The role of the choice of the physical model in the optimization of nanoelectronic device simulators

Paolo Marconcini

Abstract—In order to obtain fast and reliable simulators for nanoelectronic devices, both an algorithmic optimization and an accurate choice of the physical model adopted to describe the device properties and behavior is necessary. In this article some different alternatives which exist at the modeling level are examined, with a few examples derived from the author’s research activity, and their effect on the simulator performance and on the accuracy of the simulation results are discussed. I show how, while in many cases more detailed results can be obtained only at the expense of a larger computational cost and therefore a careful choice of the desired level of accuracy has to be made when writing a simulation code, in other situations the use of physically equivalent descriptions can determine a large variation in the complexity and efficiency of the simulation.

Keywords—computer-aided simulation, nanoelectronic devices, physical model

I. INTRODUCTION

Nanoelectronics is the branch of electronics dealing with nanometer scale devices. It includes both traditional silicon CMOS electronics scaled (thanks to the use of technological innovations, like the use of high-κ dielectrics) to a size of a few nanometers, and more innovative devices. The latter have been conceived to solve some of the problems resulting from the scaling of CMOS technology (as excessive power dissipation and insufficient electrostatic control of the channel) [1], and differ from the standard CMOS technology e.g. for the adopted materials, the physical entity used to transfer and store information, the operation principles, or the geometry of the devices. In particular, the first category includes for example devices based on semiconductor heterostructures [2-6], carbon nanotubes [7-9], and graphene [10-15], which are characterized by high charge mobilities.

At this level of miniaturization, the physical implementation of electronic devices requires quite complex, expensive and time-consuming fabrication techniques and thus the prototyping activity has to be limited as much as possible. This increases the importance of the development of computer-aided design and simulation tools for these devices, both in order to fully understand their behavior at the physical level, and to optimize their electronic performance by means of an accurate design of their structure and processing steps, before their actual fabrication, in such a way to reduce the required costs and times.

Since down at this scale quantum mechanical effects and atomistic details are generally fundamental to correctly describe the physical and electrical behavior, and thus a simple drift-diffusion model for the analysis of the transport is generally not acceptable, the study of nanoelectronic devices requires quite a large computational effort, with respect to traditional components. On the other hand, in order to reduce the simulation times and thus to allow the analysis of a large design parameter space and the application of the simulation code also to circuits containing many devices, an optimization is needed, both from an algorithmic and from a modeling point of view.

From a computational point of view, a parallelization of the simulation codes (for example, using MPI libraries) is generally necessary and a proper and optimized implementation of the simulation codes, which can benefit from it, has to be performed, identifying and applying the parallelization to the parts of the code that can be executed simultaneously.

Moreover, often complex quantities that are computed in a part of the code are then useful in different parts of the program; therefore there is the need to store this information, in order to reduce the computational burden. On the other hand, there is the necessity to limit memory occupation, in order to avoid possible paging problems, and to limit the number of read/write operations, which strongly increase the execution times. A careful trade-off between these two requirements has therefore to be achieved.

Additionally, since nanoelectronic device simulations require an intensive use of advanced numerical solution techniques, the exploitation of well-tested and optimized Fortran and C routines (such as Lapack [16] and Numerical Recipes [17] codes) is largely adopted. A proper choice of the adopted numerical routines has to be made, exploiting the symmetries and mathematical properties of the considered physical problem.

This paper, instead, is focused on the importance, for the optimization of the simulation codes, of the choice of the physical model adopted for the description of the device. Some
of the main alternatives that exist for the modeling of nanoelectronic devices will be examined. In particular, referring to the present research activity of the author, it will be stressed how, while a trade-off between the simplicity of the adopted representation (and therefore the performance of the code) and the accuracy of the obtained results is often necessary, in many cases physically equivalent treatments can lead to a different efficiency of the simulation, depending on the specific application. A proper choice of the physical model is therefore a fundamental aspect to perform optimized computer-aided simulations.

