with energy higher than the collector barrier located in the GaN base of the 1nm AlN HET. The three planes are slices of the 3D BZ1 for different values of Kz shown in the plot. The device is biased at Ib=1665 A/cm^2, Vbc=0.

noticed that the GaN band structure displays a linear dispersion in the range of energy where most of the electrons are seen in Figure 5.11, and therefore the same velocity can be associated to different energies. As a result of the optimization of the injection energy aimed to avoid satellite valley transfer, the current gain is calculated to increase to 3 as opposed to 1 in the original HET.

5.5 Emitter Barrier Scaling

In Section 5.4 it was shown that the emitter-base junction plays a pivotal role in the device performance. In this section we examine layouts where the thickness of the emitter-base barrier is changed as well as its composition, while everything else is kept

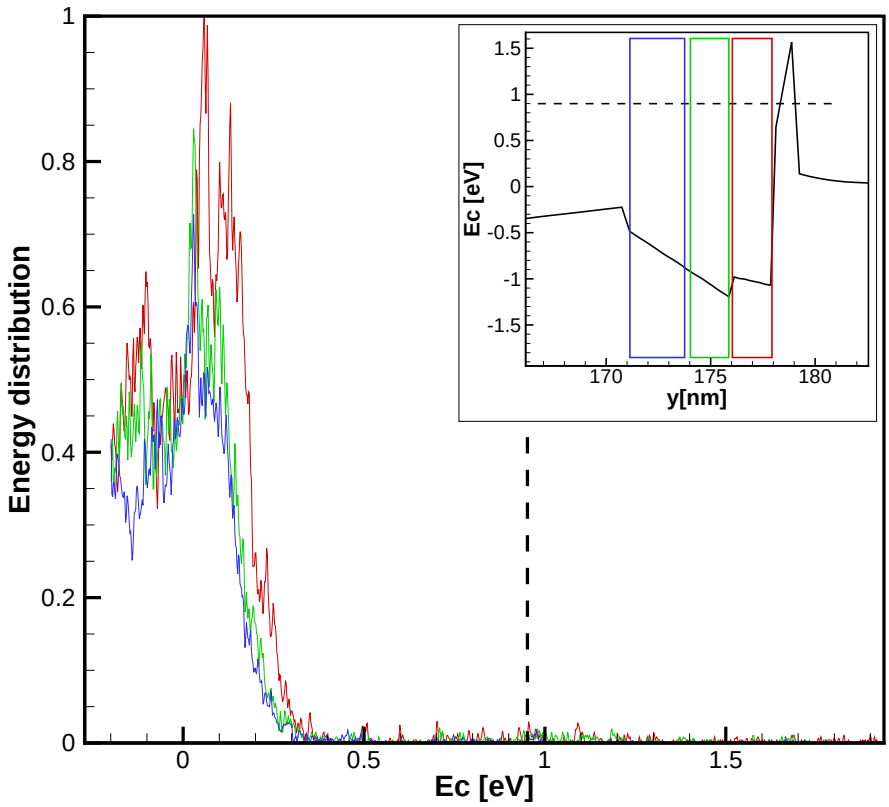

Figure 5.11: Energy distribution of the hot electrons in 3 areas of the 1nm AlN HET base shown in the band diagram in the inset. The area and related distribution corresponds to the same color. The energy in the x-axes on the main figure corresponds to the energy shown in the y-axes of the inset. The dashed line indicates the bottom of the lowest satellite valley in the GaN base. Ib=1665 A/cm^2, Vbc=0.

as the original layout. In particular, different AlGaN alloys are considered for the barrier layer in agreement with previous work on GaN HETs Gupta *et al.* (2014a). The effect of barrier thickness and alloy composition is shown in the band diagrams of Figure 5.13. In particular, two features can be identified that form the barrier: a steep increase at the GaN/AlGaN interface and a linear increase over all the thickness of the AlGaN layer. The steep increase at the interface is due to the conduction band discontinuity between the two materials, which was found to be around 70% of the difference of the bandgaps for the GaN/AlGaN system Ambacher *et al.* (2002). The linear part is due to the electric field induced by the polarization discontinuity of the

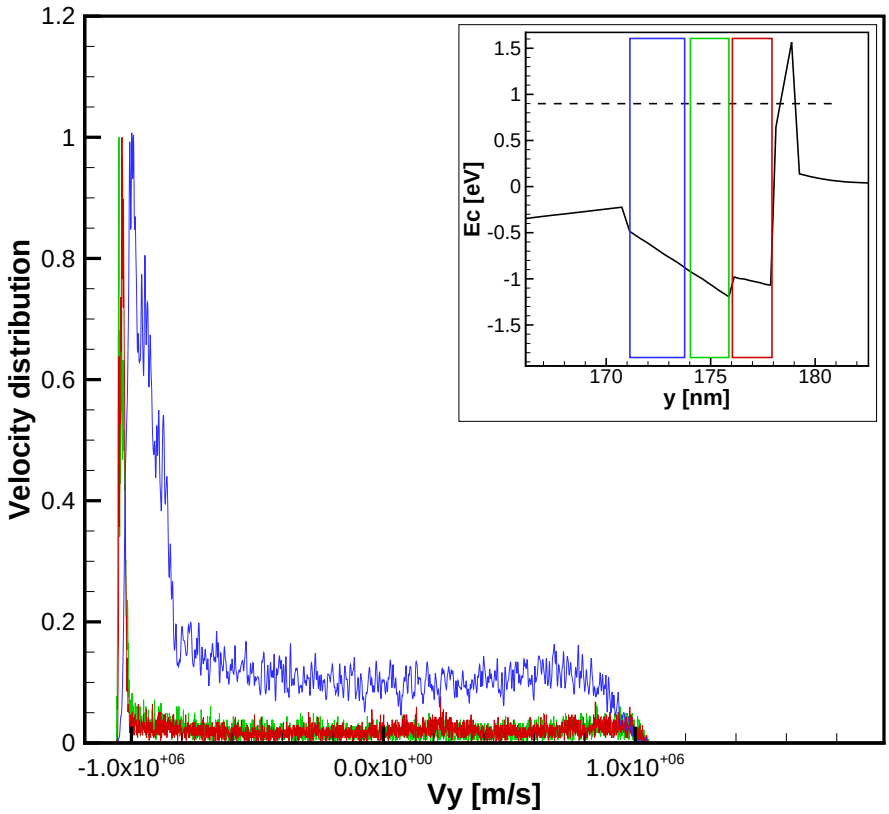

Figure 5.12: Distribution of the velocity component along the transport direction, calculated for the hot electrons in 3 areas of the 1nm AlN HET base shown in the band diagram in the inset. The area and related distribution corresponds to the same color. Ib=1665 A/cm^2, Vbc=0.

GaN/AlGaN system. Both the band discontinuity and the electric field across the barrier are increased by increasing the Al percentage in the AlGaN barrier. However, higher barrier height can be obtained also for Ga rich AlGaN barrier by exploiting the effect of the polarization-induced field increasing the thickness of the AlGaN layer.