II. FULL-QUANTUM/SEMICLASSICAL MODELS

A first choice to be made is between a full-quantum model and a semiclassical description. The first approach includes a complete quantum mechanical description of transport, based on the solution of the Schrödinger equation for the electron wave function (the square modulus of which gives the probability to find the electron at a given time in a point of the space). In a semiclassical description, instead, the Boltzmann equation is solved for the electron distribution function (which represents the probability to find an electron at a given time in a point of the phase space of position and crystal momentum) [18]. Since the second approach, although including a quantum mechanical treatment of dynamics and scattering processes, substantially describes charge carriers as classical particles characterized by distinct position and momentum and obeying Newton's law, it misses phenomena which happen on the scale of the electron wavelength, such as interference, resonances and localization.

Therefore, a semiclassical analysis, although allowing faster device simulations, gives good results only for relatively large devices, while in the case of nanometric circuits a full-quantum model is generally preferred.

For example, we can consider the case of devices based on GaAs/AlGaAs heterostructures, where an high-mobility two-dimensional electron gas (2DEG) exists at the interface between GaAs and AlGaAs. These two-dimensional devices are obtained confining the 2DEG by etching and by the electrostatic action of biased gates located on the surface of the device, at a certain distance from the 2DEG. In particular, we have considered a series of unevenly spaced tunnel barriers (obtained negatively biasing a series of gates defined on the surface) and we have computed the shot noise (the electrical noise deriving from the granularity of charge) in this device [6, 19]. For this structure, a semiclassical analysis [20] has predicted a Fano factor $F$ (ratio between the actual shot noise and the value that would be expected in the case in which charge carriers crossed the device in an independent way) approaching $1/3$ as the number of barriers increases, for every value of the barrier transparency $\Gamma$ (see Fig. 1). We have quite recently simulated the same device using a full quantum-mechanical description. Our quantum-mechanical analysis has led to a different conclusion (in agreement with existing experimental data): the Fano factor $F$ has the behavior (as a function of the number of barriers) reported in Fig. 2 for three different values of the transparency $\Gamma$, and for a high number of cascaded tunnel barriers approaches 1, as a consequence of strong localization. The reason for this discrepancy is that strong localization, deriving from interference effects, cannot be obtained with a simple semiclassical treatment, that therefore in this case would not lead to a correct result.

![Fig. 1](image1.png)

Fig. 1 Value of the Fano factor $F$ predicted by a semiclassical model for a series of barriers, as a function of the number of barriers and for two values of the barrier transparency $\Gamma$ (adapted from [20]).

![Fig. 1](image2.png)

Fig. 1 Value (averaged over 50 sets of interbarrier distances) of the Fano factor $F$ for a series of unevenly spaced tunnel barriers, as a function of the number of barriers, resulting from a quantum-mechanical simulation. The three curves correspond to three different values of the barrier transparency $\Gamma$.

III. MANY-BODY/SINGLE-ELECTRON MODELS

Another important choice that one has to make when writing a simulation code for nanoelectronic devices is that between a many-body and a single-particle model. In the first one, a complete quantum description of the system, made up of all the charge carriers in the device, is considered. In this case the system is described by a global wave function which is a function of the position of all the electrons and that can be expanded in an infinite number of Slater determinants. An exact many-body calculation is however very difficult (or impossible, due to computational limits) to perform (with the exception of systems with very few particles) and generally appropriate approximations are made.
A many-body description is essential to understand the behavior of structures where the interactions between particles play a fundamental role, as in devices based on the Coulomb-blockade phenomenon [22]. Indeed, this phenomenon, which prevents at low temperatures the transfer of electrons through a tiny conducting island (isolated from the leads by tunnel barriers) until a given voltage threshold is overcome, cannot be described with a single-particle model because is inherently a many-body phenomenon, which is related to the Coulomb interaction among the electron charges (which determines the presence of a charging energy for the island) and to charge quantization. Therefore, devices like the single-electron transistor (schematized in Fig. 3), where the tiny island is coupled to a source and drain lead through tunnel barriers and to a gate through a capacitor (in which tunneling is not allowed) and where electrons can be transferred from source to drain one by one under the control of the gate voltage (which shifts the electrostatic potential of the island), cannot be studied using a simple single-particle schematization.