In Figure 5.14 the current gain is plotted versus the barrier thickness for emitter barrier layers with increasing Ga content. Regardless the barrier material it can be seen that current amplification is achieved only when the thickness is included in a specific range of values. In fact, thinner layers creates an emitter barrier that is too low to inject carriers with enough kinetic energy to overcome the collector barrier. On the other end, thicker barriers inject the electrons directly in the satellite valleys,

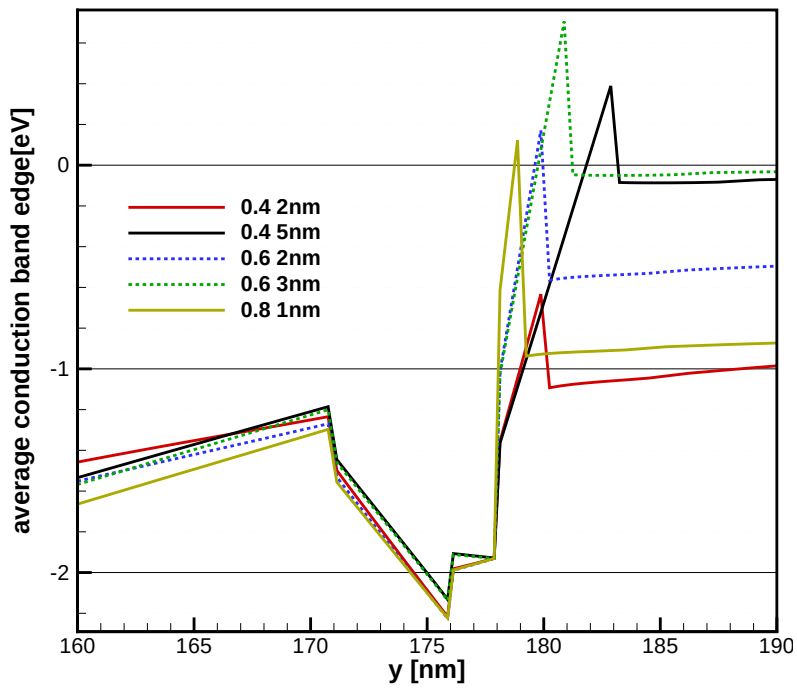

Figure 5.13: Band diagram of the HET in Figure 5.1 where the original 1.5 nm AlN emitter barrier has a thickness shown in the legend.

in analogy with what was observed in the original layout in Section 5.4. In the range of thickness where amplification is observed, the values of the current gain can be seen to be maximized by an intermediate value of barrier thickness. In fact, at lower injection energies the current gain is limited by the electrons that loose part of their energy through scattering, while higher injection energy progressively increases the population of the satellite valleys. Moreover, it can be seen that the optimal value of the barrier thickness varies with the $AlGaN$ molar fraction, because of layer band discontinuity and the polarization charge discussed before.

The effects of the emitter barrier height on output current and transconductance are shown in Figure 5.15 in relation with the HET $I_c - V_{be}$. The layouts considered in Figure 5.15 have an emitter barrier of $Al_{0.4}Ga_{0.6}N$ with thickness varying from 5 nm to 2 nm. Looking at the currents, it can be seen that the threshold voltage increases with the thickness of the barrier as the height of the emitter base diode increases, and

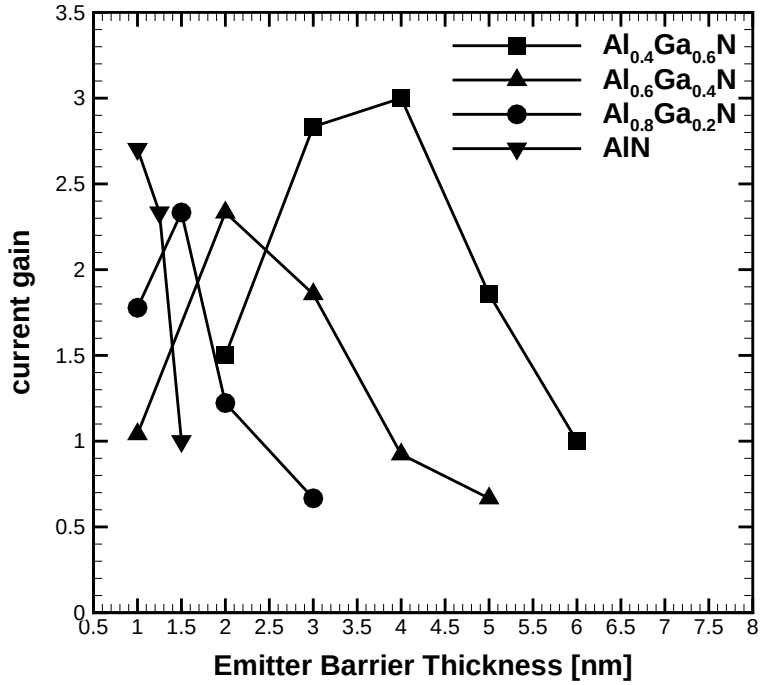

Figure 5.14: Current gain calculated for layouts as in Figure 5.1 where the emitter barrier is changed in the material according to the legend and in its thickness as shown in the x-axis. All current gains are calculated at peak g_m and Vce=5V.

a larger forward bias is required to turn on the diode. Moreover, for higher values of Vbe it can be seen that g_m starts to degrade. This is due to the second order dependence of the electron injection energy to the base-emitter voltage mentioned in Section 5.4. In particular, for a base-emitter voltage above 2.5 V, electrons start to be injected in the satellite valleys even for thinner barriers, hence degrading the current gain and g_m.

Figure 5.16 shows the HET cutoff frequency reported as a function of the emitter barrier thickness. The cutoff frequency tends to be maximized for lower values of the barrier thickness in the amplification range, i.e. $\beta > 1$. Similarly, g_m was observed in Figure 5.15 to be higher when the AlGaN barrier was closer to the collector barrier than to the satellite valleys. The similarity between g_m and f_T is in agreement with the direct proportionality between the two quantities described in Section 4.6.

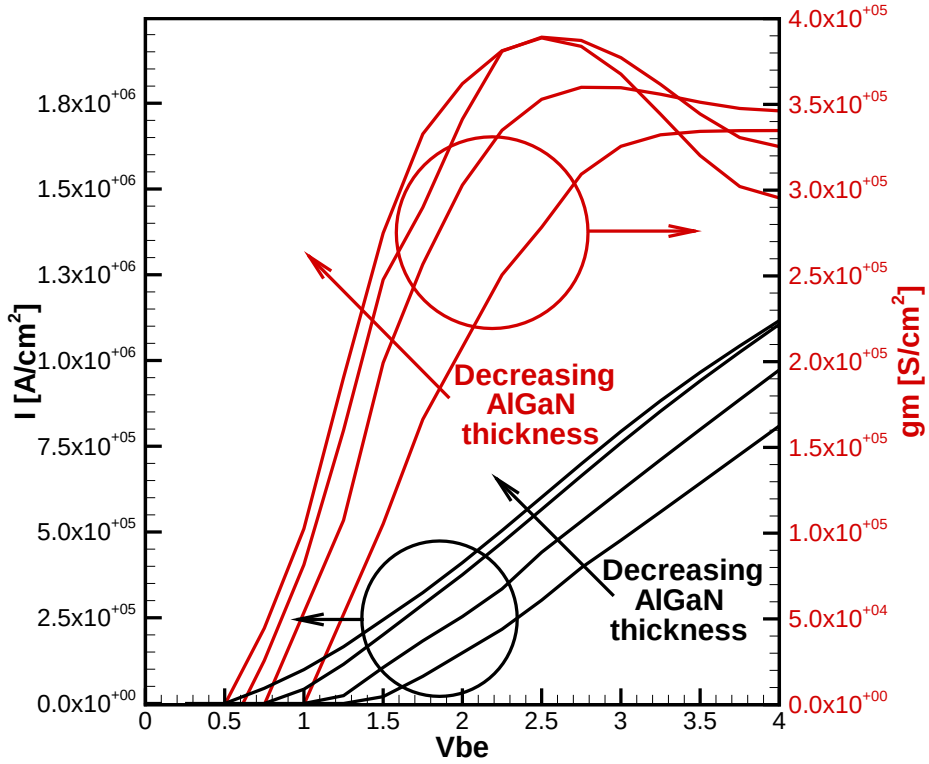

Figure 5.15: $I_c - V_{be}$ characteristic and transconductance of layouts as in Figure 5.1 but with $Al_{0.4}Ga0.6N$ emitter barrier of thickness increasing from 2nm to 5nm at 1nm steps according to the arrows in figure. All the currents are calculated at Vce=5V.

The peak f_T reported in Figure 5.16 is 270 GHz, which is obtained by biasing the specific HET layout with a base current density of $2x10^5$ A/cm^2. However, upon increase of the emitter doping to $5x10^{18}$ cm^{-3}, values of f_T above 500 GHz have been predicted. In fact, the higher doping in the emitter area increases the driving current and consequently g_m. In particular, with the reported doping of $5x10^{18}$ cm^{-3}, the peak g_m occurs at a base current about $2x10^7$ A/cm^2 and a collector current of $5x10^7$ A/cm^2 . However, for these levels of current, the performance of the device is likely to be influenced by self heating phenomena. Therefore, the inclusion of self heating, which is not yet available in the code at the current time, is needed to confirm HET theoretical cutoff frequencies above half terahertz.

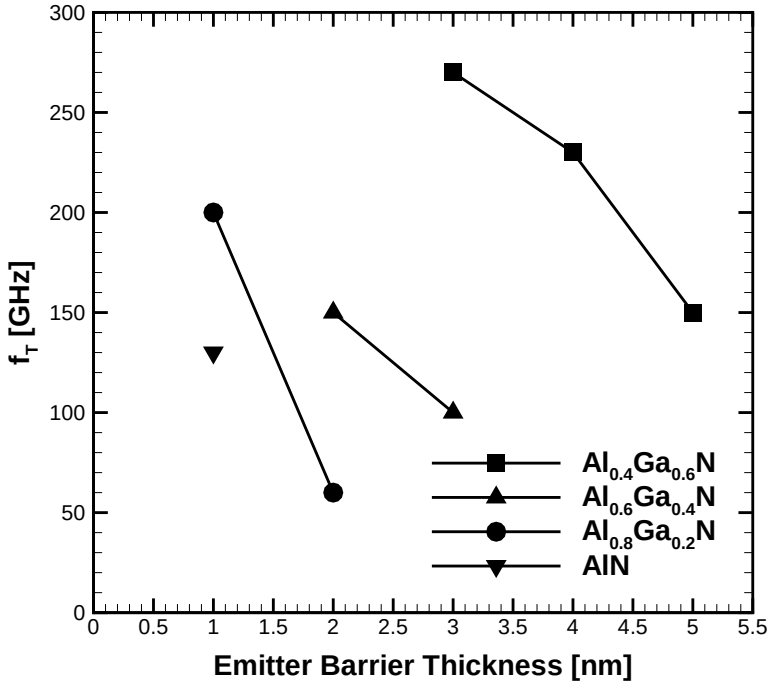

Figure 5.16: f_T calculated for devices differing from Figure 5.1 only in the material and thickness of the emitter barrier, as indicated in the plot. All the f_T are calculated at peak g_m and Vce=5V.

5.6 Large Signal Analysis

Another important aspect to evaluate the performance of HETs is the capability to perform as power amplifier. For this purpose the full band CMC device simulator is self-consistently coupled with an Harmonic Balance (HB) solver. More details on the specific implementation can be found in Guerra *et al.* (2011b). In order to perform a HB simulation, a sinusoidal input signal is applied to the input contact which, in the common emitter configuration, is represented by the base. The amplitude of the input voltage swing is increased to perform a sweep of the input power until the device is driven to a maximum current I_{max}. The DC value of the input signal is fundamental to determine the class of operation of the amplifier. The simplest class of operation is called class A where the DC bias is chosen such that the device is always ON during the whole cycle of the input signal. For this purpose, the input DC

voltage is calculated as the average between the threshold voltage and base-emitter voltage. The maximum output power can be achieved by maximizing the voltage swing on the load, or equivalently the collector terminal of the device. This voltage swing is limited by the knee voltage for small values of Vce and high currents, and by the transistor breakdown voltage for high values of Vce and small currents.

5.6.1 HET $I_c - V_{ce}$

The simulated common emitter $I_c - V_c$ of a GaN HEMT with 5 nm $Al_{0.4}Ga_{0.6}N$ emitter barrier is shown in Figure 5.17. A sudden increase of the current is seen as the emitter-collector voltage exceeds certain values, and it represents the breakdown of the device. However, it can be noticed that the ON-state breakdown seems to occur at voltages higher than for the OFF-state. This behavior is opposite to what is usually seen in HEMTs Somerville *et al.* (1999). In fact, in HEMTs the ON-state breakdown is due the avalanche generation of carriers which is increased for higher currents. However, the same $I_c - V_{ce}$ characteristic and breakdown is obtained for the simulated HETs when the electron-hole pair generation mechanism is deactivated during the simulation, therefore avalanche generation is not the mechanism responsible for the current increase. On the other hand, when tunneling across the collector barrier is turned OFF, the breakdown does not occur as shown Figure 5.18. In fact, as the collector-to-base voltage is increased, the collector barrier that confines the cold electrons in the base gets thinner and thinner until these electron start tunneling through it. The amplitude of this current only depends on the collector to base voltage and is not related to the collector current as in the avalanche breakdown. Therefore, this additional term in the collector current is the same for all the curves of the $I_c - V_{ce}$ but its contribution becomes relevant only when its value is comparable with the hot electron current. For this reason, the ON state breakdown results to be