In the simulation of most nanoelectronic devices, however, a single-electron description of transport is sufficient to correctly describe the system. Therefore this approximate description, which considers only a single electron, in the average electrostatic potential due to the other carriers, fixed charges and fields, without describing exactly the interaction with the individual electrons and holes, is often used to obtain faster nanoelectronic device simulators.

IV. SPATIAL DETAIL OF THE MODEL

Another important element that can have a strong effect on the computational burden of the simulation code is the degree of spatial detail of the physical model. Inside a full-quantum description, a possibility is to solve the Schrödinger equation for the electron wave function, considering all the atomic details of the devices, including the potential of each single atom. Another one is to use a continuum, envelope function approach [23, 24], which neglects the atomic details of the structure and replaces the electron wave function with the so-called envelope function, which is a slowly-varying function which multiplied by a function periodic with the lattice period gives the actual electron wave function. In this approach, the equation that has to be solved is the effective mass Schrödinger equation in the case of common semiconductors, while is the Dirac equation in the case of graphene.

If a device containing a very large number of atoms has to be simulated, a fully-atomistic simulation is generally unpractical from the computational point of view and the envelope function approach is preferable. On the contrary, if atomistic details are expected to play a fundamental role, especially in the case of small devices, an atomistic description has to be used.

For example, let us consider the case of graphene, which consists of a layer of $sp^2$ hybridized carbon atoms arranged in a honeycomb lattice. Field-effect transistors with a graphene channel have been theoretically proposed and experimentally realized, in order to exploit the high mobility and one-atom thickness of this material. In order to open an energy gap in graphene and to introduce an electron-hole asymmetry in its transport behavior (making it possible to use graphene-based transistors for logical applications) it has been proposed to use as a transistor channel a doped graphene nanoribbon. We have recently performed a numerical analysis of the effect of a low concentration of substitutional boron doping on the transmission of narrow ribbons of graphene [13,25]. In this case, due to the reduced size of the structure, the atomistic details in correspondence of the boron atoms are fundamental to study the transport behavior of the device. In particular, we have found that the presence of quasi-bound states localized around boron impurities introduces a strong hole backscattering and thus generates a clear electron-hole asymmetry in the characteristics of a graphene-based transistor. However, we have noticed (see Fig. 4) that while an atomistic (DFT or tight-binding) description captures this effect, a continuum model based on the solution of the Dirac equation misses it and reproduces only the minor (in the case of narrow ribbons) electrostatic effect of negatively charged dopants. A continuum model misses also, in quasi-metallic graphene ribbons, the energy gap, which in that case derives only from the edge distortion, which is included in an atomistic description but not in the less-detailed envelope function description.

In order to simulate more complex structures, capturing the most important atomistic details without too large a computational effort, a possibility is to use a hierarchical simulation approach, in which from an atomistic, low-level analysis of a sub-region or of a simplified structure a small number of parameters are extracted, which are then used in a less-detailed, higher-level description of the overall device, which can be performed faster. We have recently adopted this methodology to study silicon nanowire transistors.
V. SELF-CONSISTENT SOLUTION AND SEMI-ANALYTICAL APPROXIMATIONS

In the solution of the transport (Schrödinger or Dirac) equation we assume to know the value of the external potential. However, the value of the potential derives from the solution of the electrostatic (Poisson) equation, where also the charge density in the device (deriving from the solution of the transport equation) appears. Therefore a correct simulation of a nanoelectronic device requires the self-consistent solution of the transport and electrostatic equations, through an iterative procedure that consists in repeatedly solving the two equations till a solution that approximately satisfies both of them is found [8, 9, 26]. While necessary to obtain an accurate result, this procedure is very resource-demanding from the computational point of view.