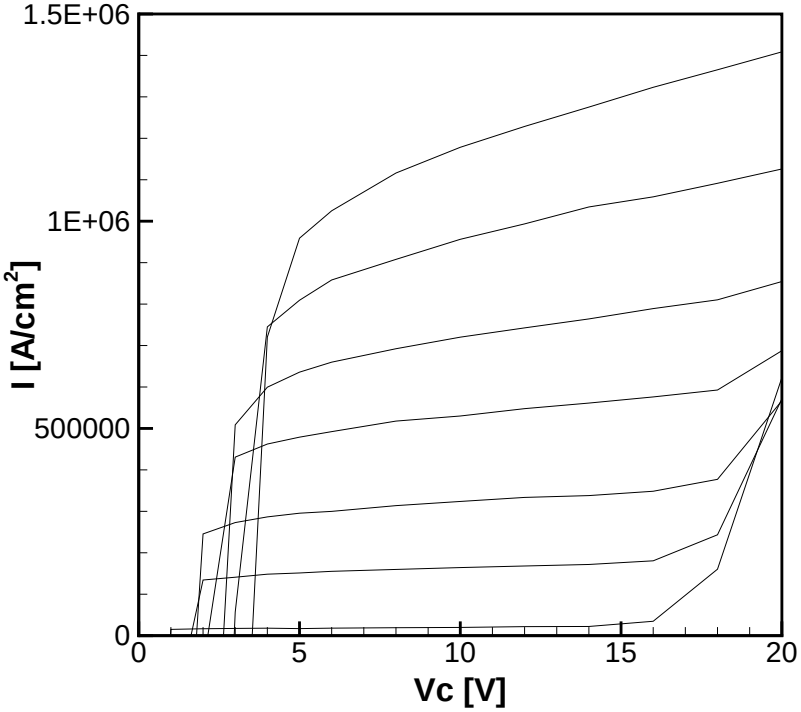

Figure 5.17: $I_c - V_{ce}$ of a 5 nm $Al_{0.4}Ga_{0.6}N$ emitter barrier HET.

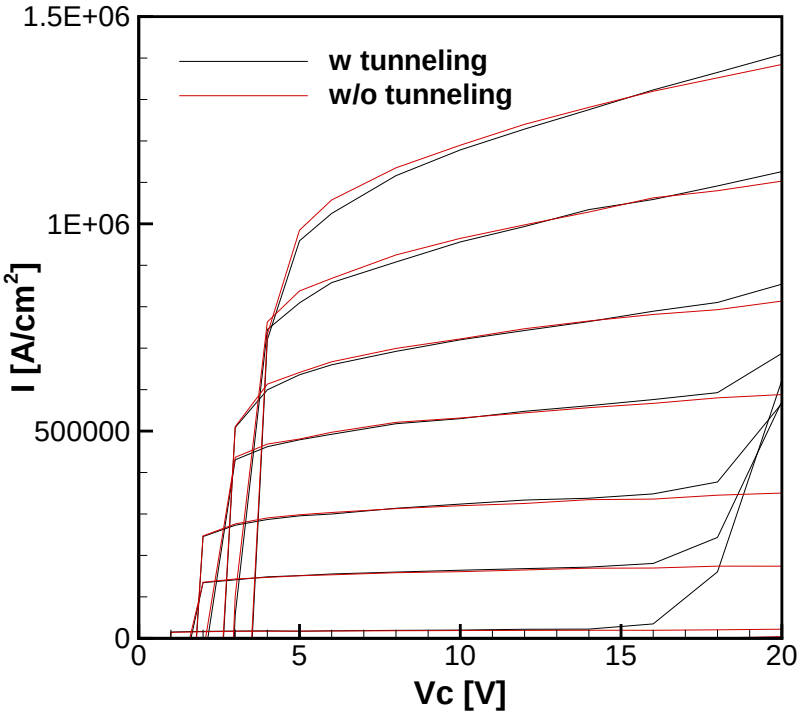

Figure 5.18: $I_c - V_{ce}$ of a 5 nm $Al_{0.4}Ga_{0.6}N$ emitter barrier HET (black) compared to the same $I_c - V_{ce}$ calculated without enabling tunneling though the collector barrier (red).

increased for higher curves of the $I_c - V_{ce}$.

As reported in Gupta *et al.* (2014a), the collector barrier is due to the polarization dipole present at the ends of the InGaN layer. Similarly to what was shown in Section 5.5, the height of the barrier can be reduced by decreasing the InGaN layer thickness. The effects of such modification have been already discussed in terms of current gain and cutoff frequency in Section 4.4 and Section 4.5, where we concluded that smaller collector barriers were beneficial for the device performance as long as the leakage could be controlled. On the other hand, Figure 5.19 shows that the OFF-state breakdown voltage is increased by thicker InGaN barrier layers, as a higher collector-base voltage is needed to thin the barrier to the point that electrons can tunnel through.

The optimal load lines for the devices are also shown in Figure 5.19. In particular, the $I_c - V_{ce}$ limiting the current to the maximum value of $1x10^6$ A/cm^2 is shown in red, while the low currents curve limiting the maxim voltage are shown in blue. It can be noticed that only one $I_c - V_{ce}$ is shown corresponding to I_{max}. This is because the knee voltage is not seen to be affected by scaling the $In_{0.1}Ga_{0.9}N$ layer, therefore only one curve is plotted. The real part of the load applied to the devices to perform the HB simulation is calculated as the inverse of the slope of the schematic load line shown in Figure 5.19. The imaginary part of the load is instead zero in order to simulate a purely resistive load. In particular, Figure 5.20 shows the good agreement between the load line calculated with the HB algorithm and the schematic one used to design the optimal load.

5.6.2 Power Characterization

The output power and power added efficiency of the devices considered in Figure 5.19 are shown in Figure 5.21, while the transducer gain is reported in Figure 5.22.

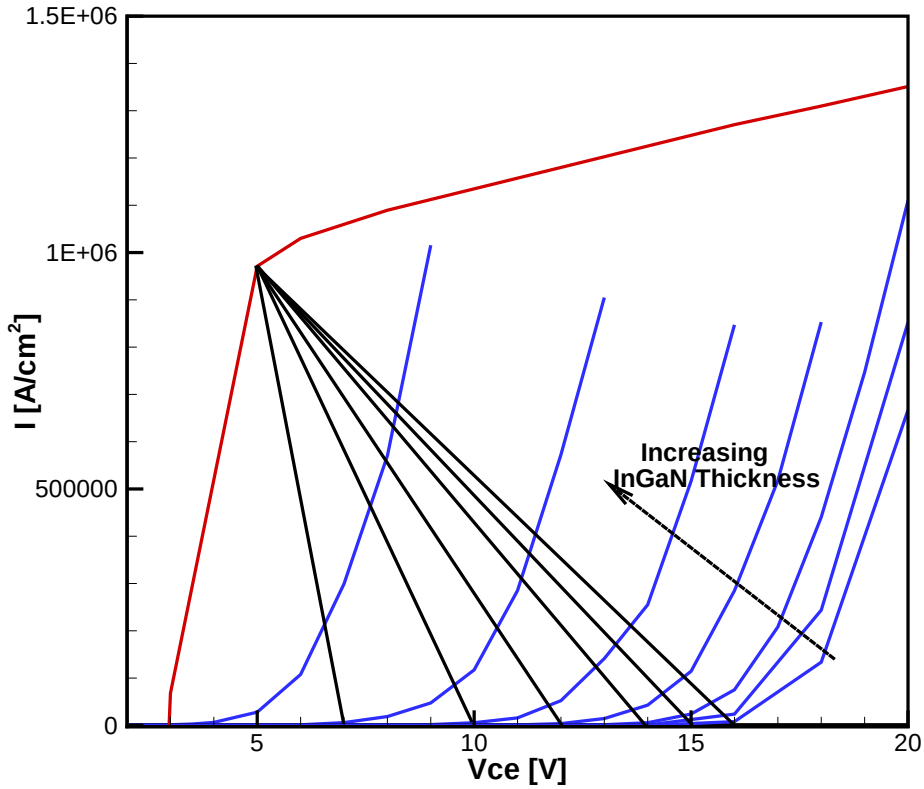

Figure 5.19: In blue the $I_c - V_{ce}$ calculated at Ib=0 is plotted for variation of the device in Figure 5.1 with emitter barrier of 5 nm $Al_{0.4}Ga_{0.6}N$ and thickness of the InGaN decreasing from 8 nm to 2 nm at 1 nm steps according to the arrow in the figure. In red the $I_c - V_{ce}$ corresponding to an I_{max} of $1x10^6$ A/cm^2 common to all devices. In black the schematic load lines used to calculate the optimum load resistance.