In many cases an approximate solution can be sufficient to express an opinion on the correct functionality of a device, especially in the cases in which a preliminary study on the dependence of the device characteristics on a large parameter space is needed. In these situations is can be useful to adopt, instead of a self-consistent solution of the transport and electrostatic equations, approximate procedures to compute the screened potential that has to be used in the transport equation.

For example, these approximate techniques have been used to study devices based on GaAs/AlGaAs heterostructures, obtained confining the 2DEG located at the interface between GaAs and AlGaAs through the electrostatic action of negatively biased gates defined on the surface. In particular, we have observed [5, 27] that the conductance of the series of two quantum dots, connected to the external leads by constrictions and reciprocally separated by a tunnel barrier (Fig. 5), strongly depends on the presence or absence of symmetry in the device. The conductance has a maximum value if the device is symmetric (i.e. if the tunnel barrier has exactly the same distance from the input and output constrictions, condition in which, due to constructive interference, the overall conductance is even higher than that of the single tunnel barrier, i.e. the presence of the two constrictions determines a tunneling enhancement effect), while strongly decreases if symmetry is destroyed. In order to simulate the behavior of the device, we have repeated the conductance calculation for several positions of the barrier inside the cavity delimited by the two constrictions, including also the effect on the 2DEG potential of the unavoidable edge roughness existing in real gates. Since the number of configurations to be simulated was very large and only a feasibility analysis was necessary, in that case we have decided to obtain the potential profile at the 2DEG level using (instead of an exact self-consistent solution of the electrostatic and transport equations) an approximate technique based on the calculation of the linear response of the 2DEG to the potential of the gates [28]. The result, for the case of a tunnel barrier deriving from a gate with edge roughness, is reported in Fig. 6.

In order to test the effect of the presence of ionized donors and charged impurities on the described conductance behavior without explicitly solving the complete Schrödinger/Poisson problem we also adopted the approximate formula proposed by Stern and Howard [29] for the screened potential produced in the 2DEG by a point charge located at a given distance from it. The adoption of these approximate techniques has allowed us to make a wide analysis and to conclude that the presence of a realistic level of edge roughness and of a low level of potential disorder in the device should not destroy the expected tunneling enhancement effect.
VI. SOLUTION DOMAIN

So far we have considered a series of choices in which an higher computational performance is obtained only adopting a more approximate description for the system and thus a careful trade-off between the speed of the device simulations and the accuracy or detail of the obtained results is required.

However, modeling a nanoelectronic device, very often it is also possible to choose between physically equivalent alternatives, which can have a different impact on the performance of the simulation code.

From this point of view, an important choice that has to be made in the physical model is the domain in which the calculation has to be performed. For example, it is possible to operate in the direct domain (the spatial domain) or in the Fourier-transformed one, i.e. the reciprocal domain (the domain of the wave vectors).

An enlightening case is the transport study of graphene nanoribbons in the presence of a generic external potential, that we are performing using an envelope-function approach. In particular, the simulation is performed subdividing the overall structure in transversal sections in which the potential is approximately longitudinally constant. In each of these sections we have to solve a one-dimensional Dirac equation in order to find the longitudinal wave vector \( \kappa_x \) and the transverse components of the envelope functions.

In this case, the solution in the direct domain is not trivial. The reason is that the Dirac equation, if solved with a standard finite-difference approach, leads (depending on the particular way in which we discretize the differential equation) either to the presence of spurious solutions (as we show, for the case of a null external potential, in Fig. 7, where the numerically found longitudinal wave vectors are compared with the exact ones), or to the problem of fermion-doubling (which corresponds to the appearance of an unrealistic degeneracy in the longitudinal wave vectors and discontinuities in the corresponding envelope functions).

An alternative approach in the direct space, which does not suffer these numerical problems, exists [30-32] and is based on the use (in the finite difference discretization of the Dirac equation) of two shifted grids for the points where the equations are evaluated and for the nodes where the values of the unknowns are obtained. However, the numerical efficiency of the solution in the direct space remains quite low.