As expected the maximum power that can be delivered to the load increases for devices with thicker collector barrier, as a bigger voltage swing is enabled by the higher breakdown voltage. However, it can be noticed in Figure 5.22 that the transducer gain follows the opposite trend. The transducer gain represents the large signal equivalent of the small signal current gain introduced in Section 2.6 to calculate the cutoff frequency. In particular, the transducer gain is calculated as the ratio between the AC output power and the AC input power. Intuitively, it can be interpreted as a power gain that can be obtained as the product of current gain and the voltage gain. Therefore, the decrease of the transducer gain with thicker InGaN barrier can be

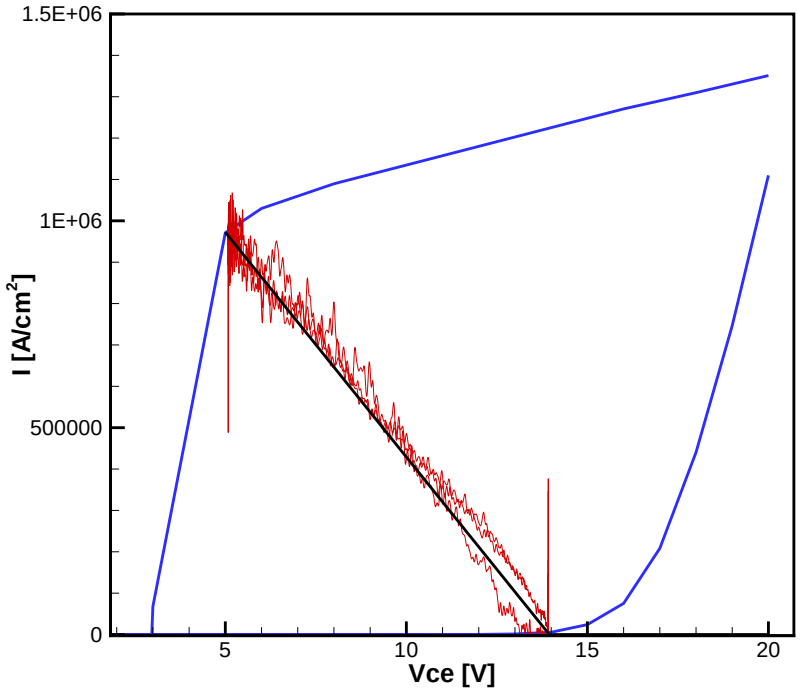

Figure 5.20: Comparison of the theoretical load line (black) and simulated one (red) for the 6 nm device calculated for peak PAE condition. The blue lines represents the $I_c - V_{ce}$ curve that limits the device large signal operation with I_{max}=$1x10^6$ A/cm^2.

attributed to the decrease of current gains for higher collector barriers as shown in Section 4.4.

The same argument can be used to explain the lower efficiency calculated for devices with thicker InGaN layers as shown in Figure 5.21. The power added efficiency is defined as the ratio of the difference between the output AC power and the input AC power over the input DC power as shown in Eq. 5.2:

$$PAE = 100 * \frac{Pout_{AC} - Pin_{AC}}{Pin_{DC}}. \tag{5.2}$$

In particular, the higher current gain of thin InGaN barrier devices allows to obtain the same current for a lower DC input voltage (V_{BE}), thus reducing the denominator in

Eq. 5.2 and increasing the PAE. To summarize, the scaling of the InGaN barrier poses a trade-off between maximum output power and efficiency, and therefore its design is important to improve one or the other to meet specific design criteria depending on the application.

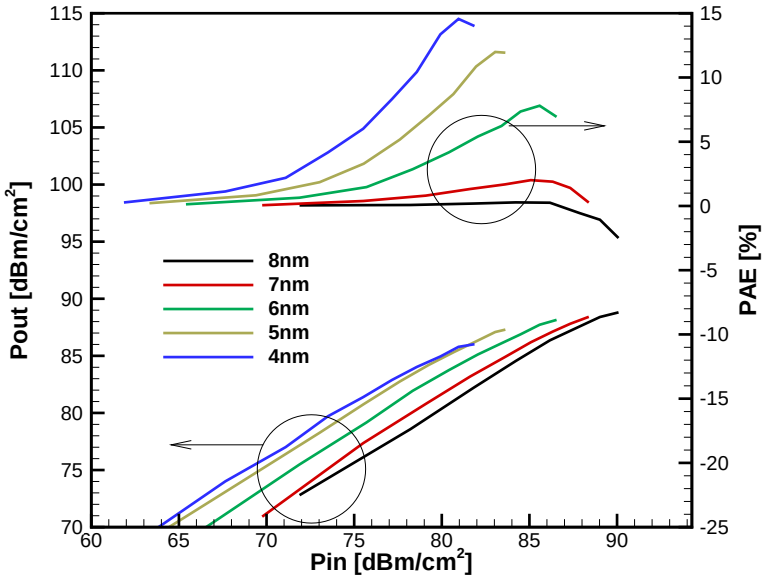

Figure 5.21: Calculation of the output power and efficiency power added efficiency for variations of the device in Figure 5.1 with emitter barrier of 5 nm $Al_{0.4}Ga_{0.6}N$ and thickness of the InGaN layer changed as shown in the legend. The fundamental frequency of the input signal is 20 GHz.

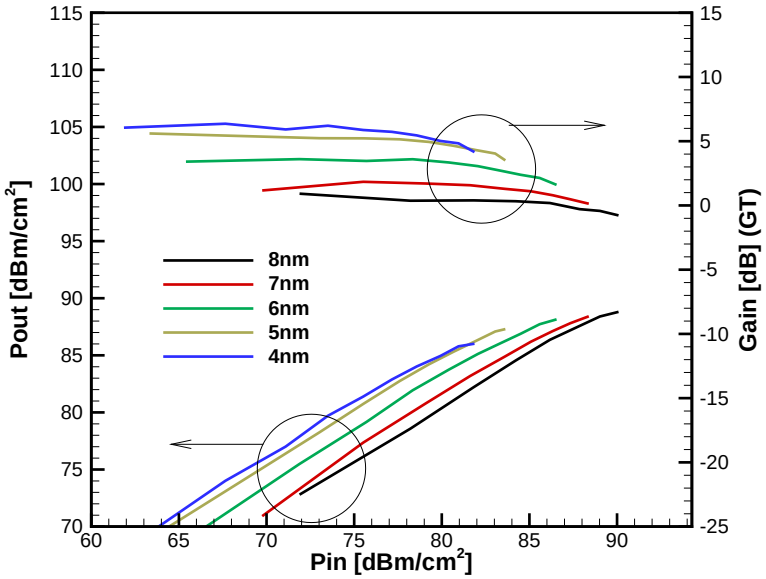

Figure 5.22: Calculation of the output power and transducer gain for variations of the device in Figure 5.1 with emitter barrier of 5 nm $Al_{0.4}Ga_{0.6}N$ and thickness of the InGaN layer changed as shown in the legend. The fundamental frequency of the input signal is 20 GHz.

5.7 Conclusions

The dynamics of the hot electrons transiting through the base of a GaN HET have been characterized in terms of their distribution in momentum space, energy and velocity. From the analysis of these distributions it was found that for the particular layout designed to reproduce the experimental device, a considerable fraction of electrons are injected into the satellite valleys. The low value of current gain of this layout is found to be considerably increased by lowering emitter barrier reducing the AlN layer thickness, which is seen to prevent electrons to be transferred into the satellite valleys. Different combinations of thickness and AlGaN with variable molar fraction have been used to realize the emitter barrier and the related performance have been evaluated in terms of current gain and f_T. Optimal values of barrier thickness have been found for different AlGaN alloy, in order to optimize the injection energy and get a current gain up to 3. f_T up to 270 GHz have been calculated for current densities of $5x10^5$ A/cm^2, and possibly over 500 GHz at $5x10^7$ A/cm^2. The large signal operation of HETs was also characterized, and it was found that a thicker $InGaN$ collector barrier is beneficial to increase to maximum output power at the price of a lower efficiency. The advantage of the thicker InGaN layer is to increase the collector barrier height and thus delay the tunneling of the cold base electrons through the collector barrier that is thinned as the collector to emitter voltage is increased. The awareness of this peculiar breakdown effect for HETs can lead to the development of new layouts for the collector barrier that can reduce the dependence of its thickness to the collector voltage and allows HETs to operate at higher power.

Chapter 6

SUMMARY

In this work the insight provided by our sophisticated Full Band Monte Carlo simulator is used to analyze the behavior of state-of-art devices like GaN High Electron Mobility Transistors and Hot Electron Transistors. Chapter 1 was dedicated to the description of the simulation tool used to obtain the results shown in this work. Within Chapter 1 a separate section was dedicated to the tunneling algorithm that has been recently implemented, and in particular to the set up of a benchmark to validate its implementation. By simulating a square potential barrier it was possible to compare the well known analytical solution with the transmission coefficient calculated from the current which is the final output of the code. This method proved to be useful debugging the code and ultimately matching the simulation results with the analytical ones.

Chapter 2 introduced High Electron Mobility Transistors, state-of-art devices characterized by highly non linear transport phenomena that require a powerful simulation program to be captured. The techniques for device modeling are applied to a recent GaN-HEMT, and validated by comparisons with experimental measurements. The characterization included the AC domain, where the main techniques were discussed along with an innovative method to determine the area of the channel where the charge is controlled by the gate. This procedure showed to reliably predict the device cutoff frequency without running any AC simulation, by integrating the velocity profile of the electrons along the area under control of the gate, also called effective gate length. Moreover, every step of this method is precisely defined, removing any ambiguity from previous definitions of the effective gate length.

Chapter 3 focuses on a common technique used to enhance HEMTs performance: the down-scaling of the device dimensions. In particular, this chapter was dedicated to the scaling of the lateral dimensions of the device, meaning the source to gate, gate to drain distance, and the gate length. It was found that reducing the source to gate and gate to drain distance (also known as access regions) increases both the DC and AC performance of the device of a factor of 3. Moreover, further improvements have been observed when this scaling technique was combined with the downscaling of the gate length. In particular, by extracting the limiting transit velocity of electrons and limiting effective gate length for vanishing access region and gate respectively, a limiting cut-off frequency of about 1.5 THz was calculated.

Finally, Chapter 4 and Chapter 5 described the modeling of Hot Electron Transistors (HETs). The simulation approach was validated by matching the current-voltage characteristics with experimental measurements, and particular attention was given to reproduce the transfer coefficient which is related to the current gain of these devices. Both the layouts that have been reproduced showed a current gain below unity, which is not sufficient to provide any signal amplification. Therefore, different techniques have been implemented the increase the fraction of the electron injected from the emitter that reaches the collector. Critical elements in the design have been found to be the thickness of the layers implementing the emitter and collector barrier as well as the thickness of the base. However, the current gain was found to be below unity even in devices with a very short base when electrons are injected in the satellite valleys. By calibrating the injection energy and reducing the base length, the current gain was increased up to 3 and values of f_T up to 250 GHz have been calculated. Large signal performance of HETs have also been calculated in Chapter 5. In this regard, the design of the collector barrier has shown to be of crucial importance to control the breakdown of the device allowing a larger power to be supplied.

A common element of all the devices that have been simulated is the extremely out of equilibrium transport, whether consequence of the reduced dimensions or intrinsic in the operation of the device. Therefore, one of the conclusions that can be drawn from this work is the importance of full band Monte Carlo simulators in the analysis and design of the new generation of high performance devices.

Future work will expand the analyses reported in this dissertation based on new experimental data. In particular, the design guidelines developed for HETs have been used to fabricate GaN HET with the aim to experimentally characterize its frequency response for the first time. Moreover, new geometries and materials could be simulated in order to increase the promising performance that have been reported, and fully exploit the potential of Hot Electron Transistors.

REFERENCES

Akis, R., J. Ayubi-Moak, N. Farelli, D. K. Ferry, S. M. Goodnick and M. Saraniti, "The upper limit of the cutoff frequency in ultrashort gate-length InGaAs/InAlAs HEMTs: A new definition of effective gate length", IEEE Electron Device Letters **29**, 306 – 308 (2008).

Ambacher, O., B. Foutz, J. Smart, J. Shealy, N. Weimann, K. Chu, M. Murphy, A. Sierakowski, W. Schaff, L. Eastman, R. Dimitrov, A. Mitchell and M. Stutzmann, "Two dimensional electron gases induced by spontaneous and piezoelectric polarization in undoped and doped AlGaN/GaN heterostructures", Journal of Applied Physics **87**, 334–344 (2000).

Ambacher, O., J. Majewski, C. Miskys, A. Link, M. Hermann, M. Eickhoff, M. Stutzmann, F. Bernardini, V. Fiorentini, V. Tilak, B. Schaff and L. Eastman, "Pyroelectric properties of Al(In)GaN/GaN hetero and quantum well structures", Journal of physics: condensed matter **14**, 3399 – 3434 (2002).

Ando, Y. and T. Itoh, "Calculation of transmission tunneling current across arbitrary potential barriers", Journal of applied physics **61**, 4, 1497–1502 (1987).

Ardaravicius, L., A. Matulionis, J. Liberis, O. Kiprijanovic, M. Ramonas, L. F. Eastman, J. R. Shealy and A. Vertiatchikh, "Electron drift velocity in AlGaN/GaN channel at high electric fields", Applied Physics Letters **83**, 19, 4038–4040 (2003).

Bengtsson, O., L. Vestling and J. Olsson, "A computational load-pull method with harmonic loading for high-efficiency investigations", Solid-State Electronics **53**, 1, 86 – 94 (2009).

Bernat, J., P. Javorka, A. Fox, M. Marso, H. Lüth and P. Kordoš, "Effect of surface passivation on performance of algan/gan/si hemts", Solid-State Electronics **47**, 11, 2097–2103 (2003).