On the contrary, if the problem is numerically transformed into an equivalent one with periodic boundary conditions [33] and an analysis in the Fourier-transformed domain is performed, no numerical problem appears and, due to the exact treatment of derivatives, to the use of optimized FFT routines (for the transformation of the potential from the direct to the reciprocal domain), and to the limited number of Fourier components that are generally needed in the case of slow-varying potentials, the speed of the code largely increases with respect to a direct solution in the space domain.

VII. BASIS SETS

The transition from the real space to the reciprocal space can be seen as a change [34] from the basis of discrete delta functions centered on the nodes of the discretization grid in the real space, to the basis of plane waves corresponding to the different discrete spatial frequencies. Therefore this alternative can be seen as a particular case of the more general problem of the choice of the basis set adopted in the transport calculation.

In general, the more similar the basis set on which we express the different quantities involved in the calculations is to the final solution, the more efficient the calculation will be. For example, in the transport study of heterostructure-based confined devices, where in each section the final electron wave function will be a proper combination of the transversal eigenmodes (instead of over a discretization grid in space), in order to simplify the numerical treatment of the problem.

The observation that equivalent physical representations can determine a different computational complexity in nanodevice simulation is, however, much more general. In the following, a couple of significant examples deriving from the research experience of the author will be presented.

VIII. EXAMPLES OF OTHER POSSIBLE CHOICES

A. Lattice unit vectors

When a solid-state-physics description of nano electronic devices is performed and we have to deal with a lattice of atoms, sometimes it can be necessary to individuate the point equivalent to a known one inside a particular region of the direct or reciprocal space. In this case, the choice of a particular set of unit vectors (among the infinite ones that can equivalently describe the lattice) can strongly reduce the times...
for this search.

For example, nanoelectronic devices which use carbon nanotubes as conducting channel have been proposed [7]. A carbon nanotube is a graphene sheet rolled, along one of its lattice vectors (called “chiral vector”) into a cylindrical shape. Therefore the dispersion relations of a carbon nanotube can be obtained from those of graphene enforcing a periodic boundary condition along the chiral vector, which corresponds to consider only discretized values of the wave vector along the direction of the chiral vector, i.e. to cross-section the two-dimensional energy bands of graphene along parallel segments in the reciprocal space. The unit cell of the graphene reciprocal space where it is more convenient, from the numerical and analytical point of view, to take these sections is a rectangular region an edge of which coincides with the reciprocal unit vector of the nanotube [35] (in Fig.8 we have represented this region in the case of the nanotube (5,2)). Since the nanotube energy bands of main interest for electronic applications are those obtained cross-sectioning the graphene dispersion relations near their maxima and minima, it can be important to know the positions of those points inside the considered rectangular region of the reciprocal space. We have shown that the easiest way to find these positions is to adopt, instead of the standard unit vectors of the graphene reciprocal space, another particular pair of unit vectors (one of which coincides which an edge of the rectangular region). In this way the positions of the desired points are immediately known and the time necessary for the search is drastically reduced [36].

Fig. 8 Representation, in the graphene reciprocal space, of the rectangular zone (shaded in the figure) where it is more convenient to cross-section the graphene dispersion relations to obtain the bands of a (5,2) carbon nanotube. \( \mathbf{K} \) represents the unit vector of the carbon nanotube reciprocal space. The vectors \( \mathbf{b}_1 \) and \( \mathbf{b}_2 \) represent the standard unit vectors of the graphene reciprocal space, while \( \mathbf{b}_x \) and \( \mathbf{b}_y \) are the unit vectors that make easier to find the extrema points \( K \) and \( K' \) of the graphene dispersion relations inside the rectangular zone.

B. Potential vector

Another example is the choice of the particular form of the potential vector in the simulation of nanoelectronic devices in the presence of a magnetic field [4-6, 19].