Berroth, M. and R. Bosch, "Broad-band determination of the fet small-signal equivalent circuit", Microwave Theory and Techniques, IEEE Transactions on **38**, 7, 891–895 (1990).

Bollaert, S., Y. Cordier, M. Zaknoune, T. Parenty, H. Happy and A. Cappy, "Hemts capability for millimeter wave applications", in "Annales des télécommunications", vol. 56, pp. 15–26 (Springer, 2001).

Chang, C., P. Asbeck, P. Zampardi and K. Wang, "Direct measurement of c_{be} and c_{bc} versus voltage for small HBTs with microwave s-parameters for scaled Gummel-Poon BJT models", Microwave Theory and Techniques, IEEE Transactions on **47**, 1, 108 (1999).

Chang, E.-Y., C.-I. Kuo, H.-T. Hsu, C.-Y. Chiang and Y. Miyamoto, "InAs thin-channel high-electron-mobility transistors with very high current-gain cutoff frequency for emerging submillimeter-wave applications", Applied Physics Express **6**, 3, 034001 (2013).

Chattopadhyay, D. and H. Queisser, "Electron scattering by ionized impurities in semiconductors", Reviews of Modern Physics **53**, 4, 745 (1981).

Choudhry, V. and V. Arora, "Field dependence of drift velocity and electron population in satellite valleys in gallium arsenide", Electronics Letters **22**, 5, 271–272 (1986).

Chung, J., E. Piner and T. Palacios, "N-face GaN/AlGaN HEMTs fabricated through layer transfer technology", IEEE Electron Device Letters **30**, 2, 113 – 116 (2009).

Deger, C., E. Born, H. Angerer, O. Ambacher, M. Stutzmann, J. Hornsteiner, E. Riha and G. Fischerauer, "Sound velocity of al x ga 1-x n thin films obtained by surface acoustic-wave measurements", Applied physics letters **72**, 19, 2400–2402 (1998).

Denton, A. R. and N. W. Ashcroft, "Vegards law", Physical Review A **43**, 6, 3161 (1991).

Diff, K. and K. F. Brennan, "Theory of electron-plasmon-scattering rate in highly doped bulk semiconductors", Journal of applied physics **69**, 5, 3097–3103 (1991).

Dingle, R., H. Störmer, A. Gossard and W. Wiegmann, "Electron mobilities in modulation-doped semiconductor heterojunction superlattices", Applied Physics Letters **33**, 7, 665–667 (1978).

Ferry, D., "The onset of quantization in ultra-submicron semiconductor devices", Superlattices and Microstructures **27**, 61–66 (2000).

Ferry, D. K., *Semiconductors* (Macmillan Publishing Company, New York, 1991).

Fischetti, M. and S. Laux, "Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects", Physical Review B **38**, 14, 9721–9745 (1988).

Forghieri, A., R. Guerrieri, P. Ciampolini, A. Gnudi, M. Rudan and G. Baccarani, "A new discretization strategy of the semiconductor equations comprising momentum and energy balance", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on **7**, 2, 231–242 (1988).

Foutz, B. E., S. K. O'Leary, M. S. Shur and L. F. Eastman, "Transient electron transport in wurtzite GaN, InN, and AlN", Journal of Applied Physics **85**, 11, 7727–7734 (1999).

Griffiths, D. J. and E. G. Harris, *Introduction to quantum mechanics*, vol. 2 (Prentice Hall New Jersey, 1995).

Guerra, D., R. Akis, F. A. Marino, D. K. Ferry, S. M. Goodnick and M. Saraniti, "Aspect ratio impact on RF and DC performance of state-of-the-art short-channel GaN and InGaAs HEMTs", Electron Device Letters, IEEE **31**, 11, 1217–1219 (2010).

Guerra, D., F. Marino, R. Akis, D. Ferry, S. Goodnick and M. Saraniti, "Terahertz-capability nanoscale InGaAs HEMT design guidelines by means of full-band monte carlo device simulations", in "IEEE 11th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF), 2011", pp. 193 –196 (2011a).

Guerra, D., F. Marino, D. Ferry, S. Goodnick, M. Saraniti and R. Soligo, "Large-signal full-band monte carlo device simulation of millimeter-wave power GaN HEMTs with the inclusion of parasitic and reliability issues", pp. 34.2.1 –34.2.4 (2011b).

Gupta, G., E. Ahmadi, K. Hestroffer, E. Acuna and U. Mishra, "Common emitter current gain 1 in III-N hot electron transistors with 7nm GaN/InGaN base", (2015a).

Gupta, G., M. Laurent, H. Li, D. Suntrup, E. Acuna, S. Keller and U. Mishra, "Design space of III-N hot electron transistors using algan and ingan polarization-dipole barriers", (2015b).

Gupta, G., M. Laurent, H. Li, D. J. Suntrup, E. Acuna, S. Keller and U. Mishra, "Common emitter operation of III-N HETs using AlGaN and InGaN polarization-dipole induced barriers", in "Device Research Conference (DRC), 2014 72nd Annual", pp. 255–256 (2014a).

Gupta, G., M. Laurent, J. Lu, S. Keller and U. K. Mishra, "Design of polarization-dipole-induced isotype heterojunction diodes for use in III-N hot electron transistors", APE **7**, 1, 014102 (2014b).

Hamaguchi, C., *Basic semiconductor physics*, vol. 212 (Springer, 2001).

Heiblum, M. and M. Fischetti, "Ballistic hot-electron transistors", IBM J. of Research and Development **34**, 4, 530–549 (1990).

Jacoboni, C. and L. Reggiani, "The Monte Carlo method for solution of charge transport in semiconductors with applications to covalent materials", Reviews of Modern Physics **55**, 3, 645–705 (1983).

Kim, D.-H. and J. del Alamo, "Lateral and vertical scaling of $in_{0.7}$ $ga_{0.3}$ As hemts for post-Si-CMOS logic applications", Electron Devices, IEEE Transactions on **55**, 10, 2546–2553 (2008a).

Kim, D.-H. and J. A. del Alamo, "Lateral and vertical scaling of $In_{0.7}Ga_{0.3}As$ HEMTs for post-Si-CMOS logic applications", IEEE Transactions on Electron Devices **55**, 10, 2546 – 2553 (2008b).

Kittel, C., *Introduction to Solid State Physics* (New York: Wiley, 1976, 5th ed., 1976).

Knezevic, I., D. Vasileska and D. Ferry, "Impact of strong quantum confinement on the performance of a highly asymmetric device structure: Monte Carlo particle-based simulation of a focused-ion-beam mosfet", IEEE Transaction Electron Devices **49**, 1019–1026 (2002).

Laux, S., “Techniques for small-signal analysis of semiconductor devices”, IEEE Transactions on Electron Devices **ED-32**, 10, 2028–2037 (1985).

Lee, D. S., J. Chung, H. Wang, X. Gao, S. Guo, P. Fay and T. Palacios, “245-GHz InAlN/GaN HEMTs with oxygen plasma treatment”, IEEE Electron Device Letters **32**, 6, 755–757 (2011).

Levi, A. and T. Chiu, “Room-temperature operation of hot-electron transistors”, Applied physics letters **51**, 13, 984–986 (1987).