For example, let us consider the case of heterostructure-based electronic devices threaded by a uniform orthogonal magnetic field \( \mathbf{B} \). If we define \( x, y \) and \( z \) as the longitudinal, transverse and orthogonal directions, respectively, we have that \( \mathbf{B} = [0 \ 0 \ B]' \). The quantity which appears inside the Schrödinger equation is the potential vector \( \mathbf{A} \), which is related to \( \mathbf{B} \) by the correspondence \( \mathbf{B} = \mathbf{\nabla} \times \mathbf{A} \). Infinite choices for the potential vector exist for a single value of the magnetic field; among them, we have considered \( \mathbf{A} = [0 \ Bx \ 0]' \) and \( \mathbf{A} = [-By \ 0 \ 0]' \).

The simulation approach we have used has been to divide the device into a number of transversal sections inside which both the potential and the potential vector can be approximately considered longitudinally constant, solve the Schrödinger equation inside each of these sections, use the obtained results to compute the scattering matrix of the transversal regions which go from the middle of each section to the middle of the next one, and compose these matrices in order to obtain at the end the overall transmission of the device.

Since, with the first choice, \( \mathbf{A} \) depends on \( x \), while this does not happen with the second choice, in the first case both the value of the magnetic field and the longitudinal variations of the potential limit the length of the sections we have to consider, while in the second case the only constraint to this length is given by the potential. In detail, the additional condition that has to be satisfied when using the first choice is that the magnetic flux threading each transversal region has to be much less than the flux quantum \( \hbar/e \) (where \( \hbar \) is the reduced Planck’s constant and \( e \) is the elementary charge).

On the other hand, if we rewrite the Schrödinger equation inside each section with longitudinally constant potential and potential vector, we observe that in the first case its solution is immediate (once the solutions in the absence of magnetic field are known) [37], while in the second case it requires the solution of a computationally expensive eigenproblem [38]. In more detail, in the first case the transverse energies (from which it is immediate to obtain the longitudinal wave vectors) are identical to those in the absence of magnetic field, while the transverse functions are multiplied by proper phase factors. In the second case, instead, for each injection energy it is necessary to solve a proper eigenproblem to find (from the eigenvalues) the longitudinal wave vectors and (from the eigenvectors) the coefficients with which we have to combine the wave functions in the absence of magnetic field to obtain the actual wave functions.

For potentials with fast variations in the longitudinal direction and low magnetic fields the length of the sections is mainly limited by the potential variations and thus the number of sections required by the two approaches is comparable. Therefore in this case the first approach is preferable, due to the smaller amount of calculations required for each section.

Instead for potentials with slow variations in the longitudinal direction and high magnetic fields the second
approach is the most convenient, since it strongly reduces the number of sections that have to be considered.

For example, in Fig. 9 we show the shape of an ideal mesoscopic cavity connected to the left and right leads by two narrow constrictions, and in Fig. 10 we show its transmission as a function of the electron energy for an orthogonal magnetic field $B=3$ T. In this case, the second technique is clearly preferable, due to the presence of a large magnetic field and of a very small number of longitudinal variations in the potential profile. This is a clear example in which a proper choice at the modeling level can dramatically impact the efficiency of the simulation code.

![Image of a mesoscopic cavity](image1)

Fig. 9 Mesoscopic cavity (7.75 μm long and 5 μm wide), connected to the left and right leads by two narrow (60 nm wide) constrictions.

![Image of transmission graph](image2)

Fig. 10 Transmission of the mesoscopic cavity represented in Fig. 8, as a function of the electron energy, computed for an orthogonal magnetic field $B=3$ T.

IX. CONCLUSION

I have listed, discussed and illustrated with examples some of the alternatives that exist in the physical description of nanoelectronic devices, clarifying the effect that the different choices have on the accuracy of the results and on the execution times of the simulation code. This not exhaustive description shows that a careful trade-off between the necessity to correctly describe, with a given level of detail, the behavior of the considered device and the need to limit the simulation times is necessary in the development of a nanodevice simulator, and that a proper choice between equivalent physical schematizations can lead to a notable improvement in the performance of the code.

REFERENCES