Mansour, N. S., S. H. Janzou and K. F. Brennan, “Theory of the electron-plasmon interaction in monte carlo calculations through the direct solution of the poisson equation”, Journal of applied physics **72**, 11, 5277–5282 (1992).

Marino, F. A., *ADVANCED SIMULATION METHODS FOR PHYSICAL INVESTIGATION OF HIGH ELECTRON MOBILITY TRANSISTORS*, Phd thesis, Department of Electrical and Energy engineering, Arizona State Univerisity (2010).

Martín, M., T. González, D. Pardo and J. Velázquez, “Monte carlo analysis of a schottky diode with an automatic space-variable charge algorithm”, Semiconductor science and technology **11**, 3, 380 (1996).

Matulionis, A., J. Liberis, I. Matulioniene, H.-Y. Cha, L. F. Eastman and M. G. Spencer, “Hot-phonon temperature and lifetime in biased 4H-SiC”, Journal of Applied Physics **96**, 11, 6439–6444 (2004).

Mead, C., “Operation of tunnel-emission devices”, Journal of Applied Physics **32**, 4, 646–652 (1961).

Mishra, U., S. Likun, T. Kazior and Y.-F. Wu, “GaN-Based RF power devices and amplifiers”, Proceedings of the IEEE **96**, 2, 287–305 (2008).

Mishra, U. and J. Singh, *Semiconductor device physics and design* (Springer Science & Business Media, 2007).

Mishra, U. K., P. Parikh, Y.-F. Wu *et al.*, “Algan/gan hemts-an overview of device operation and applications”, PROCEEDINGS-IEEE **90**, 6, 1022–1031 (2002).

Muto, S., K. Imamura, N. Yokoyama, S. Hiyamizu and H. Nishi, “Subpicosecond base transit time observed in a hot-electron transistor (HET)”, Electronics Letters **21**, 13, 555–556 (1985).

Nidhi, S. Dasgupta, Y. Pei, B. Swenson, D. Brown, S. Keller, J. Speck and U. Mishra, “f_t and f_{max} of 47 and 81 GHz, respectively, on n-polar GaN/AlN MIS-HEMT”, IEEE Electron Device Letters **30**, 6, 599 – 601 (2009).

Palacios, T., S. Rajan, A. Chakraborty, S. Heikman, S. Keller, S. P. DenBaars and U. K. Mishra, “Influence of the dynamic access resistance in the g_m and f_t linearity of AlGaN/GaN HEMTs”, Electron Devices, IEEE Transactions on **52**, 10, 2117–2123 (2005).

Pearton, S., J. Zolper, R. Shul and F. Ren, "Gan: Processing, defects, and devices", Journal of applied physics **86**, 1, 1–78 (1999).

Pop, E., R. Dutton and K. Goodson, "Detailed heat generation simulations via the monte carlo method", International Conference on Simulation of Semiconductor Processes and Device (SISPAD) (2003).

Pozar, D. M., *Microwave Engineering* (Wiley, 2005), 3rd edn.

Rowlette, J., E. Pop, S. Sinha, M. Panzer and K. Goodson, "Thermal simulation techniques for nanoscale transistors", in "ICCAD '05: Proceedings of the 2005 IEEE/ACM International conference on Computer-aided design", pp. 225–228 (IEEE Computer Society, Washington, DC, USA, 2005).

Russo, S. and A. Di Carlo, "Influence of the source–gate distance on the AlGaN/GaN HEMT performance", Electron Devices, IEEE Transactions on **54**, 5, 1071–1075 (2007).

Saraniti, M. and S. Goodnick, "Hybrid full-band Cellular Automaton/Monte Carlo approach for fast simulation of charge transport in semiconductors", IEEE Transactions on Electron Devices **47**, 10, 1909–1915 (2000).

Saraniti, M., A. Rein, G. Zandler, P. Vogl and P. Lugli, "An efficient multigrid poisson solver for device simulations", IEEE Transaction on Computer–aided Design of Integrated Circuits and Systems **15**, 2, 141–150 (1996).

Selberherr, S., *Analysis and simulation of semiconductor devices* (Springer Wien; New York, 1984).

Shannon, J. and A. Gill, "High current gain in monolithic hot-electron transistors", Electronics Letters **17**, 17, 620–621 (1981).

Shinohara, K., A. Corrion, D. Regan, I. Milosavljevic, D. Brown, S. Burnham, P. Willadsen, C. Butler, A. Schmitz, D. Wheeler, A. Fung and M. Micovic, "220 GHz f_T and 400 GHz f_{max} in 40-nm GaN DH-HEMTs with re-grown ohmic", in "Electron Devices Meeting (IEDM), 2010 IEEE International", pp. 30.1.1 –30.1.4 (2010).

Somerville, M. H., R. Blanchard, J. A. Del Alamo, K. G. Duh and P. Chao, "On-state breakdown in power HEMTs: Measurements and modeling", Electron Devices, IEEE Transactions on **46**, 6, 1087–1093 (1999).

Sze, S. and H. Gummel, "Appraisal of semiconductor-metal-semiconductor transistor", Solid-State Electronics **9**, 8, 751–769 (1966).

Sze, S. M. and K. K. Ng, *Physics of Semiconductor Devices* (John WILEY and Sons, 2007), third edn.

Tilak, V., B. Green, V. Kaper, H. Kim, T. Prunty, J. Smart, J. Shealy and L. Eastman, "Influence of barrier thickness on the high-power performance of AlGaN/GaN HEMTs", Electron Device Letters, IEEE **22**, 11, 504–506 (2001).

van Nguyen, H., J. Moreno, R. Teissier and M. Zaknoune, "Submicrometer process and RF operation of InAs quantum hot-electron transistors", Electron Device Letters, IEEE **33**, 6, 797–799 (2012).

Wong, M., Y. Pei, R.Chu, S. Rajan, B. Swenson, D. Brown, S. Keller, S. DenBaars, J. Speck and U. Mishra, "N-face metal-insulator-semiconductor high-electron-mobility transistors withAlN back-barrier", IEEE Electron Device Letters **29**, 10, 1101 – 1104 (2008).

Wood, C. and D. Jena, *Polarization Effects in Semiconductors: From Ab Initio Theory to Device Applications* (Springer Science & Business Media, 2007).

Wu, Y.-R., M. Singh and J. Singh, "Device scaling physics and channel velocities in AIGaN/GaN hfets: velocities and effective gate length", Electron Devices, IEEE Transactions on **53**, 4, 588–593 (2006).

Yamashita, Y., I. Watanabe, A. Endoh, N. Hirose, T. Mimura and T. Matsui, "Effect of source-drain spacing on dc and rf characteristics of 45 nm-gate AlGaN/GaN MIS-HEMTs", Electronics Letters **47**, 3, 211 –212 (2011).

Yokoyama, N., K. Imamura, S. Muto, S. Hiyamizu and H. Nishi, "A new functional, resonant-tunneling hot electron transistor (RHET)", Japanese journal of applied physics **24**, 11A, L853 (1985).

Yue, Y., Z. Hu, J. Guo, B. Sensale-Rodriguez, G. Li, R. Wang, F. Faria, T. Fang, B. Song, X. Gao *et al.*, "InAlN/AlN/GaN HEMTs with regrown ohmic contacts and of 370 GHz", Electron Device Letters, IEEE **33**, 7, 988–990 (2012).